Introduction à l'économétrie : Etude des Modèles Linéaires

CQLS : Jean-François Coeurjolly & Rémy Drouilhet
Jean-Francois.Coeurjolly@upmf-grenoble.fr, Remy.Drouilhet@upmf-grenoble.fr

Table des matières

Présentation du document

Lecture du document

Ce document est une introduction à l'économétrie via l'étude des modèles linéaires, problème traité dans les sections 1 à 11.

L'esprit de ce document est un peu particulier et mérite d'être souligné. L'accent ne sera pas mis sur l'obtention des résultats mathématiques (bien que leurs démonstrations soient proposées pour les gens intéressés) mais plus sur leurs interprétations. Autrement dit, nous essaierons de discuter beaucoup plus du langage mathématique que des techniques mathématiques. Par ailleurs, certaines propriétés (de type estimateurs sans biais, lois des estimateurs,...) caractérisant les variables aléatoires mises en jeu ou même la notion de modèle aléatoire étant particulièrment difficiles à appréhender (et qui constituent des résultats classiques issus de la théorie des probabilités), nous proposons à chaque étape de nous appuyer sur une **approche expérimentale des probabilités** (notée par la suite *A.E.P.*). Cette approche peut être développée dès que l'on s'intéresse à un phénomène aléatoire et sort donc complètement du cours d'économétrie où elle n'y est qu'appliquée. C'est pourquoi, nous avons créé une première annexe (annexe A) dans laquelle nous développons cette approche dans sa généralité.
Le test d'hypothèses est l'outil de base de la statistique puisqu'il est de manière générale très utilisée à tous les niveaux (en Deug, comme en Licence et dans toutes les formations nécessitant la statistique inférentielle). Il est par conséquent largement utilisé dans ce cours d'économétrie. Et pourtant, même si cela n'apparaît pas toujours aux yeux du néophyte, la démarche pour construire un test d'hypothèses est toujours la même. Du coup, nous proposons de l'intégrer au document dans l'annexe B.

Une bonne démarche pour lire et comprendre ce document consiste dès qu'une notion vous semble obscure à vous référer à l'une de ces deux annexes (A ou B).

Les Annexes D et E rassemblent les outils mathématiques et les démonstrations des différents résultats. Il est bien évident qu'il n'est pas nécessaire de les comprendre mais ils sont présentés pour les gens qui seraient intéressés par une approche plus classique des probabilités ou simplement par souci de complétude du document.

Le logiciel `R`

Parvenu à un certain niveau de statistiques, il est clair que les calculs mis en jeu sont de plus en plus lourds. Autant, il est possible (bien que peu judicieux) de ne pas utiliser un quelconque logiciel en Deug, autant il est presque indispensable (d'en utiliser un) à partir de la licence puisque les modèles linéaires requièrent des calculs matriciels conséquents.

Dans le cadre de ce cours, nous nous appuyons fortement sur le logiciel `R` car nous lui trouvons de nombreux avantages. Tout d'abord, il est gratuit (issu des logiciels libres sous licence GPL) et disponible sur n'importe quelle plate-forme ce qui d'un point de vue éthique est tout à fait satisfaisant. Mais surtout, il s'agit d'un outil graphique extrêmement intéressant et d'un outil de calcul particulièrement puissant permettant des calculs matriciels précis. De plus, il connaît toutes les lois usuelles (et bien plus encore) ce qui permet de manière précise de calculer des quantiles ou des probabilités sans avoir à se référer à des tables statistiques. Ainsi, l'utilisation de `R` dans le document de cours devrait permettre, nous l'espérons de comprendre au fur et à mesure par le détail des calculs les formules

mathématiques importantes. Nous rassemblons l'ensemble des instructions `R` utilisées dans le cours (et les devoirs) dans l'annexe C.

Cependant, nous tenons à souligner que le fait de se tourner vers tel ou tel logiciel n'est pas un choix définitif. Nous préférons `R` pour ses vertus pédagogiques mais de nombreux logiciels permettent de traiter le problème de régression linéaire via une interface souvent plus conviviale. On peut citer entre autres : SPSS, S-plus, Statistica, SAS,... et éventuellement les tableurs.

Pour ceux qui veulent tout de même en savoir plus sur le logiciel `R` (bien que ce ne soit pas requis rappelons-le), voici l'adresse internet où vous pourrez facilement le télécharger et ce quel que soit votre système d'exploitation : http ://cran.r-project.org/

Les principales notations

→ $\boldsymbol{y}$: (notation en gras) désigne un vecteur (un paquet) de réels.

→ $\underline{\boldsymbol{x}}$: (notation en gras souligné) désigne une matrice (un paquet de vecteurs).

→ Y : la notation en majuscules d'une quantité désignera une variable aléatoire (une future réalisation si ce la peut vous aider à mieux comprendre cette notion)

→ $\boldsymbol{Y}$: en intégrant les deux notations, il s'agit là d'un vecteur aléatoire i.e. d'un futur vecteur.

→ $\widehat{\theta}(\boldsymbol{y}|\underline{\boldsymbol{x}})$: estimation du paramètre θ calculée à partir des données stockées dans le vecteur $\boldsymbol{y}$.

→ $\widehat{\theta}(\boldsymbol{y}|\underline{\boldsymbol{x}})$: estimation du paramètre θ calculée à partir des futures données stockées dans le vecteur $\boldsymbol{Y}$ et des régresseurs stockés dans la matrice $\underline{\boldsymbol{x}}$. On appelle plus communément cette quantité estimateur du paramètre θ et il s'agit à nouveau d'une variable aléatoire.

Ces différentes notations même si dans une première lecture peuvent sembler barbares et particulièrement lourdes, nous semblent primordiales car très significatives. Il est important de pouvoir dissocier une quantité disponible (des données, une estimation) de données inaccessibles (futur jeu de données, variable et vecteur aléatoire, estimateur) pour lesquelles le mathématicien sait établir a priori beaucoup de résultats.

1 Introduction générale

Dans beaucoup de disciplines, nous sommes amenés à observer un phénomène d'intérêt en ayant le secret espoir de pouvoir le comprendre le mieux possible. La plupart du temps, cela se traduit par la récolte d'observations de certaines variables relatives à ce phénomène. A partir de ces données (le mot est explicite), nous espérons le décrire voire l'expliquer et par la suite proposer quelques interprétations pertinentes le concernant. Pour ce faire, nous commençons par dégager une ou plusieurs variables d'intérêt caractéristiques du phénomène étudié. Ensuite, la démarche se poursuit par la recherche d'autres variables ayant pour but d'expliquer les valeurs des premières. De cette description il se dégage un concept qui essaierait de porter les informations pertinentes d'un phénomène : c'est la notion de **modèle**!!!

En fait, il est sous-entendu que les variables observables sont quantifiables (i.e. remplaçables par des valeur numériques) et qu'il y aurait un grand intérêt à se tourner vers des disciplines à vocations quantitatives. L'économétrie fait partie de l'une d'entre elles appliquée dans le contexte des sciences économiques. Le principal centre d'intérêt de l'économétrie est précisément de proposer des candidats de modèle ayant pour but de décrire le mieux possible un certain phénomène économique.

Tentons de rentrer plus en détails dans la notion de modèle en utilisant une formulation plus quantitative. Sous sa forme générale, il établit la relation entre une variable d'intérêt (notée y) en fonction de (co)variables qui en seraient les causes. Mathématiquement, cela s'exprime par : $y \simeq f(x_1, x_2, \ldots, x_p)$ où f est une fonction de forme quelconque et généralement inconnue. Il y a fondamentalement deux types de modèles :

- → les modèles dit **déterministes** où il est supposé que la variable d'intérêt peut être caractérisée par les covariables (i.e. $y = f(x_1, x_2, \ldots, x_p)$). Cela signifie bien que la donnée de $x_1, x_2, \ldots, x_p$ et la connaissance de la fonction f permettrait de fournir la valeur exacte de y. Ces modèles ne font pas l'objet de ce cours mais il faut prendre conscience de leur existence. Au vu de cette description somme toute générale, nous pouvons nous demander s'il existe d'autre type de modèle. Par ailleurs, il existe beaucoup de personnes qui pensent que tout phénomène dans la nature peut se décrire avec les modèles précédents.
- → les modèles dit **aléatoires** où la formulation générale $y \simeq f(x_1, x_2, \ldots, x_p)$ est remplacée par : $Y = f(x_1, x_2, \ldots, x_p, U)$ que l'on se propose de simplifier tout de suite (peut-être abusivement mais ce n'est pas grave) en séparant les $x_1, x_2, \ldots, x_p$ du bruit U par la formultaion $Y = f(x_1, x_2, \ldots, x_p) + U$, avec $f(x_1, x_2, \ldots, x_p)$ décrit la composante déterministe du modèle (donc à comparer avec les modèles précédents) et U représente la composante aléatoire du modèle. Notons la notation en majuscule de la variable à expliquer Y qui est une convention admise (et utilisée dès que possible dans ce document) pour souligner sa nature aléatoire. Il y a quelques exceptions et notamment U qui n'a pas son pendant en majuscule dans l'alphabet grec. Cette formulation apparemment générale est pourtant loin de l'être puisqu'il n'est pas du tout évident que les deux composantes déterministe et aléatoire soient à additionner. On pourrait très bien penser soit à les multiplier soit à faire dépendre la composante aléatoire U des $x_1, x_2, \ldots, x_p$, soit .... Il y a beaucoup de variantes possibles mais faut-il être capable de les justifier. Bien qu'il soit possible de structurer la composante aléatoire de différentes façons, soulignons que ce cours d'introduction à l'économétrie ne traitera exclusivement que du cas le plus simple mathématiquement.

La question après une telle présentation qui nous brûle les lèvres est : comment est-il possible de différencier ces deux types de modèles ? En fait la réponse est simple mais elle repose dans beaucoup de situations pratiques sur une croyance (puisqu'il ne nous est pas possible d'en être sûr). Cela peut se formuler par une deuxième question : si vous pouviez observer à nouveau le même phénomène exactement dans les mêmes conditions, i.e. les valeurs des $x_1, x_2, \ldots, x_p$ restant inchangées, pensez-vous alors que la valeur de la variable d'intérêt doit elle-aussi rester inchangée ? Si votre réponse est affirmative alors vous devez postuler pour un modèle déterministe et dans le cas contraire pour un modèle aléatoire ? Ceci étant dit, il est à souligner qu'un modèle déterministe peut être vu comme un modèle aléatoire et que la réciproque est bien entendu fausse. En effet, il est possible d'imaginer une composante aléatoire ne variant (presque) plus (relativement à la composante déterministe) correspondant à poser U (presque) nulle.

Précisons à présent la forme des modèles étudiés dans un cours d'introduction à l'économétrie. Nous partirons du modèle "général" présenté précédemment et nous le simplifierons au fur et à mesure :

→ modèle "général" (explicite) :
modèle que nous allons tout de suite simplifier en supposant que le bruit U est additif :

$$\boxed{Y = f(x_1, x_2, \ldots, x_p) + U}$$

Soulignons qu'un tel modèle soulève une difficulté essentielle : comment déterminer f et est-ce possible ? Ce type de problème est d'un niveau bien plus avancé. Il y a principalement deux axes d'étude : soit avec des considérations théoriques décrire f comme la solution (à résoudre) d'une équation d'équilibre soit sous un angle statistique via des méthodes dites non paramétriques ou semi-paramétriques (mais c'est pas du tout le propos de ce cours). Nous devons alors opérer une première simplification à savoir fixer la forme de la fonction f. Cela pourrait être fait de manière quelconque mais nous allons encore une fois choisir la plus simple à traiter dans un contexte mathématique.

→ modèle linéaire :

$$\boxed{Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}_{f(x_1, x_2, \ldots, x_p)} + U}$$

Dans ce cas, les covariables sont aussi appelées régresseurs car ce modèle est souvent appelé modèle de régression (ici à p régresseurs). La terminologie "linéaire" associée au nom du modèle indique la nature de la fonction f ($f(x_1, \ldots, x_p)$ s'exprimant comme une combinaison linéaire des régresseurs). L'une des hypothèses qui est faite classiquement sur la composante aléatoire (appelée plus simplement "bruit" comme pour indiquer qu'elle perturbe la composante déterministe) est qu'elle suit une loi normale de moyenne nulle et d'écart-type σ inconnu (appelé aussi paramètre de nuisance). Cette hypothèse, appelée cas gaussien, n'est pas toujours vérifiable et peut être généralement levée lorsque le nombre d'observations du phénomène sera suffisamment grand. Quoi qu'il en soit, avec cette hypothèse (et quelques autres à détailler plus tard) le modèle peut se reformuler comme ci-dessous :

$$\boxed{Y \rightsquigarrow \mathcal{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \sigma) \text{ (cas gaussien)}}$$

Nous remarquons alors que la composante déterministe correspond exactement à la moyenne de la variable d'intérêt.

L'un des problèmes de tout modèle de la forme générale avec $p > 2$ covariables est que nous ne pouvons pas représenter le nuage de points associé. Il est donc un usage en économétrie de

commencer par introduire les modèles de régression faciles à représenter graphiquement.

→ modèle linéaire à un régresseur ($p = 1$) :

$$\boxed{\underbrace{Y = \beta_0 + \beta_1 x_1}_{\text{Equation d'une droite}} + U \text{ ou } Y \rightsquigarrow \mathcal{N}(\beta_0 + \beta_1 x_1, \sigma) \text{ (cas gaussien)}}$$

La dernière simplification est maintenant de se dire que l'on ne dispose d'aucun régresseur.

→ modèle sans régresseur ($p = 0$) :

$$\boxed{Y = \beta_0 + U \text{ ou } Y \rightsquigarrow \mathcal{N}(\beta_0, \sigma) \text{ (cas gaussien)}}$$

On constate donc qu'il existe une relation entre les notions introduites dans un cours de statistique inférentielle (généralement en deuxième année d'un cursus en sciences économiques) et celles relatives à un cours d'économétrie. La notion de modèle est plutôt passée sous silence préférée à la notion plus explicite de paramètre d'intérêt (ici l'écart-type σ et plus couramment la moyenne β_0 appelée dans ce contexte plus naturellement μ ou m) dans un cours de statistique alors qu'elle se révèle indispensable en économétrie.

Après cette présentation générale des modèles linéaires, nous pouvons essayer de motiver leur étude. Un modèle est capable de générer une certaine quantité de données. Ceci étant, il faut réaliser que ce modèle ne se dévoile qu'en rendant observables ces données et qu'il est donc impossible pour un observateur de tout savoir à son sujet. Pour illustrer ces propos, pour un modèle de régression les paramètres de régressions $\beta_0, \beta_1, \ldots, \beta_p$ ainsi que le paramètre de nuisance σ (avec $\sigma > 0$ pour indiquer le cas d'un modèle aléatoire) resteront toujours inconnus (exception faite de son concepteur s'il en existe un). La connaissance du modèle peut être décrite soit à partir de points de vue théoriques (par exemple économiques) soit à partir des seules informations étant reliées au modèle à savoir les données. Dans ce cours, nous décrirons les méthodes économétriques (ou plus généralement statistiques) permettant de révéler les informations portées par les données au sujet du modèle. Plus concrètement dans le cadre des modèles de régression linéaire, cela passera par la recherche de remplaçants (officiellement appelées estimations) des paramètres (théoriquement inconnus) de régressions et de nuisance calculables à partir des données. Ayant admis et pris conscience que ces modèles sont de nature aléatoire, il en sera de même pour les données récoltées et les estimations associées. Grâce à une Approche Expérimentale des Probabilités (A.E.P.), nous pourrons plus facilement appréhender des résultats statistiques (voire plutôt probabilistes) obtenus par l'Approche Classique des Probabilités (A.C.P.) plus adaptée. Les objectifs atteints dans ce cours introductif seront (dès lors que le modèle est accepté) :

→ une connaissance du modèle portée par les données observées (sensées être générées par ce dernier) avec une mesure des qualités des estimations. Il sera alors possible pour un praticien d'argumenter son discours théorique en évoquant que les données relatives au phénomène étudié confirment ou pas certaines assertions. L'utilisation de tests d'hypothèses sera dans ce cas souvent appropriée.

→ la prévision de la variable d'intérêt dès lors que l'on connaisse ou que l'on se fixe de nouvelles valeurs des régresseurs. Les intervalles de prévision à utiliser pour répondre à cet objectif intègrent (et c'est leur force) la variabilité du phénomène étudié.

Ce document sera accompagné d'une série d'exemples pour illustrer tous les points du cours dont trois sont décrit dès à présent.

Nous présentons ici une analyse très simple basée sur un exemple classique consistant en l'explication du salaire de travailleurs d'une certaine catégorie socio-professionnelle. Le problème est présenté successivement à trois étudiants plus ou moins compétents en statistique. Chacun d'entre eux explique sa démarche, recueille des données et procède à une analyse et commente brièvement ses résultats.

Exemple 1 : Le premier se dit que le salaire (variable notéé Sal par la suite) doit certainement dépendre du niveau d'expérience professionnelle de chacun. Il utilise un indicateur du niveau d'expérience à valeurs dans $[0, 1]$: plus cet indicateur, noté par la suite IndExp, est proche de 0 et moins l'individu interrogé a d'expérience professionnelle et inversement lorsque cet indicateur est proche de 1. Pour tenter d'expliquer le salaire en fonction du niveau d'expérience, ce premier étudiant envisage le modèle linéaire très simple : $Sal = \beta_0 + \beta_1 IndExp + U$. Pour appuyer son analyse il interroge $n = 5$ individus et recueille alors les salaires (en euros) et niveaux d'expérience des cinq individus. Assisté du logiciel R (précisons pour couper court à toute mutinerie que tout autre logiciel aurait fourni des résultats strictement équivalents), il obtient les résultats suivants :

```
> summary(lm(Sal~IndExp,data=ex1SalData))

Call:
lm(formula = Sal ~ IndExp, data = ex1SalData)

Residuals:
      1       2       3       4       5
-141.28  -79.75  315.05  -32.54  -61.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1599.8      266.3   6.007  0.00924 **
IndExp         279.6      491.8   0.569  0.60943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 208.5 on 3 degrees of freedom
Multiple R-Squared: 0.09727,    Adjusted R-squared: -0.2036
F-statistic: 0.3233 on 1 and 3 DF,  p-value: 0.6094
```

Commentaires de l'étudiant : *"L'expérience professionnelle n'explique en rien le niveau des salaires."*

Exemple 2 : Le deuxième étudiant reprend l'analyse effectuée par le premier étudiant et est particulièrement critique quant à ses compétences statistiques.
Commentaires de l'étudiant : *"Il est totalement dénué de sens de fonder des conclusions sur un modèle (où trois paramètres sont à estimer) en se basant sur l'analyse d'un jeu de données constitué uniquement de $n = 5$ observations. Cinq observations ne permettent pas d'appréhender la variabilité du salaire. Sur cinq autres observations j'aurais pu obtenir des résultats menant à des interprétations foncièrement différentes. En conséquence, si jamais je n'avais à disposition que cinq observations, je*

m'abstiendrai de formule une quelconque conlusion !"
Autrement dit, le deuxième étudiant ne critique pas le modèle envisagé par le premier étudiant mais émet de sérieuses réserves quant à sa démarche d'interroger un aussi faible nombre de personnes. Pour cela, il décide de compléter le jeu de données de son camarade à $n = 200$. Après recueil des nouvelles données, il effectue l'analyse suivante :

```
> summary(lm(Sal~IndExp,data=exSalData))

Call:
lm(formula = Sal ~ IndExp, data = exSalData)

Residuals:
   Min     1Q Median     3Q    Max
-519.9 -265.3  -59.9  290.6  648.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1047.83      43.86   23.89   <2e-16 ***
IndExp       1487.89      75.83   19.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 317.7 on 198 degrees of freedom
Multiple R-Squared: 0.6604,     Adjusted R-squared: 0.6587
F-statistic:   385 on 1 and 198 DF,  p-value: < 2.2e-16
```

Commentaires de l'étudiant : *"Mon jeu de données me permet de montrer que l'action de IndExp dans l'explication du salaire est très forte. De plus, toute prévision que je pourrais effectuer à l'aide de mon modèle estimé aura une précision relativement satisfaisante."*

Exemple 3 : Le troisième étudiant apprécie les résultats et interpétations obtenues par le deuxième étudiant et souhaite apporter sa pierre à l'édifice en complétant le modèle. Il a l'intuition que le salaire pourrait dépendre outre le niveau d'expérience professionnelle du niveau d'étude de chacun. De manière analogue à l'indicateur IndExp, il utilise un indicateur du niveau d'étude (noté IndEtu) qu'il contraint également à être sur $[0, 1]$ et dont l'interprétation est aisée : plus cet indicateur est proche de 0 et plus le niveau d'étude de l'individu est faible et inversement lorsque cet indicateur est proche de 1.

```
> summary(lm(Sal~IndExp+IndEtu,data=exSalData))

Call:
lm(formula = Sal ~ IndExp + IndEtu, data = exSalData)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-539.07 -265.83    1.85  273.36  644.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   928.93      55.56  16.720  < 2e-16 ***
IndExp       1489.23      73.94  20.140  < 2e-16 ***
IndEtu        238.91      71.25   3.353 0.000959 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 309.8 on 197 degrees of freedom
Multiple R-Squared: 0.6787,     Adjusted R-squared: 0.6755
F-statistic: 208.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

Commentaires de l'étudiant : *"Au vu du jeu de données, l'action de chacun des indicateurs est très significative dans l'explication du salaire individuel. En outre, le pouvoir explicatif de ce nouveau modèle n'a pas beaucoup évolué par rapport au précédent modèle."*

Ayant nous l'espérons aguiché le lecteur, les sections suivantes visent entre autres à fournir les outils théoriques nécessaires pour effectuer une analyse de régression linéaire et en comprendre et interpréter les différents résultats.

La section 2 se propose de définir le modèle linéaire sous sa forme la plus simple, la forme matricielle. Nous expliquons également plus en détail en quel sens le modèle est aléatoire. Les méthodes permettant d'estimer les paramètres du modèle nécessitent quelques hypothèses particulières sur le modèle : celles-ci y sont énoncées dans la même section. Dans la section 3, on discute de la manière dont on peut estimer les paramètres d'un modèle linéaire général. Les ordres de grandeur des paramètres inconnus du modèle y sont définis ainsi que la notion de résidus et de valeurs ajustées. La méthode sur laquelle on s'appuie est la methode des moindres carrés ordinaires. Le cas particulier où la variable expliquée ne dépend que d'un régresseur est mis en évidence. Ceci étant fait, on peut naturellement s'interroger sur la qualité des estimations (ou remplaçants) des paramètres que l'on obtient via la M.C.O. : la section 4 présente les principales propriétés des estimateurs développés. On s'intéresse à l'estimation du niveau de bruit dans la section 5 ce qui nous permet également d'estimer le niveau de précision des estimations à partir d'un unique jeu de données. Dans la section 6, on s'appuie sur une propriété géométrique de la méthode M.C.O. pour dégager un indicateur de la qualité d'ajustement du "nuage de points" autour du modèle estimé. Cet indicateur est appelé coefficient de détermination multiple ou coefficient de corrélation lorsqu'il n'y a qu'une seule variable explicative. En s'appuyant sur les hypothèses du modèle, on précise dans la section 7 les comportements aléatoires des différents estimateurs mis en jeu. La connaissance de ces comportements aléatoires nous permet d'une part de construire toute une série de tests d'hypothèses (on y retrouve entre autres le test de significativité locale) et d'autre part de construire des intervalles et régions de confiance pour les paramètres du modèle. Tout ceci est résumé dans la section 8. Enfin, le thème de la prévision est traité dans la section 9 : la construction d'une nouvelle valeur prédite ainsi que l'intervalle de prévision associé y

sont présentés. Le lecteur intéressé par la démonstration des résultats énoncés pourra se référer à la section E.

D'un point de vue général, (re)précisons que nous ferons l'effort d'illustrer chacun des résultats mathématiques présentés via l'A.E.P. et que chacun d'entre eux (lorsque ce sera possible) sera appliqué aux trois exemples présentés ci-dessus.

2 Présentation du modèle

2.1 Description générale du modèle

Le modèle linéaire général, appelé aussi modèle de régression multiple, se présente sous la forme suivante

$$Y_t = \beta_0 + \beta_1 x_t^{(1)} + \beta_2 x_t^{(2)} + \ldots + \beta_p x_t^{(p)} + U_t,\ t = 1, 2, \cdots, n, \cdots.$$

Dans cette étude, on disposera des observations jusqu'au temps n, s'exprimant par le système d'équations linéaires suivant :

$$\left\{\begin{array}{rcl} Y_1 &=& \beta_0 + \beta_1 x_1^{(1)} + \beta_2 x_1^{(2)} + \cdots + \beta_p x_1^{(p)} + U_1 \\ & \vdots & \\ Y_t &=& \beta_0 + \beta_1 x_t^{(1)} + \beta_2 x_t^{(2)} + \cdots + \beta_p x_t^{(p)} + U_t \\ & \vdots & \\ Y_n &=& \beta_0 + \beta_1 x_n^{(1)} + \beta_2 x_n^{(2)} + \cdots + \beta_p x_n^{(p)} + U_n \end{array}\right.$$

Sous forme matricielle ce système se réécrit comme suit :

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} \underbrace{1}_{\boldsymbol{x}^{(0)}} & \underbrace{x_1^{(1)}}_{\boldsymbol{x}^{(1)}} & \underbrace{x_1^{(2)}}_{\boldsymbol{x}^{(2)}} & \cdots & \underbrace{x_1^{(p)}}_{\boldsymbol{x}^{(p)}} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(p)} \end{pmatrix}}_{\underline{\boldsymbol{x}}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_n \end{pmatrix}}_{\boldsymbol{U}},$$

soit

$$\boxed{\boldsymbol{Y} = \underline{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{U}}$$

Les vecteurs $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(p)}$ seront appelés régresseurs et $\boldsymbol{U}$ le bruit. En particulier le vecteur $\boldsymbol{x}^{(0)}$ constitué uniquement de 1, noté aussi $\mathbf{1}$, est appelé régresseur constant.

De manière plus générale, notons que l'indice t représentant usuellement le temps dans les applications économétriques pourrait être substitué à un indice plus général i qui symboliserait un individu (ou une unité statistique).

Exemple de cours : Pour essayer d'illustrer la notion de modèle, ainsi que tous les concepts qui suivront, nous proposons de considérer un exemple totalement fictif et ce uniquement lorsque $p = 1$ (un seul régresseur). Le choix d'un seul régresseur dans l'étude provient de la facilité de visualisation des différents concepts. Deux régresseurs auraient impliqué une visualisation en 3D moins évidente ; au-delà de deux régresseurs toute visualisation est impossible et l'abstraction est indispensable. Revenons-en à notre exemple, et définissons notre modèle théorique qui dans le cas d'un modèle à un régresseur revient à spécifier une droite ou encore spécifier les deux paramètres β_0 et β_1. L'ordinateur est capable de choisir arbitrairement β_0 et β_1 sans que nous puissions a priori les connaître. Il faut voir le modèle comme faisant partie intégrante de l'ordinateur, et ceci est comparable (en un certain sens) à la pratique où le modèle n'est jamais (complètement) connu. À la différence du cadre pratique, nous aurons la possibilité d'interroger l'ordinateur pour nous fournir autant de données que l'on désire. De plus, l'ordinateur est programmé pour nous représenter graphiquement le modèle (sans les échelles pour ne pas nous dévoiler les vrais paramètres) comme suit :

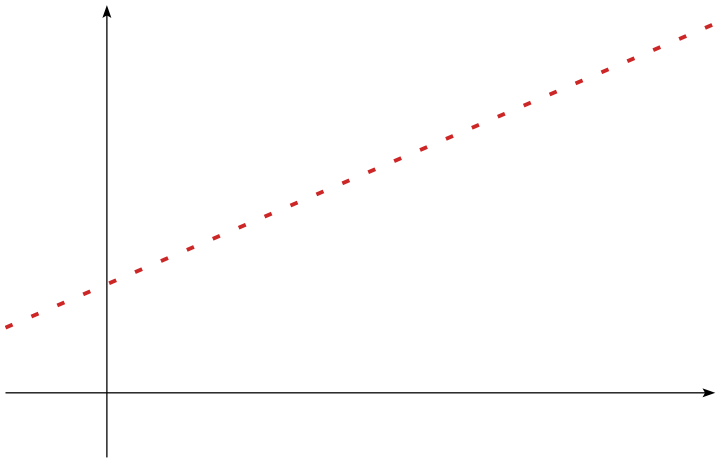


FIGURE 1 – Modèle de l'exemple de cours

À partir du modèle, nous allons générer un jeu de données (qui constituera par la suite l'unique information dont on dispose en pratique). On commence par définir un vecteur de régresseurs déterministe $\boldsymbol{x}_1$, i.e. n points sur l'axe des abscisses. Le nombre d'observations considéré dans cet exemple fictif est $n = 10$. Pour $t = 1, \ldots, 10$, on part de l'abscisse x_{t1} on définit l'ordonnée associée sur la droite théorique i.e. $\beta_0 + \beta_1 x_{t1}$, ce qui constitue la partie déterministe du modèle. Ensuite, on définit pour chacun des points sur la droite théorique une "perturbation aléatoire" (notion expliquée dans la section 2.4), notée ϵ_t. Le jeu de données est donc constitué du vecteur $\boldsymbol{x}_1$ et des $y_t = \beta_0 + \beta_1 x_{t1} + U_t$

pour $t = 1, \ldots, 10$ (notez bien la notation en minuscules !). Autrement dit, les n points générés ne sont que les n points de la droite théorique (dont les abscisses sont portées par le vecteur $\boldsymbol{x}_1$) qui ont été "perturbés" par les U_t.

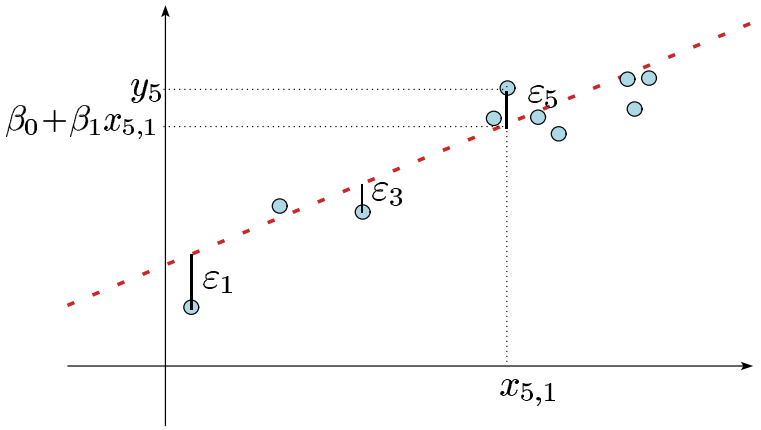


FIGURE 2 – Modèle et jeu de données généré de l'exemple de cours

La figure ci-dessous illustre tout à fait le type de problème soulevé : comment retrouver (est-ce possible ?) le modèle théorique si je ne dispose que du nuage de points, i.e. des observations $(x_{t1}, y_t)_{1 \leq t \leq 10}$ perdant ainsi toute l'information sur les "perturbations" U_t ?

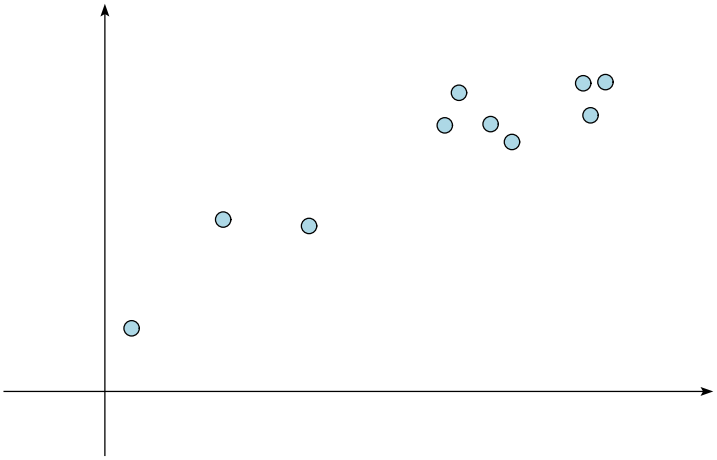


FIGURE 3 – Jeu de données de l'exemple de cours

Revenons brièvement aux exemples pour décrire avec les notations de cette section les différents jeux de données.
Exemple 1 : la matrice $\underline{\boldsymbol{x}}$ des régresseurs, de taille (5×2), est définie par $\underline{\boldsymbol{x}} = (\mathbf{1}, \boldsymbol{IndExp})$ et le vecteur des réponses $\boldsymbol{y}$ est le vecteur à $n = 5$ composantes $\boldsymbol{Sal}$.
Exemple 2 : la matrice $\underline{\boldsymbol{x}}$, de taille (200×2), est définie par $\underline{\boldsymbol{x}} = (\mathbf{1}, \boldsymbol{IndExp})$ et le vecteur des réponses $\boldsymbol{y}$ est le vecteur à $n = 200$ composantes $\boldsymbol{Sal}$.
Exemple 3 : la matrice $\underline{\boldsymbol{x}}$, de taille (200×3), est définie par $\underline{\boldsymbol{x}} = (\mathbf{1}, \boldsymbol{IndExp}, \boldsymbol{IndEtu})$ et le vecteur des réponses n'a pas changé : $\boldsymbol{y} = \boldsymbol{Sal}$.

Désormais, chacun doit avoir conscience que parmi les quatres exemples présentés ci-dessus, le modèle n'est jamais connu. Nous ne disposons que d'un nuage de points et c'est a priori notre seule information.

2.2 Modèles linéarisables

Il existe d'autres modèles que le modèle linéaire pour lesquels tout ce que nous allons raconter par la suite reste valable : il s'agit des modèles linéarisables (dont font partie le modèle log-linéaire, le modèle logarithmique, le modèle exponentiel et les modèles polynomiaux entre autres). On dira qu'un modèle est linéarisable s'il existe une transformation simple qui le ramène à un modèle linéaire. Le tableau suivant présente ces différents modèles ainsi que leur forme linéarisée (les hypothèses que nécessitent chacun des modèles ainsi que celles permettant de passer à la forme linéarisée sont volontairement omises par souci de simplicité) :

Modèle initial	Modèle "linéarisé"
Modèle log-linéaire $Y_t = \kappa \left(x_t^{(1)}\right)^{\beta_1} \left(x_t^{(2)}\right)^{\beta_2} \dots \left(x_t^{(p)}\right)^{\beta_p} \eta_t$	$\log(Y_t) = \beta_0 + \beta_1 \log(x_t^{(1)}) + \dots + \beta_p \log(x_t^{(p)}) + U_t$ avec $\beta_0 = \log(\kappa)$ et $U_t = \log(\eta_t)$
Modèle logarithmique (déjà linéarisé)	$Y_t = \beta_0 + \beta_1 \log(x_t^{(1)}) + \dots + \beta_p \log(x_t^{(p)}) + U_t$
Modèle exponentiel $Y_t = \kappa \exp(\beta_1 x_t^{(1)}) \exp(\beta_2 x_t^{(2)}) \dots \exp(\beta_p x_t^{(p)}) \eta_t$	$\log(Y_t) = \beta_0 + \beta_1 x_t^{(1)} + \dots + \beta_p x_t^{(p)} + U_t$ avec $\beta_0 = \log(\kappa)$ et $U_t = \log(\eta_t)$
Modèle polynômial $Y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \dots + \beta_p x_t^p + U_t$	$Y_t = \beta_0 + \beta_1 x_t^{(1)} + \dots + \beta_p x_t^{(p)} + U_t$ avec $x_t^{(1)} = x_t, x_t^{(2)} = x_t^2, \dots, x_t^{(p)} = x_t^p$

Parmi ces modèles linéarisables, le modèle log-linéaire connaît un franc succès auprès des économistes car ce modèle permet de modéliser des phénomènes à **élasticité constante**. Considérons deux variables x et y (quantité de bien,...), l'élasticité de y sur x est défini par : $e_{y/x} = \dfrac{dy/y}{dx/x}$. Supposons qu'elle soit constante et égale à β_1. Ceci peut alors se réécrire $\dfrac{dy}{y} = \beta_1 \dfrac{dx}{x}$, et en intégrant cette égalité, il vient : $\log(y) = \beta_0 + \beta_1 \log(x)$. En conséquence, si l'on pense que deux variables évoluent à élasticité constante, le seul modèle permettant de traduire approximativement cette relation est le modèle log-linéaire : $\log(Y_t) = \beta_0 + \beta_1 \log(x_t) + U_t$. Cette remarque est très importante car l'élasticité

constante entre deux variables possède une interprétation explicite et très informative. Par exemple, si l'élasticité de y sur x vaut $\beta_1 = 10\%$ (resp. $\beta_1 = -5\%$) une augmentation de δ de la variable x entraînera une augmentation (resp. une diminution) de $\delta \times 10\%$ (resp. $-\delta \times 5\%$). Enfin, terminons en précisant que le raisonnement tenu avec deux variables (donc un régresseur) est valable avec plusieurs variables (et donc p régresseurs). Ainsi, chacun des paramètres β_1 pour $i = 1, \ldots, p$ du modèle : $\log(Y_t) = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_p x_{tp} + U_t$ peut s'interpréter comme l'élasticité de Y sur x_i pour $i = 1, \ldots, p$.

Désormais, nous supposerons toujours que nos observations sont issues, après une éventuelle transformation, d'un modèle linéaire. Il est assez commun d'affirmer que ces modèles couvrent à l'heure actuelle un grand nombre d'applications économétriques, d'où l'intérêt à nouveau de comprendre parfaitement ce modèle.

2.3 Modèle linéaire avec covariables qualitatives

Jusqu'à présent nous n'avons pas décrit en détail la nature des variables utilisables dans les modèles linéaires. Il était même sous-entendu que les variables mises en jeu dans le modèle devaient être **quantitatives** (i.e. à valeurs numériques). Alternativement, il existe un autre type de variable dite **qualitative** ou **nominale** servant généralement à décrire littéralement une certaine caractéristique. Ce type de variable permet donc de séparer les observations en groupes (chaque groupe étant associé à une modalité). Précisons tout de suite que la variable d'intérêt dans un modèle de régression linéaire (classique) ne peut pas être de ce type. En effet, de par les hypothèses faites sur ce modèle, la variable d'intérêt doit être quantitative et en plus continue (i.e. l'ensemble de ses modalités est une réunion d'intervalles de réels). Il existe cependant certains modèles capables de traiter ce type de problème (voir cours d'économétrie très avancé).

Nous allons montrer comment à peu de frais elles peuvent être utilisées en lieu et place dans les régresseurs (covariables). Il faut auparavant signaler qu'un praticien inexpérimenté transformant une variable qualitative ordinale (par exemple le niveau d'expérience $NivExp$) par un codage numérique respectant l'ordre exprimé par ses modalités (1 pour "moyen", 2 pour "bon" et 3 "très bon") commet une erreur en l'utilisant en tant que covariable. En effet, pour que ce type de variable puisse être utilisée comme covariable il faudrait que celle-ci soit véritablement (après le codage numérique) quantitative. Sur l'exemple du niveau d'expérience $NivExp$ (à ne pas confondre avec la variable quantitative $IndExp$ des exemples du cours), on peut douter que l'écart exprimé entre "moyen" et "bon" soit le même que "bon" et "très bon" alors que c'est précisément ce que sous-entend le codage numérique puisque ces écarts sont égaux à 1 dans les deux cas (2-1 et 3-2). Il y a fort heureusement une solution à ce problème qui passe par l'étude des variables qualitatives à deux modalités (par exemple, la variable S communément appelée "Sexe" qui indique le genre "femme" ou "homme" d'un individu). Le codage général par 1 pour la première modalité (ex : "femme") et par 0 pour la seconde modalité (ex : "homme") ou vice versa, porte la même information la variable nominale d'origine puisqu'elle exprime l'appartenance à la catégorie associée à la première modalité (le 1 jouant le rôle du "oui" et le 0 celui du "non"). De plus, et c'est l'objectif visé, ce type de variable peut très facilement être utilisée dans un modèle de régression linéaire. Reprenons pour illustrer ces propos l'exemple 1 ou 2 du cours. Si nous pensions que l'étude du Salaire a toute les raisons d'avoir différents comportement selon la variable Sexe S alors nous devons postuler pour le modèle suivant à trois régresseurs :

$$\boxed{Sal = (\beta_0^{(1)} + \beta_0^{(2)} S) + (\beta_1^{(1)} + \beta_1^{(2)} S) IndExp + U}$$

qui se peut se reécrire

$$\boxed{Sal = \beta_0 + \beta_1 S + \beta_2 IndExp + \beta_3 IndExp \times S + U}$$

ou encore

$$\boxed{Sal = \beta_0 S + \beta_1 (1 - S) + \beta_2 IndExp \times S + \beta_3 IndExp \times (1 - S) + U}$$

Ce type de modèle permet de traiter simultanément dans un même modèle les deux régressions simples qu'il aurait été possible de traiter séparément sur les deux nuages de points des femmes et des hommes.

Le point important est de comprendre que les variables qualitatives permettent de découper le nuage de points en plusieurs nuages de points. Nous comprenons alors que la variable d'appartenance à une modalité (ayant uniquement deux modalités 1 et 0) appelée en économétrie variable "muette" (ou "dummy variable" en anglais) est l'outil permettant de traiter n'importe quelle variable qualitative. En effet, si nous reprenons l'exemple de la variable $NivExp$ en introduisant les variables d'appartenance à chacune des modalités $Moyen$, Bon et $TresBon$ (égales à 1 pour des individus ayant la modalité associée et 0 sinon), il nous est possible de les intégrer à un quelconque modèle linéaire mettant en jeu des covariables quantitatives. Pour étendre une régression simple d'une variable d'intérêt Y en fonction d'un quelconque régresseur x, nous pouvons alors poser le modèle à $p = 6$ régresseurs :

$$\boxed{Y = (\beta_0^{(1)} + \beta_0^{(2)} Moyen + \beta_0^{(3)} Bon) + (\beta_1^{(1)} + \beta_1^{(2)} Moyen + \beta_1^{(3)} Bon) \times x + U}$$

qui se peut se reécrire

$$\boxed{Y = \beta_0 + \beta_1 Moyen + \beta_2 Bon + \beta_3 x + \beta_4 x \times Moyen + \beta_5 x \times Bon + U}$$

ou encore

$$\boxed{Y = \beta_0 Moyen + \beta_1 Bon + \beta_2 TresBon + \beta_3 x \times Moyen + \beta_4 x \times Bon + \beta_5 x \times TresBon + U}$$

En résumé, toute variable qualitative pourra être utilisée dans un modèle linéaire après avoir introduit les variables quantitatives d'appartenance à chacune de ses modalités.

2.4 Appréhension de la composante aléatoire du modèle

Ayant accepté l'existence d'une composante aléatoire dans le modèle, nous allons de manière disciplinée (cf Annexe A), appréhender la variabilité du modèle via l'**Approche Expérimentale des Probabilités (A.E.P.)**. Introduisons alors les notations relatives à cette approche dans le cadre des modèles linéaires. Nous invitons le lecteur de ce document soucieux d'avoir une bonne compréhension de la nature aléatoire du modèle à bien lire (voire relire autant de fois que nécessaire) cette section, l'annexe A ainsi que toutes les parties du document se référant à l'**A.E.P.**. Avant d'étudier la nature aléatoire du jeu de données intéressons-nous d'abord à son procédé de "fabrication" (une à une) à savoir le modèle.

Comportement aléatoire du modèle

Une question des plus intéressantes est de savoir combien de données il faudrait disposer pour porter toute l'information caractérisant le modèle, si cela reste possible. Cette question est tout à fait adaptée à n'importe quel type de modèle générateur de données. Concentrons-nous cependant sur le modèle linéaire et essayons d'établir la relation entre ses paramètres le caractérisant et les données qu'il génère. De par la nature aléatoire intrinsèque du modèle, il faudra au moins une infinité de données (pas moins) pour y réussir. L'expérience aléatoire $\mathcal{E}$ est donc d'acquérir une future donnée Y générée par le modèle

à partir d'un vecteur quelconque $\boldsymbol{x}$ de valeurs des p régresseurs et de sa composante aléatoire U appelée bruit. Grâce à l'**A.E.P.** et un peu d'imagination, répétons une première fois cette expérience aléatoire à partir d'un premier vecteur $\boldsymbol{x_{[1]}}$ et d'une première réalisation $U_{[1]}$ du bruit pour obtenir $y_{[1]}$, puis une deuxième fois à partir d'un deuxième vecteur $\boldsymbol{x_{[2]}}$ et d'une deuxième réalisation $U_{[2]}$ du bruit pour obtenir $y_{[2]}$,..., puis une $m^{ème}$ fois à partir d'un $m^{ème}$ vecteur $\boldsymbol{x_{[m]}}$ et d'une $m^{ème}$ réalisation $U_{[m]}$ du bruit pour obtenir $y_{[m]}$, puis.... Poursuivons dans l'imaginaire et représentons toutes ces informations par un nuage d'une infinité de points dans un espace de représentation en dimension $p+1$. La nature du jeu de données (dont disposera en pratique) représentable lui-aussi par un nuage de n points dans le même espace de représentation est ainsi tout à fait comparable à celle de ce nuage virtuel d'une infinité de points virtuels, la différence résidant dans leurs tailles respectives. Soulignons qu'un nuage de points ne conserve que l'information de leurs coordonnées (à savoir tous les y et $\boldsymbol{x}$) et perd l'information de leurs distances (à savoir tous les U) à l'hyperplan (d'équation $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$) décrivant la composante déterministe du modèle (la figure 3 suivant la figure 2 illustre ce propos sur l'exemple du cours). Voilà, le cadre de l'**A.E.P.** relativement au modèle est posé mais nous n'avons pas encore tenté d'établir les éventuels liens entre les paramètres du modèle et le nuage d'une infinité de points. A vrai dire, à ce niveau du document, il nous est impossible de penser qu'il existe une quelconque relation entre ces données virtuelles et les paramètres de régression $\boldsymbol{\beta}$ d'une part et le paramètre de nuisance σ^2 d'autre part.

Pour uniquement comprendre la nature des hypothèses faites sur le modèle dans la section suivante, nous allons nous contenter d'étudier l'infinité des réalisations $U_{[1]}, \ldots, U_{[m]}, \ldots$ caractérisant la composante aléatoire du modèle et contribuant à la génération des $y_{[1]}, \ldots, y_{[m]}, \ldots$. Elles ne peuvent pas être appelées "données" puisqu'elles sont indisponibles en pratique. Elles nous permettent toutefois d'appréhender la nature aléatoire du bruit U du modèle et d'établir la relation qui les relient au paramètre de nuisance σ^2. Par définition d'un bruit, nous savons que $E(U) = 0$ (il est dit centré) qui se traduit via l'**A.E.P.** par $\lim\limits_{m\to+\infty} \dfrac{1}{m} \sum\limits_{i=1}^{m} U_{[i]} = 0$. Le paramètre de nuisance $\sigma^2 = \mathbb{V}\text{ar}(U) = E(U^2)$ mesure l'intensité de la dispersion des points du nuage au tour de l'hyperplan et s'exprime via l'**A.E.P.** par $\lim\limits_{m\to+\infty} \dfrac{1}{m} \sum\limits_{i=1}^{m} U_{[i]}^2 = \sigma^2$. Le bruit U étant une variable aléatoire, nous savons que fixer la répartition d'une infinité virtuelles de ses réalisations $U_{[1]}, \ldots, U_{[m]}, \ldots$ permet de complètement le définir. Dans ce cours introductif d'économétrie, des hypothèses simplificatrices seront faites et notamment l'une portant justement sur la loi de bruit. Il sera supposé par la suite que ce bruit suit une loi normale centrée et d'écart-type σ. Via l'**A.E.P.**, cela signifie que la répartition d'une infinité $U_{[1]}, \ldots, U_{[m]}, \ldots$ de réalisations du bruit doit satisfaire la contrainte d'être représentable par un histogramme à pas zéro ayant une forme préfixée. En y réfléchissant bien, nous pouvons comprendre le sens de la terminologie "hypothèse simplificatrice" utilisée précédemment. Comment à la seule vue d'un jeu de données (un nuage de points) pouvons-nous déterminer le comportement aléatoire du bruit U a priori déconnecté des données. Fort heureusement, dès que le nombre n de points observés est suffisamment grand, ce type d'hypothèse peut être levé et grâce à un véritable miracle nous pourrons pourtant atteindre les mêmes objectifs.

Comportement aléatoire du futur jeu de données

L'expérience aléatoire (notée $\boldsymbol{\mathcal{E}}$) consiste à "demander" au modèle de nous fournir un nouveau jeu de données de taille n, les régresseurs $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(p)}$ étant déterminés une fois pour toute (d'où la terminologie "déterministe"). Selon le schéma type de l'**A.E.P.**, imaginons alors pouvoir disposer d'un grand nombre m (en fait une infinité) de jeux de données, résultats de la même expérience aléatoire $\boldsymbol{\mathcal{E}}$. Ils seront notés $\boldsymbol{y}_{[1]}, \boldsymbol{y}_{[2]}, \ldots, \boldsymbol{y}_{[m]}, \ldots$ dans la suite du document. L'**A.E.P.** nous fait comprendre que tout utilisateur de modèles économétriques et notamment aléatoires encourt le risque d'oublier que son jeu de données n'est qu'une réalisation du modèle et qu'il aurait pu obtenir un tout autre jeu de données à savoir l'un des $\boldsymbol{y}_{[1]}, \boldsymbol{y}_{[2]}, \ldots, \boldsymbol{y}_{[m]}, \ldots$.

Exemple de cours : la figure Fig.4 illustre la variabilité du modèle puisque dix jeux de données ont été générés avec le même modèle, le dernier graphique étant supposé l'unique jeu de données disponible en pratique. Ce dernier est précisément celui qui a été présenté comme exemple de base du cours. Comprenez bien à présent que vous ne disposez que du dixième jeu de données mais que vous auriez parfaitement pu obtenir l'un des neuf autres voire d'une **infinité** d'autres non représentées car bien évidemment non représentables dans ce document.

2.5 Hypothèses sur le modèle

Nous serons amenés à faire quelques hypothèses sur le modèle dites **hypothèses classiques**. Il convient dans une première lecture de ce document de prendre conscience que tout modèle nécessite des hypothèses particulières sans pour autant les comprendre et leur prêter beaucoup d'attention :

(C1) : $\boldsymbol{Y}$ et $\underline{\boldsymbol{x}}$ sont observés sans erreur.

(C2) :
- **(C2-1)** *(Homoscédasticité)* Pour tout $n > 0$, $\boldsymbol{\mu_U} \overset{\text{Déf.}}{=} \boldsymbol{E}\left((U_1, \cdots, U_n)^T\right) = \boldsymbol{0}$ et $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{U}} \overset{\text{Déf.}}{=} \underline{\boldsymbol{V}}\left((U_1, \cdots, U_n)^T\right) = \sigma^2 \underline{\boldsymbol{I}}_n$ avec $\sigma^2 < +\infty$.
- **(C2-2)** *(Bruit gaussien)* Pour tout $n > 0$, $(U_1, \cdots, U_n)^T$ est un vecteur gaussien.

(C3) :
- **(C3-1)** *(Linéaire indépendance des $\boldsymbol{x}^{(j)}, j = 1, \cdots, p$)* La matrice $\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)$ est inversible.
- **(C3-2)** Lorsque n tend vers $+\infty$, $\frac{1}{n}\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}$ tend vers une matrice inversible.

L'hypothèse **(C2-2)** exprime entre autres choses que les $U_1, \ldots, U_n$ sont des variables aléatoires centrées, mutuellement indépendantes, toutes de loi Normale $\mathcal{N}(0, \sigma)$. Ainsi, le paramètre σ^2 quantifie à lui seul le niveau de bruit. Cette hypothèse **(C2-2)** constitue un cadre d'étude classique mais parfois restrictif ; il sera possible de s'en abstenir lorsque la taille d'échantillon est suffisamment grande.

Remarque sur l'exemple de cours : les U_t qui ont permis d'obtenir les 10 points à partir de la droite théorique ont été générés à partir d'une loi normale. La perturbation aléatoire satisfait donc l'hypothèse **(C2-2)**. Ce choix est induit par le fait que la taille d'échantillon est faible ($n = 10$).

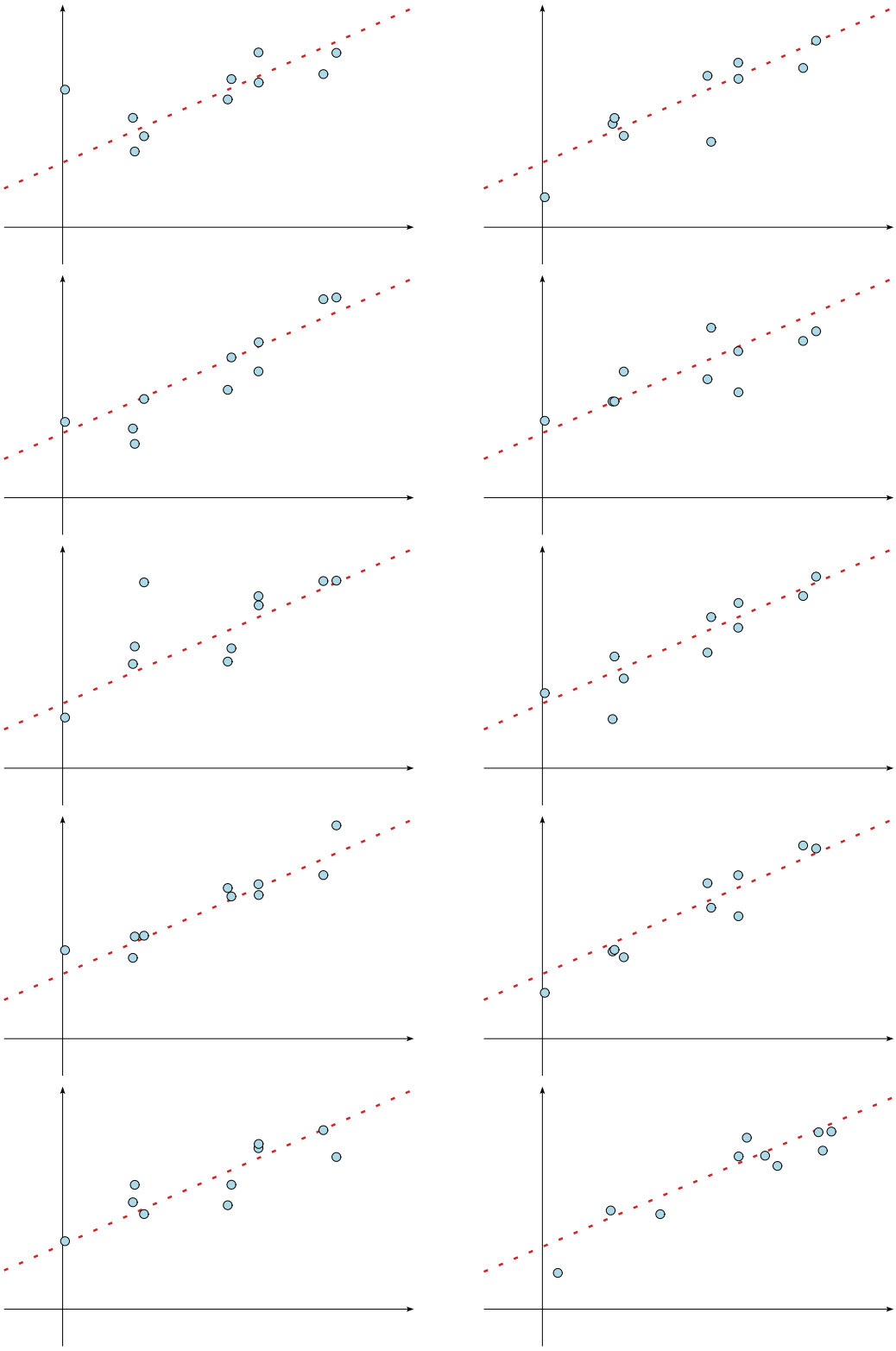


FIGURE 4 – Différentes réalisations du modèle issu de l'exemple du cours

3 Estimation des paramètres de la régression

Rappelons la position du problème via l'exemple de cours. À la vue de la figure Fig. 3 ne représentant que le nuage de points dont on dispose en pratique, nous nous demandons assez naturellement s'il est possible de retrouver une quelconque information sur le modèle théorique qui les a générés. Cela revient à proposer des remplaçants des paramètres β_0 et β_1 caractérisant la composante déterministe du modèle. Sans difficulté, nous pouvons étendre ce problème d'estimations des paramètres de régression au cadre général du modèle linéaire à p régresseurs.

3.1 Estimateur des Moindres Carrés Ordinaire (MCO)

Essayons de déterminer une méthode d'estimation des paramètres de régression en commençant par l'exemple de cours.

Exemple de cours : Chercher à estimer les paramètres β_0 et β_1 revient géométriquement à chercher une droite (dite droite ajustée) remplaçant celle caractérisant la composante déterministe du modèle (dite droite théorique). Nous savons que la droite représentant le modèle a de fortes chances de "passer à l'intérieur" des points. N'ayant plus que l'information du nuage de points et souhaitant retrouver l'information de la droite théorique associée au modèle, il semble que toute droite "passant à l'intérieur" du nuage de points puisse être un candidat potentiel pour la droite ajustée. Les points observés sont obtenus par des "perturbations aléatoires verticales" (cf figure Fig.2) autour de la droite théorique. Il semble alors que la droite la plus proche verticalement de tous les points soit la solution la plus satisfaisante. C'est précisément ce que réalise la méthode des moindres carrés ordinaires (M.C.O.) en cherchant les coefficients d'une droite qui minimisent la somme des distances verticales (en ordonnée) au carré entre les points et cette droite (cf figure Fig.6). Mathématiquement, ceci exprime que la droite ajustée sera déterminée en minimisant en β_0 et β_1 la quantité $\boxed{\min_{\beta_0,\beta_1} \sum_{t=1}^{10} (y_t - (\beta_0 + \beta_1 x_t))^2}$

Précisons tout d'abord un abus de notation. Le vecteur $\boldsymbol{\beta}$ correspond aux vraies valeurs (dites théoriques) des paramètres du modèle. Par la suite, cette même notation spécifiera aussi le vecteur des paramètres utilisé comme une variable. En utilisant cette convention, l'estimateur des moindres carrés $\widehat{\boldsymbol{\beta}}$ est la solution en $\boldsymbol{\beta}$ qui **minimise** le critère MCO (consistant en suivant l'exemple de cours à minimiser la somme des carrés des distances verticales) suivant :

$$\boxed{||\boldsymbol{Y} - \underline{\boldsymbol{x}}\boldsymbol{\beta}||^2 = \sum_{t=1}^{n} \left(Y_t - \left(\beta_0 + \beta_1 x_t^{(1)} + \beta_2 x_t^{(2)} + \cdots + \beta_p x_t^{(p)}\right)\right)^2.}$$

La solution à un tel problème peut être obtenue de manière géométrique.

Nous montrons dans le paragraphe E.1 que cet estimateur s'écrit explicitement :

$$\boxed{\widehat{\boldsymbol{\beta}} \overset{\text{Not.}}{=} \widehat{\boldsymbol{\beta}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{Y} \iff ||\boldsymbol{Y} - \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}||^2 = \min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \underline{\boldsymbol{x}}\boldsymbol{\beta}||^2}$$

Cas particulier d'un seul régresseur ($p = 1$) : le vecteur des estimateurs prend la forme explicite suivante

$$\widehat{\beta}_1 \overset{\text{Not.}}{=} \widehat{\beta}_1(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \frac{cov(\boldsymbol{Y}, \boldsymbol{x}^{(1)})}{var(\boldsymbol{x}^{(1)})} \text{ et } \widehat{\beta}_0 \overset{\text{Not.}}{=} \widehat{\beta}_0(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \overline{Y} - \widehat{\beta}_1 \overline{\boldsymbol{x}^{(1)}},$$

$$\textit{avec,} \qquad cov(\boldsymbol{Y}, \boldsymbol{x}^{(1)}) = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \overline{Y}\right)\left(x_t^{(1)} - \overline{\boldsymbol{x}^{(1)}}\right) = \text{ et } var(\boldsymbol{x}^{(1)}) = \frac{1}{n}\sum_{t=1}^{n}\left(x_t^{(1)} - \overline{\boldsymbol{x}^{(1)}}\right)^2.$$

Nour renvoyons le lecteur à l'annexe C pour comprendre comment le logiciel `R` intègre le calcul matriciel, ce qui nous permet de calculer les estimations des paramètres pour les exemples des trois étudiants :

Exemple 1 : Avec le modèle envisagé par le premier étudiant et son jeu de données associé (de taille $n = 5$), nous obtenons $\widehat{\beta}_0(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 1599.84$ et $\widehat{\beta}_1(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 279.618$:

```
################################################################################
#### REMARQUE TECHNIQUE : Pour rendre accessibles des variables d'un jeu de donnees #
#### il faut executer une et une seule fois la fonction ci-dessous                #
################################################################################
> attach(ex1SalData)  ## Rend accessibles les variables Sal et IndExp de ex1SalData
> x<-cbind(1,IndExp)
> solve(t(x)%*%x) %*% t(x) %*% Sal-> betaChapo
> betaChapo
             [,1]
        1599.840
  IndExp  279.618
> cov(Sal,IndExp)/var(IndExp)->beta1Chapo ## appliquons la formule ds le cas ou p=1
> beta1Chapo
[1] 279.618
> mean(Sal)-beta1Chapo *mean(IndExp)      ## c'est la meme chose!
[1] 1599.84
```

Exemple 2 : Les estimations sont dans cette situation (même modèle, taille d'échantillon $n = 200$) $\widehat{\beta}_0(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 1047.832$ et $\widehat{\beta}_1(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 1487.894$:

```
> attach(exSalData)  ## Rend accessibles les variables Sal et IndExp de exSalData
> x<-cbind(1,IndExp)
> solve(t(x)%*%x) %*% t(x) %*% Sal-> betaChapo
> betaChapo
             [,1]
        1047.832
  IndExp 1487.894
```

Exemple 3 : Les estimations sont dans cette situation $\widehat{\beta}_0(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 928.9317$, $\widehat{\beta}_1(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 1489.2292$ et $\widehat{\beta}_2(\boldsymbol{Sal}|\underline{\boldsymbol{x}}) \simeq 238.9072$:

```
> attach(exSalData)  ## Rend accessibles les variables Sal, IndExp et IndEtu de exSalData
> x<-cbind(1,IndExp,IndEtu)
> solve(t(x)%*%x) %*% t(x) %*% Sal -> betaChapo
> betaChapo
             [,1]
         928.9317
IndExp 1489.2292
IndEtu  238.9072
```

3.2 Vecteur des valeurs ajustées et vecteur des résidus

Définissons les deux vecteurs aléatoires suivants exprimant le vecteur des prévisions

$$\boxed{\widehat{\boldsymbol{Y}} = \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}} \Leftrightarrow \widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 x_t^{(1)} + \widehat{\beta}_2 x_t^{(2)} + \ldots + \widehat{\beta}_p x_t^{(p)} \ , \ t = 1, 2, \cdots, n,}$$

et celui des résidus utilisés plus tard

$$\boxed{\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}}$$

Ces deux définitions peuvent être facilement illustrées dans le cadre des différents exemples :

Exemple 1 :

```
> SalChapo<-x%*% betaChapo
> SalChapo
[1] 1684.065 1784.599 1735.821 1686.239 1817.726
> epsChapo<-Sal-SalChapo
> epsChapo
[1] -141.27670  -79.75178  315.05207  -32.54492  -61.47868
```

Exemple 2 et 3 :

```
> SalChapo <- x%*% betaChapo
> epsChapo <- Sal-SalChapo
```

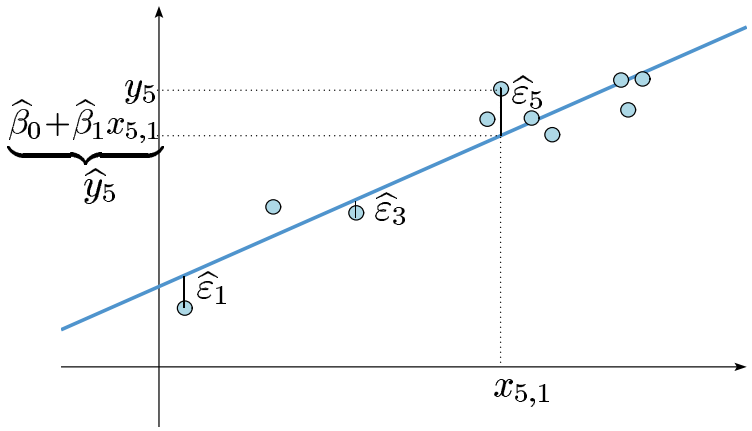


FIGURE 5 – Jeu de données et droite ajustée pour l'exemple de cours

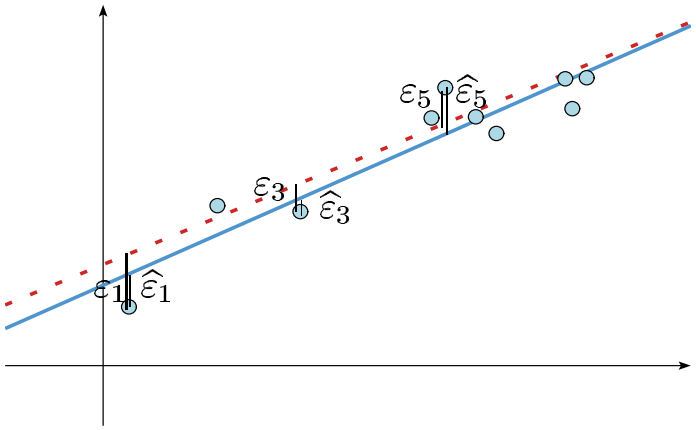


FIGURE 6 – Modèle, jeu de données et droite ajustée pour l'exemple de cours

3.3 Appréhension de la variabilité des estimations via l'A.E.P.

A la lumière de la figure 5, il est indéniable que les estimations varient en fonction des données observées. Cette figure nous propose en effet dix jeux de données possibles différents (tous générés à partir du modèle de l'exemple de cours) conduisant à autant de droites ajustées différentes. La clé de voûte des méthodes statistiques est précisément la connaissance de cette variabilité. Encore une nouvelle fois, nous nous tournons vers l'**A.E.P.**. Poursuivant le travail commencé à la section 2.4, nous imaginons disposer d'une infinité d'estimations $\widehat{\boldsymbol{\beta}}\left(\boldsymbol{y}_{[1]}\right), \ldots, \widehat{\boldsymbol{\beta}}\left(\boldsymbol{y}_{[m]}\right), \ldots$. Il ne nous reste plus qu'à essayer de caractériser leurs répartitions. Contentons-nous pour l'instant d'étudier l'exemple de cours

via une approche exclusivement assistée par ordinateur.

Exemple de cours : Comme nous ne sommes pas encore en mesure de traiter complètement cette caractérisation, nous l'abordons en combinant l'**A.E.P.** avec l'utilisation de l'ordinateur. Après un très grand nombre $m \simeq 10000$ de jeux de données générés par le modèle, nous sommes en mesure d'établir un premier histogramme en 8 classes, puis en 32 classes et enfin en 128 classes des m estimations associées (cf. figure 7). D'après l'annexe A, nous savons que l'histogramme à pas zéro d'une infinité de réalisations d'une variable aléatoire continue la caractérise. Bien que nous ne soyons pas capable avec un ordinateur d'atteindre ce cas limite, il n'est pas difficile de s'en convaincre à la vue des histogrammes de la figure 7.

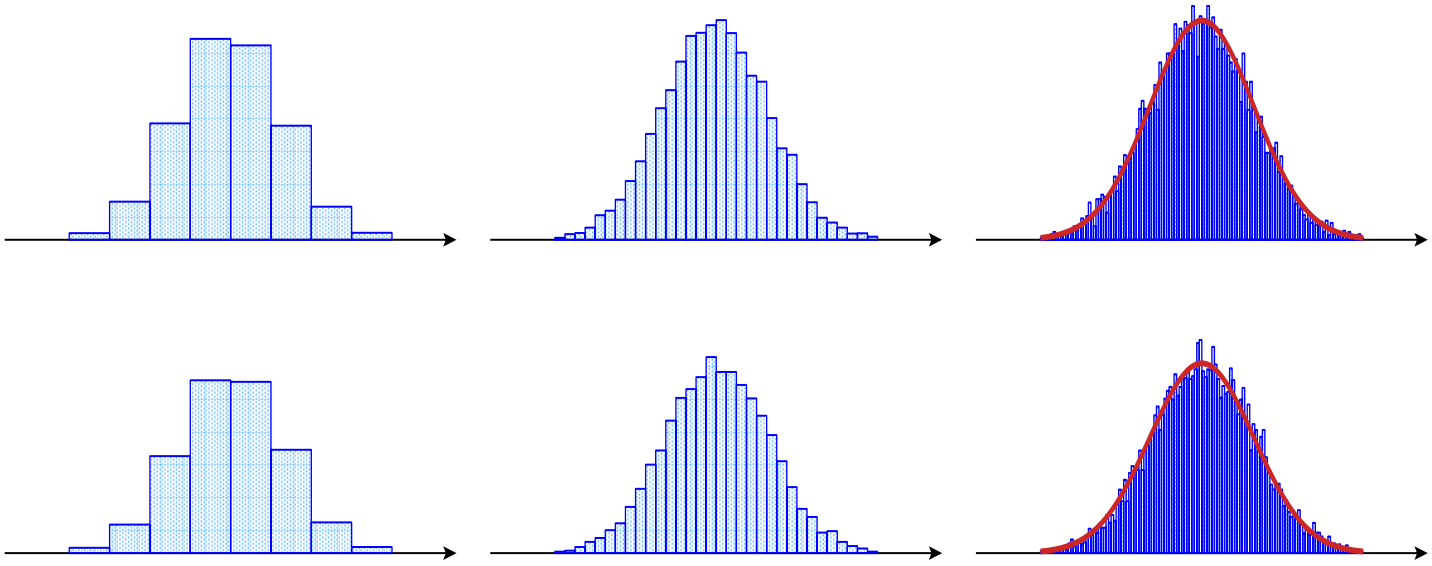


FIGURE 7 – Histogrammes d'un grand nombre $m \simeq 10000$ d'estimations de β_0 (en haut) et β_1 (en bas)

L'Approche Classique des Probabilités (**A.C.P.**) à vocation plus calculatoire nous confirme qu'avec les hypothèses classiques les estimateurs M.C.O. ont un comportement aléatoire complètement caractérisable. Les courbes lisses aposées dans les graphiques à droite représentent cette connaissance. Ces courbes n'ont pu être représentées par l'ordinateur que parce qu'il connaissait d'une part les vraies valeurs de β_0 et β_1 et d'autre part la valeur du paramètre de nuisance σ^2. Dans l'annexe B, ce type de paramètre dans le contexte d'estimation de β_0 ou β_1 est appelé paramètre parasite.

3.4 Estimateur du maximum de vraisemblance

Si l'on suppose **(C2-2)**, il est immédiat que $\boldsymbol{Y}$ est un vecteur gaussien comme transformation affine de $\boldsymbol{U}$ qui est lui-même gaussien. D'après ce qui précède, on peut exprimer la densité de probabilité de

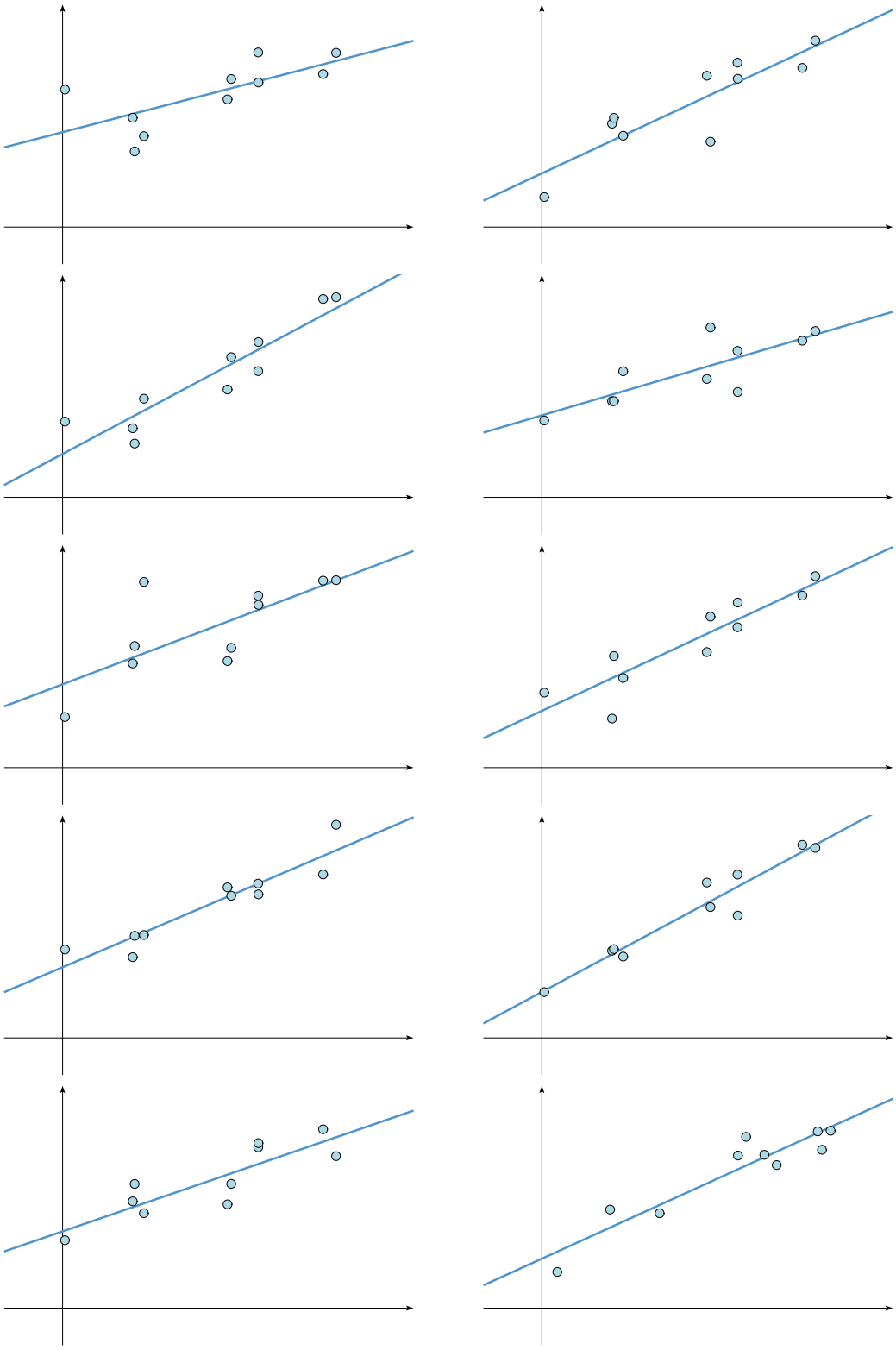


FIGURE 8 – Mise en évidence de la variabilité des estimations.

$\boldsymbol{Y}$, appelée aussi fonction de vraisemblance, par

$$f_{\boldsymbol{Y}}(\underbrace{y_1, y_2, \cdots, y_n}_{\boldsymbol{y}}) = \frac{1}{\sigma\,(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \underline{\boldsymbol{x}}\boldsymbol{\beta})^T\,(\boldsymbol{y} - \underline{\boldsymbol{x}}\boldsymbol{\beta})\right).$$

On peut montrer que l'estimateur qui maximise cette fonction de vraisemblance, appelé communément estimateur du maximum de vraisemblance, n'est rien d'autre que l'estimateur obtenu par la méthode des MCO, $\widehat{\boldsymbol{\beta}}$. C'est un fait remarquable car en statistiques (lorsque tout se passe bien), l'estimateur du maximum de vraisemblance est celui qui possède les "meilleures" performances, d'où l'intérêt à nouveau de s'intéresser à la méthode des M.C.O.

4 Propriétés de l'estimateur M.C.O.

Essayons de motiver cette section via l'exemple de cours. Nous avons pris conscience de la nature aléatoire du modèle et ainsi des données. Les estimations dépendant des données, elles seront elles-aussi intrinsèquement aléatoires. Nous allons mesurer la qualité des estimations M.C.O. en mettant l'accent sur les interprétations proposées par l'**A.E.P.**. La figure 4 nous permet en un seul coup d'oeil d'apprécier la qualité de la méthode d'estimation par les moindres carrés puisqu'il nous est possible de comparer les droites ajustées (ou estimées) avec la droite théorique caractérisant la composante déterministe du modèle.

Tous les résultats annoncés ici sont démontrés mathématiquement dans la section E.2. Notons *anecdotiquement* que l'hypothèse **(C2-2)** (hypothése gaussienne du bruit additif) n'a pas été utilisée pour mettre en avant l'ensemble de ces propriétés, et donc qu'elles restent valables dans un cadre beaucoup plus large.

4.1 Estimateur sans biais

Cette propriété exprime que la moyenne des différentes estimations M.C.O. possibles du vecteur des paramètres de régression correspond au vecteur lui-même.

$$\boxed{\boldsymbol{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}}$$

Exemple de cours : en suivant l'**A.E.P.** (cf. annexe A) cette propriété exprime simplement le fait que la moyenne d'une infinité ($m = +\infty$) d'estimations virtuelles du paramètre β_0 (resp. β_1) vaut exactement β_0 (resp. β_1). Essayons d'appréhender cette propriété en simulant un certain nombre de fois (m fois) le modèle et en évaluant pour différentes valeurs de m la moyenne $\frac{1}{m}\sum_{i=1}^{m} \widehat{\beta}_i\left(\boldsymbol{y}_{[i]}|\underline{\boldsymbol{x}}\right)$ des m estimations de β_i (pour $i = 0, 1$). Lorsque $m = +\infty$, l'**A.E.P.** précise que nous devrions retrouver les paramètres β_0 et β_1 .

Moyenne des	m=10	m=100	m=1000	m=10000	m=+∞
$\left(\widehat{\beta}_0\left(\boldsymbol{y}_{[i]}|\underline{\boldsymbol{x}}\right)\right)_{i=1,\ldots,m}$	0.9347449	0.9690022	0.9942976	0.9970814	1
$\left(\widehat{\beta}_1\left(\boldsymbol{y}_{[i]}|\underline{\boldsymbol{x}}\right)\right)_{i=1,\ldots,m}$	2.127965	2.027007	2.003907	2.003768	2

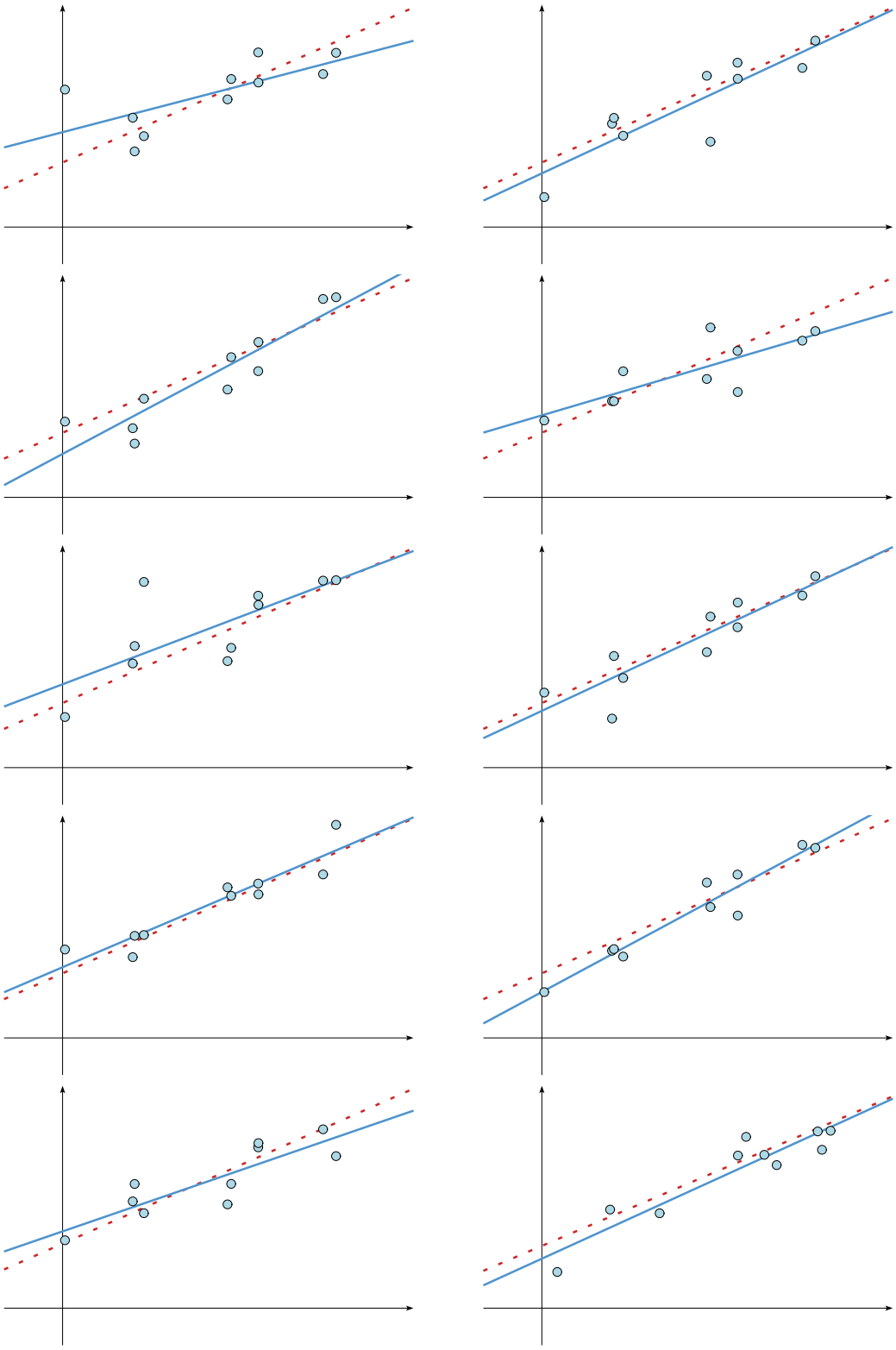


FIGURE 9 – Les différentes réalisations du modèle issu de l'exemple du cours, avec les droites "estimées".

4.2 Estimateur convergent en moyenne quadratique

Puisque l'estimateur $\widehat{\boldsymbol{\beta}}$ est sans biais, ceci se traduit plus simplement en disant que la variance des estimateurs tend vers 0 lorsque $n \to +\infty$. Cela provient d'une part du fait que la matrice de covariances $\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right)$ de $\widehat{\boldsymbol{\beta}}$ s'exprime par

$$\boxed{\underline{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}} \overset{\text{Not.}}{=} \underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) \overset{\text{Déf.}}{=} \boldsymbol{E}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\right) = \sigma^2\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}}$$

et d'autre part de l'hypothèse **(C3-2)**. Le $i^{ème}$ élément diagonal de la matrice $\underline{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}$ correspond à la variance du $i^{ème}$ estimateur $\widehat{\beta}_i$ et sera noté $\boxed{\sigma^2_{\widehat{\beta}_i}}$.

Exemple de cours : commençons par vérifier la dernière formule énoncée (explicitant la variance des estimateurs de β_0 et β_1 avec le nombre d'observations fixé à $n = 10$) : en suivant l'**A.E.P.**, on sait que la variance d'une infinité ($m = +\infty$) d'estimations virtuelles du paramètre β_0 (resp. β_1) vaut exactement $\sigma_{\widehat{\beta}_0}$ (resp. $\sigma_{\widehat{\beta}_1}$). Essayons d'appréhender cette propriété en simulant un certain nombre de fois (m fois) le modèle et en évaluant pour différentes valeurs de m la variance $\frac{1}{m}\sum_{i=1}^{m}\left(\widehat{\beta}_i\left(\boldsymbol{y}_{[i]}|\underline{\boldsymbol{x}}\right) - \frac{1}{m}\sum_{i=1}^{m}\widehat{\beta}_i\left(\boldsymbol{y}_{[i]}|\underline{\boldsymbol{x}}\right)\right)^2$ des m estimations de β_i (pour $i = 0, 1$)

Variance des	m=10	m=100	m=1000	m=10000	m=$+\infty$
$\left(\widehat{\beta}_0\left(\boldsymbol{y}_{[i]}\|\underline{\boldsymbol{x}}\right)\right)_{i=1,\ldots,m}$	0.04805025	0.03294972	0.02623897	0.02523644	$\sigma^2_{\widehat{\beta}_0}$ =0.02608868
$\left(\widehat{\beta}_1\left(\boldsymbol{y}_{[i]}\|\underline{\boldsymbol{x}}\right)\right)_{i=1,\ldots,m}$	0.1621368	0.1483871	0.1336362	0.1383188	$\sigma^2_{\widehat{\beta}_1}$ =0.1420237

En conséquence, on sait désormais que pour connaître le niveau de précision (ou variance) d'un estimateur il faut s'appuyer sur une infinité d'estimations virtuelles des paramètres. On va illustrer la convergence en moyenne quadratique uniquement pour l'estimateur $\widehat{\beta}_1$. Pour cela, nous allons évaluer la variance de m estimations ($m = 10000, +\infty$) virtuelles de β_1 pour différentes valeurs de n :

Variance des $\left(\widehat{\beta}_1\left(\boldsymbol{y}_{[i]}\|\underline{\boldsymbol{x}}\right)\right)_{i=1,\ldots,m}$ pour différentes valeurs de n	$n = 10$	$n = 100$	$n = 1000$	$n = +\infty$
$m = 10$	0.1621368	0.01717733	0.001125605	0
$m = 10000$	0.1383188	0.01219024	0.001112239	0
$m = +\infty$	$\sigma^2_{\widehat{\beta}_1}$ =0.1420237	$\sigma^2_{\widehat{\beta}_1}$ =0.01207125	$\sigma^2_{\widehat{\beta}_1}$ =0.001129981	$\sigma^2_{\widehat{\beta}_1}$ =0

En bref, les estimateurs issus de la méthode M.C.O. réalisent ce que l'on attend d'eux : en moyenne on retrouve les vraies valeurs des paramètres (propriéteé sans biais) et plus le nombre d'observations n augmente et plus les estimateurs sont précis (propriété convergent en moyenne quadratique). Par ailleurs, à n fixé on connaît de manière exacte le niveau de précision des estimateurs si on connaît le niveau de bruit σ.

Notons tout de même qu'en pratique, on ne peut connaître de manière exacte la variance des estimateurs car celle-ci dépend de σ inconnu.

4.3 Estimateur efficace

C'est le meilleur estimateur linéaire sans biais (BLUE en anglais). Cette propriété plus technique à appréhender ne fera pas l'objet d'une vérification expérimentale.

4.4 Relation entre les paramètres du modèle et les données

Cette partie ne décrit pas à proprement parler une propriété relative aux estimateurs M.C.O. des paramètres de régression. C'est plutôt une conséquence de leurs bonnes propriétés. En fait, nous sommes à présent capables d'établir la relation attendue (exprimée à la section 2.4) entre les paramètres de régression $\boldsymbol{\beta}$ et une infinité de données générées par le modèle. Compte tenu des propriétés des estimateurs M.C.O. d'être sans biais et de variance s'annulant lorsque la taille d'échantillon tend vers $+\infty$, nous pouvons avancer que l'hyperplan (i.e. la droite lorsque $p = 1$) le plus proche "verticalement" d'une infinité de points (virtuels) représentant une infinité de données (virtuelles) $y_{[1]}, \ldots, y_{[m]}, \ldots$ générées par le modèle à partir des vecteurs respectifs $\boldsymbol{x_{[1]}}, \ldots, \boldsymbol{x_{[m]}}, \ldots$ de valeurs de p régresseurs coïncident avec l'hyperplan caractérisant la composante déterministe (d'équation $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$). Tout ceci est facilement vérifiable avec l'exemple du cours expérimentalement en augmentant au fur et à mesure le nombre m de points puis en déterminant à chaque étape les estimations M.C.O. pour enfin constater que celles-ci convergent vers les vrais paramètres $\boldsymbol{\beta}$ de régression. Le dernier tableau ci-dessus conforte exactement ces propos. Si nous poursuivons ce raisonnement, cela veut même dire que les $U_{[1]}, \ldots, U_{[m]}, \ldots$ a priori inconnus (surtout lorsque nous en disposons que d'un nombre fini) coïncident avec les $\widehat{U}_{[1]}, \ldots, \widehat{U}_{[m]}, \ldots$ (conséquence directe de la coïncidence entre les deux hyperplans théoriques et ajustées) . Dès lors, nous pouvons achever cette section en établissant la relation entre le paramètre de nuisance σ^2 et une infinité de données virtuelles. En effet, puisque les $\widehat{U}_{[1]}, \ldots, \widehat{U}_{[m]}, \ldots$ se déduisent complètement de $(y_{[1]}, \boldsymbol{x_{[1]}}), \ldots, (y_{[m]}\boldsymbol{x_{[m]}},), \ldots$, nous pouvons établir que : $\lim\limits_{m \to +\infty} \dfrac{1}{m} \sum\limits_{i=1}^{m} \widehat{U}_{[i]}^2 = \sigma^2$

5 Estimation des variances du bruit et des estimateurs M.C.O.

L'objectif de cette section répond entre autres au probléme soulevé précédemment à savoir : puisque l'on ne peut connaître de manière exacte le niveau de précision des estimateurs, peut-on en avoir une idée (donc une estimation) ? Le but est également de construire des outils usuels de statistique inférentielle tels que régions de confiance et tests d'hypothèses. De manière analogue au cas d'estimation de la moyenne, il est nécessaire de connaître la variabilité des estimateurs exprimés ici par leur matrice de covariance. On vient de voir que

$$\underline{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}} = \sigma^2 \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}} \right)^{-1},$$

i.e. dépend explicitement et uniquement du paramètre de nuisance σ^2 que l'on se propose donc d'estimer.

Comment estimer ce paramètre ? Il suffit de regarder plus attentivement la figure Fig.6 : le vecteur des $\widehat{U}_t$ semble être une bonne approximation des U_t non observables. Maintenant, puisque d'après

la section 2.5, σ^2 n'est rien d'autre que la variance d'une infinité de U_t, on substitue simplement l'infinité des U_t non disponibles par les $\widehat{U}_t$ pour $t = 1, \ldots, n$. On peut par conséquent proposer comme estimateur sans biais de σ^2 (voir section E.3) :

$$\boxed{\widehat{\sigma}^2 \overset{\text{Not.}}{=} \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{1}{n-p-1}\sum_{t=1}^{n}\widehat{U}_t^2.}$$

De cet estimateur on déduit un estimateur de la matrice $\underline{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}$ de covariances des $\widehat{\boldsymbol{\beta}}$

$$\boxed{\widehat{\underline{\boldsymbol{\Sigma}}}_{\widehat{\boldsymbol{\beta}}} \overset{\text{Not.}}{=} \widehat{\underline{\boldsymbol{\Sigma}}}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}})\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}.}$$

Le $i^{ème}$ élément diagonal de la matrice $\widehat{\underline{\boldsymbol{\Sigma}}}_{\widehat{\boldsymbol{\beta}}}$ correspond à la variance estimée du $i^{ème}$ estimateur $\widehat{\beta}_i$ et sera noté $\boxed{\widehat{\sigma}^2_{\widehat{\beta}_i}}$ (et voilà défini l'estimateur du niveau de précision des estimateurs mis en jeu).

Reprenons les trois exemples pour évaluer dans chacune des situations les niveaux de précision estimés de chacun des estimateurs :

Exemple 1 :

```
> sigma2Chapo<-sum(epsChapo ^2)/(5-1-1)
> sigma2Chapo
[1] 43472.02
> sigma2Chapo * diag(solve(t(x)%*%x))
[1]  70940.91 241872.24
> sqrt(sigma2Chapo * diag(solve(t(x)%*%x)))
[1] 266.3474 491.8051
```

Exemple 2 :

```
> sigma2Chapo<-sum(epsChapo ^2)/(200-1-1)
> sigma2Chapo
[1] 100955.5
> sigma2Chapo * diag(solve(t(x) %*% x))
[1] 1923.894 5750.176
> sqrt(sigma2Chapo * diag(solve(t(x) %*%x )))
[1] 43.86221 75.82992
```

Exemple 3 :

```
> sigma2Chapo<-sum(epsChapo ^2)/(200-2-1)
> sigma2Chapo
[1] 95990.36
> sigma2Chapo * diag(solve(t(x) %*% x))
[1] 3086.839 5467.530 5077.218
> sqrt(sigma2Chapo * diag(solve(t(x) %*% x)))
[1] 55.55933 73.94275 71.25460
```

6 Analyse de la variance et coefficient de détermination multiple

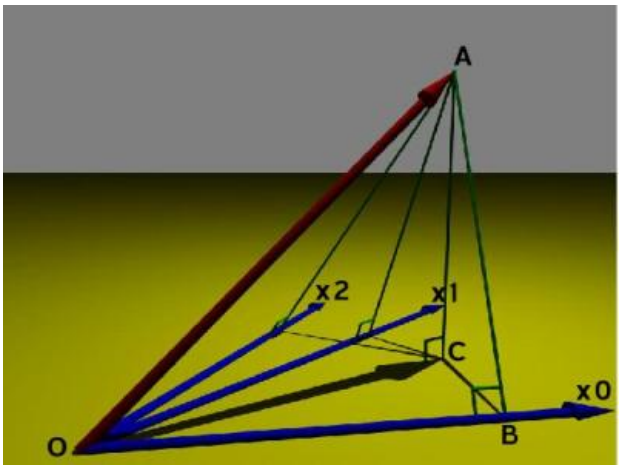


FIGURE 10 – Espace de représentation des variables : il y a correspondance entre les vecteurs $\vec{OA}$ et $\boldsymbol{Y}$, $\vec{OB}$ et $\overline{\boldsymbol{Y}}$, $\vec{OC}$ et $\widehat{\boldsymbol{Y}}$. De plus, comme nous nous sommes placés dans le cadre d'une régression multiple à $p = 2$ régresseurs, les vecteurs $\vec{OX}_0$, $\vec{OX}_1$ et $\vec{OX}_2$ correspondent respectivement aux régresseurs $\boldsymbol{x}^{(0)} = \mathbf{1}$ (constant), $\boldsymbol{x}^{(1)}$ et $\boldsymbol{x}^{(2)}$. Notons que tout vecteur $\vec{OZ}$ (non représenté dans la figure) vivant dans le même plan que les vecteurs $\vec{OX}_0$, $\vec{OX}_1$ et $\vec{OX}_2$ s'écrit comme une combinaison linéaire de ces vecteurs (correspondant alors à $\gamma_0\boldsymbol{x}^{(0)}+\gamma_1\boldsymbol{x}^{(1)}+\gamma_2\boldsymbol{x}^{(2)}$). Le vecteur $\vec{OB}$ étant le vecteur le plus proche dans ce plan du vecteur $\vec{OA}$, cette figure nous permet donc de comprendre la méthode géométrique conduisant à l'obtention des estimations des paramètres de régression qui ne sont que les coefficients de la combinaison linéaire associée à ce vecteur. En fait, pour faire cette figure avec un logiciel 3D, il suffit de mettre une lumière à l'extrémité du vecteur $\vec{OA}$ dont son ombre sur le plan représente le vecteur $\vec{OB}$.

Il découle d'une propriété géométrique de l'estimateur MCO l'équation suivante (appelée *équation de l'analyse de la variance*) :

$$\boxed{\|\boldsymbol{Y}-\overline{\boldsymbol{Y}}\|^2 = \underbrace{\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\|^2}_{\|\widehat{\boldsymbol{U}}\|^2} + \underbrace{\|\widehat{\boldsymbol{Y}}-\overline{\boldsymbol{Y}}\|^2}_{\|\widehat{\boldsymbol{Y}}-\overline{\widehat{\boldsymbol{Y}}}\|^2}}$$

Ceci s'exprime littéralement en divisant tous les termes par n : "variance totale est égale à la somme des variance résiduelle et variance expliquée". Les données seront d'autant mieux ajustées que la norme $\|\widehat{\boldsymbol{U}}\|^2$ mesurant la dispersion des erreurs est faible. Ceci permet alors d'introduire un indicateur de la qualité d'ajustement appelé *coefficient de détermination multiple* :

$$\boxed{R^2 = cos^2(\widehat{ABC}) = \frac{\left\|\widehat{\boldsymbol{Y}}-\overline{\widehat{\boldsymbol{Y}}}\right\|^2}{\left\|\boldsymbol{Y}-\overline{\boldsymbol{Y}}\right\|^2} = 1 - \frac{\left\|\widehat{\boldsymbol{U}}\right\|^2}{\left\|\boldsymbol{Y}-\overline{\boldsymbol{Y}}\right\|^2} = corr^2(\boldsymbol{Y},\widehat{\boldsymbol{Y}})}$$

Notons que ce coefficient est la généralisation du coefficient de corrélation linéaire (au carré) bien connu dans le cadre de la régression simple. Dans ce cas particulier, $R = \dfrac{cov(\boldsymbol{Y}, \boldsymbol{x}_{(1)})}{\sqrt{var(\boldsymbol{x}_{(1)})var(\boldsymbol{Y})}}$.

Ainsi, l'ajustement des données sera d'autant meilleur que R^2 est proche de 1. Il est aisé de constater que R est le cosinus de l'angle $\alpha = \widehat{ABC}$ avec A, B et C les extrémités des vecteurs $\boldsymbol{Y}$, $\overline{\boldsymbol{Y}}$ et $\widehat{\boldsymbol{Y}}$ (lorsque représenté avec l'origine comme point de départ) comme dans la figure 6. Le défaut de ce coefficient est qu'il ne tient ni compte du nombre de données ni du nombre de régresseurs. Une version ajustée du R^2 a alors été introduite. Noté, R^2_a, ce coefficient est défini par

$$\boxed{R^2_a = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right).}$$

La notation pourrait laisser croire que ce coefficient est positif mais il n'en est rien.

Exemple 1 :

```
## Analyse de la variance
> var(Sal)
[1] 36117.14    # variance totale
> var(SalChapo)
[1] 3513.128    # variance expliquee par le modele
> var(SalChapo)+var(epsChapo)
[1] 36117.14    # variance expliquee par le modele + variance residuelle
## Coeff de determination R2
> var(SalChapo)/var(Sal)
[1] 0.09727037
> 1-var(epsChapo)/var(Sal)
[1] 0.09727037
## Coeff de determination R2 ajuste
> 1-(5-1)/(5-2)*(1-var(SalChapo)/var(Sal))
[1] -0.2036395
```

Exemple 2 :

```
## Analyse de la variance
> var(Sal)
[1] 295765.2    # variance totale
> var(SalChapo)
[1] 195317.0    # variance expliquee par le modele
> var(SalChapo)+var(epsChapo)
[1] 295765.2    # variance expliquee par le modele + variance residuelle
## Coeff de determination R2
> var(SalChapo)/var(Sal)
[1] 0.6603785
> 1-var(epsChapo)/var(Sal)
[1] 0.6603785
```

```
## Coeff de determination R2 ajuste
> 1-(200-1)/(200-2)*(1-var(SalChapo)/var(Sal))
[1] 0.6586632
```

Exemple 3 :

```
## Analyse de la variance
> var(Sal)
[1] 295765.2     # variance totale
> var(SalChapo)
[1] 200739.6     # variance expliquee par le modele
> var(SalChapo)+var(epsChapo)
[1] 295765.2     # variance expliquee par le modele + variance residuelle
## Coeff de determination R2
> var(SalChapo)/var(Sal)
[1] 0.6787126
> 1-var(epsChapo)/var(Sal)
[1] 0.6787126
## Coeff de determination R2 ajuste
> 1-(200-1)/(200-3)*(1-var(SalChapo)/var(Sal))
[1] 0.6754508
```

7 Loi empirique des estimateurs

Dans cette partie, nous allons établir les comportements aléatoires des différents estimateurs introduits. Ce type de résultat est la base des outils inférentiels tels les intervalles de confiance, tests d'hypothèses et intervalles de prévisions.

Avant de présenter les différents résultats, il est important de noter qu'il existe deux cadres d'étude distincts : un cadre gaussien qui s'appuie sur l'hypothèse **(C2-2)** et un cadre dit asymptotique ne faisant pas appel à d'hypothèse sur la distribution du bruit. Le cadre gaussien (section 7.1) n'est a priori intéressant que s'il est possible (éventuellement à partir d'un outil statistique) de vérifier que le bruit U est distribué selon une Normale. Alors qu'il faudrait disposer d'un grand échantillon pour cette vérification, l'usage dans la littérature statistique est d'utiliser ce cadre d'étude même pour de petits échantillons. Les résultats de cette partie seront illustrés via l'**A.E.P.** dans le contexte de l'exemple du cours (qui rappelons satisfait par définition l'hypothèse **(C2-2)**). Les mêmes notations introduites dans la section 2.4 seront utilisées.

Le cadre asymptotique (section 7.2) comme son nom l'indique s'applique lorsque la taille d'echantillon est suffisamment grande (scolairement $n - p - 1 > 30$) et permet de s'affranchir de l'hypothèse **(C2-2)**. C'est de loin le cadre de travail le plus naturel et donc le plus intéressant. Le seul inconvénient du cadre asymptotique est que les résultats ne sont qu'approximatifs alors qu'ils sont exacts si l'on se place dans le cadre gaussien.

La section 7.3 se propose de comparer les résultats obtenus dans les deux sections précédentes et constitue donc le résumé de toute cette section.

Par la suite, nous appellerons **mesure d'écart standardisée** une variable aléatoire calculée à partir du futur jeu de données $\boldsymbol{Y}$ et des régresseurs regroupés dans la matrice $\underline{\boldsymbol{x}}$ mesurant l'écart (exprimé comme une différence ou un rapport) entre la future estimation $\widehat{\theta}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ d'un paramètre d'intérêt θ et ce paramètre d'intérêt ; elle sera notée $\delta_\theta(\boldsymbol{Y}|\underline{\boldsymbol{x}})$. La mesure d'écart est "normalisée" de telle sorte que son comportement aléatoire ne dépende plus d'aucun paramètre parasite. Nous renvoyons le lecteur à l'Annexe B pour une définition plus détaillée de ce que nous appelons mesure d'écart standardisée. Par ailleurs, le lecteur plus curieux des mathématiques sous-jacentes aux résultats qui vont suivre pourra se référer à la section E.4 pour les preuves mathématiques (via l'**A.C.P.**).

7.1 Cadre gaussien

Paramètre de régression

La quantité suivante exprime une mesure d'écart standardisée entre une future estimation M.C.O. $\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ et le paramètre de régression β_i associé. Dans le contexte de l'hypothèse **(C2-2)**, nous pouvons montrer via l'**A.C.P.** que :

$$\boxed{\delta_{\beta_i}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \beta_i}{\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{Y}|\underline{\boldsymbol{x}})} \rightsquigarrow \mathcal{S}t\left(n-p-1\right).}$$

Exemple de cours : (rappelons que l'exemple de cours s'intègre dans cette section car l'ordinateur a généré des perturbations des points du modèle théorique selon une loi Normale). En combinant l'**A.E.P.** avec l'utilisation de l'ordinateur, nous pouvons nous convaincre de la validité du résultat (cf. figure 11). Grâce à celui-ci et notamment à sa vitesse de calcul, nous pouvons obtenir $m \simeq 10000$ réalisations de cette mesure d'écart standardisée $\delta_{\beta_i}\left(\boldsymbol{y}_{[1]}|\underline{\boldsymbol{x}}\right), \ldots, \delta_{\beta_i}\left(\boldsymbol{y}_{[m]}|\underline{\boldsymbol{x}}\right)$ pour (i=0 et 1) et représenter leurs histogrammes en 8 classes, puis en 32 classes et enfin en 128 classes. Ce dernier histogramme ressemble à une courbe lisse qui correspond en fait au cas limite où m tend vers l'infini et le pas de l'histogramme devient nul (concrètement, il faudrait imaginer l'ordinateur ne jamais s'arrêter de générer des données et interactivement au fur à mesure diminuer de plus en plus la taille des classes). Cette courbe lisse n'est ni plus ni moins que le contour supérieur d'un "tas" d'une infinité $\delta_{\beta_i}\left(\boldsymbol{y}_{[1]}|\underline{\boldsymbol{x}}\right), \ldots, \delta_{\beta_i}\left(\boldsymbol{y}_{[m]}|\underline{\boldsymbol{x}}\right), \ldots$ de mesures d'écart standardisées (virtuelles) identifié par l'**A.C.P.** comme celui d'une loi $\mathcal{S}t\left(10-1-1\right)$.

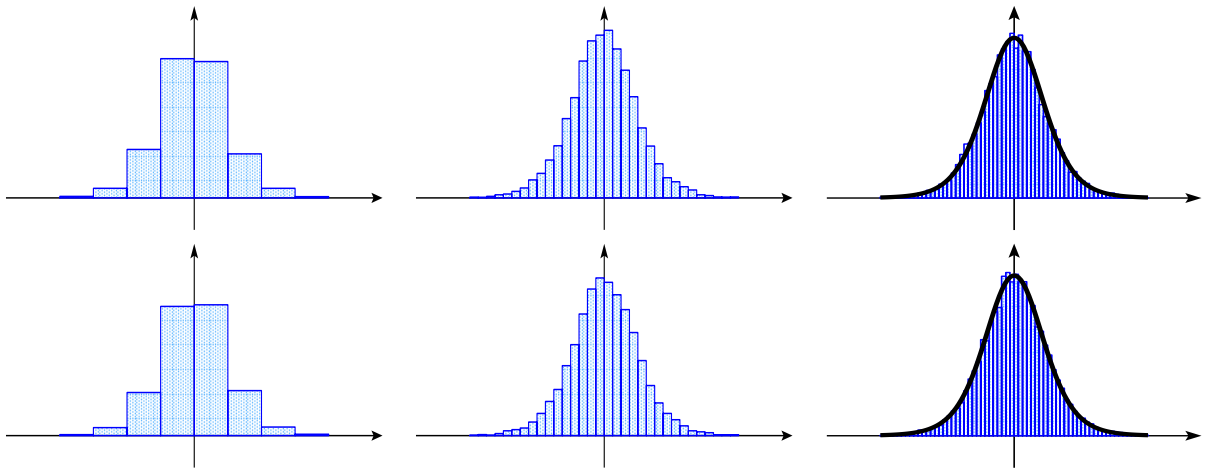


FIGURE 11 – Histogrammes d'un grand nombre $m \simeq 10000$ de mesures d'écart standardisées δ_{β_0} (en haut) et δ_{β_1} (en bas). Les courbes lisses dans les graphiques de droite représentent les cas limites lorsque le nombre m et le pas de classe de l'histogramme tendent respectivement vers l'infini et zéro.

Vecteur des paramètres de régression

Soit $Q \subset \{0, 1, \ldots, p\}$ un ensemble d'indices à q éléments, soit $\boldsymbol{\beta}_Q = (\beta_i)_{i \in Q}$ et $\widehat{\boldsymbol{\beta}}_Q = \left(\widehat{\beta}_i\right)_{i \in Q}$ les vecteurs de q paramètres de régression et de leurs estimateurs. On montre alors que :

$$\boxed{\delta_{\boldsymbol{\beta}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{1}{q}\left(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\right)^T \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\boldsymbol{\beta}}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \left(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\right) \rightsquigarrow \mathcal{F}(q, n-p-1)}$$

Ce type de résultat est le moteur du test de significativité globale.

Paramètre de nuisance

Dans ce document, la convention est de noter par la lettre "δ" toute variable aléatoire dépendant d'un futur jeu de données qui exprime une mesure d'écart entre l'estimateur et son paramètre théorique. Il en va de même ici malgré les apparences. Un écart peut aussi se mesurer par un rapport de quantités comme c'est le cas ici. Dans le cadre où l'hypothèse **(C2-2)** est supposée, nous affirmons donc

$$\boxed{\delta_{\sigma^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} (n-p-1)\,\frac{\widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}})}{\sigma^2} \rightsquigarrow \chi^2(n-p-1)\,.}$$

permettant par exemple ainsi de mesurer les chances de réalisation de $\delta_{\sigma^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ dans un intervalle quelconque.

Exemple de cours : La démarche expérimentale est identique à celle employée pour les mesures d'écart standardisées relatives aux paramètres de régression. Les résultats ne sont pas les mêmes puisque la loi est maintenant celle d'une $\chi^2(n-p-1)$.

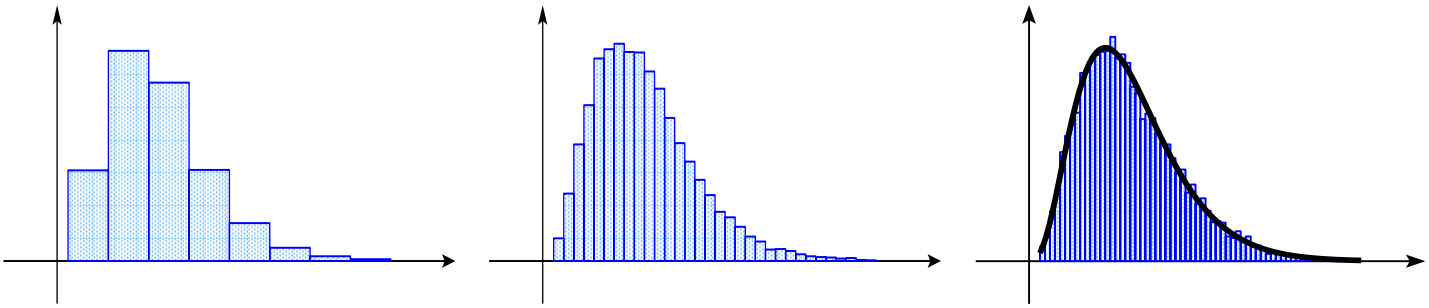


FIGURE 12 – Histogrammes d'un grand nombre $m \simeq 10000$ de mesures d'écart standardisées δ_{σ^2}. La courbe lisse dans le graphique de droite représente le cas limite lorsque le nombre m et le pas de classe de l'histogramme tendent respectivement vers l'infini et zéro.

7.2 Cadre Asymptotique

Les résultats s'appuient principalement sur l'application du théorème central limite (TCL) dont nous rappelons l'énoncé en Annexe D.6.

Paramètre de régression et paramètre de nuisance

Le cadre asymptotique a ceci de remarquable que les mesures d'écart standardisées se décrivent selon le même schéma dès que l'on s'intéresse au comportement aléatoire d'une future estimation d'un paramètre réel. Notons θ le paramètre d'intérêt (indifféremment β_i ou σ^2) et $\widehat{\theta}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ la future estimation de θ (ici $\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ ou $\widehat{\sigma^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$), la mesure d'écart standardisée s'écrit :

$$\boxed{\delta_\theta(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{\widehat{\theta}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \theta}{\widehat{\gamma}(\boldsymbol{Y}|\underline{\boldsymbol{x}})}}$$

La quantité $\widehat{\gamma}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ est une future estimation d'un paramètre γ exprimant le niveau de précision (en termes de variance) de la future estimation de θ par $\widehat{\theta}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$; elle dépend du paramètre d'intérêt et se décline comme suit :

Paramètre d'intérêt		
Paramètre d'intérêt θ	β_i	σ^2
Comportement asymptotique de la variance de la future estimation		
Quantité γ	$\sqrt{\sigma^2_{\widehat{\beta}_i}}$	$\sqrt{\dfrac{\sigma^2_{\boldsymbol{U^2}}}{n}}$
Future $\widehat{\gamma}(\boldsymbol{Y}\|\underline{\boldsymbol{x}})$	$\sqrt{\widehat{\sigma^2_{\widehat{\beta}_i}}(\boldsymbol{Y}\|\underline{\boldsymbol{x}})}$	$\sqrt{\dfrac{\widehat{\sigma^2_{\boldsymbol{U^2}}}(\boldsymbol{Y}\|\underline{\boldsymbol{x}})}{n}}$

plus précisément

$$\widehat{\sigma^2_{\boldsymbol{U^2}}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{1}{n-1}\sum_{t=1}^{n}\left((\widehat{\boldsymbol{\epsilon}}^{\boldsymbol{2}})_t - \overline{\widehat{\epsilon}^2}\right)^2,$$

où le vecteur $\boldsymbol{U^2}$ (resp. $\widehat{\boldsymbol{\epsilon}}^{\boldsymbol{2}}$) est le vecteur de composant $(U_t)^2$ (resp. $(\widehat{U}_t)^2$) pour $t = 1, \ldots, n$. Tout ceci étant défini, l'**A.C.P.** permet de montrer le très simple résultat suivant :

$$\boxed{\delta_\theta(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).}$$

Remarque : lorsque le paramètre d'intérêt est $\theta = \sigma^2$, le précédent résultat a besoin d'une petite hypothèse supplémentaire que nous noterons **(C2-2*)** :

(C2-2*) *Pour tout* $t = 1, \ldots, n, \boldsymbol{E}\left(U_t^4\right) < +\infty$, *i.e. que les* U_t *possèdent un moment d'ordre 4 fini.*

Cette hypothèse peut sembler plus technique mais croyez les auteurs elle n'est que très peu restrictive et vous n'auriez vraiment pas de chance si celle-ci n'était pas vérifiée.

Vecteur des paramètres de régression

En notant toujours $Q \subset \{0, 1, \ldots, p\}$ un ensemble d'indices à q éléments et $\boldsymbol{\beta}_Q = (\beta_i)_{i\in Q}$ et $\widehat{\boldsymbol{\beta}}_Q = \left(\widehat{\beta}_i\right)_{i\in Q}$ les vecteurs de q paramètres de régression et de leurs estimateurs. On montre (toujours via l'**A.C.P.**) que :

$$\boxed{\delta_{\boldsymbol{\beta}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \frac{1}{q}\left(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\right)^T \widehat{\underline{\boldsymbol{\Sigma}}}^{-1}_{\widehat{\boldsymbol{\beta}}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \left(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\right) \overset{approx.}{\rightsquigarrow} q \times \chi^2(q)}$$

7.3 Comparaison des cadres gaussien et asymptotique

Nous pouvons rassembler l'ensemble des résultats des deux sections précédentes dans le tableau suivant :

	Cadre gaussien	Cadre Asymptotique
Paramètre β_i		
Mesure d'écart standardisée	$\Delta_{\widehat{\beta}_i,\beta_i}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \dfrac{\widehat{\beta}_i(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) - \beta_i}{\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{Y}\|\underline{\boldsymbol{x}})}$	
Comportement aléatoire	$\Delta_{\widehat{\beta}_i,\beta_i}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \rightsquigarrow \mathcal{S}t\,(n-p-1)$	$\Delta_{\widehat{\beta}_i,\beta_i}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$
Paramètre β_Q		
Mesure d'écart standardisée	$\Delta_{\widehat{\boldsymbol{\beta}_Q},\boldsymbol{\beta}_Q}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \dfrac{1}{q}\Big(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\Big)^T \widehat{\underline{\boldsymbol{\Sigma}}}^{-1}_{\widehat{\boldsymbol{\beta}}_Q}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \Big(\widehat{\boldsymbol{\beta}}_Q(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) - \boldsymbol{\beta}_Q\Big)$	
Comportement aléatoire	$\Delta_{\widehat{\boldsymbol{\beta}_Q},\boldsymbol{\beta}_Q}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \rightsquigarrow \mathcal{F}\,(q, n-p-1)$	$\Delta_{\widehat{\boldsymbol{\beta}_Q},\boldsymbol{\beta}_Q}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{approx.}{\rightsquigarrow} q \times \chi^2(q)$
Paramètre σ^2		
Mesure d'écart standardisée	$\Delta_{\widehat{\sigma^2},\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} (n-p-1)\,\dfrac{\widehat{\sigma}^2(\boldsymbol{Y}\|\underline{\boldsymbol{x}})}{\sigma^2}$	$\Delta_{\widehat{\sigma^2},\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \dfrac{\widehat{\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) - \sigma^2}{\sqrt{\frac{\widehat{\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}})}{\underline{\boldsymbol{U}}^{\boldsymbol{2}}_n}}}$
Comportement aléatoire	$\Delta_{\widehat{\sigma^2},\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \rightsquigarrow \chi^2\,(n-p-1)$	$\Delta_{\widehat{\sigma^2},\sigma^2}(\boldsymbol{Y}\|\underline{\boldsymbol{x}}) \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1)$

Il semble intéressant de pousser la comparaison pour le cas des paramètres β_1 et $\boldsymbol{\beta}_Q$ car dans ces deux situations l'expression de la mesure d'écart standardisée est la même. Ceci n'est visiblement le cas pour le cas du paramètre σ^2 : la mesure d'écart standardisée s'exprime tantôt comme un rapport (cadre gaussien) tantôt comme une différence (cadre asymptotique) entre l'estimateur et le paramètre d'intérêt. Revenons aux paramètres β_1 et $\boldsymbol{\beta}_Q$, deux questions naturelles surviennent :

- Y a-t-il une grande différence entre une loi de $\mathcal{S}t(n-p-1)$ et une loi $\mathcal{N}(0,1)$?
- Y a-t-il une grande différence entre une loi de $\mathcal{F}(q, n-p-1)$ et une loi $p \times \chi^2(q)$?

Dans l'Annexe D.7, nous illustrons le résultat suivi énonçant le fait que lorsque $n \to +\infty$:

$$\mathcal{S}t(n-p-1) \to \mathcal{N}(0,1) \quad \text{et} \quad \mathcal{F}(q, n-p-1) \to p \times \chi^2(q)$$

Pour la petite histoire, les statisticiens dénomment ce genre de résultat, des résultats de **robustesse**. Essayons d'en comprendre le sens. La convergence en question est une convergence en loi, concept que nous ne développerons pas ici mais qui signifie entre autres que les densités de probabilité sont sensiblement égales et par conséquent tout quantile $(1-\alpha)$ d'une $\mathcal{S}t(n-p-1)$ (resp. d'une $\mathcal{F}(p, n-p-1)$) est approximativement égal à celui d'une $\mathcal{N}(0,1)$ (resp. d'une $p \times \chi^2(p)$, lorsque n est grand. Idem lorsque l'on cherchera à évaluer des probabilités des surfaces. Autant dire que pour ces paramètres et *lorsque n est grand*, nous n'avons pas besoin de faire l'hypothèse **(C2-2)** puisque l'on obtient pratiquement le même résultat si on ne se soumet pas à cette hypothèse : c'est stable ou plutôt **robuste**.

En Bref

On peut faire les deux remarques générales extrêmement importantes suivantes :

→ lorsque n est petit : quel que soit le paramètre d'intérêt (β_i, $\boldsymbol{\beta}_Q$ ou σ^2), l'hypothèse **(C2-2)** est primordiale pour obtenir un quelconque résultat sur la loi empirique des estimateurs.

→ lorsque n est grand

- et lorsque le paramètre d'intérêt est β_i ou $\boldsymbol{\beta}_Q$, les deux cadres d'étude se rejoignent : même si l'on ne suppose plus l'hypothèse **(C2-2)**, les comportements des mesures d'écart standardisée sont sensiblement équivalents à ceux obtenus dans le cadre gaussien. En conséquence, autant utiliser par la suite les résultats issus du cadre gaussien, i.e. la loi de Student (si $\theta = \beta_i$) et la loi de Fisher (si $\theta = \boldsymbol{\beta}_Q$), ceci pour des considérations pratiques.
- lorsque le paramètre d'intérêt est σ^2 : puisque la mesure d'écart standardisée ne s'exprime pas de la même manière dans les deux cadres d'étude, il est clair que pour les utilisations des résultats (tests et intervalles de confiance), il faut absolument savoir si l'on se situe dans un cadre gaussien (i.e. petit échantillon) ou dans un cadre asymptotique.

8 Tests d'hypothèses et régions de confiance

Nous sommes à présent en mesure d'énoncer les outils standards de la statistique inférentielle tels les tests d'hypothèses et les intervalles de confiance.

Remarque : dans *un cadre asymptotique* (pas d'hypothèse **(C2-2)**), l'évaluation des régions critiques, les calculs des p−valeurs (associées aux différents tests), des intervalles ou régions de confiance ne peuvent qu'être approximatifs. Ceci est simplement dû au fait que les comportements en loi des estimateurs et des mesures d'écart standardisée ne sont connus qu'approximativement dans ce cadre. **Il faut toujours avoir ceci en tête même si par abus de notation ceci ne sera pas reprécisé par la suite (i.e. la notation '=' sera utilisée indifféremment)**.

8.1 Tests d'hypothèses

Nous recommandons au lecteur de se référer à l'annexe B décrivant de manière générale les différentes étapes de construction d'un test d'hypothèses relatif à un paramètre (dit test paramétrique).

On sera alors amené à se placer sous l'hypothèse nulle $\mathbf{H}_0$ pour construire la règle de décision. Il est donc naturel d'introduire la statistique de test $\widehat{\delta}_{\theta,\theta_0}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ égale à la statistique $\delta_\theta(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ sous $\mathbf{H}_0$, i.e. $\theta = \theta_0$.

L'un des objectifs principaux (énoncés en introduction) est de savoir si l'un des régresseurs du modèle apporte de l'information dans l'explication de la variable d'intérêt. Le test de significativité locale cherchant à montrer que $\beta_i \neq 0$ c'est-à-dire que le $i^{ème}$ régresseur est informatif pour la variable d'intérêt répond statistiquement à cette question. Par statistiquement, nous entendons que si l'on ne parvient pas à montrer que $\beta_i \neq 0$ au vu des données (par application de la règle de décision) cela ne

signifie pas pour autant que le $i^{ème}$ régresseur est pour autant non significatif. Plus généralement, on rappelle qu'un non rejet de l'hypothèse nulle (i.e. une non acceptation de l'hypothèse alternative) ne signifie pas pour autant l'acceptation de l'hypothèse nulle.

<table>
<tr><th colspan="2">Test d'hypothèses sur β_i</th></tr>
<tr><td>Hypothèses de test</td><td>$\mathbf{H}_0 : \beta_i = \beta_i^0$ contre $\mathbf{H}_1 : \begin{cases} \beta_i > \beta_i^0 & \text{(cas } \mathbf{(a)}\text{)} \\ \beta_i < \beta_i^0 & \text{(cas } \mathbf{(b)}\text{)} \\ \beta_i \neq \beta_i^0 & \text{(cas } \mathbf{(c)}\text{)} \end{cases}$</td></tr>
<tr><td>Statistique de test sous $\mathbf{H}_0$</td><td>$\widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \dfrac{\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \beta_i^0}{\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{Y}|\underline{\boldsymbol{x}})} \begin{cases} \rightsquigarrow \mathcal{S}t\,(n-p-1) & \text{sous l'hypothèse } \mathbf{(C2\text{-}2)} \\ \overset{approx.}{\rightsquigarrow} \mathcal{S}t\,(n-p-1) & \text{si n est grand} \end{cases}$</td></tr>
<tr><td>Règle de décision</td><td>Rejet de $\mathbf{H}_0$ si $\begin{cases} \widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}}) > \delta^+_{lim,\alpha} & \text{(cas } \mathbf{(a)}\text{)} \\ \widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}}) < \delta^-_{lim,\alpha} & \text{(cas } \mathbf{(b)}\text{)} \\ \widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}})\,| > \delta^+_{lim,\alpha/2} & \text{(cas } \mathbf{(c)}\text{)} \end{cases}$
ou si
$p\text{-}valeur = \left\{ \begin{array}{ll} \mathbb{P}_{\beta_i=\beta_i^0}\left(\widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) > \widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right) & \text{(cas } \mathbf{(a)}\text{)} \\ \mathbb{P}_{\beta_i=\beta_i^0}\left(\widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) < \widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right) & \text{(cas } \mathbf{(b)}\text{)} \\ \mathbb{P}_{\beta_i=\beta_i^0}\left(|\widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\,| > |\widehat{\delta}_{\beta_i,\beta_i^0}(\boldsymbol{y}|\underline{\boldsymbol{x}})\,|\right) & \text{(cas } \mathbf{(c)}\text{)} \end{array} \right\} < \alpha$</td></tr>
</table>

où $\delta^-_{lim,\alpha}$ et $\delta^+_{lim,\alpha}$ sont les quantiles d'ordre α et $1-\alpha$ de la loi de la statistique de test sous $\mathbf{H}_0$. Rappelons qu'en théorie (cf section 7.3), si on ne suppose pas l'hypothèse **(C2-2)** et si n est grand la mesure d'écart standardisée de test suit approximativement une $\mathcal{N}(0,1)$ équivalent à une $\mathcal{S}t(n-p-1)$ (lorsque n est grand).
Soulignons que dans le cas **(c)** et $\beta_i^0 = 0$, c'est un **test de significativité locale** du $i^{ème}$ régresseur.

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
## test de significativite locale de beta1
> delta0 <- (betaChapo[2]-0)/sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[2]
> delta0
[1] 0.5685545
> qt(.975,5-1-1)
[1] 3.182449
## p-valeur associee
> 2*(1-pt(abs(delta0),5-1-1))
[1] 0.6094346
```

Exemple 2 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
## test de significativite locale de beta1
> delta0 <- (betaChapo[2]-0)/sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[2]
```

```
> delta0
[1] 19.62147
> qt(.975,200-1-1)
[1] 1.972017
> 2*(1-pt(abs(delta0),200-1-1))   ## p-valeur associee
[1] 0     # c'est preque zero !!!
```

Exemple 3 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
#### test de significativite locale de beta1
> delta0 <- (betaChapo[2]-0)/sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[2]
> delta0
[1] 20.1403
> qt(.975,5-1-1)
[1] 1.972079
> 2*(1-pt(abs(delta0),200-2-1))  ##  p-valeur associee
[1] 0     # c'est preque zero !!!

#### test de significativite locale de beta2
> delta0 <- (betaChapo[3]-0)/sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[3]
> delta0
[1] 3.352868
> 2*(1-pt(abs(delta0),200-2-1)) ## p-valeur associee
[1] 0.0009590043
```

Présentons ici le test de significativité globale cherchant à montrer qu'il existe au moins un régresseur informatif dans l'explication de la variable d'intérêt. En aucun cas, cela n'est pas équivalent à tester la significativité locale de chacun des régresseurs (exception faite du cas particulier d'un modèle linéaire à un régresseur ($p = 1$). Il est rare en pratique d'intégrer le régresseur constant dans le test de significativité globale. En conséquence définissons $\boldsymbol{\beta}' = \boldsymbol{\beta}_{\{1,\dots,p\}} = \big(\beta_1, \dots, \beta_p\big)$.

Test de significativité globale	
Hypothèses de test	$\mathbf{H}_0 : \boldsymbol{\beta}' = \mathbf{0}$ contre $\mathbf{H}_1 : \boldsymbol{\beta}' \neq \mathbf{0}$.
Statistique de test sous $\mathbf{H}_0$	$\widehat{\delta}_{\boldsymbol{\beta}',\mathbf{0}}(\boldsymbol{Y}\vert\underline{\boldsymbol{x}}) =$ $\dfrac{n-p-1}{p}\dfrac{\left\Vert\widehat{\boldsymbol{Y}}-\overline{\widehat{\boldsymbol{Y}}}\right\Vert^2}{\left\Vert\widehat{\boldsymbol{U}}\right\Vert^2}$ $\begin{cases} \rightsquigarrow \mathcal{F}(p, n-p-1) & \text{sous l'hypothèse } \mathbf{(C2\text{-}2)} \\ \overset{approx.}{\rightsquigarrow} \mathcal{F}(p, n-p-1) & \text{si } n \text{ est grand} \end{cases}$
Règle de décision	Rejet de $\mathbf{H}_0$ si $\boxed{\widehat{\delta}_{\boldsymbol{\beta}',\mathbf{0}}(\boldsymbol{y}\vert\underline{\boldsymbol{x}}) > q_{1-\alpha}}$ ou si $\boxed{\underbrace{\mathbb{P}_{\boldsymbol{\beta}'=\mathbf{0}}\left(\widehat{\delta}_{\boldsymbol{\beta}',\mathbf{0}}(\boldsymbol{Y}\vert\underline{\boldsymbol{x}}) < \widehat{\delta}_{\boldsymbol{\beta}',\mathbf{0}}(\boldsymbol{y}\vert\underline{\boldsymbol{x}})\right)}_{\text{p-valeur}} < \alpha}$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une variable aléatoire suivant une loi de Fisher $\mathcal{F}(p, n-p-1)$ Rappelons qu'en théorie (cf section 7.3), si on ne suppose pas l'hypothèse **(C2-2)** et si n est grand la mesure d'écart standardisée de test suit approximativement une loi $p \times \chi^2(p)$ équivalene à une $\mathcal{F}(p, n-p-1)$ (lorsque n est grand).

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
## test de significativite globale
> delta0 <- (5-1-1)/1*var(SalChapo)/var(epsChapo)
> delta0
[1] 0.3232542
> 1-pf(delta0,1,5-1-1) # equivalent au test de significativite locale de beta1
[1] 0.6094346      # en comparant les p-valeurs
```

Exemple 2 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
#### test de significativite globale
> delta0 <- (200-1-1)/1*var(SalChapo)/var(epsChapo)
> delta0
[1] 385.0019
> 1-pf(delta0,1,200-1-1) # equivalent a significativite locale de beta1
[1] 0                # c'est preque zero !!!
```

Exemple 3 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
#### test de significativite globale
> delta0 <- (200-2-1)/2*var(SalChapo)/var(epsChapo)
> delta0
[1] 208.0791
> 1-pf(delta0,2,200-2-1)
[1] 0     # c'est preque zero !!!
```

Le test d'hypothèses précédent peut être vu comme un cas particulier du test d'hypothèses plus général sur le vecteur de paramètres $\boldsymbol{\beta}_Q$:

- **Hypothèses de test** :
$$\mathbf{H}_0 : \boldsymbol{\beta}_Q = \mathbf{0} \text{ contre } \mathbf{H}_1 : \boldsymbol{\beta}_Q \neq \mathbf{0}.$$
- **Statistique de test** : sous $\mathbf{H}_0$,
$$\widehat{\delta}_{\boldsymbol{\beta}_Q,\mathbf{0}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \frac{n-p-1}{q} \frac{\left\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{Q^c}\right\|^2}{\left\|\widehat{\boldsymbol{U}}\right\|^2} \begin{cases} \rightsquigarrow \mathcal{F}(q, n-p-1) & \text{sous l'hypothèse } \textbf{(C2-2)} \\ \overset{approx.}{\rightsquigarrow} \mathcal{F}(q, n-p-1) & \text{si } n \text{ est grand} \end{cases}$$
où Q^c est le complémentaire de Q dans $\{0, 1, \ldots, p\}$ et où $\widehat{\boldsymbol{Y}}_{Q^c}$ est le vecteur des prévisions de $\boldsymbol{Y}$ sur les régresseurs indexés par Q^c.

– **Règle de décision** : à partir des observations

$$\text{Rejet de } \mathbf{H}_0 \text{ si } \widehat{\delta}_{\boldsymbol{\beta}_Q,\mathbf{0}}(\boldsymbol{y}|\underline{\boldsymbol{x}}) > q_{1-\alpha} \text{ ou si } \underbrace{\mathbb{P}\left(\widehat{\delta}_{\boldsymbol{\beta}_Q,\mathbf{0}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) > \widehat{\delta}_{\boldsymbol{\beta}_Q,\mathbf{0}}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right)}_{\text{p-valeur}} < \alpha$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une variable aléatoire suivant une loi de Fisher $\mathcal{F}(q, n-p-1)$.

Test d'hypothèses sur le paramètre de nuisance	
Hypothèses de test	$\mathbf{H}_0 : \sigma^2 = \sigma_0^2$ contre $\mathbf{H}_1 : \begin{cases} \sigma^2 > \sigma_0^2 & \text{(cas } \mathbf{(a)}\text{)} \\ \sigma^2 < \sigma_0^2 & \text{(cas } \mathbf{(b)}\text{)} \\ \sigma^2 \neq \sigma_0^2 & \text{(cas } \mathbf{(c)}\text{)} \end{cases}$
Statistique de test sous $\mathbf{H}_0$	$\widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \overset{\text{Déf.}}{=} \begin{cases} (n-p-1)\,\dfrac{\widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}})}{\sigma_0^2} \rightsquigarrow \chi^2(n-p-1) & \text{sous } \mathbf{(C2\text{-}2)} \\ \dfrac{\widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \sigma_0^2}{\sqrt{\frac{\widehat{\sigma^2_{\boldsymbol{U}^2}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1) & \text{si n est grand et sous } \mathbf{(C2\text{-}2^*)} \end{cases}$
Règle de décision	Rejet de $\mathbf{H}_0$ si $\boxed{\begin{cases} \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}}) > \delta^+_{lim,\alpha} & \text{(cas } \mathbf{(a)}\text{)} \\ \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}}) < \delta^-_{lim,\alpha} & \text{(cas } \mathbf{(b)}\text{)} \\ \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}}) < \delta^-_{lim,\alpha/2} \text{ ou } \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}}) > \delta^+_{lim,\alpha/2} & \text{(cas } \mathbf{(c)}\text{)} \end{cases}}$ ou si $\boxed{p\text{-}valeur = \begin{cases} \mathbb{P}_{\sigma^2=\sigma_0^2}\left(\widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) > \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right) & \text{(cas } \mathbf{(a)}\text{)} \\ \mathbb{P}_{\sigma^2=\sigma_0^2}\left(\widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) < \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right) & \text{(cas } \mathbf{(b)}\text{)} \\ 2\times\min\left(\mathbb{P}_{\sigma^2=\sigma_0^2}\left(\widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) < \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right),\right. & \\ \left.\mathbb{P}_{\sigma^2=\sigma_0^2}\left(\widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) > \widehat{\delta}_{\sigma^2,\sigma_0^2}(\boldsymbol{y}|\underline{\boldsymbol{x}})\right)\right) & \text{(cas } \mathbf{(c)}\text{)} \end{cases} < \alpha}$

où $\delta^-_{lim,\alpha}$ et $\delta^+_{lim,\alpha}$ sont les quantiles d'ordre α et $1-\alpha$ de la loi de la statistique de test sous $\mathbf{H}_0$.

Illustrons ce test dans le cadre des trois exemples où on cherchera à montrer que le paramètre de nuisance $\sigma < 350$ soit $\sigma^2 < 122500$.

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
> sqrt(sigma2Chapo) # au vu de son estimation, pensez-vous que sigma<350?
[1] 208.4994
## p-valeur associee au test H1:sigma<350
> pchisq(sigma2Chapo/122500*(5-1-1),5-1-1)
[1] 0.2143791       # votre conclusion?
```

Exemple 2 : les deux cadres d'étude (gaussien et asymptotique) sont illustrés

```
> sqrt(sigma2Chapo) # au vu de son estimation, pensez-vous que sigma<350?
[1] 317.735
#### p-valeur associee au test H1:sigma<350
## sous l'hypothese (C2-2)
> pchisq(sigma2Chapo/122500*(200-1-1),200-1-1)
[1] 0.03358866       # votre conclusion?
## sous l'hypothese (C2-2*)
> pnorm( (sigma2Chapo-122500)/sqrt(var(epsChapo^2)/200) )
[1] 0.0003271913     # votre conclusion?
```

Exemple 3 : les deux cadres d'étude (gaussien et asymptotique) sont illustrés

```
> sqrt(sigma2Chapo) # au vu de son estimation, pensez-vous que sigma<350?
[1] 309.8231
#### p-valeur associee au test H1:sigma<350
## sous l'hypothese (C2-2)
> pchisq(sigma2Chapo/122500*(200-2-1),200-2-1)
[1] 0.01096636       # votre conclusion?
## sous l'hypothese (C2-2*)
> pnorm( (sigma2Chapo-122500)/sqrt(var(epsChapo^2)/200) )
[1] 1.027214e-06     # votre conclusion?
```

8.2 Intervalles et régions de confiance

Rappelons que la vocation d'un intervalle de confiance est de proposer une "fourchette" contenant avec un certain niveau de confiance un paramètre inconnu. Si vous avez bien assimilé ce qui vient d'être dit, il n'est pas nécessaire d'insister sur la magie d'un tel outil. Soulignons que cet outil inférentiel (ou inductif) apporte un complément d'information à celle fournie par une simple et unique estimation (dite par opposition ponctuelle). En effet, avec un même jeu de données un intervalle de confiance contenant toujours l'estimation ponctuelle a l'avantage par rapport à cette dernière d'intégrer l'information concernant la qualité d'estimation. L'intervalle de confiance répond à la critique faite à l'outil d'estimation ponctuelle ne fournissant qu'un unique remplaçant $\widehat{\theta}$ pour donner l'ordre de grandeur d'un paramètre inconnu θ (cf. dernières elections et notamment les sondages préélectoraux). Elle peut se formuler via la question suivante : même si nous pensons qu'une première estimation obtenue à partir d'un jeu de données de taille plus grande qu'un second jeu de données doit être meilleure (plus précise) que la seconde estimation, savons-nous pour autant si cette première estimation est suffisamment précise ? La réponse à une telle critique est apportée par l'outil d'intervalle de confiance qui sera à un même niveau de confiance $1-\alpha$ donné (à préfixer) d'autant meilleur qu'il sera petit (i.e. précis). En résumé, l'intervalle de confiance est l'outil d'estimation d'un paramètre inconnu avec qualité d'estimation intégrée obtenu à partir d'un seul et simple jeu de données.

- **Paramètre de régression** :
  D'après les résultats établis dans la section 7, il est facile de construire un intervalle aléatoire $\left[\widetilde{\beta}_{i,\inf}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \, , \widetilde{\beta}_{i,\sup}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\right]$ de confiance de β_i au niveau de confiance $1-\alpha$. En effet,

$$1-\alpha=\mathbb{P}\left(\widetilde{\beta}_{i,\inf}(\boldsymbol{Y}|\underline{\boldsymbol{x}})<\beta_i<\widetilde{\beta}_{i,\sup}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\right)$$

avec

$$\boxed{\begin{array}{c}\widetilde{\beta}_{i,\inf}(\boldsymbol{Y}|\underline{\boldsymbol{x}})=\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}})-\delta^+_{lim,\frac{\alpha}{2}}\times\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\\ \text{et}\\ \widetilde{\beta}_{i,\sup}(\boldsymbol{Y}|\underline{\boldsymbol{x}})=\widehat{\beta}_i(\boldsymbol{Y}|\underline{\boldsymbol{x}})+\delta^+_{lim,\frac{\alpha}{2}}\times\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\end{array}}$$

où $\delta^+_{lim,\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de $\mathcal{S}t(n-p-1)$ sous l'hypothèse **(C2-2)** et une loi $\mathcal{N}(0,1)$ si n est grand.

Nous allons proposer l'interprétation de ce type de résultat via l'**A.E.P.** pour l'exemple de cours mais tout ce qui pourra être dit dans cet exemple restera vrai dans sa généralité.

Exemple de cours : Comme l'ordinateur est le seul à tout connaître du modèle, il peut faire ce que nous ne pouvons pas faire en pratique, à savoir, vérifier si le paramètre d'intérêt (inconnu pour nous) est dans l'intervalle de confiance fourni par la formule précédente. Affirmer que cet intervalle a (par exemple) $1-\alpha=95\%$ de confiance signifie aussi qu'il doit commettre $\alpha=5\%$ d'erreurs. Mais pour bien comprendre ceci, quelle autre solution a-t-on que de construire un grand nombre m (voire une infinité) d'intervalles de confiance sur autant de jeux de données générés par le (même) modèle ? C'est ce que propose le tableau suivant où les 10 (les 5 premiers et 5 derniers d'une série de $m=10000$) intervalles de confiance de β_1 :

m	$\widetilde{\beta}_{1,\inf}\left(\boldsymbol{y}_{[m]}\|\underline{\boldsymbol{x}}\right)$	$\widetilde{\beta}_{1,\sup}\left(\boldsymbol{y}_{[m]}\|\underline{\boldsymbol{x}}\right)$	Succès	Taux de succès
1	1.852202	2.130699	1	1/1=100%
2	1.849770	2.124234	1	2/2=100%
3	1.898553	2.188893	1	3/3=100%
4	1.876174	2.163741	1	4/4=100%
5	1.918631	2.215488	1	5/5=100%
⋮	⋮	⋮	⋮	⋮
9996	1.740761	2.027172	1	9519/9996≃ 95.228%
9997	1.921547	2.200326	1	9520/9997≃ 95.229%
9998	2.060123	2.362413	0	9521/9998≃ 95.229%
9999	1.905373	2.220442	1	9521/9999≃ 95.22%
10000	1.810335	2.113418	1	9522/10000=95.22%
⋮	⋮	⋮	⋮	⋮
$+\infty$	-	-	-	95%

où le $m^{ème}$ succès est la réponse codée par 1 ou 0 à la question : le paramètre β_1 (ici égal à 2, révélation faite par l'ordi) appartient-il ou non à l'intervalle $\left[\widetilde{\beta}_{1,\inf}\left(\boldsymbol{y}_{[m]}|\underline{\boldsymbol{x}}\right),\widetilde{\beta}_{1,\sup}\left(\boldsymbol{y}_{[m]}|\underline{\boldsymbol{x}}\right)\right]$?
Le $m^{ème}$ taux de succès est la fréquence des succès parmi les m premiers intervalles.

Parmi $m=10000$ intervalles de confiance de β_1 obtenus via la même formule (proposée ci-dessus grâce à l'**A.C.P.**) sur autant de jeux de données générés par le modèle (il est unique!!!), nous observons bien (via l'**A.E.P.**) qu'il y en a bien environ $1-\alpha=95\%$ qui contiennent le vrai

paramètre à estimer. Finalement, il est facile de se persuader que la formule d'obtention de l'intervalle de confiance ci-dessus est ainsi établie de sorte que parmi une infinité (m tendant vers $+\infty$) d'intervalles de confiance de β_1 construits une proportion $1-\alpha = 95\%$ (resp. $\alpha = 5\%$) contiendraient (resp. ne contiendraient pas) β_1. Par conséquent, en pratique, lorsque nous ne disposons que d'un seul et unique jeu de données nous ne pouvons simplement qu'espérer que notre intervalle de confiance fait partie des $1-\alpha$ qui ne se trompent pas et surtout pas des α qui se trompent. Autrement dit, il ne nous reste plus qu'à parier que notre intervalle de confiance ne se trompe pas et contient bien le paramètre **inconnu** β_1 sinon tant pis pour nous!!!!

Avec le `R` utilisé comme une calculette un peu plus perfectionnée (fournissant en plus les quantiles), nous pouvons sans trop de difficulté obtenir ces intervalles de confiance à 95% de niveau de confiance. A titre de comparaison, nous sommes incapables de dire si oui ou non ils contiennent le vrai paramètre à estimer. Cependant, nous pouvons le parier s'il ne nous reste que cette possibilité.

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
## le [1] extrait le premier element d'un vecteur (ex: beta0Chapo=betaChapo[1])
> betaChapo[1]+c(-1,1)*qt(.975,5-1-1)* sqrt(sigma2Chapo * diag(solve(t(x)%*%x)))[1]
[1]  752.203 2447.477    # i.e. intervalle de confiance de beta0

## le [2] extrait le second element d'un vecteur (ex: beta1Chapo=betaChapo[2])
> betaChapo[2]+c(-1,1)*qt(.975,5-1-1)* sqrt(sigma2Chapo * diag(solve(t(x)%*%x)))[2]
[1] -1285.527  1844.763  # i.e. intervalle de confiance de beta1
```

Exemple 2 : plus besoin del'hypothèse **(C2-2)** car $n = 200$

```
## intervalle de confiance de beta0
> betaChapo[1]+c(-1,1)*qt(.975,200-1-1)*sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[1]
[1]  961.3347 1134.3288
## intervalle de confiance de beta1
> betaChapo[2]+c(-1,1)*qt(.975,200-1-1)*sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[2]
1338.356 1637.432
```

Exemple 3 : les deux cadres d'étude (gaussien et asymptotique) sont illustrés

```
## intervalle de confiance de beta0
> betaChapo[1]+c(-1,1)*qt(.975,200-2-1)*sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[1]
[1]  819.3643 1038.4991
## intervalle de confiance de beta1
> betaChapo[2]+c(-1,1)*qt(.975,200-2-1)*sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[2]
[1] 1343.408 1635.050
## intervalle de confiance de beta2
> betaChapo[3]+c(-1,1)*qt(.975,200-2-1)*sqrt(sigma2Chapo*diag(solve(t(x)%*%x)))[3]
[1]  98.38755 379.42694
```

- **Vecteur des paramètres de régression** :
  On peut également construire une région de confiance de $\boldsymbol{\beta}_Q$ au niveau de confiance $1-\alpha$. En effet, soit $\mathcal{D}_\alpha$ l'intérieur de l'ellipsoïde définie par

$$\delta_{\boldsymbol{\beta}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = q_{1-\alpha}$$

  où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une variable aléatoire suivant une loi de Fisher $\mathcal{F}(q, n-p-1)$ sous l'hypothèse **(C2-2)** et une loi du $\chi^2(q)$ si n est grand. Alors, on a

$$1-\alpha = \mathbb{P}\left(\boldsymbol{\beta}_Q \in \mathcal{D}_\alpha\right) = \mathbb{P}\left(\delta_{\boldsymbol{\beta}_Q}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) < q_{1-\alpha}\right)$$

  ce qui traduit le fait que $\mathcal{D}_\alpha$ est une région de confiance de $\boldsymbol{\beta}_Q$ au niveau de confiance $1-\alpha$.

- **Paramètre de nuisance** :
  Les explications pour comprendre la notion d'intervalle de confiance ont été fournies à la section (à consulter) précédente relative à l'estimation d'un paramètre de régression. Il est également envisageable de construire un intervalle aléatoire $\left[\widetilde{\sigma}^2_{\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}), \widetilde{\sigma}^2_{\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\right]$ (non optimal) de confiance de σ^2 au niveau de confiance $1-\alpha$. En effet,

$$1-\alpha = \mathbb{P}\left(\widetilde{\sigma}^2_{\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) < \sigma^2 < \widetilde{\sigma}^2_{\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})\right)$$

  les bornes étant définies

  - sous l'hypothèse **(C2-2)** par :

$$\widetilde{\sigma}^2_{\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \times \frac{n-p-1}{\delta^-_{lim,\frac{\alpha}{2}}}$$

$$\text{et}$$

$$\widetilde{\sigma}^2_{\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) \times \frac{n-p-1}{\delta^+_{lim,\alpha/2}}$$

    où $\delta^-_{lim,\frac{\alpha}{2}}$ et $\delta^+_{lim,\frac{\alpha}{2}}$ sont les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ d'une loi du $\chi^2(n-p-1)$.

  - si n est grand et sous l'hypothèse **(C2-2*)** par :

$$\widetilde{\sigma}^2_{\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) - \delta^+_{lim,\frac{\alpha}{2}} \times \sqrt{\frac{\widehat{\sigma^2_{\boldsymbol{U^2}}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})}{n}}$$

$$\text{et}$$

$$\widetilde{\sigma}^2_{\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}}) = \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}}) + \delta^+_{lim,\frac{\alpha}{2}} \times \sqrt{\frac{\widehat{\sigma^2_{\boldsymbol{U^2}}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})}{n}}$$

    où $\delta^+_{lim,\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi du $\mathcal{N}(0,1)$.

Exemple de cours : La procédure ci-dessous est la même que celle décrite dans la section précédente relative à l'estimation d'un paramètre de régression.

m	$\widetilde{\sigma}^2_{\text{inf}}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)$	$\widetilde{\sigma}^2_{\text{sup}}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)$	Succès	Taux de succès
1	0.07068328	0.1049155	1	1/1=100%
2	0.06865046	0.1018982	1	2/2=100%
3	0.0768226	0.1140282	1	3/3=100%
4	0.07536241	0.1118608	1	4/4=100%
5	0.08030967	0.1192040	1	5/5=100%
⋮	⋮	⋮	⋮	⋮
9996	0.07475763	0.1109631	1	9498/9996$\simeq$ 95.018%
9997	0.07082663	0.1051283	1	9499/9997$\simeq$ 95.019%
9998	0.08327662	0.1236079	1	9500/9998$\simeq$ 95.019%
9999	0.09046674	0.1342802	0	9500/9999$\simeq$ 95.019%
10000	0.08371438	0.1242577	1	9501/10000=95.01%
⋮	⋮	⋮	⋮	⋮
$+\infty$	-	-	-	95%

où le $m^{ème}$ succès est la réponse codée par 1 ou 0 à la question : le paramètre σ^2 (ici égal à 0.09) appartient-il ou non à l'intervalle $\left[\widetilde{\sigma}^2_{\text{inf}}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right), \widetilde{\sigma}^2_{\text{sup}}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)\right]$? Le $m^{ème}$ taux de succès est la fréquence des succès parmi les m premiers intervalles.

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
> (5-1-1)*sigma2Chapo/c(qchisq(.975,5-1-1),qchisq(.025,5-1-1))
[1]  13950.62 604350.82    # i.e. intervalle de confiance de sigma2
```

Exemple 2 : les deux cadres d'étude (gaussien et asymptotique) sont illustrés

```
#### intervalle de confiance de sigma2
## sous l'hypothese (C2-2)
> (200-1-1)*sigma2Chapo/c(qchisq(.975,200-1-1),qchisq(.025,200-1-1))
[1]  83685.4 124214.7
## sous l'hypothese (C2-2*)
> sigma2Chapo+c(-1,1)*qnorm(.975)*sqrt(var(epsChapo^2)/200)
[1]  88565.22 113345.87
```

Exemple 3 : les deux cadres d'étude (gaussien et asymptotique) sont illustrés

```
#### intervalle de confiance de sigma2
## sous l'hypothese (C2-2)
> (200-2-1)*sigma2Chapo/c(qchisq(.975,200-2-1),qchisq(.025,200-2-1))
[1]  79533.57 118170.90
## sous l'hypothese (C2-2*)
> sigma2Chapo+c(-1,1)*qnorm(.975)*sqrt(var(epsChapo^2)/200)
[1]  85047.23 106933.49
```

9 Prévision

Envisageons maintenant la prévision de la valeur de Y_τ pour un instant τ (pour lequel y_τ est) non observé, en supposant connu le vecteur $\boldsymbol{x}_\tau$ des valeurs des p régresseurs à cet instant τ. La prévision se fait naturellement à partir du modèle ajusté :

$$\boxed{\widehat{Y}_\tau \overset{\text{Not.}}{=} \widehat{Y}_\tau(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) \overset{\text{Déf.}}{=} \boldsymbol{x}_\tau{}^T\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}|\underline{\boldsymbol{x}})}$$

On peut alors considérer l'erreur de prédiction

$$\boxed{\widehat{U}_\tau \overset{\text{Déf.}}{=} Y_\tau - \widehat{Y}_\tau(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)}$$

d'espérance nulle et de variance

$$\boxed{\sigma^2_{\widehat{U}_\tau} \overset{\text{Not.}}{=} V\left(\widehat{U}_\tau\right) = \sigma^2\left(1 + \boldsymbol{x}_\tau{}^T\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\boldsymbol{x}_\tau\right)}$$

naturellement estimée par

$$\boxed{\widehat{\sigma}^2_{\widehat{U}_\tau}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) \overset{\text{Déf.}}{=} \widehat{\sigma}^2(\boldsymbol{Y}|\underline{\boldsymbol{x}})\left(1 + \boldsymbol{x}_\tau{}^T\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\boldsymbol{x}_\tau\right)}$$

On peut également montrer que (voir section E.6) :

$$\boxed{\Delta_{Y_\tau,\widehat{Y}_\tau} \overset{\text{Déf.}}{=} \Delta_{Y_\tau,\widehat{Y}_\tau}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) = \frac{Y_\tau - \widehat{Y}_\tau(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)}{\widehat{\sigma}_{\widehat{U}_\tau}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)} \begin{cases} \rightsquigarrow \mathcal{S}t\,(n-p-1) & \text{sous l'hypothèse } \textbf{(C2-2)} \\ \overset{approx.}{\rightsquigarrow} \mathcal{S}t\,(n-p-1) & \text{si } n \text{ est grand} \end{cases}}$$

Ce résultat permet de construire un intervalle de prévision $\left[\widetilde{Y}_{\tau,\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)\,,\widetilde{Y}_{\tau,\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)\right]$ de Y_τ au niveau $1-\alpha$. En effet,

$$1-\alpha = \mathbb{P}\left(\widetilde{Y}_{\tau,\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) < Y_\tau < \widetilde{Y}_{\tau,\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau)\right)$$

avec

$$\boxed{\begin{array}{rl} \widetilde{Y}_{\tau,\text{inf}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) & = \widehat{Y}_\tau(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) - \delta^+_{lim,\frac{\alpha}{2}} \times \widehat{\sigma}_{\widehat{U}_\tau}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) \\ & \text{et} \\ \widetilde{Y}_{\tau,\text{sup}}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) & = \widehat{Y}_\tau(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) + \delta^+_{lim,\frac{\alpha}{2}} \times \widehat{\sigma}_{\widehat{U}_\tau}(\boldsymbol{Y}|\underline{\boldsymbol{x}};\boldsymbol{x}_\tau) \end{array}}$$

où $\delta^+_{lim,\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi de probabilités de $\Delta_{Y_\tau,\widehat{Y}_\tau}$. Bien que très ressemblant à un intervalle de confiance, l'intervalle de prévision est fondamentalement différent puisqu'il cherche un encadrement d'une variable aléatoire (en l'occurrence Y_τ) et non pas un paramètre réel. Comme vous en avez l'habitude, nous illustrons cette notion avec l'exemple de cours.

Exemple de cours : Comparativement à la section concernant les intervalles de confiance, il faut pour appréhender les résultats via l'**A.E.P.** disposer à chaque $m^{ème}$ expérience répétée non seulement

de l'intervalle de prévision mais aussi de la valeur théorique (puisqu'inconnue en pratique) $y_{\tau,[m]}$ à prédire (i.e. la $m^{ème}$ réalisation généralement virtuelle sauf ici grâce à l'ordinateur de Y_τ). Nous sommes intéressés par savoir combien de fois l'intervalle de prévision (fourni en pratique) contient la valeur à prédire. Le tableau à interpréter comme ceux relatifs aux intervalles de confiance nous propose ce type d'information.

m	$y_{\tau,[m]}$	$\widetilde{y}_{\tau,\inf}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)$	$\widetilde{y}_{\tau,\sup}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)$	Succès	Taux de succès
1	3.447568	2.257075	3.417652	0	0/1=0%
2	2.715760	2.30412	3.447886	1	1/2=50%
3	2.974731	2.245604	3.455533	1	2/3$\simeq$ 66.67%
4	2.92217	2.279384	3.477759	1	3/4=75%
5	2.430905	2.269779	3.506864	1	4/5=80%
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
9996	2.624872	2.237856	3.431413	1	9533/9996$\simeq$ 95.368%
9997	3.205067	2.310475	3.472228	1	9534/9997$\simeq$ 95.369%
9998	2.689362	2.292099	3.551828	1	9535/9998$\simeq$ 95.369%
9999	2.771527	2.240646	3.553632	1	9536/9999$\simeq$ 95.37%
10000	2.451935	2.216719	3.479754	1	9537/10000=95.37%
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$+\infty$	-	-	-	-	95%

où le $m^{ème}$ succès est la réponse codée par 1 ou 0 à la question : la $m^{ème}$ valeur générée (par le modèle) pour le même régresseur x_τ (ici arbitrairement choisi à 0.933) appartient-il ou non à l'intervalle $\left[\widetilde{y}_{\tau,\inf}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right),\widetilde{y}_{\tau,\sup}\left(\boldsymbol{y}_{[m]}\vert\underline{\boldsymbol{x}}\right)\right]$? Le $m^{ème}$ taux de succès est la fréquence des succès parmi les m premiers intervalles. Les interprétations sont tout à fait semblables à celles proposées pour l'intervalle de confiance.

Comme pour la comparaison faite entre l'estimation ponctuelle et l'estimation par intervalle de confiance, nous pouvons en faire de même entre la prévision ponctuelle $\widehat{Y}_\tau$ de Y_τ et l'intervalle de prévision $\left[\widetilde{Y}_{\tau,\inf}(\boldsymbol{Y}\vert\underline{\boldsymbol{x}}),\widetilde{Y}_{\tau,\sup}(\boldsymbol{Y}\vert\underline{\boldsymbol{x}})\right]$ de Y_τ. En résumé, nous proclamons que ce dernier est l'outil de prévision avec qualité de prévision intégrée. A un niveau de confiance (à préfixer) donné, plus l'intervalle de prévision est "petit", meilleure sera la prévision. Nous pouvons le constater sur les trois exemples suivants.

Exemple 1 : Etant donné le faible nombre de données ($n = 5$), on supposera **abusivement** l'hypothèse **(C2-2)** vraie

```
> newIndExp<-3/4
> betaChapo[1]+betaChapo[2]*newIndExp ## prevision a partir de la droite ajustee
[1] 1809.553
> xTau<-c(1,newIndExp)
> t(xTau)%*%betaChapo+c(-1,1)*qt(.975,5-1-1)*sqrt(sigma2Chapo)
                                    *sqrt(1+t(xTau)%*%solve(t(x)%*%x)%*%xTau)
[1]  989.410 2629.697     # i.e. intervalle de prevision du futur YTau
```

Exemple 2 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
> newIndExp<-3/4
> t(betaChapo)%*%c(1,newIndExp)    ## prevision a partir de la droite ajustee
[1] 2163.753
> xTau<-c(1,newIndExp)
## Intervalle de prevision du futur Ytau
> t(xTau)%*%betaChapo+c(-1,1)*qt(.975,200-1-1)*sqrt(sigma2Chapo)
                                      *sqrt(1+t(xTau)%*%solve(t(x)%*%x)%*%xTau)
[1] 1534.469 2793.036
```

Exemple 3 : plus besoin de l'hypothèse **(C2-2)** car $n = 200$

```
> newIndExp <- newIndEtu <- 3/4
>  t(betaChapo)%*%c(1,newIndExp,newIndEtu)  ## prevision a partir du plan ajuste
[1] 2225.034
> xTau <- c(1,newIndExp,newIndEtu)
## Intervalle de prevision du futur Ytau
> t(xTau)%*%betaChapo+c(-1,1)*qt(.975,200-2-1)*sqrt(sigma2Chapo)
                                      *sqrt(1+t(xTau)%*%solve(t(x)%*%x)%*%xTau)
[1] 1610.343 2839.725
```

10 Epilogue

Avant de conclure nous vous recommandons un excellent exercice personnel qui consiste à retrouver chacun des résultats obtenus avec la formule mathématique (estimation, écart-type estimé, p−valeur du test de significativité locale,...) avec la sortie de la fonction `summary(lm( ))` présentée en introduction.

Il est temps à présent de revenir sur les trois exemples sur lesquels avaient travaillé les trois étudiants et de proposer et/ou critiquer chacun de ces exemples.

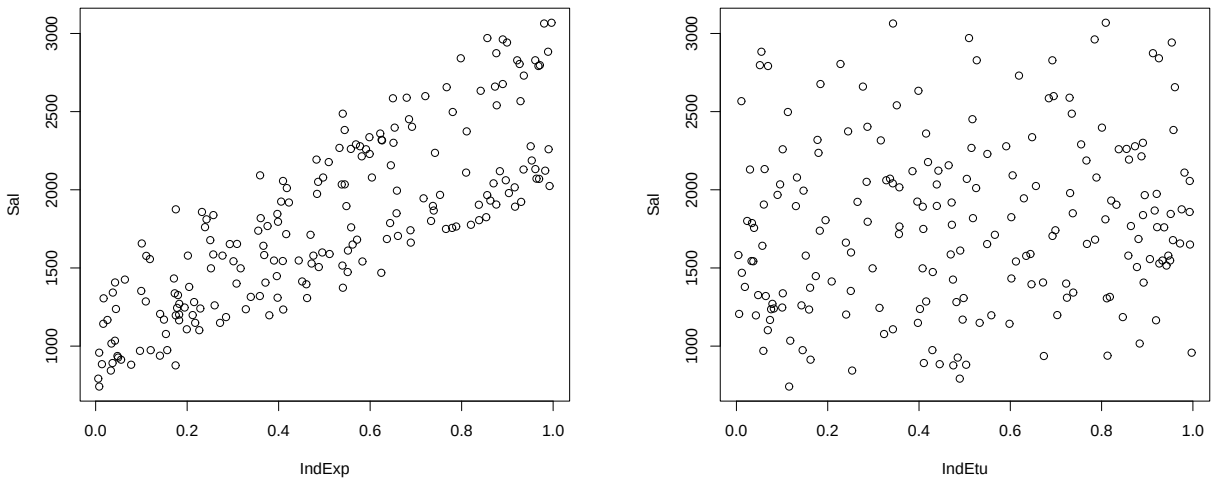
Exemple 1 : Il est clair que le premier des trois n'avait qu'un objectif pédagogique. Les cinq observations ne permettent en aucun cas d'appréhender la variabilité du salaire. En conséquence, le niveau de précision de l'estimateur est assez important ($\widehat{\sigma}_{\widehat{\beta_1}} = 491.8$) compte tenu de l'estimation ($\widehat{\beta_1} = 279.6$). Et de par ce manque de précision, il est impossible (à moins d'être particulièrement imprudent) de parvenir à montrer au vu de ces cinq observations que l'expérience professionnelle possède une quelconque information dans l'explication du salaire individuel. Et pourtant, il ne faut pas conclure qu'il n'y a pas de relation linéaire entre ces deux quantités. Il se peut très bien que le niveau du bruit σ qui a généré les données était trop grand pour que les cinq observations ne traduisent de manière évidente la relation linéaire.

Exemple 2 : La démarche est intéressante, et elle suit en tous points les lacunes exposées ci-dessus. Le jeu de données de taille $n = 200$ offre de biens meilleurs résultats. Le paramètre associé au régresseur `IndExp` est estimé à 1487.89. On peut accorder pas mal de confiance à l'estimation puisque son niveau de précision est estimé à 75.83. Du coup, on parvient à montrer sans aucun problème à un risque

d'erreur de première espèce de l'ordre de 10^{-16} (risque plus qu'extrêmement faible) que le régresseur `IndExp` est très significatif. C'est déjà un excellent résultat en soi. Pour avoir une idée du niveau de confiance que nous pouvons accorder au modèle estimé pour l'appliquer à un problème de prévision, il faut avoir une idée du "niveau de dispersion" du nuage de points autour de la droite estimée : ceci est exprimé par le coefficient de détermination multiple évalué à 0.6604 lui-même exprimant que le régresseur `IndExp` explique 66.04% de la variance de la variable expliquée `Sal`, ce qui est assez relativement satifaisant. Seulement "relativement satisfaisant" car on peut se demander si on ne peut pas faire mieux étant donnée la relative simplicité du problème, classique et largement traité par les économistes.

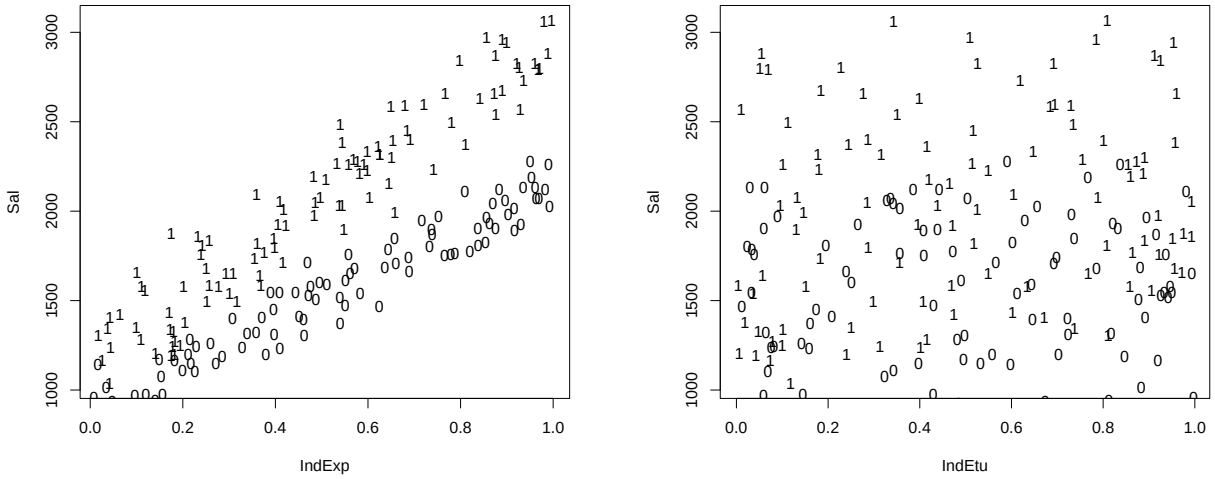
Exemple 3 : À nouveau, la démarche est très justifiée. On espère en introduisant le nouveau régresseur `IndEtu` apporter une information non redondante avec `IndExp` qui permettrait de remonter de manière significative le niveau du coefficient de détermination multiple. Après analyse des résultats, il faut noter que les deux régresseurs sont tous deux très significatifs. Notons par la même occasion que l'estimation du paramètre β_1 associé au régresseur `IndExp` n'a pas beaucoup évolué, traduisant le fait que l'introduction du nouveau régresseur n'a pas détérioré le premier. Autrement dit, la variable expliquée `Sal` parvient à tirer des deux régresseurs l'information qui leur est propre sans "s'emmêler les pinceaux". On peut simplement être déçus par le fait que le R^2 n'a pas beaucoup évolué puisqu'il est passé de 0.6604 à 0.6787. En conclusion, l'étude n'est certainement que trop partielle et il doit certainement manquer dans le modèle une ou des variables informatives.

Il est temps de pousser la réflexion un peu plus profondément en visualisant graphiquement le jeu de données `exSalData`, i.e. en traçant l'évolution du salaire en fonction de `IndExp` et du salaire en fonction de `IndEtu`.



On comprend mieux pourquoi le régresseur `IndEtu` apporte peu d'information dans l'explication : le nuage de points est particulièrement dispersé. À présent en regardant de plus près le premier graphique, on parvient visuellement à dissocier plus ou moins deux nuages de points. Par conséquent, on est en droit de se dire que si la distribution de ces points correspond à une caractéristique particulière des individus, nous avons tout intérêt si cela est possible à intégrer cette caractéristique dans le modèle.

Il n'est pas très difficile de deviner la variable qui discriminerait l'ensemble des individus au vu de la problématique envisagée. Il est très naturel de penser que le salaire dépend du sexe des individus. Supposososns que l'on puisse recontacter les 200 individus pour connaître leur sexe. Cette variable notée `S` est une variable qualitative M/F qu'il convient de coder numériquement en 1 et 0 selon que l'individu soit un homme ou une femme. Pour essayer de vérifier graphiquement l'intuition initiale, reprenons les deux graphiques précédents et codons chaque point en fonction de la variable `S` :



Tout semble concorder. Les deux régresseurs sont assez bien discriminés par la variable sexe. Ce genre de variables binaires en statistiques est appelé variable indicatrice, variable muette (ou dummy variable). Elle peut être intégrée dans le modèle du troisième étudiant comme suit :

$$(Sal)_t = \beta_0 + \beta_1 1_{(S)_t=1} + \beta_2 (IndExp)_t + \beta_3 (IndEtu)_t + \beta_4 (IndExp)_t \times 1_{(S)_t=1} + \beta_5 (IndEtu)_t \times 1_{(S)_t=1} + U_t$$

Il est assez facile de voir que le modèle ainsi spécifié permet de définir deux plans estimés passant chacun d'eux au mieux par le nuage de points associés aux caractéristiques des hommes respectivement des femmes. Sans plus de détails voilà comment `R` traite ce genre de modèles :

```
> summary(lm(Sal~S+IndExp*S+IndEtu*S))

Call:
lm(formula = Sal ~ S + IndExp * S + IndEtu * S)

Residuals:
     Min       1Q   Median       3Q      Max
-225.057  -77.273   -2.623   72.424  261.169

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   819.70      27.73  29.562  < 2e-16 ***
S             155.63      38.53   4.039 7.73e-05 ***
```

```
IndExp       1275.40      35.25  36.178  < 2e-16 ***
IndEtu        108.22      36.53   2.963  0.00343 **
S:IndExp      530.59      51.25  10.354  < 2e-16 ***
S:IndEtu      265.97      49.48   5.375 2.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 107.1 on 194 degrees of freedom
Multiple R-Squared: 0.9622,     Adjusted R-squared: 0.9612
F-statistic:   987 on 5 and 194 DF,  p-value: < 2.2e-16
```

Et voilà tout est dit, la variable discriminante sexe rend tous les régresseurs très significatifs. Par ailleurs, le pouvoir prédictif de ce nouveau modèle est excellent $R^2 = .9622$, bien supérieur au modèle ne tenant pas compte du sexe des individus.

11 Découverte de la colinéarité via une application pratique

Concentrons-nous sur le jeu de données suivant décrivant le prix de 10 voitures en fonction de leur âge du nombre de km. Pour tenter d'expliquer le prix d'une voiture, on envisage un modèle linéaire en intégrant les deux régresseurs. On fera l'hypothèse certainement abusive que le bruit est gaussien :

```
> voiture
   age   km prix
1    1  8.1 5.45
2    2 17.0 4.80
3    2 12.6 5.00
4    3 18.4 4.00
5    3 19.5 3.70
6    4 29.2 3.20
7    6 40.4 3.15
8    7 51.6 2.69
9    8 62.6 1.90
10  10 80.1 1.47
> attach(voiture)
> summary(lm(prix~age+km))

Call:
lm(formula = prix ~ age + km, data = voiture)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59142 -0.21317  0.08918  0.28038  0.38024
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.56844    0.26768  20.803 1.49e-07 ***
age         -0.66388    0.37462  -1.772    0.120
km           0.03009    0.04663   0.645    0.539
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3898 on 7 degrees of freedom
Multiple R-Squared: 0.9317,     Adjusted R-squared: 0.9122
F-statistic: 47.75 on 2 and 7 DF,  p-value: 8.326e-05
```

A première vue, aucun des deux régresseurs ne semble être significatif au seuil de 5% (pas même à 10% d'ailleurs) ce qui n'est pas très encourageant quant au caractère informatif de chacun des régresseurs. Pourtant à y regarder de plus près, le modèle semble assez prédictif puisque le R^2 est de l'ordre de 93%. Continuons l'analyse avec les deux régressions simples suivantes :

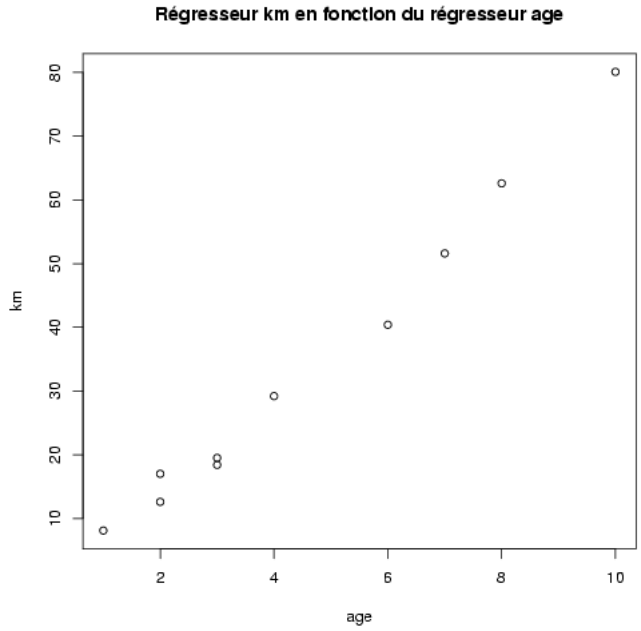
```
> summary(lm(prix~age,data=voiture))

Call:
lm(formula = prix ~ age, data = voiture)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5903 -0.2093  0.1666  0.2189  0.3882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.48562    0.22615   24.26 8.90e-09 ***
age         -0.42383    0.04185  -10.13 7.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3753 on 8 degrees of freedom
Multiple R-Squared: 0.9276,     Adjusted R-squared: 0.9186
F-statistic: 102.6 on 1 and 8 DF,  p-value: 7.724e-06

> summary(lm(prix~km,data=voiture))

Call:
lm(formula = prix ~ km, data = voiture)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.58739 -0.29500  0.01059  0.34880  0.56982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.301368   0.249062  21.285 2.50e-08 ***
km          -0.051999   0.006092  -8.536 2.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4388 on 8 degrees of freedom
Multiple R-Squared: 0.9011,     Adjusted R-squared: 0.8887
F-statistic: 72.86 on 1 and 8 DF,  p-value: 2.731e-05
```

En conséquence, les deux variables prises séparément apportent de l'information dans l'explication du prix et semblent ne plus en apporter lorsqu'elles sont présentes ensemble dans le modèle. On remarquera facilement que ceci est dû au fait que les écarts-types estimés de chaque estimateur ont fortement augmenté lorsque les deux régresseurs sont intégrés dans le modèle comme le rappelle le petit tableau suivant :

	Le modèle à deux régresseurs prix˜ age+km		Les deux modèle à un régresseur prix˜ age	prix˜ km
	`age`	`km`	`age`	`km`
Paramètre estimé	-0.664	0.030	-0.424	-0.052
Ecart-type estimé	0.375	0.047	0.042	0.006

Avec le graphique suivant, on comprend un peu plus d'où pourrait provenir le problème : les deux régresseurs sont pratiquement linéairement liés entre eux.



Pour essayer de comprendre pourquoi la liaison linéaire entre régresseurs (ce que l'on définit en statistique par colinéarité) influe sur la précision estimée des estimateurs il suffit d'analyser précisément la formule suivante qui est une alternative (beaucoup plus riche en information) à celle qui a été présentée dans la section 5. Avec les notations usuelles depuis le début du cours, cette formule établit que la variance théorique de $\widehat{\beta}_j(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ estimateur de β_j (pour $j = 1, \ldots, p$) s'écrit également :

$$\sigma^2_{\widehat{\beta}_j} := \left(\Sigma_{\widehat{\boldsymbol{\beta}}}\right)_{jj} = \frac{\sigma^2}{n \times s_j^2} \times \frac{1}{1 - R_j^2}, \tag{1}$$

où $s_j^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_{j,i} - \overline{x})^2$ et où R_j^2 ne désigne rien d'autre que le coefficient de corrélation linéaire carré lorsque l'on régresse $\boldsymbol{x}_j$ sur l'ensemble des autres régresseurs. Nous précisions un peu avant que cette formule était riche en information car on comprend aisément quels sont les **acteurs** qui influent sur la précision des estimateurs :

- σ^2 : plus le niveau du bruit est élevé et moins les estimateurs seront précis.
- n : plus la taille d'échantillon est grande et plus la variance est faible, jusqu'à tendre vers 0 lorsque $n \to +\infty$ (ce qui fait que les estimateurs sont consistants).
- s_j^2 : ce terme ne dépend que du j-ème régresseur. Il exprime le fait que plus le support de ce régresseur est étendu, plus sa variance est élevée et plus les estimateurs seront précis.
- $\frac{1}{1-R_j^2}$: dans la communauté statistique ce terme est appelé "variance inflation factor" (notée dans les logiciels VIF). Et l'on comprend aisément pourquoi. Plus $\boldsymbol{x}_j$ est colinéaire aux autres

régresseurs, plus R_j^2 est proche de 1, donc plus le terme $\frac{1}{1-R_j^2}$ est élevé ; la variance de l'estimateur $\widehat{\beta}_j$ est alors très élevée. A l'inverse, plus R_j^2 est proche de 0 plus le VIF associé est proche de 1 (le minimum). Ainsi, plus $\boldsymbol{x}_j$ est "indépendant" des autres régresseurs, et moins les estimateurs seront détériorés. La précision ne dépend alors que du support, du niveau du bruit et de la taille d'échantillon.

Ainsi la colinéarité entre régresseurs se répercute inévitablement dans la précision des estimateurs. Pour pouvoir la détecter plusieurs moyens s'offrent à nous : raisonnement théorique, graphiques en deux ou trois dimensions. Mais le plus simple consiste à évaluer les VIF associés à chaque régresseur. Ceci est toujours possible en pratique car rappelons-le, cette quantité ne dépend que de la connaissance des régresseurs. On estime (mais ceci est très objectif) qu'il y a forte colinéarité lorsque $\text{VIF}_j > 10$ (i.e. $R_j^2 > 0.9$). Appliquons ceci à notre exemple initial :

```
> vif(lm(prix~age+km))              ## clalcul des VIF_ j   j=1,...,2
     age       km
74.27853 74.27853
> 1-1/vif(lm(prix~age+km))          ## calcul des R^ 2_j associé
      age        km
0.9865372 0.9865372
```

Remarquons que les VIF sont identiques car il n'y a que deux régresseurs. C'est un exemple très simple (peut-être même trop) car nous n'avions que deux régresseurs et nous aurions pu analyser cette colinéarité graphiquement mais l'on comprend aisément son intérêt lorsque l'on a un grand nombre de régresseurs.

Face à la découverte d'un problème de colinéarité, on peut observer notamment trois stratégies différentes :

- sélectionner un modèle par une méthode pas à pas ascendante ou descendante selon un certain critère comme celui de la significativité locale des régresseurs (mais il en existe bien d'autres!!! Il ne s'agit donc là que d'une recette de cuisine).
- L'inconvénient de la précédente stratégie est de potentiellement éliminer des variables qui ont de l'information pour expliquer Y. Une stratégie pourrait être la suivante : à partir de deux (ou plus) variables colinéaires, on en construit une qui est combinaison linéaire de ces variables et on effectue la régression en utilisant cette nouvelle variable. Le problème revient alors de définir correctement la combinaison linéaire.
- Si le spécialiste n'est pas satisfait par ces stratégies, on lui laisse les résultats tels qu'ils sont en lui précisant que vous suspectez un problème de colinéarité et qu'il vous est difficile d'interpréter les estimations ainsi que les différents tests mis en place. Néanmoins si son objectif n'est que d'entreprendre une prévision ceci peut être amplement suffisant.

Annexe A

Approche expérimentale

Le cadre du rappel se limite à la caractérisation d'une variable aléatoire réelle notée Z relative à une expérience aléatoire $\mathcal{E}$ (ex : $\mathcal{E}$='lancer d'un dé' et Z='face du dé'). Rappelons que la notion de variable aléatoire est plus conceptuelle qu'autre chose et pour aider à sa compréhension nous conseillons de remplacer la notion d'aléatoire par celle de **futur** : ainsi Z peut être vue comme un futur résultat relatif à l'expérience aléatoire $\mathcal{E}$, résultat non encore accessible puisque futur. Il est à noter que ce type de concept n'est pas sans intérêt puisqu'il peut souvent se révéler malgré les apparences porteur d'informations insoupçonnées.

A.1 Notion d'expérience aléatoire

Comme exprimé ci-dessus, il faut concevoir une **expérience aléatoire** (notée $\mathcal{E}$) comme une **future expérience**. Nous ne devons alors pas confondre cette notion avec celle d'expérience (en un mot). En effet, une **expérience aléatoire** exprime, comme le font une "recette de cuisine" ou "une notice d'utilisation", le procédé de construction d'une expérience. La terminologie "aléatoire" est quant à elle justifiée dès que ce procédé est susceptible de générer (dans les mêmes conditions) au moins deux expériences ayant des résultats différents (ex : deux lancers de dé peuvent conduire à deux résultats différents). Par opposition, un procédé de construction d'expérience conduisant systématiquement à un unique résultat est dit expérience **déterministe**.

Avant et afin d'étudier une variable aléatoire Z, nous nous devons de prendre le temps de décrire l'**expérience aléatoire** $\mathcal{E}$ dont dépend la variable aléatoire étudiée. Cette étape est primordiale puisqu'elle permet de bien identifier la nature aléatoire du phénomène étudié.

A.2 Démarche pour la caractérisation d'une variable aléatoire Z

La première étape consiste à déterminer les valeurs possibles de Z, appelées **modalités de** Z (ex : $\{1, \ldots, 6\}$ pour un dé). Cependant, ceci ne permet pas de la caractériser car il peut exister d'autres variables aléatoires ayant les mêmes modalités que Z mais dont le comportement aléatoire

est différent (ex : deux (faces de) dés dont un seul est pipé). Il nous faut alors à présent être capable de mesurer la variabilité de Z. Dans l'approche alternative plus classique des probabilités (appelée par la suite **Approche Classique des Probabilités** ou plus simplement **A.C.P.**), la question serait exprimée par la mesure des chances qu'une future réalisation de Z soit comprise entre deux réels a et b arbitrairement choisis (notée mathématiquement, $\mathbb{P}(a < Z \leq b)$). La réponse apportée par l'**A.E.P.** se décompose en les étapes suivantes :

- Imaginons pouvoir répéter un grand nombre de fois m (en réalité une infinité de fois) la même expérience aléatoire $\mathcal{E}$ indépendamment les unes des autres. Notons alors $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ la suite des valeurs réalisées de Z issues de ces expériences. Comme elles ne sont généralement jamais accessibles en pratique, nous qualifierons ces réalisations par la suite de **virtuelles** pour exprimer que chacune d'entre elles serait susceptible d'être la prochaine réalisation Z.
- La répartition des valeurs $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ caractérise la variabilité de Z.
- La proportion (ou fréquence) des $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ compris entre a et b correspond aux chances que le futur résultat Z soit lui-même compris entre a et b.

Le rôle joué par le nombre de répétitions de l'expérience aléatoire $\mathcal{E}$ est central dès lors que l'on cherche à caractériser une variable aléatoire : on le notera donc toujours m et par opposition avec la taille du jeu de données n dans une problématique statistique il pourra être aussi grand que l'on veut voire infini (si on a suffisamment d'imagination).

A.3 Relation entre l'A.C.P. et l'A.E.P. :

Cette relation est caractérisée par la formule suivante avec comme précédemment a et b deux réels choisis arbitrairement :

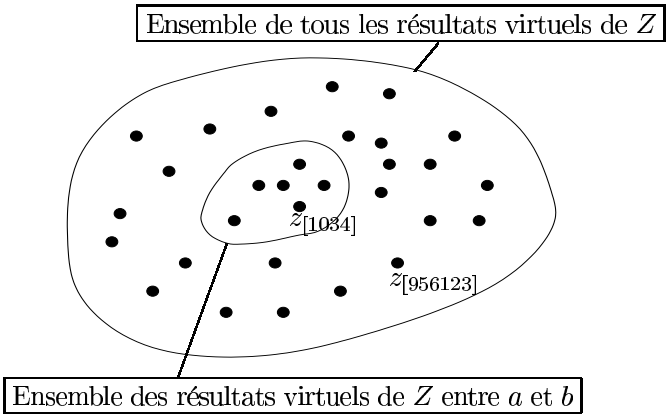
$$\underbrace{\mathbb{P}(a < Z \leq b)}_{\textbf{A.C.P.}} = \underbrace{\lim_{m \to +\infty} \frac{\text{Nombre de fois où } (a < z_{[j]} \leq b) \text{ pour } j = 1, 2, \ldots, m}{m}}_{\textbf{A.E.P.}}$$

Notons aussi que l'**A.C.P.** (non développée ici) décrit un cadre mathématique conduisant à des méthodes d'évaluation de quantité relative à Z (ex : $\mathbb{P}(1 < Z \leq 1.5) = 3/8$ pour Z égale à la somme de réels au hasard entre 0 et 1). L'**A.E.P.** quant à elle n'a aucune prétention de réussir à faire la même chose. Elle permet simplement de mieux visualiser et ainsi d'appréhender des concepts probabilistes pas toujours évident. Ces deux approches sont donc complètement complémentaires.

A.4 Représentation de la répartition d'une infinité de résultats virtuels de Z

Nous avons vu précédemment qu'une variable aléatoire était caractérisée quand nous étions capable d'évaluer toutes les probabilités de la forme $\mathbb{P}(a < Z \leq b)$. Intuitivement, une future réalisation Z peut être vue comme l'une (choisie au hasard) parmi une infinité des réalisations $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$. En un sens, l'**A.E.P.** (et c'est ce qu'essaye d'exprimer la figure ci-dessous où il faut imaginer qu'il y

a une infinité de points représentant toute une infinité de résultats virtuels de Z) vous permet de visualiser tous les résultats possibles (ici matérialisés par l'infinité des $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$) d'une future réalisation Z.



On comprend alors que plus il y a de valeurs parmi $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ comprises entre a et b plus les chances sont grandes pour qu'un futur résultat de Z soit compris entre a et b.

Cependant, cette représentation graphique (en "patate") bien qu'assez intuitive reste trop abstraite car elle ne permet pas de représenter la répartition des $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ porteuse des informations caractérisant la variable aléatoire Z. En effet, elle n'induit aucune localisation des valeurs $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ alors que nous savons (puisqu'elles représentent des valeurs numériques) les ordonner. Une représentation sur **la droite des réels** semblerait solutionner ce problème mais il surgirait un nouveau problème non rencontrée dans la représentation en "patate" : comment représenter les valeurs identiques ? Nous savons alors que la statistique descriptive (la bien nommée) résoud tous ces problèmes dès lors que nous distinguons par avance la nature **discrète** ou **continue** de la variable étudiée. Le mot "discret" (pas vraiment explicite pour un non-matheux) signifie que les **modalités** peuvent être toutes identifiées une à une (en mathématiques, on dit qu'elles sont dénombrables). Par opposition, une variable est dite "continue" si son ensemble des **modalités** représente un continuum (i.e. une réunion d'intervalles de réels). Dans ce cas, il est illusoire de chercher à déterminer le voisin supérieur le plus proche d'une modalité à l'intérieur d'un intervalle (ex : celui de $1/2$ dans l'intervalle $[0, 1]$).

Ceci étant dit,la statistique descriptive nous fournit différents types de représentation graphique plus ou moins adaptés à ces deux types de variables. Pour les variables **discrètes**, nous avons l'embarras du choix avec une préférence pour le **diagramme en bâton** où la hauteur de chaque bâton mesure la fréquence des valeurs représentées de la variable. Soulignons que cette dernière représentation est le prolongement naturel de la représentation des valeurs sur **la droite des réels** évoquée précédemment en levant le problème de représentation de valeurs égales puisque toutes représentées par un même "bâton".

Cependant, nous savons que ce type de représentation fort agréable n'est malheureusement pas adaptée aux variables **continues**. Pour s'en convaincre, il suffit de s'imaginer en train de représenter les fréquences associées au continuum (sans possibilité d'identification des voisins) de l'infinité des modalités. Bien heureusement, il existe une représentation graphique spécialement adaptée à une variable continue appelée **histogramme de fréquences**. La répartition en classes d'intervalles (arbitrairement choisies) d'une série de valeurs de cette même variable est représentée par autant de rectangles qu'il

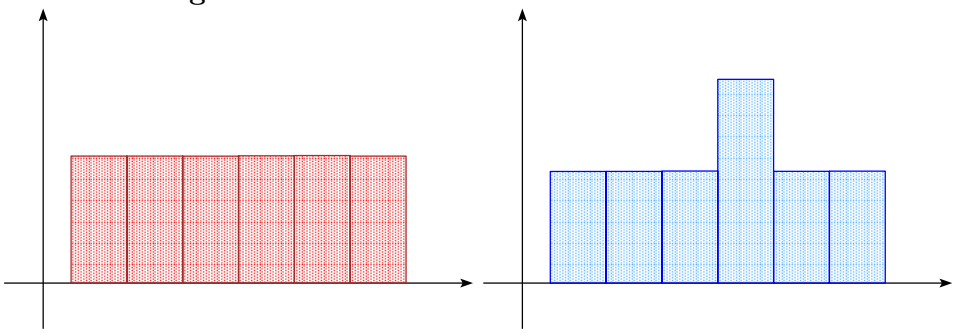
y a d'intervalles et où chaque rectangle (ayant pour base, i.e. côté sur l'axe des abscisses, l'intervalle associé) a une surface égale à la fréquence des valeurs de la variable "tombant" dans un intervalle. Il est à remarquer que les hauteurs (et c'est la différence majeure avec le **diagramme en bâton**) des rectangles n'ont a priori de sens qu'à travers leurs surfaces.

Dans un souci de rendre possible la comparaison entre répartitions de variables discrète et continue, nous allons définir la notion d'**histogramme discret** pour une **variable discrète**. Par opposition, l'histogramme classiquement associée à une **variable continue** sera appelée par la suite **histogramme continu**. La différence notoire entre ces deux histogrammes est que chaque rectangle d'un **histogramme discret** sera non pas associé à l'intervalle de sa base (comme c'est le cas pour le classique **histogramme continu**) mais uniquement à son centre de classe représentant une modalité de la **variable discrète** représentée. En revanche, le point commun aux deux histogrammes (qui justifient cette appellation) est que les fréquences associées aux rectangles sont toutes mesurées en surface (ce qui permet notamment la comparaison de variables discrète et continue).

A.5 Deux exemples

Il est maintenant temps d'illuster ces notions sur des exemples de variables aléatoires discrète et continue.

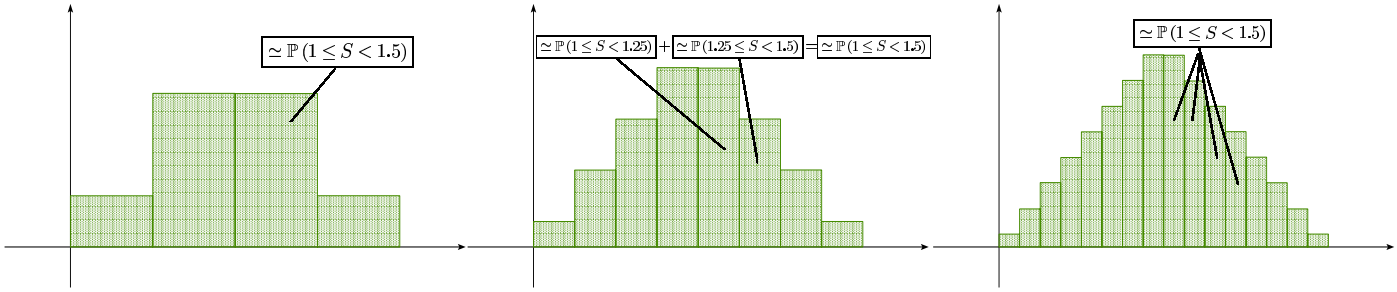
Exemple du dé : Soit Z_1 et Z_2 les faces respectives de deux dés relativement à deux expériences aléatoires $\mathcal{E}_1$ et $\mathcal{E}_2$ consistant au lancers respectifs des deux dés. La caractérisation de **variable aléatoire discrète** est facile à appréhender via l'**A.E.P.** en représentant graphiquement la répartition d'un très grand nombre m puis une infinté de résultats virtuels. Sauriez-vous identifier lequel des deux dés est pipé à partir de leurs **histogrammes discrets** ?



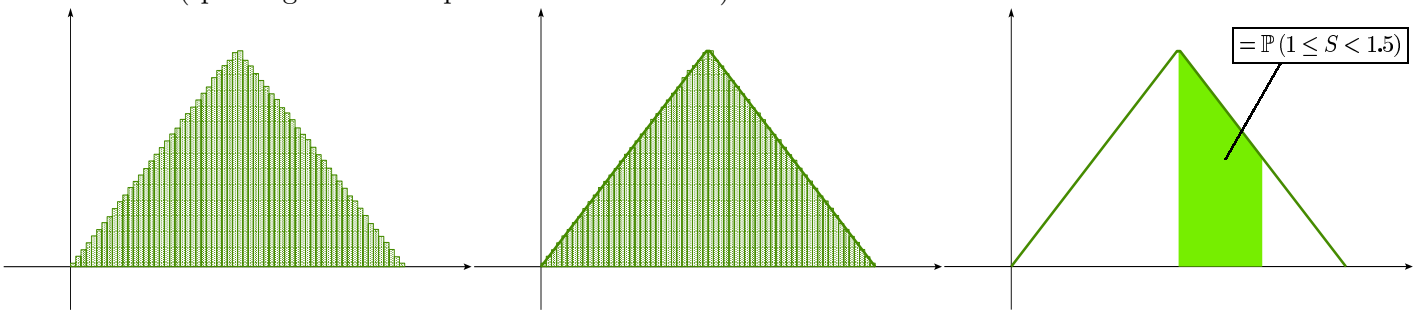
Souvenez-vous que chaque surface d'un rectangle mesure la fréquence (ou proportion) des résultats d'un grand nombre de lancers du même dé. Grâce à l'ordinateur, nous pouvons simuler le lancer d'un dé et représenter un très grand nombre de fois de résultats de cette expérience aléatoire. A partir d'un très très très ... grand nombre, nous avons quasiment l'illusion que nous en disposons d'une infinité. Il est clair que chaque variable aléatoire est caractérisée par la donnée de toutes ces fréquences représentant (d'après la relation entre l'**A.E.P.** et et l'**A.C.P.**) les probabilités d'obtention des faces associées.

Exemple de la somme de deux réels choisis au hasard entre $[0,1]$ *:* Notons S cette variable aléatoire continue dont l'ensemble des modalités est l'intervalle $[0,2]$. Pour caractériser (i.e. "tout dire" sur) la variable aléatoire S, nous devons être capables d'évaluer toutes les probabilités de réalisation de S dans un intervalle quelconque. Commençons par l'intervalle $]1,1.5]$ et intéressons-nous à $\mathbb{P}(1 < S \leq 1.5)$. Via l'**A.E.P.**, nous imaginons disposer d'un très grand nombre m voire une infinité de réalisations vir-

tuelles $s_{[1]}, s_{[2]}, \ldots, s_{[m]}, \ldots$ de S. Comme pour l'exemple du dé, l'ordinateur (sachant choisir au hasard un réel entre 0 et 1) peut nous assister en nous permettant de générer un très très très ... grand nombre d'expériences. Il ne nous reste plus qu'à lui demander de les représenter par un **histogramme continu**. Mais comment lui fixer les classes ? A priori, puisque nous nous intéressons à $\mathbb{P}(1 < S \leq 1.5)$, nous commençons par représenter un histogramme en 4 classes de même longueur (i.e., $[0, 0.5]$,$]0.5, 1]$,$]1, 1.5]$ et $]1.5, 2]$). Mais comment faire avec cet histogramme pour visualiser par exemple $\mathbb{P}(1 < S \leq 1.25)$, probabilité sur un intervalle de longueur deux fois plus petite que le précédent. Mais pourquoi deux fois plus petit et pourquoi pas quatre fois plus petit comme par exemple pour évaluer $\mathbb{P}(1 < S \leq 1.125)$ et ainsi de suite... ? En effet, la difficulté d'une variable continue est que son ensemble de modalités forme un continuum et qu'il faut donc pouvoir évaluer (ou simplement visualiser pour l'**A.E.P.**) la probabilité que la variable aléatoire tombe dans n'importe lequel des intervalles (i.e. dans n'importe quel intervalle de la forme $]a, b]$). A partir des histogrammes en 4, puis 8 et puis 16 intervalles d'amplitudes égales, voyez-vous poindre la solution à ce problème.



Et oui ! Nous constatons que lorsque le nombre de classes est multiplié par 2 nous ne perdons jamais les surfaces représentant des probabilités acquises (en fait approximativement acquises car m est trèe trèe très... grand mais pas infini dans les graphiques) à l'étape précédente et tout cela grâce à de simples regroupements deux à deux. Puisque nous disposons d'autant que l'on veut de réalisations virtuelles (en fait une infinité avec un peu d'imagination) il semble alors possible de diviser encore et encore les amplitudes des intervalles. La prochaine série des trois graphiques met en valeur l'existence de ce que nous appelerons (compte tenu de ce que l'**A.E.P.** nous permet de visualiser) l'**histogramme à "pas zéro"** ("pas" signifiant l'amplitude des intervalles) d'une infinité de résultats virtuels de S.



En clair, c'est le contour supérieur (ici de forme triangulaire) de l'**histogramme à "pas zéro"** qui caractérise la variabilité de S et qui permet donc de mesurer toutes les probabilités relatives à S. Dans le cadre de l'**A.C.P.**, ce contour supérieur de l'**histogramme à "pas zéro"** est appelé **densité de probabilité** de S. Rappelons-le l'axe des ordonnées sur de tels graphiques n'a pas de réelle signification en soi, la terminologie "densité de probabilité" indique simplement que celle-ci est

forte et plus S a des chances de "tomber" autour de la modalité associée : il faut la voir comme une mesure de l'intensité de la probabilité de "tomber" autour d'un point. Constatez que la probabilité que S "tombe" sur une modalité quelconque est nulle puisque la fréquence associée via l'**A.E.P.** serait représentée par une "très très très ..." fine droite dont on sait qu'elle est de surface nulle.
En résumé, grâce à l'**A.E.P.** nous avons appris (et c'est vrai pour toute variable aléatoire Z continue) que :

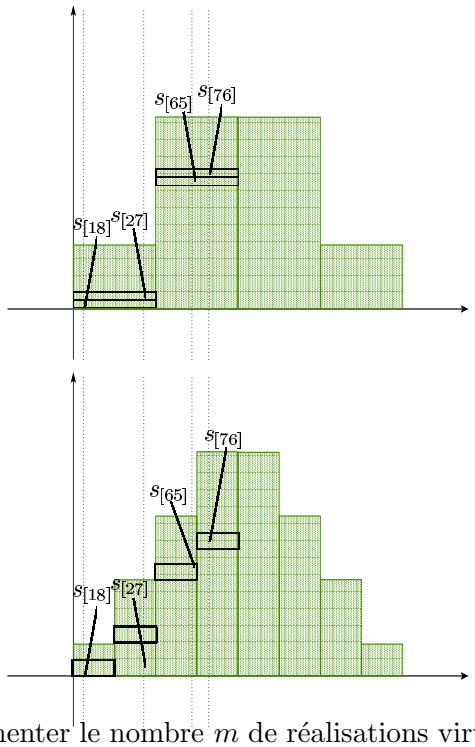
1. $\mathbb{P}(S = a) = 0$ pour toute valeur a même dans $[0, 2]$
2. La probabilité que S soit comprise entre a et b est égale à l'aire de la surface sous la densité de probabilité (notée $f_S(s)$ délimitée entre les bornes a et b. Cela s'écrit plus synthétiquement en mathématique via l'**A.C.P.** à l'aide d'une intégrale (de forme "S" et signifiant "sommation" continue) :

$$\boxed{\mathbb{P}(a < S \leq b) = \mathbb{P}(a < S < b) = \int_a^b f_S(s)ds}$$

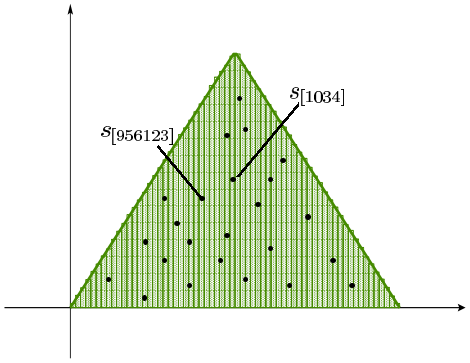
Le dernier graphique de la figure précédente illustre bien ce dernier résultat nous rappelant qu'à partir de l'**histogramme à pas zéro** nous avons conservé l'information que l'histogramme à 4 classes nous avait (approximativement) fournie à savoir $\mathbb{P}(1 < S < 1.5)$. Il faut encore insister sur la différence majeure entre l'**A.E.P.** et l'**A.C.P.** : l'**A.E.P.** nous permet de visualiser et ainsi mieux comprendre les concepts probabilistes introduits par l'**A.C.P.** qui eux permettent d'évaluer des quantités relatives à une variable aléatoire. En l'occurence, sur l'exemple, l'**A.C.P.** nous aurait permis par des calculs mathématiques (pas si difficiles dès lors que nous atteignons une licence ou maîtrise de mathématiques) de déterminer à l'avance l'expression de sa densité de probabilité de forme triangulaire. C'est un résultat que nous n'avons même pas démontré avec l'**A.E.P.** mais simplement (et c'est aussi énorme) deviné.

Il nous reste un point à développer car il permet de répondre au problème soulevé lors de la présentation du graphique en "patate". Pour une **variable aléatoire discrète**, le **diagramme en bâton** était l'une des alternatives, l'**histogramme discret** en est une autre. En revanche, pour une **variable aléatoire continue**, nous allons insister sur la manière dont l'**A.E.P.** permet de visualiser son **histogramme continu à pas zéro** comme le **"tas"** d'une infinité de ses réalisations virtuelles.

Pour cela, avant d'atteindre une infinité de réalisations virtuelles, observons l'étape où nous disposons seulement des $m = 100$ premières i.e. $s_{[1]}, s_{[2]}, \ldots, s_{[100]}$. Essayons, tout d'abord, de mieux comprendre ce que représentent les 4 rectangles de l'histogramme à 4 classes. Puisque la surface totale est égale à 1 et que l'histogramme représente la totalité des m (ici 100) valeurs, chacune de ces valeurs doit être représentée par une "brique" (en fait un "petit rectangle") de surface $1/m$. En somme, un rectangle associé à une classe (ou intervalle) se décompose en un certain nombre k de ces "briques" et la règle est que chacune d'elles est attribuée à une valeur $z_{[i]}$ comprise dans l'intervalle. Les figures ci-dessous illustrent ce propos. Le deuxième histogramme est celui à 8 classes des mêmes $m = 100$ premières valeurs $s_{[1]}, s_{[2]}, \ldots, s_{[100]}$. Sur ces deux histogrammes quatre mêmes valeurs (parmi les $m = 100$) y sont représentées sous forme de "briques" toutes de surface $1/m$.



Essayez alors d'imaginer d'augmenter le nombre m de réalisations virtuelles : les "briques" deviennent alors de plus en plus fines tout en conservant la même largeur que l'amplitude des intervalles. Si de plus, vous imaginez augmenter de nombre de classes ces "briques" deviennent de plus en plus courtes. A la limite (comme le disent les "matheux") puisqu'il faut y aller (en effet m doit tendre vers l'infini et le "pas " de l'histogramme doit tendre vers zéro), vous devez arriver à visualiser cet histogramme comme le **"tas"** de cette infinité des valeurs $s_{[1]}, s_{[2]}, \ldots, s_{[m]}, \ldots$. Il n'est pas si différent de la représentation en "patate" excepté qu'il nous informe exceptionnellement mieux sur la répartition de ces valeurs. Nous espérons que cette dernière représentation assez imagée vous permettra d'appréhender plus sereinement la variabilité de toute variable aléatoire continue (ici illustrée pour S). Bien entendu, vous comprendez que nous n'avons pas essayé de les représenter toutes comme cela avait été le cas pour la représentation en "patate" !



Ainsi, dès lors que la répartition des $z_{[1]}, z_{[2]}, \ldots, z_{[m]}, \ldots$ est complètement identifiée (partie **A.E.P.**), nous dirons que la **loi (de probabilités)** de Z est complètement connue (partie **A.C.P.**) puisque nous savons évaluer n'importe quelle probabilité de réalisation d'une assertion mettant en jeu Z (ex :

.$\mathbb{P}(1 < Z \leq 2)$). Très souvent, cette caractérisation étant faite, l'identification s'achève en proposant un nom de loi complétée d'une série de paramètres (en nombre variable) remplacés dès que possible par des valeurs (ex : $\mathcal{N}\left(p^A, \sqrt{\frac{pa(1-p^A)}{n}}\right)$).

A.6 Espérance et Variance d'une variable aléatoire

L'**A.E.P.** permet également d'appréhender plus aisément d'autres notions probabilistes que sont l'espérance et la variance d'une variable aléatoire Z :

- on définit l'**espérance de** Z, notée $\mathbb{E}(Z)$ (lorsqu'elle existe) comme étant la moyenne empirique d'une infinité de réalisations de Z.
- on définit la **variance de** Z, notée $\mathbb{V}ar(Z)$ (lorsqu'elle existe) comme étant la variance d'une infinité de réalisations de Z.

Avec les notations introduites, on peut faire le lien entre l'**A.E.P.** et l'**A.C.P.** comme suit :

$$\boxed{\mathbb{E}(Z) = \lim_{m\to+\infty} \frac{1}{m}\sum_{i=1}^{m} z_{[j]} \text{ et } \mathbb{V}ar(Z) = \lim_{m\to+\infty} \frac{1}{m}\sum_{i=1}^{m}\left(z_{[j]} - \frac{1}{m}\sum_{i=1}^{m} z_{[j]}\right)^2}$$

Autrement dit, dans l'exemple de la variable aléatoire S (somme de deux réels au hasard sur $[0,1]$), la quantité $\mathbb{E}(S)$ (resp. $\mathbb{V}ar(S)$) représente la moyenne (resp. variance) de l'infinité des points "sous" la densité de probabilité triangulaire.

Annexe B

Construction d'un test d'hypothèses

Dans ce document, on se propose de décrire le schéma de construction d'un test d'hypothèses paramétrique. Chacune des étapes sera subdivisée en trois parties :

→ l'exemple du produit A servant de fil conducteur

→ la démarche générale pour chacune des étapes.

→ quelques questions pemettant de passer la même étape pour une autre problématique.

Un document annexe sera fourni pour répondre aux questions relatives à une problématique quelconque.

Précisons plus en détail, l'histoire de la problématique nous servant de fil conducteur.
Exemple produit A : *un industriel souhaite lancer sur le marché un produit A. Ce produit de consommation peu courante, peut être acheté au plus une fois par mois. Avant même de lancer le produit sur le marché, l'industriel s'informe auprès des services financiers de son entreprise pour savoir quelles sont les contraintes de rentabilité pour un éventuel lancement du produit sur le marché. Ceux-ci lui indiquent qu'il lui faudra vendre au moins 300000 exemplaires de son produit et lui rappellent que la population ciblée est de taille* $N = 2000000$.

B.1 Présentation de la problématique

B.1.1 But de l'étude

Cette section est consacrée à l'expression littérale (en français) de ce que l'on désire montrer.
Exemple produit A : *le produit sera rentable pour l'industriel si la proportion d'acheteurs potentiels du produit A sur la population totale de taille N est supérieure à* 15%.
Généralités : On désire savoir ou pouvoir décider si une **assertion d'intérêt** est vraie.
Question : *Comment s'exprime (littéralement) l'**assertion d'intérêt** dans votre problématique ?*

B.1.2 Paramètre d'intérêt

Dans un but de formalisation mathématique de l'assertion d'intérêt, on souhaite dégager le paramètre d'intérêt de la problématique.

Exemple produit A : *Très simplement, ici le* ***paramètre d'intérêt*** *est la proportion* p^A *d'acheteurs potentiels (parmi* $N = 2000000$*) du produit A. Le lancement du produit A se traduit par une relation de rentabilité* $p^A > 15\%$

Généralités : L'**assertion d'intérêt** s'exprime en fonction d'un **paramètre d'intérêt** θ (pas toujours évident à mettre en avant). Celle-ci est de l'une des formes suivantes $\theta > \theta_0$, $\theta < \theta_0$ ou $\theta \neq \theta_0$ avec θ_0 une **valeur de référence** relative à la problématique étudiée.

Question : *Dans votre problématique, quels sont le* ***paramètre d'intérêt*** *et* ***la valeur de référence*** *ainsi que l'expression de l'****assertion d'intérêt*** *?*

Pour comprendre qu'il n'est pas toujours facile de répondre à une problématique donnée, intéressons-nous aux informations requises à l'évaluation du paramètre d'intérêt.

Exemple produit A : *L'évaluation de* p^A *passe simplement par la connaissance de tous les choix d'achats des* $N = 2000000$ *individus. Cependant, il serait trop coûteux (en temps et donc en argent) d'acquérir une telle information.*

Généralités : L'évaluation du paramètre θ peut dépendre soit d'un très grand nombre soit d'une infinité d'observations.

Question : *Dans votre problématique, de quelles et de combien d'observations vous faudraient-ils disposer pour l'évaluer ?*

En conclusion, le paramètre d'intérêt, puisqu'inévaluable en pratique, sera considéré par la suite totalement **inconnu**, et ainsi toute décision prise ne pourra être fiable à 100%.

B.1.3 Construction d'une règle de décision

Il ne nous reste plus qu'à déterminer une solution alternative réalisable.

Généralités : Acquisition d'une partie de l'information que l'on appelle **(jeu de) données**. La taille des données est notée n et les données elles-même sont notées $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. A partir de ces données $\boldsymbol{y}$, la première étape consiste à proposer un "remplaçant" du paramètre d'intérêt inconnu θ. Cette quantité, notée $\widehat{\theta}(\boldsymbol{y})$ (signifiant remplaçant de θ calculé à partir de $\boldsymbol{y}$), est appelée en statistiques **estimation** de θ.

Exemple produit A : *Ici, cela consiste à interroger une partie (*$n << N$*, i.e. n très petit par rapport à* N*) des acheteurs potentiels du produit A. Les réponses des n individus constituera notre unique* ***jeu de données y*** *sur lequel on peut évaluer la proportion d'individus qui achèteraient le produit A, quantité notée* $\widehat{p^A}(\boldsymbol{y})$*.*

Question : *Décrire le* ***jeu de données*** *associé à votre problématique et proposer l'****estimation du paramètre d'intérêt*** *en précisant sa formule d'obtention.*

A partir de cette information acquise, il nous faut pouvoir prendre une décision à propos de l'assertion d'intérêt.

Généralités : Une **règle de décision** est une proposition logique (en fait une implication) laissant penser que l'assertion d'intérêt est vraie. Dès lors qu'une relation principalement exprimée à partir des **données** et de la **valeur de référence** est vérifiée, on décide (sans pour autant en être sûr) que l'**assertion d'intérêt** doit être vraie.

Exemple produit A : *Cette proposition logique s'exprime naturellement par le lancement du produit A sur le marché si* $\widehat{p^A}(\boldsymbol{y})$ *est au moins supérieur à 15%. Autrement dit, on décide de lancer le produit A si* $\boxed{\widehat{p^A}(\boldsymbol{y}) > 15\% + C}$ *(où C reste une constante à définir). Envisageons alors différentes situations en fonction des valeurs de cette constante C :*

- → *cas où* $C = 0$ *: la règle de décision associée est appelée* ***règle de décision "naïve"*** *laissant penser que le remplaçant* $\widehat{p^A}(\boldsymbol{y})$ *est aussi "bon" que le titulaire* p^A*.*
- → *cas où C est positive : la règle de décision associée est a priori d'autant plus prudente que C est grande laissant penser que la confiance portée au remplaçant* $\widehat{p^A}(\boldsymbol{y})$ *est d'autant plus faible.*
- → *cas où C est négative : après analyse des deux premiers cas, la règle de décision bien qu'envisageable est a priori inintéressante (excepté pour ceux qui auront jugé la règle de décision naïve trop bonne ! ! !).*

Question : *En suivant le schéma général, construisez les règles de décision pour votre problématique, y compris la "naïve".*

B.1.4 Mesure du niveau de fiabilité

De par l'impossibilité d'évaluer le paramètre d'intérêt, nous prenons conscience que toute décision est entachée d'erreur. Il serait alors plus que souhaitable de construire des règles de décision ayant chacune un niveau de fiabilité (ou confiance) préfixé à la convenance de l'utilisateur. Cet objectif n'est pas chose si aisée et presque inenvisageable pour un néophyte en statistiques. Dans cette optique, on commencera par décrire tous les types d'erreurs de décision possibles.

Généralités : Les erreurs de décision sont de deux natures différentes :

- → **erreur de première espèce :** imaginons la situation (non souhaitée et ainsi pessimiste) où l'**assertion d'intérêt** est fausse en réalité (sans que personne ne puisse le savoir). Malgré cette réalité, soulignons qu'il est tout de même possible de décider (en appliquant une quelconque **règle de décision**) qu'au vu des **données** l'**assertion d'intérêt** doit être vraie. Il est clair que dans ce cas précis, nous commettons une erreur de décision relativement grave (par rapport à ce que l'on souhaitait montrer) dite de **première espèce**.
- → **erreur de seconde espèce :** imaginons la situation (souhaitée et ainsi optimiste) où l'**assertion d'intérêt** est vraie en réalité. Malgré cette réalité, soulignons qu'il est tout de même possible de décider (en appliquant une quelconque **règle de décision**) qu'au vu des **données** l'**assertion d'intérêt** ne doit pas être vraie. Il est clair que dans ce cas précis, nous commettons une erreur de décision vraiment moins importante(relativement à nos attentes) dite de **seconde espèce**.

Exemple produit A :

- → ***erreur de première espèce*** *: décider à partir du jeu de données (en pratique) de lancer le produit A qu'en réalité il n'y a pas le marché : on risque de devenir pauvre.*
- → ***erreur de seconde espèce*** *: décider à partir du jeu de données (en pratique) de ne pas lancer le produit A alors qu'en réalité il y a le marché : on risque de ne pas devenir riche.*

Il n'y a ici aucune ambiguïté sur la plus grande importance du premier risque par rapport au second.

Question : *Tentez, dans votre problématique, de décrire littéralement les deux erreurs de décision (en s'appuyant sur l'expression littérale de l'assertion d'intérêt). Précisez pour quelle(s) valeur(s) du paramètre d'intérêt θ inconnu chacune de ces deux erreurs intervient. Quelle vous semble être la plus grave de ces deux erreurs ? A votre avis autour de quelles valeurs de θ la décision vous semblera-t-elle difficile à prendre ?*

A notre connaissance, pour être capable de mesurer le niveau de confiance d'une règle de décision, il est nécessaire que les données soient issues d'un phénomène aléatoire. Pour illustrer ces propos, poursuivons l'histoire du produit A .

Exemple produit A : *L'industriel est maintenant convaincu de l'intérêt d'une étude préliminaire plus approfondie. Il prend ainsi conseil auprès d'un statisticien confirmé qui lui apprend qu'il n'est pas nécessaire de s'empresser de récolter l'unique jeu de données $\boldsymbol{y}$. L'industriel surpris par de tels propos lui demande s'il insinue qu'une quelconque information (liée à sa problématique) est disponible avant même l'acquisition des données. A cette question, le statisticien acquiesce en précisant que le procédé d'obtention des données doit absolument respecter les contraintes aléatoires suivantes :*

→ *tous les n individus constituant le jeu de données devront être choisis au hasard,*

→ *tous les N consommateurs potentiels devront avoir les mêmes chances d'être choisis.*

Un tel jeu de données respectant ces contraintes pourra être appelé ***échantillon*** *de taille n.*

[Commentaire : notons que sans aucune contrainte la solution la moins onéreuse aurait été pour l'industriel d'interroger les n premières personnes de son entourage susceptibles d'acheter son produit. Avec une telle démarche, il est très difficile (voire impossible) de mesurer le niveau de confiance d'une quelconque règle de décision. Une première tentative de justification est de penser que l'entourage de l'industriel est constitué d'individu de même niveau social et peut-être non représentatif de tous les acheteurs potentiels du produit A. En un certain sens, le procédé d'échantillonnage aléatoire décrit précédemment permet d'éviter cette critique de non-représentatitivité.]

L'industriel et le statisticien planifient le recueil effectif des données à une certaine date. Nous l'appelerons par la suite ***le jour J****. L'industriel ayant assimilé ce qui précède, demande au statisticien de préciser la nature des informations dont il peut disposer* ***avant le jour J*** *: le statisticien lui apprend alors qu'étant donnée la procédure d'échantillonnage qui sera mise en place :*

- *il est capable de mesurer le risque de devenir pauvre (ou le niveau de confiance) associé à toute règle de décision,*
- *plus fort encore si l'industriel pense a priori (de par son expérience) que le paramètre d'intérêt p^A se situe dans une plage de valeurs favorables, il peut alors évaluer le risque de ne pas devenir riche et si celui-ci est trop important il peut même lui conseiller de ne pas construire l'échantillon.*

Les deux parties suivantes fournissent les outils nécessaires à la compréhension des propos du statisticien.

B.2 Caractérisations des comportements aléatoires relatifs à la problématique

Afin d'étudier les variables aléatoires relatives à nos problématiques de test d'hypothèses, nous nous appuierons sur ce que nous appelons l'**Approche Expérimentale des Probabilités** (**A.E.P.**)

qui sera la clé de la compréhension. Nous renvoyons donc le lecteur au document traitant spécialement de ce sujet et ce avant de poursuivre la lecture de cette section. Notez aussi que quelques points de comparaison avec l'**Approche Classique des Probabilités** (**A.C.P.**) y sont établis.

B.2.1 Relation entre le paramètre d'intérêt et la notion de modèle (variable d'intérêt)

Généralités : Nous avons admis que la mesure de la fiabilité d'une règle de décision n'est possible que si les données sont issues d'un phénomène aléatoire. Nous nous intéressons donc au procédé aléatoire de fabrication de données. Cela se décompose en :

- → la description de l'expérience aléatoire $\mathcal{E}$
- → la description de la variable aléatoire d'intérêt Y relative à $\mathcal{E}$ complètement identifiable à une future donnée. En fait, cette variable aléatoire d'intérêt est communément appelée en statistiques **modèle**. Cette terminologie exprime qu'un **modèle** aléatoire peut être vu comme un générateur (ou procédé de fabrication) aléatoire de donnée(s).

Puisqu'un **modèle** Y est ici une variable aléatoire, il est judicieux d'avoir le reflexe d'utiliser l'**A.E.P.** : la caractérisation du **modèle** est alors portée par la répartition d'une infinité de données **virtuelles** $y_{[1]}, \ldots, y_{[m]}, \ldots$ générées indépendamment les unes des autres par le **modèle** lui-même. Les questions primordiales auxquelles nous devrons répondre sont les suivantes :

- → Que connaissez-vous de cette répartition ?
- → Le **paramètre d'intérêt** θ peut-il s'exprimer en fonction de l'infinité de données **virtuelles** $y_{[1]}, \ldots, y_{[m]}, \ldots$?

Les réponses à la première question peuvent être diverses et variées. Il est par ailleurs assez fréquent de n'avoir que très peu d'informations sur le **modèle**. Malheureusement, dans certains cadres d'études statistiques (en général quand le nombre de données n est trop faible) il est nécessaire de faire des hypothèses quasiment invérifiables en pratique. Malgré tout, la seconde question précédente intimement liée à la première doit généralement conduire à une réponse affirmative. En effet, il serait incongru d'espérer trouver un remplaçant du **paramètre d'intérêt** à partir d'un **jeu de données** générées par un **modèle** qui serait sans aucune relation avec **le paramètre d'intérêt**. En d'autres termes, la réponse à cette seconde question permet d'établir la nécessaire relation entre le **paramètre d'intérêt** θ et le **modèle** Y ?

Exemple produit A : *L'**expérience aléatoire** $\mathcal{E}$ consiste ici à choisir au hasard un individu au sein de la population ciblée. Le **modèle** (i.e. la variable aléatoire d'intérêt) est alors la future réponse Y de cet individu. Les modalités de Y sont* 0 *(pour une réponse de non achat du produit A) ou* 1 *(pour une réponse d'achat). Via l'**A.C.P.**, il est possible d'évaluer la probabilité que la future réponse Y soit égale à* 1 *et on obtient le résultat très intuitif :* $\mathbb{P}(Y=1) = p^A$ *(et par conséquent* $\mathbb{P}(Y=0) = 1-p^A$*). Via l'**A.E.P.** on peut interpréter ce résultat en affirmant que la répartition en* 0 *et* 1 *d'une infinité de réalisations **virtuelles** $y_{[1]}, y_{[2]}, \ldots, y_{[m]}, \ldots$ de Y est décrite par une proportion p^A de réponse* 1 *(et par conséquent* $1-p^A$ *de* 0*). Cela veut donc dire que, dans cette problématique (et c'est un fait exceptionnel), nous connaissons presque tout du **modèle** Y hormis la valeur de l'unique paramètre qui le caractérise à savoir le **paramètre d'intérêt** p^A. Soulignons au passage que l'**A.C.P.** permet également de montrer que* $\mathbb{E}(Y) = p^A$*. Ainsi la relation entre le **paramètre d'intérêt** p^A et **modèle***

s'écrit de deux façons : $\boxed{p^A = \mathbb{P}(Y=1)}$ *ou* $\boxed{p^A = \mathbb{E}(Y)}$ *(en exercice, on pourra exprimer ces relations via l'**A.E.P.**).*

Question : *Décrire l'expérience aléatoire relative à votre problématique et établir dans votre problématique, la relation liant l'infinité de réalisations* $y_{[1]}, y_{[2]}, \ldots, y_{[m]}, \ldots$ *au paramètre d'intérêt.*

B.2.2 Future estimation

L'étude de la variabilité d'une future estimation est véritablement la clé de le réussite dans la mesure du niveau de confiance d'une règle de décision.

Généralités : Commençons par rappeler que deux cas se présentent à nous :

Cas (I) : soit nous ne savons qu'établir la relation entre **le paramètre d'intérêt** et le **modèle**

Cas (II) : soit nous en savons plus sur la variabilité du **modèle**.

Il est concevable qu'une quelconque connaissance sur la variabilité du **modèle** (*Cas (II)*) va avoir une répercussion dans la caractérisation de la variabilité d'une future estimation. En revanche, dans le *Cas (I)* où nous n'avons aucun a priori sur la variabilité du modèle Y, le statisticien semble se transformer en un véritable magicien puisqu'il prétend assez bien connaître la variabilité d'une future estimation obtenues. La surprise est d'autant plus grande que cette estimation ne dépend que de données générées uniquement à partir du **modèle**. Comme dans n'importe quel tour de magie, il y a un truc!!! Dans la plupart des problématiques, l'explication réside dans la manière d'évaluer l'estimation (sa structure) couplée au nombre n suffisamment grand de données pour le déterminer.

Afin maintenant de tenter d'expliciter le comportement aléatoire d'une future estimation, décrivons le cadre d'étude :

→ L'expérience aléatoire $\boldsymbol{\mathcal{E}}$ (remarquez la notation en 'gras") se décompose en une suite de n expériences $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n$ aléatoires identiques à l'expérience de base $\mathcal{E}$ introduite dans la section précédente.

→ Chaque expérience de base $\mathcal{E}_i$ $(i = 1, \ldots, n)$ est associée à l'obtention d'une future donnée Y_i de la même nature que le **modèle** Y. Le "paquet" (i.e. le vecteur $(Y_1, Y_2, \ldots, Y_n)$) de ces n variables aléatoires constitue un $\boxed{\textbf{futur jeu de données}}$, noté plus synthétiquement $\boldsymbol{Y}$.

La notation en "gras" spécificie que la quantité désignée est un regroupement ("paquet") d'entités de base : $\boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n)$ et $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$. A la différence du nombre m que nous devons imaginer aussi grand que possible voire infini, la taille de n doit être vue comme un nombre de taille raisonnable compte tenu de la difficulté à récolter les données.

Désormais, le procédé de construction d'une estimation $\widehat{\theta}(\boldsymbol{y})$ obtenue à partir de l'unique **jeu de données** $\boldsymbol{y}$ est applicable au **futur jeu de données** $\boldsymbol{Y}$ pour obtenir une nouvelle variable aléatoire $\widehat{\theta}(\boldsymbol{Y})$ correspondant en fait à **une future estimation**. Il ne nous reste plus qu'à chercher à savoir ce que l'on peut dire au sujet de sa variabilité. Une nouvelle fois, nous nous tournons vers l'**A.E.P.** pour décrire de manière simple la variabilité de $\widehat{\theta}(\boldsymbol{Y})$. Par conséquent, on s'intéresse à la répartition d'une infinité de réalisations **virtuelles** (indépendantes entre elles) $\widehat{\theta}(\boldsymbol{y_{[1]}}), \widehat{\theta}(\boldsymbol{y_{[2]}}) \ldots, \widehat{\theta}(\boldsymbol{y_{[m]}}), \ldots$. Notons au passage qu'il nous a fallu auparavant imaginer répéter une infinité de fois notre expérience aléatoire $\boldsymbol{\mathcal{E}}$ pour obtenir une suite d'une infinité de **jeux de données virtuels** $\boldsymbol{y_{[1]}}, \boldsymbol{y_{[2]}} \ldots, \boldsymbol{y_{[m]}}, \ldots$. Il ne nous reste plus qu'à répondre à la simple question :

Sait-on (exactement ou approximativement) caractériser cette répartition même à partir de paramètres clairement identifiés sans pour autant être connus ?

Encore une fois, la réponse à cette question doit toujours être **"OUI"** si l'on a l'espoir de mesurer le niveau de confiance d'une quelconque règle de décision. Cependant, nous devons mettre l'accent sur l'existence éventuelle de paramètres autre que le **paramètre d'intérêt** dans la caractérisation de la loi de $\widehat{\theta}(\boldsymbol{Y})$. Ce type de paramètre (s'il en existe) sera appelé par la suite **paramètre parasite**.
Exemple produit A : *Cette problématique est techniquement l'une des plus simple car via l'**A.C.P.** nous pouvons affirmer que la variabilité de $\widehat{p^A}(\boldsymbol{Y})$ est complètement identifiée. Nous ne l'exprimerons que pour une taille n du jeu de données raisonnablement grande ($n \geq 30$) par souci de simplification mais nous savons aussi le faire pour toute valeur de n. En clair, nous pouvons affirmer que la loi de $\widehat{p^A}(\boldsymbol{Y})$ est approximativement celle d'une loi Normale de moyenne p^A et d'écart-type $\sqrt{\frac{p^A(1-p^A)}{n}}$ (la moyenne et l'écart-type sont les paramètres associés à une telle loi mais ce n'est bien évidemment pas toujours le cas). Mathématiquement, cela s'écrit :*

$$\widehat{p^A}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(p^A, \sqrt{\frac{p^A(1-p^A)}{n}}\right)$$

*Il en découle immédiatement qu'il n'y a pas de **paramètre parasite** puisque seul de paramètre p^A permet à lui seul de caractériser le comportement aléatoire de la **future estimation** $\widehat{p^A}(\boldsymbol{Y})$.*
Question : *Uniquement si cela reste assez facile à exprimer, proposez pour votre problématique la caractérisation de la variabilité (ou via l'**A.C.P.** la loi de probabilités) d'une future estimation (accompagnée d'un dessin à main levée d'une représentation graphique). Existe-t-il des paramètres parasites ? Si oui, proposez-en une interprétation via l'**A.E.P.**.*

Nous avons introduit précédemment la notion de **paramètre parasite**. Dans les problématiques où il en existe, cela signifiera que la caractérisation de la variabilité de $\widehat{\theta}(\boldsymbol{Y})$ est inutilisable en l'état pour la mesure de niveau de confiance d'une règle de décision. Bien heureusement, une dernière étape permet de lever tous ces soucis.

B.2.3 Mesure d'écart standardisée entre la future estimation et le paramètre d'intérêt

Il faut une fois de plus souligner que c'est via l'**A.C.P.** que les mathématiciens réussissent à expliciter les lois de certaines variables aléatoires et notamment celles qui concernent nos problématiques. L'**A.E.P.** quant à elle nous permet d'appréhender plus facilement ces résultats jusqu'à parfois nous faire apprécier leur nature magique.

Cependant, cette section va mettre l'accent sur quelques aspects techniques qui sont pour certaines problématiques des compléments des résultats énoncés dans la section précédente et pour d'autres des passages obligés. La bonne nouvelle sera toutefois qu'à la fin de cette section nous serons en mesure de proposer un cadre d'étude unique adaptable à tout type de problématique (ce qui n'était pas le cas dans la section précédente).

Un deuxième aspect peut aussi être mis en avant pour motiver cette section. Même dans le cas idéal où la loi d'une **future estimation** est complètement explicitée et ne dépend d'aucun **paramètre parasite**, il peut se révéler que ce résultat reste inutilisable en l'état si nous ne disposons pas

d'ordinateur ou alternativement d'aucun logiciel adapté aux probabilités(par exemple `R` et quelques tableurs,...). C'était par ailleurs très souvent le cas il y a quelques années. Les résultats avancés ici remédient à ce type de problème c'est déjà une belle satisfaction.

Exemple produit A : *L'origine de la caractérisation de la loi de* $\widehat{p^A}(\boldsymbol{Y})$ *(proposée dans la section précédente) est tout autre. En fait, un mathématicien pour obtenir ce résultat passe par l'étude d'une variable aléatoire dépendant de la* ***future estimation*** *et ainsi du* ***futur jeu de données*** *décrite ci-dessous ainsi que son comportement aléatoire :*

$$\delta_{p^A}(\boldsymbol{Y}) = \frac{\widehat{p^A}(\boldsymbol{Y}) - p^A}{\sqrt{\frac{p^A(1-p^A)}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

Cette variable aléatoire $\delta_{p^A}(\boldsymbol{Y})$ *mesure l'écart (cf. numérateur) entre la* ***future estimation*** $\widehat{p^A}(\boldsymbol{Y})$ *et le* ***paramètre d'intérêt*** p^A *standardisée (cf. dénominateur) grâce à un changement d'échelle des* ***modalités*** *de* $\widehat{p^A}(\boldsymbol{Y})$ *(un "zoom" des abscisses). Elle est alors tout simplement appelée* ***mesure d'écart standardisée****.*

*Pour comprendre cette transformation, il nous faut passer soit par des considérations mathématiques (via l'****A.C.P.*** *mais pas forcément simples) soit encore une fois via l'****A.E.P.****. Imaginons la répartition d'une infinité d'estimations* ***virtuelles*** $\widehat{p^A}(\boldsymbol{y}_{[1]}), \ldots, \widehat{p^A}(\boldsymbol{y}_{[m]}), \ldots$ *sous forme de "tas"(l'histogramme à pas zéro de cette infinité d'estimations). Puisque chaque valeur* $\delta_{p^A}(\boldsymbol{y}_{[j]})$ $(j = 1, \ldots, m, \ldots)$ *est obtenue par la même transformation décrite ci-dessus (soustraction par un chiffre puis division par un autre) de la* $j^{\text{ème}}$ *estimation virtuelle associée, le placement dans le tas des estimations virtuelles resteront inchangées par cette transformation et ainsi le nouveau "tas" des* $\delta_{p^A}(\boldsymbol{y}_{[1]}), \ldots, \delta_{p^A}(\boldsymbol{y}_{[m]}), \ldots$ *aura la même forme que le "tas" initial des estimations virtuelles* $\widehat{p^A}(\boldsymbol{y}_{[1]}), \ldots, \widehat{p^A}(\boldsymbol{y}_{[m]}), \ldots$*. Par placement dans le "tas", nous entendons que l'ordre de n'importe quelles estimations virtuelles est conservé pour leurs transformées respectives.*

Il est à noter que la ***loi*** *de* $\delta_{p^A}(\boldsymbol{Y})$ *ne dépend d'aucune information inconnue et plus précisément d'aucune information dépendant de la problématique. Par conséquent, ce résultat est général à toute problématique où le paramètre d'intérêt est une proportion.*

Généralités : Soulignons que l'intérêt d'introduire une transformation d'une future estimation réside dans le fait que la **loi** de la transformée de la future estimation, notée $\delta_\theta(\boldsymbol{Y})$, doit être d'une part standart (identifiée par un nom bien connu des statisticiens) et d'autre part ne doit dépendre d'aucune information propre à la problématique. Il est aussi évident que cette caractérisation ne pourra être découverte par soi-même et qu'il faudra se réferer à un tableau récapitulatif pour des problématiques classiques (paramètre d'intérêt étant de type proportion, moyenne, variance, différence de moyenne, rapport de variance). Pour les autres problématiques, il faudra en général retrouver ce type de résultats dans des ouvrages spécialisés.

Proposons un schéma général (en au plus deux étapes) d'obtention de cette transformée $\delta_\theta(\boldsymbol{Y})$:

- → opérer une première transformation sur **la future estimation** $\widehat{\theta}(\boldsymbol{Y})$ pour obtenir une transformée dont la **loi** est connue et ne dépend d'aucune information spécifique à la problématique.
- → si cette transformée ne s'exprime en fonction d'aucun **paramètre parasite** alors c'est fini. Dans le cas contraire, il faut modifier cette transformée en remplaçant les paramètres parasites inconnus par leurs futures estimations obtenues à partir du même **futur jeu de données** $\boldsymbol{Y}$.

La deuxième étape est absolument nécessaire dans la situation où le **paramètre d'intérêt** est par exemple une moyenne. Bien entendu, il faut aussi après cette deuxième étape être capable de caratériser (exactement ou approximativement) la variabilité de la variable aléatoire $\delta_\theta(\boldsymbol{Y})$ ainsi transformée. Heureusement, c'est le cas pour les problématiques classiques évoquées ci-dessus. Généralement, $\delta_\theta(\boldsymbol{Y})$ correspond à une sorte d'écart entre la future estimation $\widehat{\theta}(\boldsymbol{Y})$ et le paramètre d'intérêt θ. Le fait que que sa **loi** soit standard (en statistiques) conduit à sa terminologie dans tout ce document : **mesure d'écart standardisée**.

Question : *Exprimez la future mesure d'écart standardisée relative à votre problématique. Prenez le temps de comprendre sa relation avec la future estimation. Sa loi dépend-elle d'information propre à la problématique ? En est-il de même pour la loi de future estimation ?*

B.3 Vers la construction d'une règle de décision bien contrôlée (avant obtention des données)

Il faudrait s'émerveiller de savoir qu'une règle de décision associée à un niveau de confiance préfixé est construite avant même l'obtention des données. Et pourtant, qui aurait pu dire que nous disposions de quelques informations et non des moindres avant l'acquisition des données. N'était-il pas raisonnable de penser que seules les données (en tant que valeurs numériques) portaient de l'information. Pour rendre plus vivant cet état de fait, retournons à l'exemple de la problématique du produit A.

Exemple produit A : *L'industriel fait le point sur la situation et s'adresse au statisticien en lui posant les questions suivantes : "De quelles informations puis-je disposer **avant le jour J** alors que je n'ai pas encore obtenu mes données ? Par quelle magie puis-je construire une règle de décision m'assurant au moins* 95% *de chances de ne pas devenir pauvre ?"*

A de telles questions, le statisticien lui répond : "Dès lors que nous nous sommes fixé de choisir au hasard les individus de notre échantillon, nous avons introduit annodinement de l'information nous permettant de disposer de tous les ingrédients nécessaires à la construction d'une telle règle de décision. Certes, en vous présentant la situation de la sorte, je ne vous dis pas qu'il s'est véritablement produit un tour de magie. Pour être plus précis, je vous avouerai que de savoir mesurer les chances d'obtention d'une future estimation du paramètre d'intérêt reste et restera toujours une magnifique surprise. Plus encore, ce résultat exprime comment ces chances évoluent en fonction du nombre de données disponibles."

*Le statisticien, après avoir marqué un temps d'arrêt, poursuit sa réponse : "En revanche, vous commencez à entrevoir que puisqu'**avant le jour J** je peux évaluer les écarts raisonnablement probables (i.e. avec une bonne probabilité) entre ma future estimation $\widehat{p^A}(\boldsymbol{Y})$ et la valeur du paramètre d'intérêt p^A, **le jour J** où nous aurons les données $\boldsymbol{y}$ nous pourrons évaluer l'estimation $\widehat{p^A}(\boldsymbol{y})$ et déduire s'il est raisonnable de penser qu'il y ait le marché"*

Le reste de la section décrit les différentes étapes pour la construction d'une règle de décision contrôlant le **risque de première espèce**. Le contrôle du risque de second espèce n'est étudiée qu'à la fin du document. En général, cela ne pose pas de problème car l'**assertion d'intérêt** doit être énoncée de sorte que le **risque de première espèce** doit être plus important à contrôler que **le risque de second espèce**.

B.3.1 La pire des situations

L'un des points qui peut sembler génant dans les résultats exprimant la **loi** d'**une future estimation** est de voir qu'elle dépend toujours du **paramètre d'intérêt** dont on a souligné qu'il était inconnu. En fait, ceci est tout à fait normal car en un certain sens **une future estimation** essaie de "viser sa cible" qui est ici le **paramètre d'intérêt**. Donc comment pourrait-elle être indépendante du **paramètre d'intérêt** ?

Généralités : Les situations à traiter en urgence sont celles où l'**assertion d'intérêt** est fausse (même s'il est impossible de le savoir en réalité nous pouvons les envisager). Dans ce cas, il nous faut savoir déterminer s'il n'y a pas une unique valeur du **paramètre d'intérêt** conduisant à un risque maximal de première espèce, i.e. de décider au vu du **jeu de données** que l'**assertion d'intérêt** doit être vraie et ce par erreur puisque nous avons envisagé que la réalité était tout le contraire. S'il en existe une, nous l'appellerons assez naturellement **la pire des situations**. En fait, nous nous sommes déjà demandés quelle était la valeur du **paramètre d'intérêt** autour de laquelle la décision semblerait la plus difficile à prendre. Il est assez intuitif de penser que c'est lorsque le **paramètre d'intérêt** est égal à la **valeur de référence** (i.e. $\theta = \theta_0$).

Exemple produit A : *Pour mettre ici en évidence une telle situation, nous allons d'abord nous concentrer sur la **règle de décision naïve**. Nous l'appliquons alors à une **future estimation** $\widehat{p^A}(\boldsymbol{Y})$ en envisageant différentes réalités pessimistes (puisqu'il n'y a pas le marché) mais tout à fait possibles : $p^A = 10\%$, $p^A = 12\%$, $p^A = 14\%$ et $p^A = 15\%$ (en fait toutes les valeurs de $p^A \leq 15\%$). Sachant que nous savons "tout" sur une **future estimation** $\widehat{p^A}(\boldsymbol{Y})$ exprimée, faut-il le rappeler, par :*

$$\widehat{p^A}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(p^A, \sqrt{\frac{p^A(1-p^A)}{n}}\right)$$

*nous sommes capables de représenter via l'**A.E.P.** le "tas" d'une infinité d'estimations virtuelles $\widehat{p^A}(\boldsymbol{y}_{[1]}), \ldots, \widehat{p^A}(\boldsymbol{y}_{[m]}), \ldots$ pour chaque situation envisagée ($p^A = 10\%$, $p^A = 12\%$, $p^A = 14\%$ et $p^A = 15\%$). Il est alors immédiat que **la pire des situations** associée au risque maximal de devenir pauvre (ici $\simeq 50\%$) est celle où $\boxed{p^A = 15\%}$.*

*Poursuivons en affirmant que cette **pire des situations** est la même pour toutes les règles de décision de la forme envisagée précédemment (i.e. décider de lancer le produit A si $\widehat{p^A}(\boldsymbol{y}) > 15\% + C$)*

Question : *Pour votre problématique, déterminez et mettez en évidence **la pire des situations** en l'appliquant dans un premier temps sur la **règle de décision naïve***

B.3.2 Règle de décision associée à un risque maximum de décider à tort que l'assertion d'intérêt est vraie

Nous voilà maintenant dans la dernière ligne droite !!!

Généralités : D'après la section précédente, pour limiter les risques de décision à tort de l'**assertion d'intérêt** nous n'avons plus qu'à envisager la **pire des situations** (pessimiste) en réalité . [Attention : cela ne veut pas dire que nous connaissons la véritable valeur du paramètre d'intérêt mais ce n'est qu'une hypothèse que nous espérons fausse. En effet, l'utilisateur d'une règle de décision (et donc d'un test d'hypothèses) doit plutôt désirer que l'**assertion d'intérêt** soit en réalité vraie]. Par la suite, pour construire toute règle de décision nous nous placerons dans cette situtation hypothétique

(nous le souhaitons). Nous aurons alors au moins l'une des informations suivantes relatives à un futur échantillon $\boldsymbol{Y}$:

- → la loi de $\widehat{\theta}(\boldsymbol{Y})$ sous la **pire des situations** (i.e. $\theta = \theta_0$) excepté si celle-ci dépendait de **paramètre(s) parasite(s)**
- → la loi de la **mesure d'écart standardisée** $\delta_\theta(\boldsymbol{Y})$ toujours sous la **pire des situations** (i.e. $\theta = \theta_0$). Rappelons que cette dernière ne dépend jamais de **paramètre parasite**. La notation de la **mesure d'écart standardisée** devra donc être changée en remplaçant le paramètre d'intérêt θ par la valeur de référence θ_0. La **mesure d'écart standardisée** utilisée pour la construction de la règle de décision est appelée classiquement **statistique de test** nous l'appelerons aussi ici **mesure d'écart standardisée de test** et sera donc notée $\delta_{\theta_0}(\boldsymbol{Y})$. Puisque nous savions "tout" sur la **mesure d'écart standardisée** $\delta_\theta(\boldsymbol{Y})$ exception faite du paramètre d'intérêt θ, nous saurons donc (forcément sans exception) "tout" sur la **mesure d'écart standardisée de test** $\delta_{\theta_0}(\boldsymbol{Y})$.

Puisque l'on dispose toujours de la seconde information (quelle que soit la problématique envisagée), on se proposera de raisonner uniquement à partir de la **mesure d'écart standardisée** pour construire la règle de décision.

Exemple produit A : *Dans la **pire des situations** (ici $p^A = 15\%$), on sait désormais que*

$$\widehat{p^A}(\boldsymbol{Y}) \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(15\%, \sqrt{\frac{15\% \times 85\%}{1000}}\right)$$

et que

$$\delta_{15\%}(\boldsymbol{Y}) = \frac{\widehat{p^A}(\boldsymbol{Y}) - 15\%}{\sqrt{\frac{15\% \times 85\%}{1000}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

On privilégiera la dernière information pour construire la règle de décision.

Question : *Dans votre problématique, quelle est la **mesure d'écart standardisée de test** ainsi que sa loi ?*

Généralités : L'étape suivante consiste à construire la règle de décision au vu de $\delta_{\theta_0}(\boldsymbol{y})$ ayant une **erreur de première espèce maximale**, noté α, préfixée à sa convenance par l'utilisateur. Il s'agit de construire la règle de décision au vu de $\delta_{\theta_0}(\boldsymbol{y})$ ayant un **risque maximal de devenir pauvre**, noté α, préfixé à sa convenance par l'utilisateur. Elle est de la forme :

on décide d'accepter l'**assertion d'intérêt** $\begin{cases} \theta > \theta_0 & \textbf{si } \delta_{\theta_0}(\boldsymbol{y}) > q_{1-\alpha} \\ \theta < \theta_0 & \textbf{si } \delta_{\theta_0}(\boldsymbol{y}) < q_{\alpha} \\ \theta \neq \theta_0 & \textbf{si } \delta_{\theta_0}(\boldsymbol{y}) < q_{\alpha/2} \text{ ou } \delta_{\theta_0}(\boldsymbol{y}) > q_{1-\alpha/2} \end{cases}$

Via l'**A.E.P.**, q_α (resp. $q_{1-\alpha}$) peut être vu comme le réel séparant une infinité de réalisations $\delta_{\theta_0}(\boldsymbol{y}_{[1]}), \delta_{\theta_0}(\boldsymbol{y}_{[2]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[m]}), \ldots$ en deux, une proportion α (resp. $1-\alpha$) à gauche de q_α (resp. $q_{1-\alpha}$) et par conséquent une proportion $1-\alpha$ (resp. α) à droite. Rappelons que chaque valeur de cette infinité représente une possible mesure d'écart standardisée associée à un jeu de données que l'on pourrait obtenir si en réalité

la pire des situations se réalisait. Nous sommes alors bien assurés de ne pas encourir un risque d'erreur de première espèce supérieur α.

Il suffit alors d'utiliser (par exemple) le logiciel `R` pour évaluer le(s) quantile(s) associé(s) à loi de la **mesure d'écart standardisée de test** $\delta_{\theta_0}(\boldsymbol{Y})$.

Exemple produit A : *En suivant le cadre général, il s'agit de construire la règle de décision au vu de* $\delta_{15\%}(\boldsymbol{y})$ *ayant un* ***risque maximal de devenir pauvre****, noté* α*, préfixé à sa convenance par l'utilisateur. Elle est de la forme :*

on décide de lancer le produit A sur le marché si $\delta_{15\%}(\boldsymbol{y}) > q_{1-\alpha}$

où $q_{1-\alpha}$ *est par définition le quantile d'ordre* $1-\alpha$ *associée à la loi* $\mathcal{N}(0,1)$ *(de la* ***mesure d'écart standardisée de test****)*

Le logiciel `R` *permet d'évaluer ce quantile à partir de la syntaxe :* `qnorm(0.95)` *si par exemple on choisit* $\alpha = 5\%$.

Question : *Dans votre problématique, quelle est la règle de décision au vu de* $\delta_{\theta_0}(\boldsymbol{y})$ *ayant un risque d'erreur de première espèce préfixé à* α *? Quelle est l'instruction* `R` *permettant d'évaluer le quantile définissant la règle de décision ?*

B.4 Mesure de l'intensité de la décision (après obtention des données)

Jusqu'à présent la seule information utilisée pour construire la règle de décision a été la loi de la **mesure d'écart standardisée de test** $\delta_{\theta_0}(\boldsymbol{Y})$. Il ne nous reste plus qu'à récolter les données et conclure en les confrontant à la règle de décision fraî chement établie.

Exemple produit A : *Après avoir pris conscience des atouts de la règle de décision ainsi construite et notamment du contrôle du* ***risque maximal de devenir pauvre****, l'industriel décide, le* ***jour J****, de finalement récolter les données. Parmi les individus interrogés, 172 sont susceptibles d'acheter le produit A, ce qui revient à une proportion d'acheteurs potentiels de* $\widehat{p^A}(\boldsymbol{y}) = \frac{172}{1000}$ *et donc* $\delta_{15\%}(\boldsymbol{y}) = \frac{17.2\%-15\%}{\sqrt{15\%\times 85\%/1000}} \simeq 1.948$*. Le statisticien rappelle alors à l'industriel qu'il est le seul à encourir véritablement le risque de devenir pauvre et qu'il est donc le seul maître dans le choix de la valeur de ce risque. L'industriel propose alors de fixer (par superstition)* α *à 8% (son chiffre favori). En utilisant le logiciel* `R`*, il évalue le quantile d'ordre 92% d'une loi Normale standard (centrée réduite) :* `qnorm(.92)` $\simeq 1.405$*. À votre avis quelle décision va-t-il prendre ?*

Généralités : L'étape suivante consiste à récolter les données i.e. construire le vecteur $\boldsymbol{y}$, évaluer tour à tour l'estimation $\widehat{\theta}(\boldsymbol{y})$ de θ puis la mesure d'écart standardisée $\delta_{\theta_0}(\boldsymbol{y})$. Il est possible que dans certaines problématiques, le choix de la valeur du risque maximal d'erreur de première espèce n'appartienne à personne. Dans ce cas il existe un consensus admis par la plupart des utilisateurs de tests d'hypothèses fixant cette valeur à $\alpha = 5\%$ (et parfois 1% ou 10%).

Question : *Conclure au vu de votre jeu de données.*

Nous allons proposer une mesure d'intensité de la décision prise à partir d'un jeu de données.

Généralités : Lorsque nous ne disposions pas de données, nous avons été contraints pour construire la règle de décision de se fixer au préalable la valeur du risque maximal d'erreur de première espèce. Maintenant que nous disposons d'un vrai jeu de données, on peut directement se demander :

quel est le plus petit risque d'erreur de première espèce me permettant au vu des données de pouvoir accepter l'assertion d'intérêt ?

La solution à une telle question est appelée **p−valeur** (appelé dans certains logiciels **significativité** ou **p−value** en anglais). Rappelez-vous que notre objectif voire notre souhait est de réussir, au vu du **jeu de données**, à montrer que l'**assertion d'intérêt** semble plutôt vraie et donc de prendre la décision d'accepter l'**assertion d'intérêt** ! Il est donc normal de chercher à évaluer le **risque minimum de première espèce** à encourir pour prendre une telle décision : cette valeur correspond à la **p−valeur**. Ainsi, pour toute règle de décision associée à un risque maximal α de première espèce tolérable fixé à une valeur très très très ... légèrement supérieure à la **p−valeur** (i.e. avec α strictement supérieur à la p−valeur mais aussi proche d'elle que l'on veut ; par exemple, pour une $p-valeur = 2.5686\%$ on peut choisir $\alpha = 2.57\%$ ou $\alpha = 2.569\%$) la décision prise au vu de notre jeu de données est, de par la définition de la **p−valeur**, d'accepter l'**assertion d'intérêt**.

En somme, plus la **p−valeur** est faible et plus on a le "sourire" puisqu'on a d'autant moins de **risque de première espèce** à encourir pour décider l'acceptation de l'**assertion d'intérêt** (i.e. ce que l'on souhaitait). Nous venons de justifier que la **p−valeur** est une mesure d'intensité du niveau de confiance de la décision d'acceptation de l'**assertion d'intérêt** prise à partir d'un **jeu de données**.

Plus fort encore, nous venons tout simplement (quelques lignes au-dessus) de proposer une reformulation de la règle de décision (associée à un risque maximal α d'erreur de première espèce) **strictement équivalente** à la première :

on décide d'accepter l'**assertion d'intérêt** si **p−valeur** $< \alpha$

Il reste maintenant à exprimer mathématiquement cette $p-valeur$ selon les trois formes d'assertion d'intérêt possibles, afin de pouvoir l'évaluer.

assertion d'intérêt	**p-valeur**
$\theta > \theta_0$	$\mathbb{P}_{\theta_0}\left(\delta_{\theta_0}(\boldsymbol{Y}) > \delta_{\theta_0}(\boldsymbol{y})\right)$
$\theta < \theta_0$	$\mathbb{P}_{\theta_0}\left(\delta_{\theta_0}(\boldsymbol{Y}) > \delta_{\theta_0}(\boldsymbol{y})\right)$
$\theta \neq \theta_0$	$2 \times \min\left(\mathbb{P}_{\theta_0}\left(\delta_{\theta_0}(\boldsymbol{Y}) < \delta_{\theta_0}(\boldsymbol{y})\right), \mathbb{P}_{\theta_0}\left(\delta_{\theta_0}(\boldsymbol{Y}) > \delta_{\theta_0}(\boldsymbol{y})\right)\right)$

Via l'**A.E.P.**, la **p−valeur** peut être vue dans le cas où l'**assertion d'intérêt** est $\theta > \theta_0$ (resp. $\theta < \theta_0$) comme la proportion parmi l'infinité des réalisations $\delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{1}]}), \delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{2}]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[\boldsymbol{m}]}), \ldots$ qui sont supérieures (resp. inférieures) à $\delta_{\theta_0}(\boldsymbol{y})$. En règle générale, puisqu'on se place hypothétiquement dans **la pire des situations** (i.e. $\theta = \theta_0$) la **p−valeur** représente la proportion parmi l'infinité des réalisations $\delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{1}]}), \delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{2}]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[\boldsymbol{m}]}), \ldots$ qui conduisent à une décision plus erronnée que celle que l'on prendrait avec $\delta_{\theta_0}(\boldsymbol{y})$ obtenu à partir de notre unique jeu de données. Graphiquement, si l'on représente le "tas" de l'infinité des réalisations $\delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{1}]}), \delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{2}]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[\boldsymbol{m}]}), \ldots$ la **p−valeur** correspond à la surface couverte par les réalisations parmi $\delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{1}]}), \delta_{\theta_0}(\boldsymbol{y}_{[\mathbf{2}]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[\boldsymbol{m}]}), \ldots$ plus écartées du "centre du tas" que $\delta_{\theta_0}(\boldsymbol{y})$ et ce dans le(s) côté(s) extrême(s) représentant des erreurs de décision lorsque **la**

pire des situations est supposée. Ainsi cette surface sera représentée sur la droite pour une **assertion d'intérêt** de la forme $\theta > \theta_0$, sur la gauche pour une **assertion d'intérêt** de la forme $\theta > \theta_0$ et enfin sur les deux côtés (i.e. à la fois sur la droite et sur la gauche) pour une **assertion d'intérêt** de la forme $\theta \neq \theta_0$. Un problème reste à élucider si **assertion d'intérêt** est de la dernière forme (i.e. $\theta \neq \theta_0$) : la valeur $\delta_{\theta_0}(\boldsymbol{y})$ étant soit à gauche soit à droite du "centre du tas", comment représenter la surface opposée représentant pourtant des réalisations parmi $\delta_{\theta_0}(\boldsymbol{y}_{[1]}), \delta_{\theta_0}(\boldsymbol{y}_{[2]}), \ldots, \delta_{\theta_0}(\boldsymbol{y}_{[m]}), \ldots$ encore plus écartées que $\delta_{\theta_0}(\boldsymbol{y})$? Lorsque le "tas" est symétrique, le "tas" est généralement centré en 0 et il est facile de représenter la surface des réalisations parmi l'infinité encore plus écartées de 0 que $-\delta_{\theta_0}(\boldsymbol{y})$. Mais, le raisonnement général est de représenter une surface opposée égale à celle des réalisationsparmi l'infinité encore plus écartées du "centre du tas" que $\delta_{\theta_0}(\boldsymbol{y})$. Cela explique la multiplication par 2 dans le tableau ci-dessous fournissant l'expression de la **p−valeur** dans le cas où l'**assertion d'intérêt** est de la forme $\theta \neq \theta_0$. Le minimum, quant à lui, indique que nous ne sommes intéressés que par la surface des réalisations parmi l'infinité les plus éloignées du "centre".

Enfin, d'un point de vue pratique puisque l'on connaît la loi de la **mesure d'écart standardisée de test**, on peut évaluer la p−valeur en utilisant le logiciel `R`.

Exemple produit A : *Rappelons qu'à un risque de devenir pauvre (maximal) fixé par l'industriel à 8%, nous lui conseillons de lancer le produit A sur le marché (puisque* $\delta_{\theta_0}(\boldsymbol{y}) \simeq 1.948 > q_{0.92} \simeq 1.644$*). L'industriel décidément très joueur se demande alors s'il peut prendre moins de risque (de devenir pauvre) de décider au lancement de son produit. Sans le recours du statisticien, il décide de construire une nouvelle règle de décision avec cette fois-ci un risque de devenir pauvre maximal fixé à 1%. Il évalue le quantile d'ordre 99% associé à la loi* $\mathcal{N}(0,1)$ *à* $q_{0.99} \simeq 2.33$ *et conclut au non lancement de son produit. Comprenant bien qu'il doit exister entre 1% et 8%, des valeurs du risque de devenir pauvre qui lui permettraient de pouvoir encore lancer son produit, il se pose alors la question naturelle de savoir quelle la plus petite valeur du risque de devenir pauvre parmi toutes celles qui lui permettent de pouvoir décider au lancement. Par un simple raisonnement, il comprend alors que pour évaluer ce risque, il lui suffit de placer la règle de décision "au niveau" de* $\delta_{\theta_0}(\boldsymbol{y})$*. Fort de ce constat, il recontacte le statisticien pour lui faire part de sa découverte. Le statisticien, peu surpris, lui apprend que la quantité recherchée par l'industriel porte le nom bien spécifique de* ***p−valeur****. En utilisant le logiciel* `R`*, il lui précise finalement que cette quantité vaut* `1-pnorm(u)` $\simeq 2.57\%$ *et s'en va comme un prince laissant l'industriel interpréter cette valeur avec le paragraphe précédent.*

Question : *Dans votre problématique, définissez la p−valeur. Quelle est alors la règle de décision associée à un risque d'erreur de première espèce* α *? Quelle est l'instruction* `R` *permettant d'évaluer la p−valeur ? Interprétez alors sa valeur.*

B.5 Rédaction standard d'un test d'hypothèses

Nous allons maintenant proposer une rédaction plus synthétique d'un test d'hypothèses. Rappelons les ingrédients indispensables qui devront figurer dans cette dernière :

→ l'**assertion d'intérêt** exprimée en fonction du **paramètre d'intérêt** (inconnu). Elle sera notée $\mathbf{H_1}$.

→ la **pire des situations** exprimant la valeur du paramètre d'intérêt θ (toujours égale à la **valeur de référence** θ_0) pour laquelle on commet le plus d'erreur (de première espèce) d'accepter

l'**assertion d'intérêt**. La règle de décision est construite dans cette situation qui sera notée $\mathbf{H_0}$.

→ la **mesure d'écart standardisée de test** dite plus classiquement **statistique de test** dont on sait caractériser le comportement aléatoire dans la **pire des situations** (i.e. sous $\mathbf{H_0}$)

→ la **règle de décision** sous l'une des deux formulations proposées précédemment. On préfèrera la formulation basée sur la très informative **p−valeur**.

Classiquement, les situations $\mathbf{H_0}$ et $\mathbf{H_1}$ sont appelées **Hypothèses de test**. La rédaction standard sera donc de la forme :

Hypothèses de test :

$$\mathbf{H_0} : \theta = \theta_0 \text{ contre } \mathbf{H_1} : \begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

Statistique de test : Sous $\mathbf{H_0}$,

$$\delta_{\theta_0}(\boldsymbol{Y}) \rightsquigarrow \mathcal{L}_0$$

où $\mathcal{L}_0$ est une loi standard à préciser.

Règle de décision :

$$\text{on accepte } \mathbf{H_1} \text{ si } p - valeur < \alpha.$$

Conclusion : Application de la règle de décision au vu des données $\boldsymbol{y}$.

Annexe C

Instructions R utilisées

C.1 Les instructions de base

Le tableau suivant résume les fonctions `R` de base utilisées dans le document.

Fonction `R`	Effet
`<-` ou `->`	affectation à une variables
`c(1,-1,4)`	définition du vecteur à trois composantes $(1, -1, 4)$
`sqrt()`	racine carrée d'un réel positif ou d'un vecteur à composantes positives.
`^2`	carré d'un réel ou d'un vecteur réel.
`length()`	longueur d'un vecteur.
`sum()`	somme des composantes d'un vecteur réel.
`mean()`	moyenne des composantes d'un vecteur réel.
`sd()`	écart-type d'un vecteur réel positif.
`var()`	variance des composantes d'un vecteur réel.

La preuve par l'exemple :

```
> c(1,-1,4)          ## exemple de definition d'un vecteur
[1]  1 -1  4
> x1<-c(1,-1,4)      ## exemple d'affectation a gauche
> c(1,-1,4)->x2      ## exemple d'affectation a droite
> x1
[1]  1 -1  4
> x2
[1]  1 -1  4        ## c'est la meme chose!
> 2*c(1,-1,4) + 3   ## les calculs se font composante par composante.
[1]  5  1 11
```

```
> 0:20->y           ## vecteur des entiers de 0 a 20.
> y
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
> length(y)         ## longueur du vecteur y
[1] 21
> sum(y)            ## somme des composantes du vecteur y
[1] 210
> mean(y)           ## moyenne du vecteur y
[1] 10
> sum(y)/21
[1] 10
> var(y)            ## variance du veccteur y
[1] 38.5
> sd(y)             ## ecart-type du vecteur y
[1] 6.204837
> sqrt(var(y))      ## meme chose que la racine carree du vecteur y
[1] 6.204837
> sum(  (y-mean(y))^2  )/21
[1] 36.66667
> sum(  (y-mean(y))^2  )/20  ## meme chose que var(u)
[1] 38.5
```

Une petite explication du dernier point : soit $\boldsymbol{y} = (y_1, ..., y_n)$ un vecteur à n composantes réelles. La variance du vecteur $\boldsymbol{y}$ est **très** légèrement différente selon que l'on se situe dans un contexte de statistique descriptive ou d'un contexte de statistique de statistique inférentielle (en particulier dans le cadre de tests d'hypothèses).

$\rightarrow$ en statistique descriptive on définit la variance de $\boldsymbol{y}$ via la formule :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad \text{avec} \qquad \overline{y} = \widehat{\mu}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{C.1}$$

$\rightarrow$ en statistique inférentielle on définit la variance de $\boldsymbol{y}$ via la formule :

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{C.2}$$

C.2 Calculs de quantiles et d'aires associés à une loi de probabilité

Soit α un réel appartenant à $]0, 1[$, on définit le quantile d'ordre $1-\alpha$ associée à une loi de probabilité le réel qui via l'approche expérimentale peut être vu comme le réel qui sépare l'infinité des observations (associée à la loi de probabilité) en deux, une proportion $1-\alpha$ à gauche et une proportion α à droite. On définit également la fonction de répartition en un réel u, la proportion parmi l'infinité

des observations qui se situent avant u. Ces deux notions sont illustrées dans la figure C.1.

Certaines lois de probabilités (et c'est le cas pour celles que nous considèrerons dans ce cours) ne permettent d'obtenir des formules analytiques pour le calculs de quantiles et de fonctions de répartitions. Il faut alors avoir recours soit à des tables statistiques (non fournies dans ce document) soit à un logiciel de statistique. Notre choix s'est tourné vers le logiciel `R`.

Le tableau suivant résume les différentes lois de probabilités considérées dans ce cours de deuxième année ainsi que les instructions `R` permettant d'évaluer les quantiles et fonctions de répartitions associés à ces lois de probabilités.

lois de probabilités	raccourci `R`	quantile d'ordre $1-\alpha$	fonction de répartition en `u`
Normale $\mathcal{N}(\mu,\sigma)$	`norm`	`qnorm`$(1-\alpha,\mu,\sigma)$	`pnorm(u,`μ,σ`)`
Normale $\mathcal{N}(0,1)$	`norm`	`qnorm`$(1-\alpha)$	`pnorm(u)`
Chisquare $\chi^2(n)$	`chisq`	`qchisq`$(1-\alpha$`,n)`	`pchisq(u,n)`
Fisher $\mathcal{F}(n_1,n_2)$	`f`	`qf`$(1-\alpha,n_1,n_2)$	`pf(u,`n_1,n_2`)`
Student $\mathcal{S}t(n)$	`t`	`qt`$(1-\alpha$`,n)`	`pt(u,n)`

C.3 Calcul matriciel

Le tableau suivant résume les principales instructions liées au calcul matriciel utilisées.

Fonction `R`	Effet
`cbind`	permet de "coller" des réels ou vecteurs en colonne
`rbind`	permet de "coller" des réels ou vecteurs en ligne
`%*%`	produit matriciel (lorsdque celui-ci est posdsible)
`t()`	transposé d'un vecteur ou d'une matrice
`solve()`	calcul de la matrice inverse (lorsqu'elle existe) d'une matrice carrée

À nouveau la preuve par l'exemple :

```
> A<-cbind(c(1,3),c(-2,4))          # definition d'une matrice (2,2)
> A
     [,1] [,2]
[1,]    1   -2
[2,]    3    4
> A[1,1]                            # element de ligne 1, colonne 1 de A
[1] 1
> A[1,2]
[1] -2
> A[1,]                             # premiere ligne de A
```

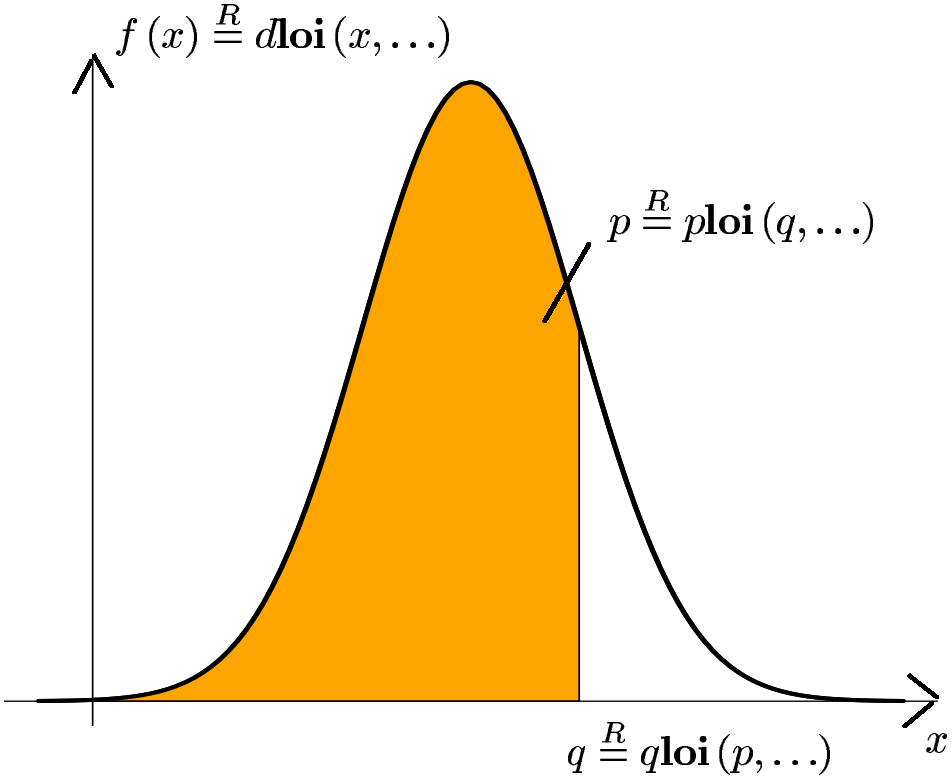


FIGURE C.1 – Si $X \rightsquigarrow \mathbf{loi}(\ldots)$ (v.a. continue), alors $f(x) \overset{R}{=} d\mathbf{loi}(x, \ldots)$ représente sa densité de probabilité, $p = F(q) = p[X \leq q] \overset{R}{=} p\mathbf{loi}(q, \ldots)$ sa fonction de répartition et $q = F^{-1}(p) \overset{R}{=} q\mathbf{loi}(p, \ldots)$ son quantile d'ordre p.

```
[1]  1 -2
> A[,2]                              # deuxieme colonne de A
[1] -2  4
> t(A)                               # matrice transposee de la matrice A
     [,1] [,2]
[1,]    1    3
[2,]   -2    4
> B<-cbind(c(1,-1),c(-1,0))
> B
     [,1] [,2]
[1,]    1   -1
[2,]   -1    0
> A * B                              # multiplication composantes par composantes
     [,1] [,2]
[1,]    1    2
[2,]   -3    0
> A %*% B                            # multiplication matricielle
     [,1] [,2]
[1,]    3   -1
[2,]   -1   -3
> C<-rbind(B,c(2,3))                 # une matrice a trois lignes et deux colonnes
> C
     [,1] [,2]
[1,]    1   -1
[2,]   -1    0
[3,]    2    3
> A %*%C                                          # la mult. matric. pas definie
Error in A %*% C : non-conformable arguments      # (2,2) * (3,2) -> impossible
> A%*%t(C)                                        # (2,2) * (2,3) -> matrice (2,3)!
     [,1] [,2] [,3]
[1,]    3   -1   -4
[2,]   -1   -3   18
> solve(A)                           # matrice inverse de la matrice A
     [,1] [,2]
[1,]  0.4  0.2
[2,] -0.3  0.1
> solve(A)%*%A                       # en theorie on obtient la matrice identite
              [,1]          [,2]
[1,] 1.000000e-00 -1.110223e-16
[2,] 1.110223e-16  1.000000e+00
```

C.4 Fonctions associées au traitement des modèles linéaires

Considérons dans cette section le jeu de données (de taille $n = 200$) servant de fil conducteur du cours où l'on tente d'expliquer linéairement le salaire individuel (variable Sal) en fonction du niveau d'expérience (variable IndExp) et/ou du niveau d'étude (variable IndEtu). La fonction `lm()` (abrégé de "linear models" en anglais) est une fonction interne au logiciel `R` qui implémente la méthode des moindres carrés ordinaires pour les modèles linéaires qu'ils soient simples ou multiples. L'argument principal de cette fonction est une formule qui précise le modèle linéaire envisagé. Voici quelques exemple de syntaxe de formules :

Modèle envisagé	Formule `R`
M1 : $Sal = \beta_0 + \beta_1 IndExp + U$	`Sal~IndExp`
M2 : $Sal = \beta_0 + \beta_1 IndExp + \beta_2 IndEtu + U$	`Sal~IndExp+IndEtu`
M3 :$Sal = \beta_1 IndExp + \beta_2 IndEtu + U$ (on force β_0 à 0)	`Sal~IndExp+IndEtu-1`

Ainsi, le vecteur des estimations des paramètres du modèle **M1** basé sur le jeu de données est obtenu comme suit :

```
> lm(Sal~IndExp)

Call:
lm(formula = Sal ~ IndExp)

Coefficients:
(Intercept)       IndExp
       1048         1488
```

Ainsi, le paramètre β_0 (resp. β_1) est approximativement estimé à 1048 (resp. 1488). Pour les modèles **M2** et **M3**, on obtient facilement :

```
 lm(Sal~IndExp+IndEtu)

Call:
lm(formula = Sal ~ IndExp + IndEtu)

Coefficients:
(Intercept)       IndExp       IndEtu
      928.9       1489.2        238.9

> lm(Sal~IndExp+IndEtu-1)

Call:
lm(formula = Sal ~ IndExp + IndEtu - 1)

Coefficients:
```

```
IndExp  IndEtu
2310.8   999.3
```

Tentons un exercice très intéressant et très pertinent avec le logiciel R, celui de retrouver les estimations des paramètres en utilisant la formule mathématique (et donc le calcul matriciel de R). On se propose de le faire uniquement pour le modèle **M2** :

```
> x<-cbind(1,IndExp,IndEtu)
> solve(t(x)%*%x) %*% t(x) %*% Sal

             [,1]
         928.9317
  IndExp 1489.2292
  IndEtu  238.9072
```

Et voilà tout est dit, on peut faire confiance à R puisqu'il implémente correctement la méthode des moindres carrés ordinaires ! Précisons pour terminer une dernière commande extrêmement intéressante pour les problèmes de régression qui propose un résumé des calculs liés au jeu de données, il s'agit de la fonction `summary()` que l'on va s'efforcer de commenter en précisant ce à quoi correspondent la plupart des résultats sans les vérifier :

```
> summary(lm(Sal~IndExp+IndEtu))           # exemple dans le cas du modele (M2)

Call:
lm(formula = Sal ~ IndExp + IndEtu)

Residuals:
    Min      1Q  Median      3Q     Max
-539.07 -265.83    1.85  273.36  644.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   928.93      55.56  16.720  < 2e-16 ***
IndExp       1489.23      73.94  20.140  < 2e-16 ***
IndEtu        238.91      71.25   3.353 0.000959 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 309.8 on 197 degrees of freedom
Multiple R-Squared: 0.6787,     Adjusted R-squared: 0.6755
F-statistic: 208.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

En notant plus simplement $\boldsymbol{y} = \boldsymbol{Sal}$ et $\underline{\boldsymbol{x}} = (\mathbf{1}, \boldsymbol{IndExp}, \boldsymbol{IndEtu})$ le tableau suivant précise les correspondances avec le langage mathématique :

Dénomination `R`	Formulation mathématique
colonne `Estimate`	estimations $\widehat{\beta}_i(\boldsymbol{y}\|\underline{\boldsymbol{x}})$ pour $i = 0, 1, 2$
colonne `Std. Error`	écarts-types estimés des estimateurs de β_i : $\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{y}\|\underline{\boldsymbol{x}})$ pour $i = 0, 1, 2$
colonne `t value`	statistique du test $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$
	de significativité locale, i.e. $\delta_0(\boldsymbol{y}\|\underline{\boldsymbol{x}}) = \dfrac{\widehat{\beta}_i(\boldsymbol{y}\|\underline{\boldsymbol{x}})}{\widehat{\sigma}_{\widehat{\beta}_i}(\boldsymbol{y}\|\underline{\boldsymbol{x}})}$ pour $i = 0, 1, 2$
colonne `Pr(>|t|)`	p-valeur du test de significativité locale
	du $i^{\text{ème}}$ régresseur pour $i = 0, 1, 2$
`Residual standard error`	estimation du niveau de bruit : $\widehat{\sigma}(\boldsymbol{y}\|\underline{\boldsymbol{x}})$
`Multiple R-Squared`	coefficient de détermination multiple
`Adjusted R-squared`	coefficient de détermination multiple ajusté
	(en fonction du nombre de régresseurs)
`F-statistic`	valeur de la stat. dite de Fisher du test de significativité globale,
	i.e. du test d'hypothèses
	$H_0 : (\beta_1, \beta_2) = (0, 0)$ contre $H_1 : \exists \in \{1, 2\}$ tq $\beta_i \neq 0$
`p-value` (ici $< 2.2e-16$)	correspond à la p-valeur du test de significativité globale.

Vous l'aurez compris, la fonction `summary(lm())` permet de répondre (en partie) au premier objectif de la modélisation : tel ou tel régresseur a-t-il de l'information dans l'explication de la variable d'intérêt. Précisons à présent une dernière commande très utile, permettant d'utiliser le modèle estimé pour effectuer une prévision. Dans l'exemple du modèle **M2**, les résultats étant satisaisants, on se propose d'utiliser le modèle estimé pour prévoir le salaire d'un individu qui aurait par exemple un niveau d'expérience professionnelle à 0.75 et un niveau d'étude à disons 0.6 et par la même occasion de calculer un intervalle de prévision (de cette valeur prévue) à disons 95% :

```
> xTau<-data.frame(IndExp=0.75,IndEtu=0.6)
> predict.lm(lm(Sal~IndExp+IndEtu),xTau)
[1] 2189.198                          ## la prevision de xTau
> predict.lm(lm(Sal~IndExp+IndEtu),xTau,interval="prediction")
          fit      lwr      upr
[1,] 2189.198 1575.382 2803.014     ## avec son intervalle de prevision
```

Ainsi, le salaire prévu pour l'individu considéré est (approximativement) de 2189.2 euros, et l'intervalle de prévision associé est $[1575.4, 2803]$.

Annexe D

Algèbre linéaire et vecteurs aléatoires

D.1 Rappels sur les matrices

D.1.1 Notation et résultats généraux

1. Une matrice de taille $(m \times n)$, notée $\underline{\boldsymbol{x}}$, est représentée par $(x_{ij})_{\substack{i=1,\cdots,m \\ j=1,\cdots,n}}$ où $x_{ij} = (\underline{\boldsymbol{x}})_{ij}$ est l'élément de $\underline{\boldsymbol{x}}$ en $i^{ème}$ ligne et $j^{ème}$ colonne.
2. Un vecteur de $\mathbb{R}^n$, noté $\boldsymbol{x}$, est représenté par $(x_1, \cdots, x_n) = (x_i)_{i=1,\cdots,m}$ où x_i est sa $i^{ème}$ coordonnée. On convient aussi de représenter un vecteur de $\mathbb{R}^n$ comme une matrice de taille $(n \times 1)$.
3. Lorsque $\underline{\boldsymbol{x}}$ désigne une matrice de taille $(m \times n)$, $\underline{\boldsymbol{x}}^T$ désigne la *matrice transposée* de $\underline{\boldsymbol{x}}$ de taille $(n \times m)$ définie par
$$\left(\underline{\boldsymbol{x}}^T\right)_{ij} = x_{ji}, \forall i \in \{1, \cdots, n\}, \forall j \in \{1, \cdots, m\}.$$
4. Nous adopterons les notations suivantes : une matrice $\underline{\boldsymbol{x}}$ de taille $(m \times n)$ s'écrit aussi
$$\underline{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x}_1{}^T \\ \boldsymbol{x}_2{}^T \\ \vdots \\ \boldsymbol{x}_m{}^T \end{pmatrix} = \left(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \cdots, \boldsymbol{x}_{(n)}\right) \text{ où } \boldsymbol{x}_i{}^T = (x_{i1}, x_{i2}, \cdots, x_{in}) \text{ et } \boldsymbol{x}_{(j)} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{pmatrix}.$$
5. **Multiplication :** si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont deux vecteurs de $\mathbb{R}^n$ alors
$$\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i = \boldsymbol{y}^T\boldsymbol{x}.$$
Cette quantité définit le produit scalaire dans $\mathbb{R}^n$ de $\boldsymbol{x}$ et $\boldsymbol{y}$. Il sera noté $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ tandis que la norme d'un vecteur $\boldsymbol{x}$ sera notée $\|\boldsymbol{x}\|$. On rappelle que
$\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$.

Si $\underline{\boldsymbol{x}}$ et $\underline{\boldsymbol{y}}$ sont deux matrices de tailles respectives $(m \times n)$ et $(n \times p)$ alors $\underline{\boldsymbol{x}}\underline{\boldsymbol{y}}$ est la matrice de taille $(m \times p)$ définie par

$$\underline{\boldsymbol{x}}\underline{\boldsymbol{y}} = \left(\boldsymbol{x}_i^T \boldsymbol{y}_{(j)}\right)_{\substack{i=1,\cdots,m \\ j=1,\cdots,p}}.$$

6. **Transposition :** si $\underline{\boldsymbol{x}}$, $\underline{\boldsymbol{y}}$, $\underline{\boldsymbol{u}}$ et $\underline{\boldsymbol{v}}$ sont des matrices de tailles respectives $(n \times n)$, $(n \times n)$, $(m \times n)$ et $(n \times p)$ alors

$$\left(\underline{\boldsymbol{x}}^T\right)^T = \underline{\boldsymbol{x}} \ , \ \left(\underline{\boldsymbol{x}} + \underline{\boldsymbol{y}}\right)^T = \underline{\boldsymbol{x}}^T + \underline{\boldsymbol{y}}^T \text{ et } \left(\underline{\boldsymbol{u}}\underline{\boldsymbol{v}}\right)^T = \underline{\boldsymbol{v}}^T \underline{\boldsymbol{u}}^T.$$

7. **Trace :** la trace $tr\left(\underline{\boldsymbol{x}}\right)$ d'une matrice carrée $\underline{\boldsymbol{x}}$ de taille $(n \times n)$ est définie par

$$tr\left(\underline{\boldsymbol{x}}\right) = \sum_{i=1}^{n} x_{ii}.$$

Si $\underline{\boldsymbol{x}}$, $\underline{\boldsymbol{y}}$, $\underline{\boldsymbol{u}}$ et $\underline{\boldsymbol{v}}$ sont des matrices de tailles respectives $(n \times n)$, $(n \times n)$, $(m \times n)$ et $(n \times p)$, et α un réel, alors

$$tr\left(\underline{\boldsymbol{x}} + \underline{\boldsymbol{y}}\right) = tr\left(\underline{\boldsymbol{x}}\right) + tr\left(\underline{\boldsymbol{y}}\right), \ tr\left(\alpha\underline{\boldsymbol{x}}\right) = \alpha tr\left(\underline{\boldsymbol{x}}\right) \text{ et } tr\left(\underline{\boldsymbol{u}}\underline{\boldsymbol{v}}\right) = tr\left(\underline{\boldsymbol{v}}\underline{\boldsymbol{u}}\right).$$

D.1.2 Matrice par bloc

Une matrice écrite en termes de sous-matrices, appelées blocs, est appelée matrice par bloc. Par exemple, si $\underline{\boldsymbol{x}}_{11}$, $\underline{\boldsymbol{x}}_{12}$, $\underline{\boldsymbol{x}}_{21}$ et $\underline{\boldsymbol{x}}_{22}$ sont des matrices de tailles respectives $(m_1 \times n_1)$, $(m_1 \times n_2)$, $(m_2 \times n_1)$ et $(m_2 \times n_2)$ alors la matrice

$$\underline{\boldsymbol{x}} = \begin{pmatrix} \underline{\boldsymbol{x}}_{11} & \underline{\boldsymbol{x}}_{12} \\ \underline{\boldsymbol{x}}_{21} & \underline{\boldsymbol{x}}_{22} \end{pmatrix}$$

est une matrice par bloc. En posant $m = m_1 + m_2$ et $n = n_1 + n_2$, cette matrice de taille $(m \times n)$ sera appelée matrice par bloc de taille $((m_1, m_2) \times (n_1, n_2))$.

En particulier, les règles usuelles de multiplication de matrices se généralisent aux matrices par bloc lorsque les blocs sont vus comme des éléments de matrices. Par exemple, si $\underline{\boldsymbol{x}}$ et $\underline{\boldsymbol{y}}$ sont deux matrices par bloc de tailles respectives $((m_1, m_2) \times (n_1, n_2))$ et $((n_1, n_2) \times (p_1, p_2))$

$$\underline{\boldsymbol{x}}\underline{\boldsymbol{y}} = \begin{pmatrix} \underline{\boldsymbol{x}}_{11} & \underline{\boldsymbol{x}}_{12} \\ \underline{\boldsymbol{x}}_{21} & \underline{\boldsymbol{x}}_{22} \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{y}}_{11} & \underline{\boldsymbol{y}}_{12} \\ \underline{\boldsymbol{y}}_{21} & \underline{\boldsymbol{y}}_{22} \end{pmatrix} = \begin{pmatrix} \left(\underline{\boldsymbol{x}}_{11}\underline{\boldsymbol{y}}_{11} + \underline{\boldsymbol{x}}_{12}\underline{\boldsymbol{y}}_{21}\right) & \left(\underline{\boldsymbol{x}}_{11}\underline{\boldsymbol{y}}_{12} + \underline{\boldsymbol{x}}_{12}\underline{\boldsymbol{y}}_{22}\right) \\ \left(\underline{\boldsymbol{x}}_{21}\underline{\boldsymbol{y}}_{11} + \underline{\boldsymbol{x}}_{22}\underline{\boldsymbol{y}}_{21}\right) & \left(\underline{\boldsymbol{x}}_{21}\underline{\boldsymbol{y}}_{12} + \underline{\boldsymbol{x}}_{22}\underline{\boldsymbol{y}}_{22}\right) \end{pmatrix}.$$

qui est une matrice par bloc de taille $((m_1, m_2) \times (p_1, p_2))$.

Si $\underline{\boldsymbol{x}}$ est une matrice par bloc de taille $((m_1, m_2) \times (n_1, n_2))$ alors

$$\underline{\boldsymbol{x}}^T = \begin{pmatrix} \underline{\boldsymbol{x}}_{11} & \underline{\boldsymbol{x}}_{12} \\ \underline{\boldsymbol{x}}_{21} & \underline{\boldsymbol{x}}_{22} \end{pmatrix}^T = \begin{pmatrix} \underline{\boldsymbol{x}}_{11}^T & \underline{\boldsymbol{x}}_{21}^T \\ \underline{\boldsymbol{x}}_{12}^T & \underline{\boldsymbol{x}}_{22}^T \end{pmatrix}$$

et si toutes les matrices nécessaires ci-dessous sont inversibles alors

$$\underline{\boldsymbol{x}}^{-1} = \begin{pmatrix} \underline{\boldsymbol{x}}_{11} & \underline{\boldsymbol{x}}_{12} \\ \underline{\boldsymbol{x}}_{21} & \underline{\boldsymbol{x}}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \underline{\boldsymbol{x}}^{11} & \underline{\boldsymbol{x}}^{12} \\ \underline{\boldsymbol{x}}^{21} & \underline{\boldsymbol{x}}^{22} \end{pmatrix}$$

avec

$$\begin{aligned}
\underline{\boldsymbol{x}}^{11} &= \left(\underline{\boldsymbol{x}}_{11} - \underline{\boldsymbol{x}}_{12}\underline{\boldsymbol{x}}_{22}^{-1}\underline{\boldsymbol{x}}_{21}\right)^{-1} \\
\underline{\boldsymbol{x}}^{22} &= \left(\underline{\boldsymbol{x}}_{22} - \underline{\boldsymbol{x}}_{21}\underline{\boldsymbol{x}}_{11}^{-1}\underline{\boldsymbol{x}}_{12}\right)^{-1} \\
\underline{\boldsymbol{x}}^{12} &= -\underline{\boldsymbol{x}}^{11}\underline{\boldsymbol{x}}_{12}\underline{\boldsymbol{x}}_{22}^{-1} = -\underline{\boldsymbol{x}}_{11}^{-1}\underline{\boldsymbol{x}}_{12}\underline{\boldsymbol{x}}^{22} \\
\underline{\boldsymbol{x}}^{21} &= -\underline{\boldsymbol{x}}_{22}^{-1}\underline{\boldsymbol{x}}_{21}\underline{\boldsymbol{x}}^{11} = -\underline{\boldsymbol{x}}^{22}\underline{\boldsymbol{x}}_{21}\underline{\boldsymbol{x}}_{11}^{-1}T.
\end{aligned}$$

D.1.3 Matrice de projection orthogonale

Rappelons au préalable quelques définitions.

a) On dit qu'une matrice carrée $\underline{\boldsymbol{x}}$ est *symétrique* si $\underline{\boldsymbol{x}}^T = \underline{\boldsymbol{x}}$.
b) On dit qu'une matrice carrée $\underline{\boldsymbol{x}}$ est *idempotente* si $\underline{\boldsymbol{x}}^2 = \underline{\boldsymbol{x}}$.

Soit $\boldsymbol{x}_{(j)}, j = 1, \cdots, p$, p vecteurs de $\mathbb{R}^n$. Soit $Q \subset \{1, \cdots, p\}$ un ensemble d'indices à q éléments, notons $\mathcal{L}_Q$ l'espace vectoriel engendré par $\boldsymbol{x}_{(j)}, j \in Q$ (i.e. l'ensemble de toutes les combinaisons linéaires de ces vecteurs). Le projeté orthogonal $\widehat{\boldsymbol{y}}_Q$ d'un vecteur $\boldsymbol{y}$ sur l'espace $\mathcal{L}_Q$ est obtenu après multiplication par la matrice $\underline{\boldsymbol{\mathcal{P}}}_Q$, dite de projection orthogonale sur $\mathcal{L}_Q$. Rappelons quelques résultats bien connus sur ce type de matrice. Auparavant, introduisons la matrice $\underline{\boldsymbol{x}}_Q = \left(\boldsymbol{x}_{(j_1)}, \boldsymbol{x}_{(j_2)}, \cdots, \boldsymbol{x}_{(j_q)}\right)$ en posant $Q = \{j_1, j_2, \cdots, j_q\}$ et supposons que les $\boldsymbol{x}_{(j_1)}, \boldsymbol{x}_{(j_2)}, \cdots, \boldsymbol{x}_{(j_q)}$ soient linéairement indépendants (non colinéaires).

1. Comme toute matrice de projection, $\underline{\boldsymbol{\mathcal{P}}}_Q$ est idempotente. De plus, $\underline{\boldsymbol{\mathcal{P}}}_Q$ est symétrique. En particulier, on a
$$\underline{\boldsymbol{\mathcal{P}}}_Q\underline{\boldsymbol{x}}_Q = \underline{\boldsymbol{x}}_Q.$$
2. Si $Q \subset Q' \subset \{1, \cdots, p\}$ alors
$$\underline{\boldsymbol{\mathcal{P}}}_Q\underline{\boldsymbol{\mathcal{P}}}_{Q'} = \underline{\boldsymbol{\mathcal{P}}}_{Q'}\underline{\boldsymbol{\mathcal{P}}}_Q = \underline{\boldsymbol{\mathcal{P}}}_Q.$$
En particulier,
$$\underline{\boldsymbol{\mathcal{P}}}_{Q'}\underline{\boldsymbol{x}}_Q = \underline{\boldsymbol{x}}_Q.$$
3. La matrice $\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp} = \underline{\boldsymbol{I}}_n - \underline{\boldsymbol{\mathcal{P}}}_Q$ est aussi une matrice de projection orthogonale sur $\mathcal{L}_Q^{\perp}$ représentant l'orthogonal de $\mathcal{L}_Q$. Il est alors immédiat que
$$\underline{\boldsymbol{\mathcal{P}}}_Q\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp} = \underline{\boldsymbol{\mathcal{P}}}_Q^{\perp}\underline{\boldsymbol{\mathcal{P}}}_Q = \underline{\mathbf{0}}.$$
En particulier,
$$\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp}\underline{\boldsymbol{x}}_Q = \underline{\mathbf{0}}.$$
4. Géométriquement et en utilisant des arguments d'orthogonalité, on montre que
$$\boxed{\underline{\boldsymbol{\mathcal{P}}}_Q = \underline{\boldsymbol{x}}_Q\left(\underline{\boldsymbol{x}}_Q{}^T\underline{\boldsymbol{x}}_Q\right)^{-1}\underline{\boldsymbol{x}}_Q{}^T.}$$

D.2 Complément sur les vecteurs aléatoires

Soit $\boldsymbol{X}$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$. On définit son espérance $\boldsymbol{E}(\boldsymbol{X})$ et sa matrice de covariances $\underline{\boldsymbol{V}}(\boldsymbol{X})$ par

$$\boldsymbol{E}(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} \text{ et } \underline{\boldsymbol{V}}(\boldsymbol{X}) = \boldsymbol{E}\left((\boldsymbol{X} - \boldsymbol{E}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{E}(\boldsymbol{X}))^T\right) = (Cov(X_i, X_j))_{i,j=1,\cdots,n}.$$

De plus, on dit que deux vecteurs $\boldsymbol{X}$ et $\boldsymbol{Y}$ (pas nécessairement de même taille) sont *indépendants* (en probabilité) si, pour tous i, j, X_i et Y_j sont indépendants (en probabilité).

Soit $\boldsymbol{\theta}$ un paramètre de $\mathbb{R}^n$ et $\widehat{\boldsymbol{\theta}}$ un estimateur de $\boldsymbol{\theta}$. On dit que

– $\widehat{\boldsymbol{\theta}}$ est un *estimateur sans biais* de $\boldsymbol{\theta}$ si le biais $\boldsymbol{E}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ est le vecteur nul, et

– $\widehat{\boldsymbol{\theta}}$ est un *estimateur convergent en moyenne quadratique* de $\boldsymbol{\theta}$ si la matrice carrée $E\left(\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T\right)$ converge vers la matrice nulle où on dit qu'une matrice $\underline{\boldsymbol{x}}_n$ converge vers la matrice $\underline{\boldsymbol{y}}$ quand n tend vers $+\infty$ si

$$\lim_{n\to+\infty} \left\|\underline{\boldsymbol{x}}_n - \underline{\boldsymbol{y}}\right\| = 0 \text{ avec } \|\underline{\boldsymbol{z}}\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}^2.$$

En particulier, si $\widehat{\boldsymbol{\theta}}$ est un estimateur sans biais de $\boldsymbol{\theta}$ et de matrice de covariances tendant vers la matrice nulle lorsque T tend vers $+\infty$, alors $\widehat{\boldsymbol{\theta}}$ est un estimateur convergent en moyenne quadratique de $\boldsymbol{\theta}$.

D.3 Rappels sur les vecteurs gaussiens

On appelle vecteur aléatoire toute variable aléatoire à valeurs dans $\mathbb{R}^n$. La notion de vecteur (aléatoire) gaussien généralise alors la notion de variable aléatoire réelle ayant une loi de probabilité normale.

Soit $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T$ un vecteur aléatoire. On dit que $\boldsymbol{X}$ est gaussien si toute combinaison linéaire de ses coordonnées $X_1, X_2, \cdots, X_n$ suit une loi normale :

$$\forall \alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}, Z = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n \rightsquigarrow \mathcal{N}\left(E(Z), \sqrt{Var(Z)}\right).$$

Dans ce cas, les paramètres de la loi de probabilité de $\boldsymbol{X}$ sont son espérance $\boldsymbol{E}(\boldsymbol{X})$ et sa matrice de covariances $\underline{\boldsymbol{V}}(\boldsymbol{X})$. On notera alors

$$\boldsymbol{X} \rightsquigarrow \mathcal{N}_n\left(E(\boldsymbol{X}), \underline{\boldsymbol{V}}(\boldsymbol{X})\right).$$

Attention : nous attirons l'attention du lecteur sur la convention prise pour décrire les paramètres de la loi normale. En dimension 1, il est d'usage d'utiliser les paramètres espérance et *écart-type* alors

qu'en dimension supérieure on préfèrera l'espérance et *la matrice de covariances.*

On peut alors remarquer que si $X_1, X_2, \cdots, X_n$ sont n variables aléatoires réelles (mutuellement) indépendantes et identiquement distribuées comme une loi normale $\mathcal{N}\left(0, \sigma^2\right)$, le vecteur aléatoire $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T$ est un vecteur gaussien d'espérance $\mathbf{0}$ et de matrice de covariances $\sigma^2 \underline{\boldsymbol{I}}_n$. N'oubliant pas que la somme de deux variables aléatoires normales n'est pas nécessairement normale, il apparaît que la définition précédente de vecteur gaussien est assez restrictive.

Nous exprimons maintenant la densité de probabilité $f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)$ du vecteur aléatoire $\boldsymbol{X}$ d'espérance $\boldsymbol{\mu}$ et de matrice de covariances $\underline{\boldsymbol{\Gamma}}$, qui s'interprète comme l'intensité de probabilité que $\boldsymbol{X}$ prenne une valeur autour du vecteur $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$:

$$\boxed{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n) = \frac{\sqrt{\det\left(\underline{\boldsymbol{\Gamma}}^{-1}\right)}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \underline{\boldsymbol{\Gamma}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).}$$

Soit un vecteur $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T \rightsquigarrow \mathcal{N}_n(\boldsymbol{\mu}, \underline{\boldsymbol{\Gamma}})$. Enonçons alors quelques propriétés bien connues :

1. Tout vecteur marginal du vecteur gaussien $\boldsymbol{X}$ est encore un vecteur gaussien. Par exemple, $(X_1, X_2, X_n)^T$ est un vecteur gaussien.
2. Toute transformation affine du vecteur gaussien $\boldsymbol{X}$ est encore un vecteur gaussien. Si $\underline{\boldsymbol{b}}$ est une matrice de taille $(m \times n)$ et $\boldsymbol{c}$ est un vecteur de $\mathbb{R}^n$ alors
$$\boldsymbol{Z} = \underline{\boldsymbol{b}}\boldsymbol{X} + \boldsymbol{c} \rightsquigarrow \mathcal{N}_m\left(\underline{\boldsymbol{b}}\boldsymbol{\mu} + \boldsymbol{c}, \underline{\boldsymbol{b}}\underline{\boldsymbol{\Gamma}}\underline{\boldsymbol{b}}^T\right).$$
3. Si la matrice de covariances est une matrice diagonale alors les variables aléatoires $X_1, X_2, \cdots, X_n$ sont (mutuellement) indépendantes.
4. Soit $\boldsymbol{X} \rightsquigarrow \mathcal{N}_n(\mathbf{0}, \underline{\boldsymbol{I}}_n)$, $\boldsymbol{Z}_1 = \underline{\boldsymbol{b}}_1\boldsymbol{X}$ et $\boldsymbol{Z}_2 = \underline{\boldsymbol{b}}_2\boldsymbol{X}$ (avec $\underline{\boldsymbol{b}}_1$ et $\underline{\boldsymbol{b}}_2$ deux matrices de tailles respectives $(m_1 \times n)$ et $(m_2 \times n)$). On sait d'après ce qui précède que le vecteur
$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} = \begin{pmatrix} \underline{\boldsymbol{b}}_1 \\ \underline{\boldsymbol{b}}_2 \end{pmatrix} \boldsymbol{X} \rightsquigarrow \mathcal{N}_{m_1+m_2}\left(\mathbf{0}, \begin{pmatrix} \underline{\boldsymbol{b}}_1{}^T\underline{\boldsymbol{b}}_1 & \underline{\boldsymbol{b}}_1{}^T\underline{\boldsymbol{b}}_2 \\ \underline{\boldsymbol{b}}_2{}^T\underline{\boldsymbol{b}}_1 & \underline{\boldsymbol{b}}_2{}^T\underline{\boldsymbol{b}}_2 \end{pmatrix}\right).$$
Ainsi, les vecteurs $\boldsymbol{Z}_1$ et $\boldsymbol{Z}_2$ sont indépendants si $\underline{\boldsymbol{b}}_2\underline{\boldsymbol{b}}_1{}^T = 0$ (exprimant que les covariances entre coordonnées de $\boldsymbol{Z}_1$ et de $\boldsymbol{Z}_2$ sont nulles).

D.4 Résultats sur la loi du khi-deux

Les résultats proposés dans cette section sont des conséquences des théorèmes de Cochran et de Craig-Lancaster.

Si

(i) $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T \rightsquigarrow \mathcal{N}_n(\mathbf{0}, \underline{\boldsymbol{I}}_n)$,

(ii) $\underline{\boldsymbol{h}}$ est une matrice symétrique et idempotente,

alors

$$\boxed{\|\underline{\boldsymbol{h}}\boldsymbol{X}\|^2 = \boldsymbol{X}^T\underline{\boldsymbol{h}}\boldsymbol{X} \rightsquigarrow \chi^2\left(tr\left(\underline{\boldsymbol{h}}\right)\right).}$$

De plus, si

(iii) $\underline{\boldsymbol{h}}'$ est une matrice symétrique et idempotente telle que $\underline{\boldsymbol{h}}\underline{\boldsymbol{h}}' = 0$,

alors

$$\boxed{\|\underline{\boldsymbol{h}}\boldsymbol{X}\|^2 = \boldsymbol{X}^T\underline{\boldsymbol{h}}\boldsymbol{X} \text{ et } \left\|\underline{\boldsymbol{h}}'\boldsymbol{X}\right\|^2 = \boldsymbol{X}^T\underline{\boldsymbol{h}}'\boldsymbol{X} \text{ sont indépendantes.}}$$

La démonstration de ces résultats s'appuie sur la diagonalisation de matrice symétrique et sur le fait qu'une matrice idempotente ne peut avoir comme valeur propre que 0 et 1.

D.5 Lois de Student et de Fisher-Snedecor

- Soit $X \rightsquigarrow \mathcal{N}(0,1)$ et $Y \rightsquigarrow \chi^2(\nu)$ deux v.a. indépendantes alors la v.a. $\dfrac{X}{Y/\nu}$ suit une loi standard appelée loi de Student à ν degrés de liberté, et on note alors

$$\boxed{\frac{X}{Y/\nu} \rightsquigarrow \mathcal{S}t\left(\nu\right).}$$

- Soit $X_1 \rightsquigarrow \chi^2(\nu_1)$ et $X_2 \rightsquigarrow \chi^2(\nu_2)$ deux v.a. indépendantes alors la v.a. $\dfrac{X_1/\nu_1}{X_2/\nu_2}$ suit une loi standard appelée loi de Fisher-Snedecor à ν_1 et ν_2 degrés de liberté, et on note alors

$$\boxed{\frac{X_1/\nu_1}{X_2/\nu_2} \rightsquigarrow \mathcal{F}\left(\nu_1,\nu_2\right).}$$

D.6 Théorème central limite (TCL)

Nous présentons ici une version (simple) du théorème central limite, résultat fondamental en statistique (donc largement utilisé en économétrie) démontré il y a déjà plus de deux siècles. Il exprime simplement le fait que l'estimateur de la moyenne calculé à partir d'un échantillon de variables aléatoires indépendantes (et identiquement distribuées) est asymptotiquement distribué selon une loi Normale. Plus précisément :

Theorem 1 *Soient $Y_1, \ldots, Y_n$ n variables aléatoires indépendantes et identiquement distribuées de moyenne μ et de variance σ^2 alors*

$$\boxed{\frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).}$$

D.7 Résultats de robustesse

Dans cette section, on va présenter deux résultats de convergence associés communément en statistiques à des résultats de robustesse.

D.7.1 Comparaison entre une loi $\mathcal{S}t(n)$ et une loi $\mathcal{N}(0,1)$

La théorie des probabilités (i.e. l'**A.C.P.**) permet de montrer que :

$$\mathcal{S}t(n) \to \mathcal{N}(0,1), \text{ lorsque } n \to +\infty$$

La convergence en question étant une convergence en loi (i.e. convergence ponctuelle des fonctions de répartition ou si elles existent des densités de probabilités). C'est un concept que nous ne développerons pas ici mais que nous pouvons illustrer très facilement

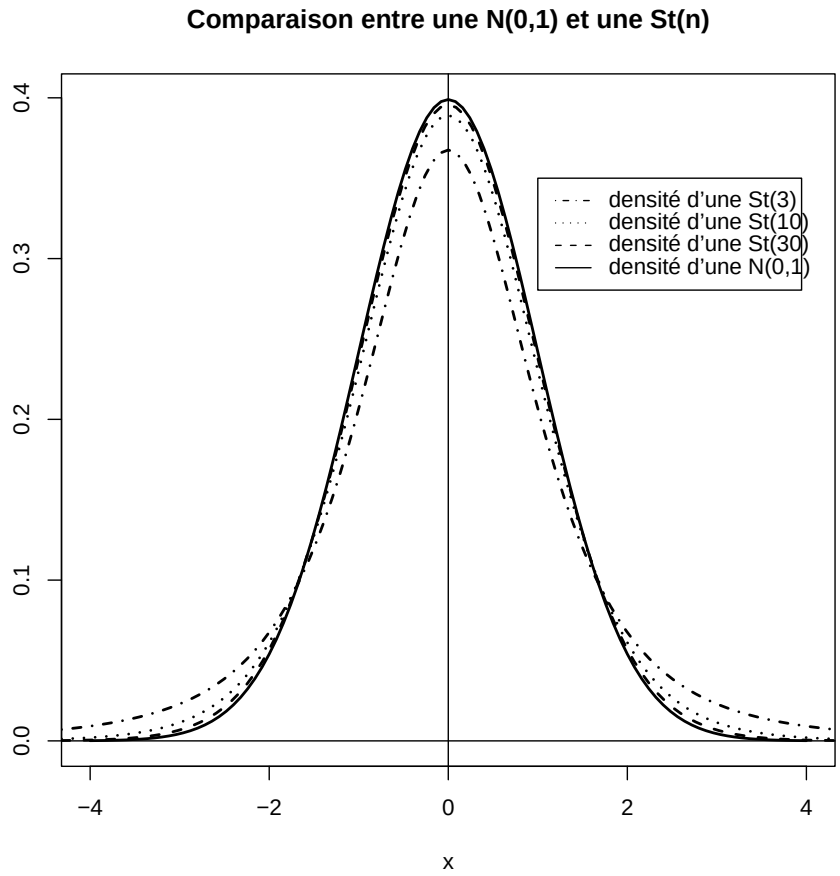


FIGURE D.1 – Comparaison entre une loi $\mathcal{S}t(n)$ (pour $n = 3, 10, 30$) et une loi $\mathcal{N}(0,1)$

Comme on peut le voir sur la précédente figure, plus n devient grand et plus la densité de probabilité d'une $\mathcal{S}t(n)$ se rapproche de celle d'une $\mathcal{N}(0,1)$. Ceci a la conséquence importante que lorsque n est grand les quantiles des deux lois seront proches (idem si on cherche à évaluer une surface). Par exemple, observez comment évolue le quantile d'une loi de Student d'ordre 95% :

quantile d'ordre 95 % d'une	$\mathcal{S}t(3)$	$\mathcal{S}t(10)$	$\mathcal{S}t(30)$	$\mathcal{S}t(100)$	$\mathcal{N}(0,1)$
expression `R`	`qt(0.95,3)`	`qt(0.95,10)`	`qt(0.95,30)`	`qt(0.95,100)`	`qnorm(0.95)`
valeur	$\simeq 2.353$	$\simeq 1.812$	$\simeq 1.697$	$\simeq 1.660$	$\simeq 1.645$

D.7.2 Comparaison entre une loi de Fisher et une loi Khi-deux

On peut également montrer que :

$$\mathcal{F}(p,n) \to p \times \chi^2(p), \text{ lorsque } n \to +\infty$$

La convergence étant à nouveau une convergence en loi. Nous pouvons illustrer à nouveau ce résultat :

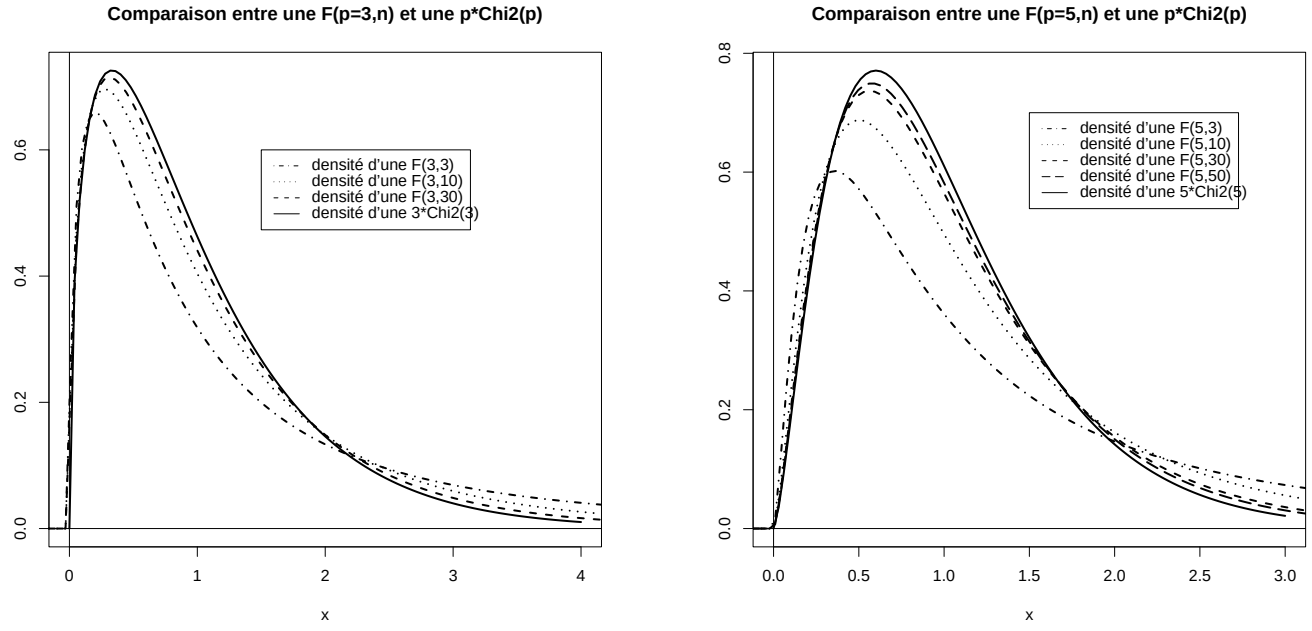


FIGURE D.2 – Comparaison entre une loi $\mathcal{F}(p,n)$ (pour $n = 3, 10, 30$) et une loi $p \times \chi^2(p)$ pour deux valeurs de p : $p = 3$ (figure de gauche) et $p = 5$ (figure de droite).

Ainsi, comme le montrent les figures plus n augmente et plus la densité de probabilité d'une $\mathcal{F}(p,n)$ se rapproche de celle d'une $p \times \chi^2(p)$. On pourra noter (et ceci est tout de même important en pratique) que l'approximation est d'autant meilleure que p est petit devant n. Illustrons les conséquences de cette convergence avec un petit calcul de quantiles d'ordre 95% (par exemple) :

Exemple pour $p = 3$					
quantile d'ordre 95 % d'une	$\mathcal{F}(3,3)$	$\mathcal{F}(3,10)$	$\mathcal{F}(3,30)$	$\mathcal{F}(3,100)$	$3 \times \chi^2(3)$
expression `R`	`qf(0.95,3,3)`	`qf(0.95,3,10)`	`qf(0.95,3,30)`	`qf(0.95,3,100)`	`qchisq(0.95,3)/3`
valeur	$\simeq 9.277$	$\simeq 3.708$	$\simeq 2.922$	$\simeq 2.696$	$\simeq 2.604909$

Exemple pour $p = 5$					
quantile d'ordre 95 % d'une	$\mathcal{F}(5,3)$	$\mathcal{F}(5,10)$	$\mathcal{F}(5,30)$	$\mathcal{F}(5,100)$	$5 \times \chi^2(5)$
expression `R`	`qf(0.95,5,3)`	`qf(0.95,5,10)`	`qf(0.95,5,30)`	`qf(0.95,5,100)`	`qchisq(0.95,5)/5`
valeur	$\simeq 9.013$	$\simeq 3.326$	$\simeq 2.533$	$\simeq 2.306$	$\simeq 2.214$

Annexe E

Démonstrations des différents résultats du cours (via l'ACP)

Avant la détermination des estimateurs des paramètres de régression, proposons quelques notations utilisées par la suite. Soit Q un ensemble d'indices, $Q \subset \{0, 1, \ldots, p\}$ de cardinal q, $\mathcal{L}_Q$ l'espace vectoriel engendré par les q régresseurs notés $\underline{\boldsymbol{x}}_Q$. Notons encore $\underline{\boldsymbol{\mathcal{P}}}_Q$ (resp. $\underline{\boldsymbol{\mathcal{P}}}_{Q^c}$) la matrice de projection orthogonale sur $\mathcal{L}_Q$ (resp. $\mathcal{L}_{Q^c}$). En consultant l'Annexe D concernant les rappels sur les matrices, on peut proposer les expressions de ces matrices

$$\underline{\boldsymbol{\mathcal{P}}}_Q = \underline{\boldsymbol{x}}_Q \left(\underline{\boldsymbol{x}}_Q{}^T \underline{\boldsymbol{x}}_Q\right)^{-1} \underline{\boldsymbol{x}}_Q{}^T \text{ (resp. } \underline{\boldsymbol{\mathcal{P}}}_{Q^c} = \underline{\boldsymbol{x}}_{Q^c} \left(\underline{\boldsymbol{x}}_{Q^c}{}^T \underline{\boldsymbol{x}}_{Q^c}\right)^{-1} \underline{\boldsymbol{x}}_{Q^c}{}^T)$$

En particulier si $Q = \{0, 1, \ldots, p\}$, on a $\underline{\boldsymbol{x}}_Q = \underline{\boldsymbol{x}}$ et $\underline{\boldsymbol{x}}_{Q^c}$ ne pourra être considéré. De plus, pour alléger les notations on notera alors $\mathcal{L}$ l'espace vectoriel engendré par les $p+1$ régresseurs, et $\underline{\boldsymbol{\mathcal{P}}}$ la matrice de projection orthogonale sur $\mathcal{L}$. De plus, pour alléger les notations, $\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp}$ (resp. $\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}$) les matrices de projection orthogonale sur les espaces vectoriels $\mathcal{L}_Q^{\perp}$ (resp $\mathcal{L}_{Q^c}^{\perp}$) qui s'expriment par

$$\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp} = \underline{\boldsymbol{I}}_n - \underline{\boldsymbol{\mathcal{P}}}_Q \text{ (resp. } \underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp} = \underline{\boldsymbol{I}}_n - \underline{\boldsymbol{\mathcal{P}}}_{Q^c})$$

Attention : dans les différentes preuves qui vont suivre, nous avons pris la convention de ne pas expliciter la dépendance en fonction de $\boldsymbol{Y}$ et $\underline{\boldsymbol{x}}$ des estimateurs, et des statistiques de tests. Ainsi, par exemple $\widehat{\beta}(\boldsymbol{Y}|\underline{\boldsymbol{x}})$ sera noté plus simplement $\widehat{\beta}$.

E.1 Détermination des estimateurs MCO

Précisons tout d'abord un abus de notation. Le vecteur $\boldsymbol{\beta}$ correspond aux vraies valeurs (dites théoriques) des paramètres du modèle. Par la suite, cette même notation spécifiera aussi le vecteur des paramètres utilisé comme une variable. En utilisant cette convention, l'estimateur des moindres carrés $\widehat{\boldsymbol{\beta}}$ est la solution en $\boldsymbol{\beta}$ qui minimise

$$\|\boldsymbol{U}\|^2 = \boldsymbol{U}^T\boldsymbol{U} = \sum_{t=1}^{n} U_t^2 = \sum_{t=1}^{n} \left(Y_t - \left(\beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_p x_{tp}\right)\right)^2.$$

La solution à un tel problème peut être obtenue de manière géométrique. En effet, la détermination du projeté orthogonal $\widehat{\boldsymbol{Y}} = \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{Y}$ sur $\mathcal{L}$ permet d'obtenir une erreur $\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ de norme minimale dans l'équation précédente. Ceci se traduit en terme d'orthogonalité par

$$\left\langle \widehat{\boldsymbol{U}}, \boldsymbol{x}_{(j)} \right\rangle = \left\langle \boldsymbol{Y} - \widehat{\boldsymbol{Y}}, \boldsymbol{x}_{(j)} \right\rangle = \left\langle \boldsymbol{Y} - \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}, \boldsymbol{x}_{(j)} \right\rangle = 0, \ j = 1, \cdots, p,$$

qui devient en exprimant le produit scalaire de $\mathbb{R}^n$

$$\boldsymbol{x}_{(j)}{}^T \boldsymbol{Y} = \boldsymbol{x}_{(j)}{}^T \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}, \ j = 1, \cdots, p,$$

et qui s'exprime sous forme matricielle par

$$\begin{pmatrix} \boldsymbol{x}_{(0)}{}^T \\ \boldsymbol{x}_{(1)}{}^T \\ \boldsymbol{x}_{(2)}{}^T \\ \vdots \\ \boldsymbol{x}_{(p)}{}^T \end{pmatrix} \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{x}_{(0)}{}^T \\ \boldsymbol{x}_{(1)}{}^T \\ \boldsymbol{x}_{(2)}{}^T \\ \vdots \\ \boldsymbol{x}_{(p)}{}^T \end{pmatrix} \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}} \Leftrightarrow \boxed{\underline{\boldsymbol{x}}^T \boldsymbol{Y} = \underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}}.$$

Compte tenu des hypothèses faites sur le modèle et en particulier l'hypothèse **(C3-1)** on en déduit l'expression

$$\widehat{\boldsymbol{\beta}} = \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{Y}.$$

Puisque $\widehat{\boldsymbol{Y}} = \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}$ est le projeté orthogonal de $\boldsymbol{Y}$ sur $\mathcal{L}$, on a

$$\widehat{\boldsymbol{Y}} = \underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}} = \underline{\boldsymbol{x}} \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{Y} = \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y}.$$

E.2 Propriétés de l'estimateur MCO

Il est facile d'exprimer $\widehat{\boldsymbol{\beta}}$ en fonction de $\boldsymbol{\beta}$ de la façon suivante

$$\widehat{\boldsymbol{\beta}} = \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{Y} = \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \left(\underline{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{U}\right) = \boldsymbol{\beta} + \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{U}.$$

Ainsi, puisque le bruit $\boldsymbol{U}$ est un vecteur centré, il vient que $\widehat{\boldsymbol{\beta}}$ est un *estimateur sans biais* de $\boldsymbol{\beta}$. En effet,

$$\boldsymbol{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{E}\left(\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{U}\right) = \left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T \boldsymbol{E}\left(\boldsymbol{U}\right) = \boldsymbol{0}.$$

La matrice de covariances $\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right)$ de $\widehat{\boldsymbol{\beta}}$ s'exprime par

$$\begin{aligned}
\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) &= \boldsymbol{E}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{T}\right) \\
&= \boldsymbol{E}\left(\left(\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^{T}\boldsymbol{U}\right)\left(\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^{T}\boldsymbol{U}\right)^{T}\right) \\
&= \boldsymbol{E}\left(\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^{T}\boldsymbol{U}\boldsymbol{U}^{T}\underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\right) \\
&= \left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^{T}\boldsymbol{E}\left(\boldsymbol{U}\boldsymbol{U}^{T}\right)\underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1} \\
&= \sigma^{2}\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}, \qquad \text{d'après l'hypothèse } \textbf{(C2-1)} \\
&= \sigma^{2}\left(\underline{\boldsymbol{x}}^{T}\underline{\boldsymbol{x}}\right)^{-1}.
\end{aligned}$$

L'hypothèse **(C3-2)** implique en particulier que $\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right)$ tend vers la matrice nulle lorsque n tend vers $+\infty$. Ainsi, $\widehat{\boldsymbol{\beta}}$ est un *estimateur convergent en moyenne quadratique* de $\boldsymbol{\beta}$.

On peut de plus montrer que $\widehat{\boldsymbol{\beta}}$ est le *meilleur estimateur linéaire sans biais* (BLUE en anglais). On dit qu'un estimateur $\widehat{\widehat{\boldsymbol{\beta}}}$ est linéaire s'il est de la forme

$$\widehat{\widehat{\boldsymbol{\beta}}} = \underline{\boldsymbol{z}}\boldsymbol{Y} \text{ avec } \underline{\boldsymbol{z}} \text{ une matrice carrée de taille } ((p+1)\times(p+1)).$$

Puisque $\boldsymbol{E}\left(\widehat{\widehat{\boldsymbol{\beta}}}\right) = \underline{\boldsymbol{z}}\underline{\boldsymbol{x}}\boldsymbol{\beta}$, $\widehat{\widehat{\boldsymbol{\beta}}}$ sera un estimateur sans biais si $\underline{\boldsymbol{z}}\underline{\boldsymbol{x}} = \underline{\boldsymbol{I}}_{p+1}$. De plus, sous cette contrainte, on peut montrer que $\widehat{\boldsymbol{\beta}}$ est le meilleur parmi ces estimateurs au sens de la minimalité de la matrice de covariances. En effet, puisque

$$\boldsymbol{E}\left(\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{T}\right) = \boldsymbol{E}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)^{T}\right) = \underline{\boldsymbol{0}}$$

on en déduit que

$$\begin{aligned}
\underline{\boldsymbol{V}}\left(\widehat{\widehat{\boldsymbol{\beta}}}\right) &= \boldsymbol{E}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{T}\right) + \boldsymbol{E}\left(\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)^{T}\right) \\
&= \underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}\right) + \boldsymbol{E}\left(\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\widehat{\boldsymbol{\beta}}}-\widehat{\boldsymbol{\beta}}\right)^{T}\right) \geq \underline{\boldsymbol{V}}(\widehat{\beta}).
\end{aligned}$$

Ceci nous assure que $\widehat{\boldsymbol{\beta}}$ est l'estimateur linéaire sans biais qui a la plus petite matrice de covariances et qui justifie le terme "meilleur".

E.3 Estimation de la variance du bruit

Exprimons $\widehat{\boldsymbol{U}}$ sous la forme suivante

$$\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} = \underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{Y} = \underline{\boldsymbol{\mathcal{P}}}^{\perp}\left(\underline{\boldsymbol{x}}\boldsymbol{\beta}+\boldsymbol{U}\right) = \underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{U}.$$

On rappelle que la matrice $\underline{\boldsymbol{\mathcal{P}}}^{\perp}$ est symétrique et idempotente. De ce fait, il vient

$$\sum_{t=1}^{n} \widehat{U}_t^2 = \widehat{\boldsymbol{U}}^T \widehat{\boldsymbol{U}} = \boldsymbol{U}^T \underline{\boldsymbol{\mathcal{P}}}^{\perp} \boldsymbol{U} = \sum_{i,j=1}^{n} \left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right)_{ij} U_i U_j$$

et ainsi

$$\boldsymbol{E}\left(\sum_{t=1}^{n} \widehat{U}_t^2\right) = \sum_{i,j=1}^{n} \left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right)_{ij} \boldsymbol{E}\left(U_i U_j\right) = \sigma^2 \sum_{i=1}^{n} \left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right)_{ii} = \sigma^2 tr\left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right).$$

Puisque

$$\begin{aligned} tr\left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right) &= tr\left(\underline{\boldsymbol{I}}_n - \underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T\right) \\ &= tr\left(\underline{\boldsymbol{I}}_n\right) - tr\left(\underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1} \underline{\boldsymbol{x}}^T\right) \\ &= tr\left(\underline{\boldsymbol{I}}_n\right) - tr\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\left(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}}\right)^{-1}\right) \\ &= tr\left(\underline{\boldsymbol{I}}_n\right) - tr\left(\underline{\boldsymbol{I}}_{p+1}\right) \\ &= n - p - 1 \end{aligned}$$

On peut alors proposer comme estimateur sans biais de σ^2 :

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} \sum_{t=1}^{n} \widehat{U}_t^2.$$

E.4 Lois empiriques des estimateurs

E.4.1 Cadre gaussien

- **Paramètre de régression avec σ^2 connu** :
  Lorsque le bruit est supposé gaussien, on sait que $\widehat{\boldsymbol{\beta}}$ est aussi un vecteur gaussien comme tranformation affine de $\boldsymbol{U}$. En particulier,

  $$\boxed{\frac{\widehat{\beta}_i - \beta_i}{\sigma_{\widehat{\beta}_i}} \rightsquigarrow \mathcal{N}(0,1)}$$

  où σ_{β_i} est le i^{ème} élément diagonal de la matrice $\sigma^2(\underline{\boldsymbol{x}}^T \underline{\boldsymbol{x}})^{-1}$.

- **Paramètre de nuisance σ^2** :
  On se propose maintenant de déterminer la loi de probabilité de $(n-p-1)\,\widehat{\sigma}^2/\sigma^2$. En fait, puisque la matrice $\underline{\boldsymbol{\mathcal{P}}}^{\perp}$ est symétrique et idempotente il vient, en posant $\boldsymbol{U}' = \dfrac{1}{\sigma}\boldsymbol{U} \rightsquigarrow \mathcal{N}_n(\boldsymbol{0}, \underline{\boldsymbol{I}}_n)$, que

  $$\delta_{\sigma^2} = (n-p-1)\,\frac{\widehat{\sigma}^2}{\sigma^2} = \left\|\frac{1}{\sigma}\widehat{\boldsymbol{U}}\right\|^2 = \left\|\underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{U}'\right\|^2 \rightsquigarrow \chi^2\left(tr\left(\underline{\boldsymbol{\mathcal{P}}}^{\perp}\right)\right) = \chi^2(n-p-1). \tag{E.1}$$

  Il s'agit là d'une simple application du Théorème de Cochran (voir section D.4).

- **Paramètres de régression avec σ^2 inconnu** :
  Ainsi (cf. section D.5),

$$\delta_{\beta_i} = \frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}_{\widehat{\beta}_i}} = \frac{\dfrac{\widehat{\beta}_i - \beta_i}{\sigma_{\widehat{\beta}_i}}}{\sqrt{\left((n-p-1)\dfrac{\widehat{\sigma}^2}{\sigma^2}/(n-p-1)\right)}} \rightsquigarrow \mathcal{S}t\,(n-p-1)$$

  puisque $\dfrac{\widehat{\beta}_i - \beta_i}{\sigma_{\widehat{\beta}_i}}$ et $(n-p-1)\,\dfrac{\widehat{\sigma}^2}{\sigma^2}$ sont indépendants (comme conséquence de l'indépendance de $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^T\boldsymbol{U}$ et $\widehat{\boldsymbol{U}} = \underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{U}$, puisque $\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\underline{\boldsymbol{x}}^T\underline{\boldsymbol{\mathcal{P}}}^{\perp} = \underline{\mathbf{0}}$).

- **Vecteur des paramètres de régression avec σ^2 connu**
  A une permutation près sur la matrice $\underline{\boldsymbol{x}}$, on peut supposer (sans perte de généralité) que $Q = \{0, 1, \ldots, q-1\}$. On peut alors montrer à partir de la formule de l'inverse d'une matrice par bloc que

$$\left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right)^T \underline{\boldsymbol{V}} \left(\widehat{\boldsymbol{\beta}}_Q\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right) = \left\| \frac{1}{\sigma}\left(\underline{\boldsymbol{\mathcal{P}}}_Q^{\perp}\underline{\boldsymbol{x}}_Q\widehat{\boldsymbol{\beta}}_Q\right)\right\|^2 \tag{E.2}$$

$$= \left\|\underline{\boldsymbol{\mathcal{P}}}_{Q'}\boldsymbol{U}'\right\|^2 \rightsquigarrow \chi^2\,(q) \quad \text{, cf section D.4.} \tag{E.3}$$

  avec

$$\underline{\boldsymbol{\mathcal{P}}}_{Q'} = \left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\left(\left(\underline{\boldsymbol{\mathcal{P}}}_{Q}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\right)^{-1}\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T$$

  et puisque cette matrice est symétrique, idempotente et telle que

$$\begin{aligned} tr\left(\underline{\boldsymbol{\mathcal{P}}}_{Q'}\right) &= tr\left(\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\left(\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\right)^{-1}\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T\right) \\ &= tr\left(\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\left(\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)^T\left(\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\right)\right)^{-1}\right) \\ &= tr\left(\underline{\boldsymbol{I}}_q\right) = q. \end{aligned}$$

- **Vecteur des paramètres de régression avec σ^2 inconnu**
  Nous renvoyons le lecteur à la section D.5 pour montrer que

$$\delta_{\beta_Q} = \frac{1}{q}\left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right)^T\widehat{\underline{\boldsymbol{V}}}\left(\widehat{\boldsymbol{\beta}}_Q\right)^{-1}\left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right) \tag{E.4}$$

$$= \frac{\frac{1}{q}\left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right)^T\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}_Q\right)^{-1}\left(\widehat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q\right)}{\left(\dfrac{1}{n-p-1}\,(n-p-1)\,\dfrac{\widehat{\sigma}^2}{\sigma^2}\right)} \rightsquigarrow \mathcal{F}\,(q, n-p-1)\,. \tag{E.5}$$

De plus, le résultat attendu se déduit alors du fait que les variables aléatoires au numérateur et au dénominateur (voir (E.5)) sont indépendantes puisqu'on vérifie que $\underline{\boldsymbol{\mathcal{P}}}_{Q'}\underline{\boldsymbol{\mathcal{P}}}_{Q}^{\perp} = \underline{\mathbf{0}}$.

E.4.2 Cadre Asymptotique

Vos serviteurs n'ont pas pris le temps de compléter cette section pourtant très intéressante.

E.5 Tests d'hypothèses

E.5.1 Test de significativité

Dans le cas du modèle de régression $\boldsymbol{Y}$ sur la matrice de régression $\underline{\boldsymbol{x}}_{Q^c}$, on pose $\widehat{\boldsymbol{Y}}_{Q^c} = \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\boldsymbol{Y}$ le projeté orthogonal de $\boldsymbol{Y}$ dans $\mathcal{L}_{Q^c}$ et $\widehat{\boldsymbol{U}}_{Q^c} = \underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\boldsymbol{Y}$.
Montrons alors que la statistique δ_{β_Q} s'exprime comme ci-dessous sous l'hypothèse nulle. Sous $\mathbf{H}_0$, on a

$$\begin{aligned}\delta_{\beta_Q} &= \frac{\frac{1}{q}\widehat{\boldsymbol{\beta}}_Q^T\underline{\boldsymbol{V}}\left(\widehat{\boldsymbol{\beta}}_Q\right)^{-1}\widehat{\boldsymbol{\beta}}_Q}{\left(\frac{1}{n-p-1}(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2}\right)} \\ &= \frac{n-p-1}{q}\frac{\left\|\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\widehat{\boldsymbol{\beta}}_Q\right\|^2}{\left\|\widehat{\boldsymbol{U}}\right\|^2}\end{aligned}$$

en utilisant la relation (E.2). De plus, on a, d'une part,

$$\begin{aligned}\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{x}}_Q\widehat{\boldsymbol{\beta}}_Q &= \underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\left(\underline{\boldsymbol{x}}_Q\widehat{\boldsymbol{\beta}}_Q + \underline{\boldsymbol{x}}_{Q^c}\widehat{\boldsymbol{\beta}}_{Q^c}\right) \\ &= \underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\left(\underline{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}\right) \\ &= \underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} \\ &= \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} \\ &= \left(\underline{\boldsymbol{\mathcal{P}}} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\right)\boldsymbol{Y}\end{aligned}$$

et d'autre part,

$$\begin{aligned}\left\|\underline{\boldsymbol{\mathcal{P}}}_{Q^c}^{\perp}\boldsymbol{Y}\right\|^2 &= \left\|\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\boldsymbol{Y}\right\|^2 \\ &= \left\|(\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y}) + (\underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\boldsymbol{Y})\right\|^2 \\ &= \left\|\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y}\right\|^2 + \left\|\underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\boldsymbol{Y}\right\|^2\end{aligned}$$

en appliquant le théorème de Pythagore puisque $\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} = \underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{Y} \in \mathcal{L}^{\perp}$ et

$\underline{\boldsymbol{\mathcal{P}}}\boldsymbol{Y} - \underline{\boldsymbol{\mathcal{P}}}_{Q^c}\boldsymbol{Y} \in \mathcal{L}$. Il vient donc que

$$\begin{aligned}\left\|\underline{\boldsymbol{\mathcal{P}}}^{\perp}_{Q^c}\underline{\boldsymbol{x}}_Q\widehat{\boldsymbol{\beta}}_Q\right\|^2 &= \left\|\underline{\boldsymbol{\mathcal{P}}}^{\perp}_{Q^c}\boldsymbol{Y}\right\|^2 - \left\|\underline{\boldsymbol{\mathcal{P}}}^{\perp}\boldsymbol{Y}\right\|^2 \\ &= \left\|\widehat{\boldsymbol{U}}_{Q^c}\right\|^2 - \left\|\widehat{\boldsymbol{U}}\right\|^2\end{aligned}$$

ce qui conduit au résultat.

Précisons pour terminer que si l'on choisit $Q = \{1, \ldots, p\}$ le test de significativité associé porte le nom de test de significativité globale.

E.6 Prévision

Avec les notations introduites en (9), considérons l'erreur de prédiction

$$\begin{aligned}\widehat{U}_\tau \overset{\text{Not.}}{=} Y_\tau - \widehat{Y}_\tau &= U_\tau - \boldsymbol{x}_\tau{}^T\widehat{\boldsymbol{\beta}} + \boldsymbol{x}_\tau{}^T\boldsymbol{\beta} \\ &= \boldsymbol{x}_\tau{}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - U_\tau\end{aligned}$$

d'espérance

$$\boldsymbol{E}\left(\widehat{U}_\tau\right) = \boldsymbol{x}_\tau{}^T\boldsymbol{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{E}\left(U_\tau\right) = 0,$$

et de variance

$$\begin{aligned}\sigma^2_{\widehat{U}_\tau} &= V(U_\tau) + \boldsymbol{E}\left(\boldsymbol{x}_\tau{}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\boldsymbol{x}_\tau\right) \text{ (puisque } U_\tau \text{ et } \widehat{\boldsymbol{\beta}} \text{ sont indépendants)} \\ &= \sigma^2 + \boldsymbol{x}_\tau{}^T\boldsymbol{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right)\boldsymbol{x}_\tau \\ &= \sigma^2\left(1 + \boldsymbol{x}_\tau{}^T\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\boldsymbol{x}_\tau\right).\end{aligned}$$

naturellement estimée par $\widehat{\sigma}^2\left(1 + \boldsymbol{x}_\tau{}^T\left(\underline{\boldsymbol{x}}^T\underline{\boldsymbol{x}}\right)^{-1}\boldsymbol{x}_\tau\right) = \widehat{\sigma}_{\widehat{U}_\tau}$. Puisque

$$(n-p-1)\frac{\widehat{\sigma}_{\widehat{U}_\tau}}{\sigma_{\widehat{U}_\tau}} = (n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \rightsquigarrow \chi^2(n-p-1)$$

et

$$\frac{\widehat{Y}_\tau - Y_\tau}{\sigma_{\widehat{U}_\tau}} \rightsquigarrow \mathcal{N}(0,1)$$

il vient que

$$\frac{\widehat{Y}_\tau - Y_\tau}{\widehat{\sigma}_{\widehat{U}_\tau}} == \frac{\dfrac{\widehat{Y}_\tau - Y_\tau}{\widehat{\sigma}_{\widehat{U}_\tau}}}{\sqrt{\left((n-p-1)\dfrac{\widehat{\sigma_{\widehat{U}_\tau}}}{\sigma_{\widehat{U}_\tau}}\right)/(n-p-1)}} \rightsquigarrow \mathcal{S}t(n-p-1).$$

L'indépendance des variables au numérateur et au dénominateur de la fraction précédente est laissée en exercice (Indication : U_τ et $\boldsymbol{U}$ sont indépendants).