# Régression linéaire et tests statistiques

Romain CAPLIEZ

## Théorie

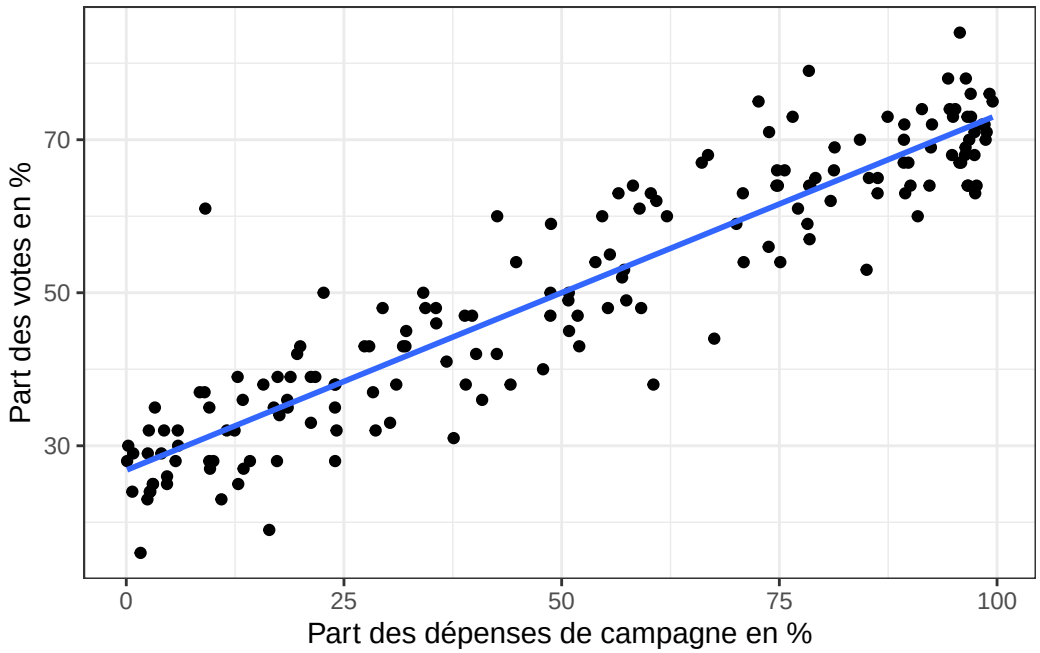
### Principe d'une régression linéaire simple

Si l'on dispose d'un échantillon de deux variables $X$ et $Y$ pour $N$ individus où chaque individu est noté $i = 1, \cdots, N$, alors le principe d'une régression linéaire est de trouver la meilleure droite possible pour décrire le couple de variables $(Y, X)$. On cherche ainsi une droite de la forme :

$$Y_i = \alpha + \beta X_i + u_i$$

Où :

- $Y_i$ : représente la variable expliquée/endogène
- $X_i$ : représente la variable explicative/exogène
- $\alpha$ : représente la constante de la droite : *l'ordonnée à l'origine*
- $\beta$ : représente le paramètre de pente de la droite *: le coefficient directeur*
- $u_i$ : représente le terme d'erreur : ce qui n'est pas expliqué par le modèle




Le modèle $Y_i = \alpha + \beta X_i + u_i$ correspond au modèle *théorique* dans la population. On fait l'hypothèse que la variable $Y_i$ est expliquée par la variable $X_i$ de manière linéaire. Le terme $u_i$ représente les erreurs du modèle, c'est à dire tous les éléments que l'on ne capte pas et qui font varier (de manière non systématique) la variable $Y_i$.

Ce modèle n'est pas estimable puisque l'on ne dispose jamais de données pour l'entièreté de la population : on ne dispose que d'échantillons de cette population. On cherche donc des estimateurs $\hat{\alpha}$ et $\hat{\beta}$ afin de décrire le mieux possible cette relation hypothétique entre $Y_i$ et $X_i$.

La méthode classique consiste à utiliser la méthode des **Moindres Carrés Ordinaires (MCO / OLS)** qui consiste à trouver les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ qui vont minimiser la distance entre les vraies valeurs $Y_i$ et les valeurs estimées de la droite $\hat{Y}_i$ au carré.

Si on définit les valeurs estimées de la variable endogène et les résidus à partir du modèle de la population comme suit et notés respectivement $\hat{Y}_i$ et $\hat{u}_i$ :

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$
$$\hat{u}_i = Y_i - \hat{Y}_i$$

Alors on cherche :

$$\underset{\hat{\alpha},\hat{\beta}}{Min} \sum_{i=1}^{N} \hat{u}_i^2$$

Ce qui revient à annuler les dérivées suivantes :

$$\begin{cases} \frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\alpha}} = 0 \\ \\ \frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}} = 0 \end{cases}$$

Une fois ce système résolu, on trouve :

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\hat{\beta} = \frac{\mathrm{Cov}\,(X, Y)}{\mathrm{V}\,(X)}$$

Avec :

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

On interprète les paramètres de la façon suivante :

- $\hat{\alpha}$ : Lorsque $X_i = 0$, $Y$ est en moyenne égal à $\hat{\alpha}$.
- $\hat{\beta}$ : Lorsque $X_i$ augmente de 1 unité, alors en moyenne $Y$ augmente de $\hat{\beta}$ unités.

La variance des résidus est donnée par :

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N - 2}$$

La variance des estimateurs est donnée par :

$$\hat{\sigma}_{\hat{\beta}}^2 = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} = \frac{\hat{\sigma}_u^2}{N \times V(X)}$$

$$\hat{\sigma}_{\hat{\alpha}}^2 = \hat{\sigma}_u^2 \frac{\sum_{i=1}^{N} X_i^2}{N^2 \times V(X)}$$

## Régression linéaire multiple

### Notation

La régression linéaire multiple reprend le même principe que la régression linéaire simple, mais est définie dans un cadre général où $Y_i$ est expliquée par $K$ variables explicatives $X_{ik}$ avec $k = 1, \cdots, K$.

Ce modèle peut s'écrire sous forme matricielle :

$$\mathbf{Y} = \mathbf{X}\beta + u$$

- $\mathbf{Y}$ : un vecteur de taille $(N \times 1)$ pour la variable endogène/expliquée.
- $\mathbf{X}$ : une matrice de taille $(N \times K)$ qui comprend les variables explicatives/exogènes. En cas d'ajout d'une constante, la taille passe à $(N \times K + 1)$.
  - $\beta$ : un vecteur de taille $(K \times 1)$ qui comprend la valeur du coefficient associé à chaque variable. En cas d'ajout de constante (notée $\alpha$), la taille passe à $(K \times 1 + 1)$.
- $u$ : un vecteur de taille $(N \times 1)$ qui contient les erreurs du modèle.

Sous forme développée cela donne :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

On peut également écrire le modèle sous forme semi-développée en une ligne :

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + u_i$$

L'interprétation se fait alors **toute chose égale par ailleurs**. Pour la variable $X_{i1}$, on dira que lorsque $X_{i1}$ augmente de 1 unité toute chose égale par ailleurs (les autres variables restant constantes), alors en moyenne $Y_i$ augmente de $\beta_1$ unités.

L'idée de la régression linéaire multiple est d'expliquer au mieux la variable $Y_i$ en incorporant l'ensemble des facteurs susceptibles d'affecter $Y_i$. Inclure une variable dans une régression linéaire multiple permet d'obtenir son effet sur la variable $Y_i$ tout en purgeant l'effet des autres variables.

### Estimateurs

Tout comme dans le cas de la régression simple, on cherche à minimiser l'écart au carré entre les vraies valeurs $Y_i$ et les valeurs estimées $\widehat{Y}i$ :

$$\underset{\hat{\beta}}{Min}\left(\hat{u}'\hat{u}\right)$$

Une fois le système résolu on trouve :

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

La variance estimée des résidus est donnée par :

$$\hat{\sigma}_u^2 = \frac{\hat{u}'\hat{u}}{N-K-1}$$

La variance estimée des estimateurs est donnée par :

$$\hat{\sigma}_{\hat{\beta}}^2 = \hat{\sigma}_u^2 \left(X'X\right)^{-1}$$

### Hypothèses

Si les hypothèses suivantes sont respectées, alors les estimateurs MCO seront des estimateurs non biaisés et de variance minimale. On dit que ce sont des estimateurs **BLUE : Best Linear Unbiased Estimator**.

- H1 : $E(u) = 0$. L'espérance de l'erreur est nulle. En moyenne le modèle ne commet pas d'erreur. Il ne sous-estime ni ne surestime la valeur de la variable endogène de manière systématique. Cette hypothèse est nécessaire pour les tests d'hypothèse et prouver l'absence de biais des estimateurs MCO.
- H2 : $X$ est une matrice certaine, c'est à dire non aléatoire, elle est observée sans erreur. On note cette hypothèse : $E\left(u_i|x_i\right) = 0$. Cela signifie qu'il n'y a aucun lien statistique entre les erreurs et la variable explicative. On n'a oublié aucune variable (qui se retrouverait donc dans les erreurs) qui soit corrélée de quelque manière que ce soit avec une variable explicative. Si cette hypothèse est violée, l'estimateur MCO n'est plus convergent et devient biaisé.

- H3 : $\text{Rang}\,(X) = K + 1$. Les variables explicatives sont linéairement indépendantes. Aucune variable ne doit pouvoir être expliquée parfaitement par une combinaison liénaire d'autres variables. Par exemple, on ne peut pas inclure le poids en Kg et le poids en gramme, puisque ces deux variables sont parfaitement liées par une combinaison linéaire : 1Kg = 1000g. Cette hypothèse si elle est violée empêche l'identification du modèle et donc de trouver des estimateurs uniques.

  Pour cette troisième hypothèse, il faut également satisfaire la condition suivante : $N > K + 1$. Le nombre d'observations doit être supérieur au nombre de variables explicatives (ici $K + 1$ si l'on considère que l'on a une constante, $K$ sinon). Si cette condition n'est pas remplie, le modèle n'est pas estimable.

- H4 : $E\,(uu') = \sigma_u^2 I_T$. Les erreurs ne doivent pas être autocorrélées dans le temps (dans le cas de séries temporelles) ou entre les individus (dans le cas d'une coupe transversale). Les erreurs doivent également avoir une variance constante, et cette variance ne doit pas dépendre des variables explicatives. Si cette hypothèse est violée, les estimateurs MCO ne sont plus de variance minimale.

  De manière générale, la matrice de variance-covariance des erreurs s'écrit :

$$E\,(uu') = \begin{bmatrix} \text{Var}(u_1) & \text{Cov}(u_1, u_2) & \cdots & \text{Cov}(u_1, u_N) \\ \text{Cov}(u_2, u_1) & \text{Var}(u_2) & \cdots & \text{Cov}(u_2, u_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(u_N, u_1) & \text{Cov}(u_N, u_2) & \cdots & \text{Var}(u_N) \end{bmatrix}$$

Dans le cas où les erreurs sont sphériques, la variance des erreurs est constante et égale à $\sigma_u^2$, tandis que la covariance entre les différents termes d'erreur est nulle. On a donc :

$$E\,(uu') = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix} = \sigma_u^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma_u^2 I_N$$

- H5 : Cette hypothèse suppose que les résidus sont normalement distribués : $u_i \sim N(0, \sigma_u^2)$. Cette hypothèse permet de faire des tests de faire de l'inférence statistique.

## Tests de significativité des coefficients

Comme on ne dispose que d'un échantillon de la population, nous n'avons qu'une estimation de $\beta_k$ notée $\widehat{\beta}_k$. S'il est aisé de voir si $\widehat{\beta}_k$ est différent de 0 (ou de n'importe quelle autre constante $a$) ou non, ce n'est pas le cas de $\beta_k$ le vrai coefficient dans la population. Le test de Student va

chercher à savoir si pour un certain niveau de risque, le vrai coefficient de la population est différent de 0 (ou d'une constante $a$) à partir de l'estimation obtenue et de sa dispersion.

Ce qui nous intéresse généralement est de savoir si la variable $X_{ik}$ a un effet ou non dans la population soit : $\beta_k = 0$. Mais on peut aussi s'intéresser à des cas différents.

Pour tester cela, on utilise le test de Student avec le jeu d'hypothèse suivant :

$$\begin{cases} H_0 : \beta = a \\ H_1 : \beta \neq a \end{cases}$$

La statistique de Student est définie de la manière suivante :

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}}$$

Sous l'hypothèse nulle $H_0$, avec un risque de $\alpha\%$ on sait que cette statistique suit une loi de Student à $N - K - 1$ degrés de libertés ($N - K$ si pas de constante) : $t_{\hat{\beta}} \sim t_{\alpha/2}\,(N - K - 1)$. Le degré de confiance du test $\alpha$ correspond au risque de première espèce (Faux-positif) c'est à dire au risque de rejeter $H_0$ alors que cette hypothèse est vraie.

La règle de décision est la suivante :

- Si $\left|t_{\hat{\beta}}\right|$ est supérieur à la valeur critique tabulée, alors on rejette $H_0$. Le coefficient $\beta$ est statistiquement différent de $a$.
- Si $\left|t_{\hat{\beta}}\right|$ est inférieur à la valeur critique tabulée, alors on ne peut pas rejeter l'hypothèse nulle. Le coefficient $\beta$ n'est pas statistiquement différent de $a$.

Il est également possible de faire un test unilatéral sous la forme :

$$\begin{cases} H_0 : \beta = a \\ H_1 : \beta < a \text{ou} \beta > a \end{cases}$$

Dans ce cas la loi suivie par la statistique est : $t_{\hat{\beta}} \sim t_{\alpha}\,(N - K - 1)$ .

Il est également possible de calculer le niveau de significativité à partir duquel l'hypothèse nulle ne peut plus être rejetée. C'est ce que l'on appel la **p-value**. Plus cette probabilité est faible, plus le risque de commettre un Faux-positif est faible. On rejettera alors l'hypothèse nulle dès lors que la p.value est inférieure à un seuil donné. Si la p.value est supérieure à ce seuil, on ne pourra pas rejeter l'hypothèse nulle (règle de décision inversée par rapport aux valeurs critiques).

### Significativité du modèle

On peut décomposer la variance de la variable endogène $Y_i$ comme étant la somme de la variance expliquée par le modèle et de la variance du terme d'erreur :

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2+\frac{1}{N}\sum_{i=1}^{N}u_i^2$$

$$\mathrm{SCT}=\mathrm{SCE}+\mathrm{SCR}$$

Cette équation de la variance nous indique que la Somme des Carrés Totaux (SCT) est égale à la Somme des Carrés Expliqués (SCE) plus la Somme des Carrés des Résidus (SCR). Autrement dit, la variation de notre variable endogène se décompose en une partie expliquée par le modèle et une partie que l'on explique pas et qui se retrouve dans les résidus.

Le $R^2$ ou coefficient d'ajustement est une mesure qui permet de déterminer le pourcentage de variance de la variable endogène qui est expliqué par le modèle. Il est donné par :

$$R^2=\frac{\mathrm{SCE}}{\mathrm{SCT}}=1-\frac{\mathrm{SCR}}{\mathrm{SCT}}\in[0,1]$$

Ce coefficient est cependant biaisé en cas de régression linéaire multiple. L'ajout d'une variable supplémentaire fait forcément augmenter le $R^2$ le rendant inutile pour la comparaison entre deux modèles ayant un nombre de variables explicatives différent.

Pour cela on utilise le coefficient de détermination ajusté $\overline{R}^2$ :

$$\overline{R}^2=1-\frac{N-1}{N-K-1}\left(1-R^2\right)$$

Ce coefficient s'interprète de la même manière que le $R^2$ et permet de comparer des modèles avec un nombre de variables explicatives différent. Il est à noter que ces modèles doivent avoir la même variable endogène et le même nombre d'observations.

On peut tester si le modèle dans son ensemble a un pouvoir explicatif en utilisant un test de Fisher défini comme suit :

$$\begin{cases} H_0 : R^2=0 \\ H_1 : R^2>0 \end{cases}$$

Avec la statistique de test qui sous l'hypothèse nulle suit une loi de Fisher :

$$F = \frac{R^2/K}{(1-R^2)/(N-K-1)} \sim \mathcal{F}(K, N-K-1)$$

La règle de décision est :

- Si $F >$ valeur critique, alors on rejette $H_0$. Le modèle est donc globalement statistiquement significatif.
- Si $F <$ valeur critique, alors on ne peut pas rejeter $H_0$. Le modèle n'est donc pas globalement statistiquement significatif.

On peut également calculer une p.value :

- Si la p.value est inférieure à un seuil donné, alors on rejette l'hypothèse nulle.
- Si la p.value est supérieure à un seuil donné, alors on ne peut pas rejeter l'hypothèse nulle.

### Tests de significativité groupé des coefficients

Il est possible de vouloir tester si de manière jointe plusieurs coefficients ont un effet sur la variable endogène. Lorsque différentes variables sont fortement corrélées entre elles (comme les mêmes modalités d'une variable qualitative ou les variables qualitatives avec leurs termes d'interactions), la variance des estimateurs augmente. On aura ainsi tendance à rejeter la significativité individuelle de ces variables. Cependant on peut être intéressé par le fait de savoir si cet ensemble de variables, pas forcément significatives de manière individuelle, a un impact sur la variable expliquée une fois considérées toutes ensembles.

Pour cela on va utiliser le test de Fisher sur un sous-groupe de variables.

Supposons que l'on estime le modèle suivant :

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Supposons que les variables $X_2$ et $X_3$ ne soient pas significatives prises individuellement et que ces variables présentent un fort degré de corrélation entre elles. On peut vouloir tester si, prises ensemble, ces variables ont un pouvoir explicatif sur $Y_i$ ou si elles sont "inutiles".

On cherche donc à tester si :

$$\begin{cases} H_0 : \beta_2 = \beta_3 = 0 \\ H_1 : \beta_2 \neq 0 \ \mid \ \beta_3 \neq 0 \end{cases}$$

Pour cela on va estimer en plus du premier modèle (que l'on appelle le modèle non-contraint) un second modèle que l'on nomme **modèle contraint** et qui va intégrer les contraintes imposées dans l'hypothèse nulle. On estime donc le modèle suivant :

$$Y_i = \alpha + \beta_1 X_{1i} + e_i$$

Sur chacun des modèles, on va calculer la somme des carrés des résidus : $\text{SCR}_{\text{NC}} = \sum_{i=1}^{N} \hat{u}_i^2$ et $\text{SCR}_{\text{C}} = \sum_{i=1}^{N} \hat{e}_i^2$.

On calcule ensuite la statistique de test qui sous l'hypothèse nulle suit la loi :

$$F = \frac{(\text{SCR}_{\text{C}} - \text{SCR}_{\text{NC}})/q}{\text{SCR}_{\text{NC}}/(T-K-1)} \sim \mathcal{F}(q, T-K-1)$$

Avec $q$ le nombre de contraintes imposées (ici 2).

La règle de décision est :

- Si $F >$ valeur critique, alors on rejette l'hypothèse nulle de non significativité jointe. Les variables prises ensembles ont un pouvoir explicatif.
- Si $F <$ valeur critique, alors on ne peut pas rejeter l'hypothèse nulle de non significativité jointe. Les variables prises ensembles n'ont pas de pouvoir explicatif significatif.

Tous comme pour les autres tests, on peut utiliser la p.value comme règle de décision.

Ce test est généralisable avec davantage de contraintes et avec des formes de contraintes différentes.

## Tests des hypothèses et corrections

### Hétéroscédasticité des erreurs

L'hypothèse d'homoscédasticité des erreurs signifie que la variance des erreurs est constante au cours du temps / des individus et constante avec les observations des variables explicatives. Lorsque cette hypothèse est violée, on parle d'erreurs hétéroscédastiques. Dans un tel cas, la variance des estimateurs MCO n'est plus minimale.

Parmi les sources les plus courantes d'hétéroscédasticité, on peut citer :

- l'hétérogénéité de l'échantillon
- l'oubli d'une variable explicative
- l'asymétrie dans la distribution de certaines variables explicatives

- une mauvaise forme fonctionnelle
- la nature des données (ex : moyennes d'observations issues d'échantillons de tailles différentes)

Pour tester la présence d'hétéroscédasticité, différents tests sont disponibles comme :

- le test de Breusch-Pagan (1979)
- le test de White (1980)
- le test ARCH

### Test de Breusch-Pagan (1979)

Supposons le modèle de régression multiple :

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

On suppose que le terme d'erreur suit une loi normale de variance :

$$\sigma^2_{u_i} = f(a_0 + a_1 Z_{1i} + \cdots + a_p Z_{pi})$$

Où $f$ est une fonction quelconque, les coefficients $a_j$ ne sont pas liés aux coefficients du modèle de régression et $Z_{1i}, \cdots, Zpi$ sont des variables susceptibles d'être à la source de l'hétéroscédasticité. Certaines de ces variables, voir toutes, peuvent être des variables explicatives du modèle de régression.

Tester l'hypothèse nulle d'homoscédasticité revient à tester :

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_p = 0 \\ H_1 : \exists \ \ a_j \neq 0 \ \ j = 1, \cdots, p \end{cases}$$

En effet, sous l'hypothèse nulle la variance des erreurs est constante :

$$\sigma^2_{u_i} = f(a_0)$$

Pour réaliser ce test, on suit les étapes suivantes :

- Estimer la régression $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$ par les MCO et en déduire la série des résidus $\widehat{u}_i$.
- On calcule l'estimateur du maximum de vraisemblance de la variance du terme d'erreur : $\widehat{\sigma}^2_{MV} = \frac{1}{N} \sum_{i=1}^{N} \widehat{u}_i$

- On calcule la quantité pour $i = 1, \cdots, N$ :

$$h_i = \frac{\hat{u}_i^2}{\hat{\sigma}_{MV}^2}$$

- Après avoir spécifié les variables $Z_{1i}, \cdots, Z_{pi}$, on régresse $h_i$ sur ces variables

$$h_i = a_0 + a_1 Z_{1i} + \cdots + a_p X_{pi} + e_i$$

- On calcule la somme des carrés expliqués (SCE) de cette régression auxiliaire
- On calcule la statistique de test $BP$ qui sous $H_0$ suit une loi du Khi-deux à $p$ degrés de liberté :

$$BP = \frac{1}{2} SCE \sim \chi_p^2$$

La règle de décision est la suivante :

- Si $BP > \chi_p^2$ (ou p.value inférieure à un seuil donné) : on rejette l'hypothèse nulle d'homoscédasticité. Les erreurs sont hétéroscédastiques, la variance des estimateurs MCO n'est donc pas minimale.
- Si $BP < \chi_p^2$ (ou p.value supérieure à un seuil donné) : on ne peut pas rejeter l'hypothèse nulle d'homoscédasticité. Les erreurs sont homoscédastiques.

Le test de Breusch-Pagan est un test général qui couvre un grand nombre de cas d'hétéroscédasticité. Il s'agit d'un test asymptotique qui n'est valable que pour des échantillons de taille suffisamment importante.

### Test de White (1980)

Le test de White est un test très général qui ne repose pas sur l'hypothèse de normalité du terme d'erreur. La procédure de test est la suivante :

- On estime le modèle de régression multiple :

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- On en déduit la série des résidus $\hat{u}_i$
- On estime la régression auxiliaire suivante :

$$\hat{u}_i^2 = a_0 + a_1 X_{1i} + b_1 X_{1i}^2 + a_2 X_{2i} + b_2 X_{2i}^2 + \cdots + a_k X_{ki} + b_k X_{ki}^2 + e_t$$

Cette régression traduit l'existence possible d'une relation entre le carré des résidus de la régression (autrement dit la variance comme la moyenne des résidus est égale à 0) et les variables explicatives (au carré). Il est également possible d'introduire des termes croisés/d'interaction.

- On calcule le coefficient de détermination $R^2$ de la régression auxiliaire
- On teste :

$$\begin{cases} H_0 : a_1 = b_1 = a_2 = b_2 = \cdots = a_k = b_k = 0 \\ H_1 : \exists \ c_j \neq 0 \quad j = 1, \cdots, k \quad c \in \{a, b\} \end{cases}$$

La statistique de test suit sous l'hypothèse nulle une loi du Khi-deux dont le nombre de degrés de liberté correspond au nombre de paramètres estimés hors constante dans la régression auxiliaire. Elle est donnée par :

$$N \times R^2 \sim \chi_{2k}^2$$

La règle de décision est la suivante :

- Si $NR^2 > \chi_{2k}^2$ (ou p.value inférieure à un seuil donné) : on rejette l'hypothèse nulle d'homoscédasticité. Les erreurs sont hétéroscédastiques, la variance des estimateurs MCO n'est donc pas minimale.
- Si $NR^2 < \chi_{2k}^2$ (ou p.value supérieure à un seuil donné) : on ne peut pas rejeter l'hypothèse nulle d'homoscédasticité. Les erreurs sont homoscédastiques.

### Test ARCH

Dans certaines séries temporelles (particulièrement les séries financières), la variance des erreurs (aussi appelée volatilité) peut dépendre de ses valeurs passées. Il s'agit d'un cas d'hétéroscédasticité conditionnelle. Le test ARCH permet de détecter ce type d'hétéroscédasticité. La procédure de test est la suivante :

- On estime le modèle de régression multiple :

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t$$

Où $t$ indique une période temporelle.

- On déduit la série des résidus $\hat{u}_t$

- On calcule la série des résidus au carrés $\widehat{u}_t^2$
- On régresse la série des résidus au carré sur ses $l$ valeurs passées et une constante :

$$\widehat{u}_t^2 = a_0 + \sum_{i=1}^{l} a_i \widehat{u}_{t-i}^2 + e_t$$

- On calcule le $R^2$ de cette régression auxiliaire
- On teste :

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_l = 0 \\ H_1 : \exists \ a_j \neq 0 \quad j = 1, \cdots, l \end{cases}$$

La statistique de test est donnée par :

$$T \times R^2 \sim \chi_l^2$$

La règle de décision est la suivante :

- Si $TR^2 > \chi_l^2$ (ou p.value inférieure à un seuil donné) : on rejette l'hypothèse nulle d'homoscédasticité. Les erreurs sont conditionnellement hétéroscédastiques, la variance des estimateurs MCO n'est donc pas minimale.
- Si $TR^2 < \chi_l^2$ (ou p.value supérieure à un seuil donné) : on ne peut pas rejeter l'hypothèse nulle d'homoscédasticité. Les erreurs sont homoscédastiques.

#### Estimation

Si la forme de l'hétéroscédasticité est connue, il est possible d'utiliser des méthodes telles que les **Moindres Carrés Pondérés** pour la prendre en compte. Généralement en pratique, on ne connaît pas la forme de l'hétéroscédasticité.

Dans ce cas, il est possible d'utiliser l'estimateur de la matrice de variance-covariance de White ou encore celui de Newey et West. Il s'agit de corrections apportées à la matrice de variance-covariance pour tenir compte de l'hétéroscédasticité. Les coefficients du modèle de régression ne changent pas. Seuls les écarts-types estimés changent.

#### Normalité des erreurs

L'hypothèse de normalité des erreurs est nécessaire afin de procéder à une analyse d'inférence statistique des coefficients estimés. Cette hypothèse permet de dériver la loi suivie par les estimateurs et donc par les statistiques de test. Pour tester la normalité des erreurs on peut utiliser le test de Jarque et Bera (1980).

### Test de Jarque et Berra (1980)

Ce test se base sur la définition des coefficients de skewness $S$ (asymétrie de la distribution) et de kurtosis $K$ (épaisseur des queues de distribution) :

$$S = \frac{\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^3\right]^2}{\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2\right]^3}$$

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2\right]^3}$$

Une loi normale dispose d'un coefficient d'asymétrie (skewness) nul et d'un coefficient d'aplatissement (kurtosis) égal à 3.

La procédure de test est la suivante :

- Estimer la régression :

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- Récupérer les résidus $\widehat{u}_i$
- Tester :

$$\begin{cases} H_0 : S = 0 \text{ et } K = 3 \\ H_1 : S \neq 0 \text{ ou } K \neq 3 \end{cases}$$

La statistique de test $JB$ est donnée par :

$$JB = \frac{N}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right] \sim \chi^2_2$$

La règle de décision est :

- Si $JB > \chi^2_2$ (ou p.value inférieure à un seuil donné) alors on rejette l'hypothèse nulle de normalité des résidus. Les résidus ne suivent pas une loi normale.
- Si $JB < \chi^2_2$ (ou p.value supérieure à un seuil donné), alors on ne peut pas rejeter l'hypothèse nulle de normalité des résidus. Il semble que les résidus suivent une loi normale.

### Estimation

Asymptotiquement les estimateurs MCO disposent de bonnes propriétés. Asymptotiquement (donc pour des échantillons assez grands), les estimateurs MCO sont efficaces. De plus, l'ensemble des méthodes basiques d'inférence visant à la mise en oeuvre des tests statistiques et à la construction d'intervalles de confiance sont approximativement valides sans recourir à l'hypothèse de normalité des erreurs.

## Formes fonctionnelles

### Modèles quadratiques

On rappelle que l'aspect "linéaire" de la régression linéaire implique la linéarité dans les *paramètres*. Il n'existe aucune restriction sur la forme que doit prendre $Y$ ou $X$.

Ainsi le modèle suivant ne peut être conceptualisé sous la forme d'un modèle de régression linéaire :

$$Y_i = \frac{1}{\alpha + \beta X_i} + u_i$$

En revanche on peut très bien imaginer un modèle de la forme

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

Une telle forme fonctionnelle permet de prendre en compte de manière simple des non-linéarités. Cela permet de capter des effets marginaux croissants ou décroissants et non plus constants. Cela permet d'avoir une relation en cloche telle que $X_i$ va faire augmenter $Y_i$ jusqu'à un certain point, puis va le faire diminuer.

Dans un tel modèle, on ne peut plus estimer l'impact global qu'une augmentation de $X_i$ entraîne sur $Y_i$. Il faut regarder la variation de $Y_i$ pour une valeur donnée de $X_i$.

Pour cela on peut utiliser l'approximation :

$$\frac{\Delta \widehat{Y_i}}{\Delta X_i} \approx \hat{\beta}_1 + 2\hat{\beta}_2 X_i$$

La pente de la relation entre $X_i$ et $Y_i$ dépend de la valeur de $X_i$. Généralement, on insère des valeurs d'intérêts dans cette approximation telles que la moyenne, médiane… afin d'obtenir une rapide description de la relation.

Pour trouver le point de retournement de la relation on utilise :

$$X^* = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$$

La forme de la relation quadratique dépend du signe des deux coefficients.

### Modèles à élasticités

Il est possible de transformer certaines variables en logarithme afin d'obtenir des interprétation en terme d'élasticités (changement en pourcentage) plutôt que des changements en unités. Par exemple, il est souvent plus intéressant d'avoir une interprétation en pourcentage lorsque l'on parle d'augmentation de salaire plutôt que d'augmentation en dollars.

| Nom | Modèle | Interprétation de $\beta$ |
|---|---|---|
| Niveau-Niveau | $Y_i = \alpha + \beta X_i + u_i$ | $\Delta Y_i = \beta_1 \Delta X_i$ |
| Niveau-log | $Y_i = \alpha + \beta \times log(X_i) + u_i$ | $\Delta Y_i = (\frac{\beta}{100})\%\Delta X_i$ |
| Log-Niveau | $log(Y_i) = \alpha + \beta X_i + u_i$ | $\%\Delta Y_i = (100\beta)\Delta X_i$ |
| Log-Log | $log(Y_i) = \alpha + \beta \times log(X_i) + u_i$ | $\%\Delta Y_i = \beta\%\Delta X_i$ |

- Dans un modèle en `Niveau-Log`, augmenter $X_i$ de 1% fait augmenter $Y_i$ de $\frac{\beta}{100}$ unités en moyenne.
- Dans un modèle `Log-Niveau`, augmenter $X_i$ de 1 unité, fait augmenter $Y_i$ de $(100 \times \beta)\%$.
- Dans un modèle `Log-Log`, augmenter $X_i$ de 1%, fait augmenter $Y_i$ de $\beta\%$ .

L'augmentation n'est plus linéaire.

### Variables indicatrices indépendantes

Il est possible d'inclure dans un modèle des variables dites binaires ou "dummy" afin de décrire une information qualitative à deux modalités telle que le sexe d'une personne.

Un tel modèle est noté :

$$Y_i = \alpha + \beta X_i + \delta \times \mathbf{1}_i + u_1$$

Avec $\mathbf{1}_i = 1$ si l'individu possède la caractéristique voulue (homme) et $\mathbf{1}_i = 0$ sinon (femme).

Dans un tel modèle, le paramètre $\delta$ correspond à la différence moyenne de $Y_i$ entre les personnes ayant la caractéristique et celles qui ne l'ont pas pour un même niveau de $X_i$. Il s'agit d'une différence systématique entre les deux groupes. On décale la constante entre les deux groupes.

La relation pour la variable $X_i$ reste la même pour les deux groupes, mais il existe une différence systématique entre eux pour tous les niveaux de $X_i$.

La constante pour le groupe de référence (donné par le groupe pour lequel $\mathbf{1}_i = 0$) est $\alpha$ tandis que la constante pour le groupe étudié (donné par le groupe pour lequel $\mathbf{1}_i = 1$) est donnée par $\alpha + \delta$.

Ce modèle avec constante est équivalent à

$$Y_i = \beta X_i + \delta_1 \times \mathbf{1}_{1i} + \delta_2 \times \mathbf{1}_{2i} + u_1$$

où :

- $\mathbf{1}_{1i} = 1$ pour le groupe étudié (homme) et 0 sinon (femme)
- $\mathbf{1}_{2i} = 1$ pour le groupe de référence (femme) et 0 sinon.

Si le modèle inclut les deux variables indicatrices ainsi que la constante, il n'est plus identifiable puisque ces trois variables sont parfaitement colinéaires entre elles : elles s'expliquent parfaitement par une combinaison linéaire : $\alpha = \mathbf{1}_{1i} + \mathbf{1}_{2i}$.

Si la qualité que l'on souhaite étudier a de multiples modalités (catégorie socio-professionnelle) , alors on intègre une variable par modalité (-1 s'il y a une constante). L'interprétation des coefficients se fera toujours par rapport à la variable de référence.

Pour tester si une variable qualitative à $j$ modalités est intéressante à inclure, on peut effectuer un test de Fisher avec l'hypothèse nulle : $\delta_1 = \delta_2 = \cdots = \delta_{j-1} = 0$.

#### Variables d'interactions

L'ajout de variables qualitatives permet d'introduire une différence systématique entre deux groupes d'individus. Mais la relation de pente reste la même entre les deux groupes. Il est possible de relâcher cette hypothèse d'homogénéité des pentes en intégrant des variables d'interactions telles que :

$$Y_i = \alpha + \delta_1 \mathbf{1}_i + \beta X_i + \delta_2 X_i \mathbf{1}_i + u_i$$

Dans ce cas, il existe une différence systématique entre les personnes du groupe étudié et les personnes du groupe de référence : $\delta_1$. Mais la relation est également différente entre les deux groupes. Pour le groupe de référence la relation est donnée par $\beta$ tandis que pour le groupe étudié la relation de pente est donnée par $\beta + \delta_2$.

La constante ainsi que la pente de la droite diffèrent selon les groupes.

Pour tester si $Y_i$ suit le même modèle pour les différents groupes étudiés, on va effectuer un test de Fisher dans lequel l'hypothèse nulle sera : $\delta_1 = \delta_2 = 0$.

### Prédiction

Un modèle de régression linéaire peut-être utilisé afin de prédire la variable endogène pour des valeurs définies des variables exogènes.

Ainsi pour un modèle :

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i$$

On souhaite trouver la valeur moyenne de $Y_i$ étant donné que les régresseurs prennent des valeurs spécifiques $c_1, c_2, \cdots, c_k$ :

$$\theta_0 = E(Y_i | X_{i1} = c_1, X_{i2} = c_2, \cdots, X_{ik} = c_k) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

Le point estimé est donc :

$$\hat{\theta}_0 = \hat{\alpha} + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

## Codes

### Setup

Charger les librairies et tous les objets de setup. Comme rien n'est chargé pour l'instant il faut :

- Utiliser la syntaxe `here::here()` pour dire que l'on utilise la fonction `here()` du package `here`
- Indiquer le chemin en entier jusqu'au script `setup.R`

```
# Exécute le script setup.R pour charger tous les éléments importants
source(here::here("02-codes", "utils", "setup.R"))
```

## Régression linéaire simple

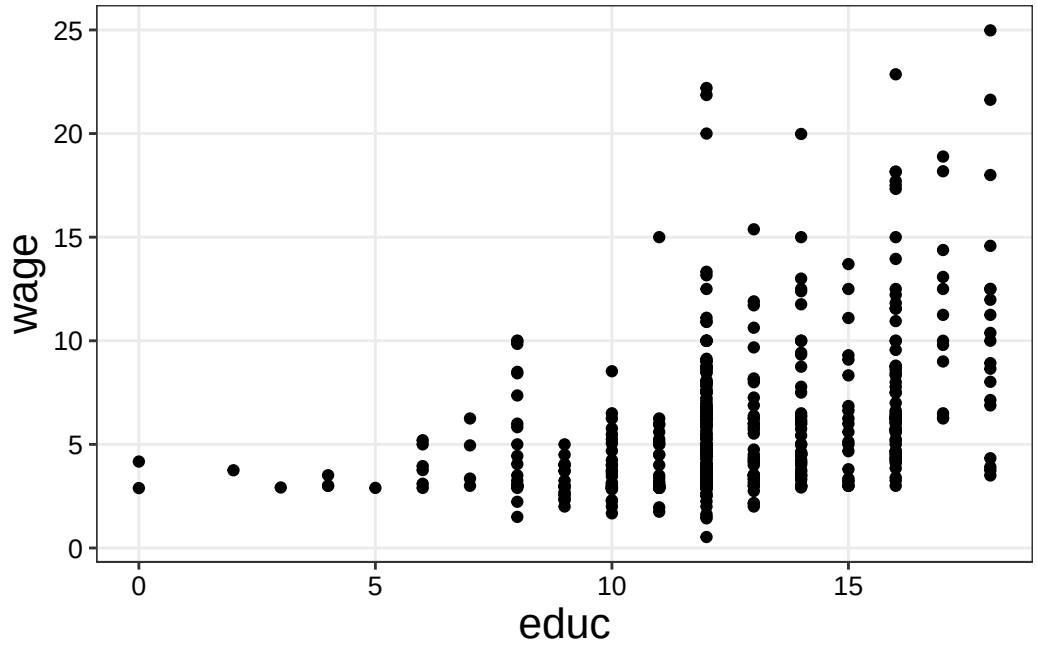
### Estimation et interprétation des coefficients

On cherche à expliquer le salaire d'un individu. On suppose que le salaire horaire moyen d'un individu dépend de son niveau d'éducation de la façon suivante :

$$\text{wage}_i = \alpha + \beta \times \text{educ}_i + u_i$$

Ce modèle suppose une relation dans la population linéaire entre l'éducation d'un individu et son salaire. Le terme $u_i$ comprend tous les facteurs pouvant affecter le salaire d'un individu et n'étant pas compris dans cette relation.

```
# Représenter graphiquement les données sous forme d'un nuage de points
wooldridge::wage1 |>
  ggplot(aes(x = educ, y = wage)) +
  geom_point() +
  theme()
```




Il semble graphiquement y avoir une relation croissante entre l'éducation et le salaire horaire moyen d'un individu.

On va estimer cette relation à partir d'une régression linéaire avec la fonction `lm()`. Cette fonction prend au minimum deux arguments : une formule indiquant la relation à estimer, et le jeu de données.

```
# Estimation de la régression linéaire
reg1 <- lm(wage ~ educ, data = wage1)

# Montrer les résultats
reg1_summary <-
  summary(reg1) |>
  print()
```

```
Call:
lm(formula = wage ~ educ, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3396 -2.1501 -0.9674  1.1921 16.6085

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.90485    0.68497  -1.321               0.187
educ         0.54136    0.05325  10.167 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.378 on 524 degrees of freedom
Multiple R-squared:  0.1648,    Adjusted R-squared:  0.1632
F-statistic: 103.4 on 1 and 524 DF,  p-value: < 0.00000000000000022
```

La ligne`(Intercept)` correspond à la constante, tandis que les lignes suivantes correspondent aux variables explicatives du modèle. La colonne `estimate` donne la valeur des estimateurs $\hat{\beta}$. La colonne `Std. Error` donne la valeur des écarts-types des estimateurs $\hat{\sigma}_{\hat{\beta}}$. La colonne `t value` indique la statistique du test de Student $t_{\hat{\beta}}$. La colonne `Pr(|t|)` donne la p.value du test de Student.

La valeur de $\hat{\alpha}$ nous indique que dans l'échantillon lorsqu'un individu a un niveau d'éducation nul, alors son revenu horaire moyen estimé est de $-0.90485$ dollars environ ce qui n'est pas réaliste, un salaire ne pouvait pas être négatif. On remarque que la p.value du test de Student est de $0.187 > 0.05$. On ne peut donc rejeter l'hypothèse selon laquelle la constante serait différente de 0.

La valeur de $\hat{\beta}$ nous indique que dans l'échantillon, lorsqu'un individu augmente son nombre d'années d'éducation de 1, alors en moyenne, son salaire horaire moyen augmente de 0.54136 dollars. La p.value du test de Student est extrêmement faible et largement inférieure à 5%. Ces estimation nous laissent penser que l'éducation a bien un effet positif sur le salaire dans la population.

Nous pouvons voir que $R^2 = 0.1648$ ce qui signifie que ce modèle permet d'expliquer 16.48% de la variance du salaire horaire moyen dans notre échantillon.

La toute dernière ligne `F-statistic` donne la statistique de test et la p.value du test de Fisher de significativité du modèle. On remarque que cette statistique de test est extrêmement faible et largement inférieure à 5%. On rejette donc l'hypothèse nulle de non significativité du modèle. Notre modèle a un certain pouvoir explicatif sur notre variable.

La fonction `lm()` retourne une liste contenant différents objets. Pour connaître la teneur de ces objets, il faut se référer à la documentation en ligne ou bien en exécutant `?lm` dans la console afin de faire apparaître la documentation associée à la fonction. Il arrive (de manière trop fréquente) que la documentation laisse à désirer. Dans ce cas, il est possible de regarder les différents objets contenus dans notre modèle de régression grâce à la fonction `names()` qui retourne le nom des objets. Si les noms sont explicites cela peut être d'une grande aide. Si les noms ne sont pas explicites, on peut regarder dans le code source de la fonction en exécutant `lm` (sans parenthèse) dans la console. Il faut alors se débrouiller pour comprendre le code. Parfois cela est inévitable.

### Manipuler les résultats

```
# Déterminer les différents objets renvoyés par le modèle de régression
names(reg1)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

On peut voir que notre objet `reg1` contenant notre régression linéaire comprend 12 objets aux noms assez explicites que l'on va pouvoir appeler pour différents usages.

```
# Extraire les coefficients estimés
reg1$coefficients
```

```
(Intercept)        educ
 -0.9048516   0.5413593
```

L'objet `reg1_summary` dispose lui aussi d'un certain nombre d'objets pouvant être utiles.

```
names(reg1_summary)
```

```
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
# Extraire les coefficients estimés
reg1_summary$coefficients
```

```
             Estimate Std. Error   t value                         Pr(>|t|)
(Intercept) -0.9048516 0.68496782 -1.321013 0.18707350076244425030004758009
educ         0.5413593 0.05324804 10.166746 0.00000000000000000000002782599
```

Ici ce sont les coefficients ainsi que les écarts-types, t-stat et p.values qui sont retournées dans un objet de classe `matrix`. Lorsque l'on effectue des manipulations, des tests ou des procédures en tour genre nécessitant d'utiliser les estimations, il est possible de simplement afficher les coefficients puis de les noter à la main. Cette méthode est déconseillée puisqu'elle implique d'avoir à modifier manuellement au moindre changement, sans compter que cela devient très vite un calvaire si le nombre d'estimations et de coefficients grandit.

Etant un objet `matrix`, les règles de manipulation classiques sont faisables :

```
# Extraire les p.values des coefficients : toutes les lignes, 4ème colonne
reg1_summary$coefficients[,4]
```

```
                  (Intercept)                                  educ
0.18707350076244425030004758009 0.00000000000000000000002782599
```

Par exemple, on peut chercher à définir automatiquement si les coefficients sont statistiquement significatif au seuil de 5% ou non.

```
# Case_when() permet de tester des conditions sur des objets multiples et de
manière imbriquée
case_when(
  # Si les p.valuues sont inférieures à 0.05 alors le coefficient est
  significatif
  reg1_summary$coefficients[,4] <= 0.05 ~ "significatif",
  # Sinon le coefficient n'est pas significatif
  .default = "Non significatif"
) |>
  # Pour plus de clarté rajoute les noms des coefficients
  setNames(rownames(reg1_summary$coefficients))
```

```
       (Intercept)              educ
"Non significatif"     "significatif"
```

Une méthode simple de manipulation consiste à transformer l'output en un `tibble` afin de pouvoir le manipuler avec les techniques classiques de manipulation de dataframe.

```
df_reg1_coef <-
  reg1_summary$coefficients |>
  as_tibble() |>
  # Clean_names permet de nettoyer les noms des variables afin qu'ils soient
  conforme avec une syntaxe standard
  clean_names() |>
  mutate(
    # Ajouter les noms des coefficients
    coef = rownames(reg1_summary$coefficients),
    # Arrondir à 3 chiffres après la virgule pour gagner en lisibilité
    across(
      .cols = c(estimate, std_error, t_value, pr_t),
      .fns = \(variable) round(variable, 3)
    ),
    # Déterminer si le coefficient est significatif ou non
    signif =
      case_when(
        pr_t <= 0.05 ~ "significatif",
        .default = "non-significatif"
      )
  ) |>
  # Placer le nom des coef en première position
  relocate(coef) |>
  print()
```

```
# A tibble: 2 x 6
  coef        estimate std_error t_value  pr_t signif
  <chr>          <dbl>     <dbl>   <dbl> <dbl> <chr>
1 (Intercept)   -0.905     0.685   -1.32 0.187 non-significatif
2 educ           0.541     0.053   10.2  0     significatif
```

### Tests de Student

La p.value reportée dans le tablau de régression correspond à un test de Student bilatéral dans le quel hypothèse nulle testée est $\beta = 0$. On peut tester d'autres hypothèses de manière un peu plus manuelle.

On veut dans un premier temps tester si le coefficient d'éducation est statistiquement supérieur à 0. Il s'agit donc d'un test unilatéral :

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta > 0 \end{cases}$$

```r
# Extraire la valeur du coefficient de l'éducation
educ_coef_reg1 <-
  df_reg1_coef |>
  filter(coef == "educ") |>
  pull(estimate)

# Extraire l'erreur standard du coefficient de l'éduction
educ_ste_reg1 <-
  df_reg1_coef |>
  filter(coef == "educ") |>
  pull(std_error)

# Définir le risque à 5%
alpha <- 0.05

# Récupérer les degrés de liberté de la régression : N-K-1
freedom_degree <- reg1_summary$df[2]

# Calculer la statistique de test
t_stat <- educ_coef_reg1 / educ_ste_reg1

# Obtenir les valeurs critiques avec un niveau de confiance de 1-alpha (en
unilatéral on ne divise pas alpha par deux)
crit_val <- qt(1 - alpha, freedom_degree)

# Calculer la p_value (en unilatéral il s'agit de la p_value bilatérale
divisée par 2)
p_value <- (2*pt(-abs(t_stat), freedom_degree))/2

# Déterminer si le coefficient est significatif ou pas
result <- if_else(p_value <= alpha, "significatif", "non-significatif")

# Afficher une phrase de résultats
print(glue("le résultat du test de student unilatéral est : {result} avec une
p.value de {p_value} pour un risque de {alpha * 100}%"))
```

```
le résultat du test de student unilatéral est : significatif avec une p.value
de 0.00000000000000000000000980228196202429 pour un risque de 5%
```

On va maintenant chercher à déterminer si le coefficient de l'éducation est significativement différent de 0.5 ou pas :

$$\begin{cases} H_0 : \beta = 0.5 \\ H_1 : \beta \neq 0.5 \end{cases}$$

```r
# Extraire la valeur du coefficient de l'éducation
educ_coef_reg1 <-
  df_reg1_coef |>
  filter(coef == "educ") |>
  pull(estimate)

# Extraire l'erreur standard du coefficient de l'éduction
educ_ste_reg1 <-
  df_reg1_coef |>
  filter(coef == "educ") |>
  pull(std_error)

# Définir le risque à 5% et 1%
alpha <- c(0.05, 0.01)

# Récupérer les degrés de liberté de la régression : N-K-1
freedom_degree <- reg1_summary$df[2]

# Définir la valeur de beta que l'on veut tester : ici 0.5
mu <- 0.5

# Calculer la statistique de test
t_stat <- (educ_coef_reg1 - mu) / educ_ste_reg1

# Obtenir les valeurs critiques avec un niveau de confiance de 1-alpha/2
crit_val <- qt(1 - alpha/2, freedom_degree)

# Calculer la p_value
p_value <- 2*pt(-abs(t_stat), freedom_degree)

# Afficher les résultats
print(glue("La valeur critique à {alpha*100}% est {crit_val}"))
```

```
La valeur critique à 5% est 1.96450151697793
```

```
La valeur critique à 1% est 2.58524428855796
```

```
print(glue("La statistique de test est {abs(t_stat)}"))
```

```
La statistique de test est 0.773584905660378
```

```
print(glue("La p.value du test est {round(p_value, 3)}"))
```

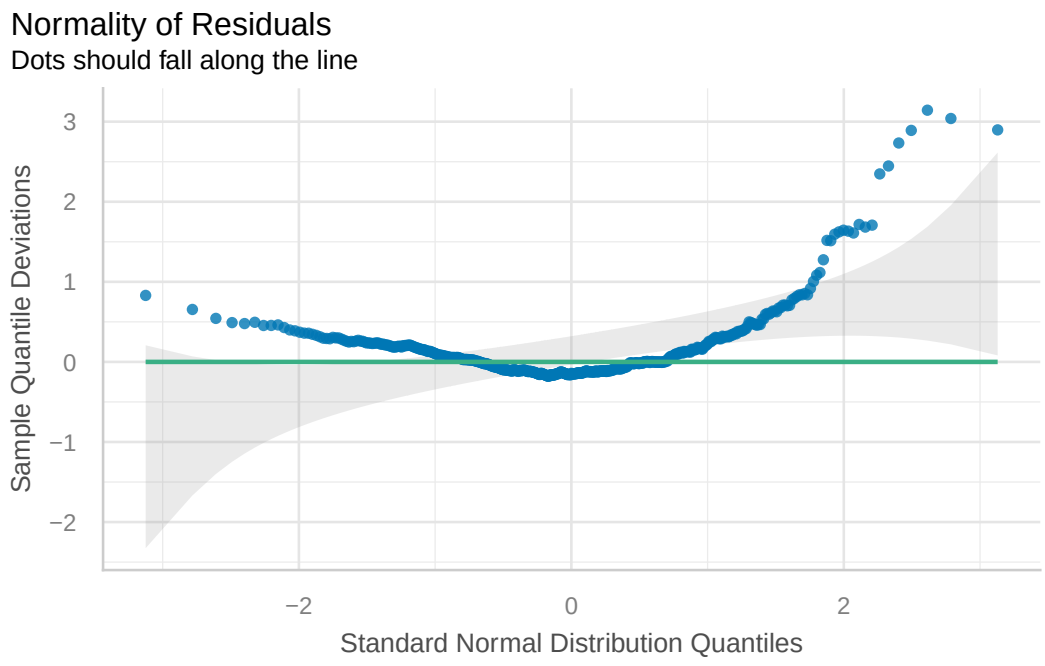
```
La p.value du test est 0.44
```

On ne peut pas dire que le coefficient de l'éducation soit significativement différent de 0.5. Si dans votre projet il est fréquent que vous ayez à effectuer des tests de Student avec d'autres types d'hypothèse, créez votre propre fonction afin de ne pas avoir à répéter ce type de code en permanence.

### Tester la normalité des résidus

Avant de tester statistiquement la normalité des résidus, il peut être intéressant de regarder graphiquement s'ils semblent suivre ou non une loi normale.

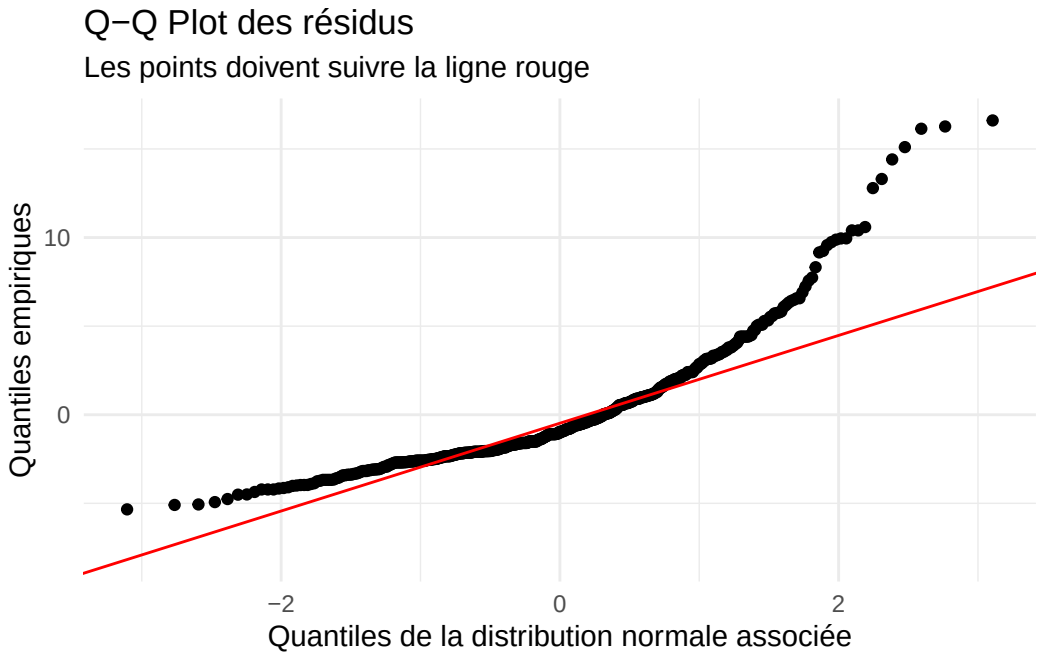
```
performance::check_normality(reg1) |>
  plot()
```



Ce graphique est un `QQ-plot`. Il compare la distribution des quantiles d'une loi normale (ligne verte) avec la distribution des quantiles des résidus (points bleus). Si les résidus suivent une loi normale, alors la distribution des quantiles des résidus doit se superposer à la ligne verte. Ici on remarque que les quantiles des résidus ne suivent absolument pas ceux d'une loi normale.

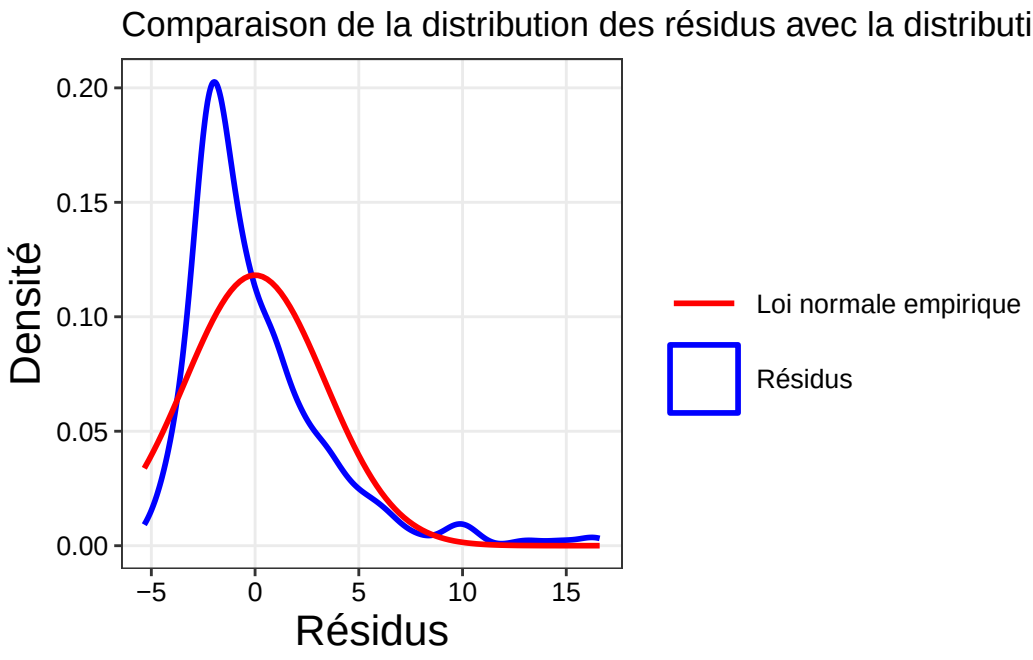
On peut Représenter le même type de graphique de manière plus manuelle :

```
tibble(residuals = reg1$residuals) |>
  ggplot(aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  theme_minimal() +
  labs(
    title = "Q-Q Plot des résidus",
    x = "Quantiles de la distribution normale associée",
    y = "Quantiles empiriques",
    subtitle = "Les points doivent suivre la ligne rouge"
    )
```




On peut également représenter la distribution des résidus et la comparer avec celle d'une loi normale ayant la même variance et la moyenne moyenne.

```r
# Représenter la distrution empirique des résidus ainsi que la distribution
de la loi normale qu'ils devraient suivre
  tibble(residuals = reg1$residuals) |>
    ggplot(aes(x = residuals)) +
    geom_density(aes(color = "Résidus"), linewidth = 1) +
    stat_function(
      aes(color = "Loi normale empirique"),
      fun = dnorm,
      args = list(mean = mean(reg1$residuals), sd = sd(reg1$residuals)),
      linewidth = 1
      ) +
    scale_color_manual(
      values = c("Résidus" = "blue", "Loi normale empirique" = "red")
    ) +
    labs(
      x = "Résidus",
      y = "Densité",
      color = "",
      title = "Comparaison de la distribution des résidus avec la
      distribution normale"
    )
```




Il semble, au vu des graphiques précédents, que nos résidus ne suivent pas une loi normale. Ils

semblent être caractérisés par une asymétrie à droite ainsi que des queues de distribution plus épaisses impliquant des valeurs extrêmes plus fréquentes. On rappelle que la loi normale est caractérisée par un coefficient de skewness égal à 0 et un coefficient de kurtosis égal à 3.

```
moments::skewness(reg1$residuals) # Coefficient de skewness des résidus
```

```
[1] 1.860679
```

```
moments::kurtosis(reg1$residuals) # Coefficient de kurtosis des résid
```

```
[1] 7.797006
```

On remarque que les résidus sont caractérisés par une skewness positive et une kurtosis plus élevée que celle de la loi normale ce qui semble nous indiquer que nos résidus ne suivent pas une loi normale. On peut tester individuellement si la skewness est statistiquement différente de 0 et si la kurtosis est statistiquement différente de 3 avec les tests d'Agostino et d'Anscombe :

```
# Tester si les résidus ont de la skewness
moments::agostino.test(reg1$residuals)
```

```

    D'Agostino skewness test

data:  reg1$residuals
skew = 1.8607, z = 12.2099, p-value < 0.00000000000000022
alternative hypothesis: data have a skewness
```

```
# Tester si les résidus ont de la kurtosis en excès
moments::anscombe.test(reg1$residuals)
```

```

    Anscombe-Glynn kurtosis test

data:  reg1$residuals
kurt = 7.797, z = 7.967, p-value = 0.000000000000001626
alternative hypothesis: kurtosis is not equal to 3
```

Statistiquement il semble que pour un seuil inférieur à 1%, les résidus soient caractérisés par une skewness positive et un excès de kurtosis par rapport à la loi normale, ce qui confirme un peu plus que les résidus ne semblent pas suivre une loi normale.

On peut également tester si conjointement les coefficients de skewness et de kurtosis sont égaux à 0 et 3 respectivement avec le test de Jarque et Bera

```
moments::jarque.test(reg1$residuals)
```

```
    Jarque-Bera Normality Test

data:  reg1$residuals
JB = 807.84, p-value < 0.00000000000000022
alternative hypothesis: greater
```

Le test nous indique que les résidus ne semblent pas avoir des coefficients de skewness et de kurtosis égaux à ceux de la loi normale pour un niveau de risque très faible. On peut en conclure que nos résidus ne suivent pas une loi normale.

On peut également tester l'hypothèse nulle que les données proviennent d'une population normalement échantillonnée avec le test de Shapiro-Wilk. On peut utiliser la fonction `check_normality()` du package `performance` pour faire cela, mais cela ne renvoit qu'un message et on ne peut pas extraire le résultat du test ce qui peut être handicapant dans certains cas.

```
performance::check_normality(reg1)
```

```
Warning: Non-normality of residuals detected (p < .001).
```

```
shapiro.test(reg1$residuals)
```

```
    Shapiro-Wilk normality test

data:  reg1$residuals
W = 0.84606, p-value < 0.00000000000000022
```

Encore une fois, ce test rejette la normalité des résidus pour un niveau de risque très faible. On peut conclure que nos résidus ne suivent pas une loi normale.

Cela n'est cependant pas un gros problème pour notre estimation puisque asymptotiquement nos estimateurs disposent de bonnes propriétés et que les tests statistiques usuels restent valides asymptotiquement. Or nous avons 526 observations ce qui est suffisant généralement pour utiliser les propriétés asymptotiques.

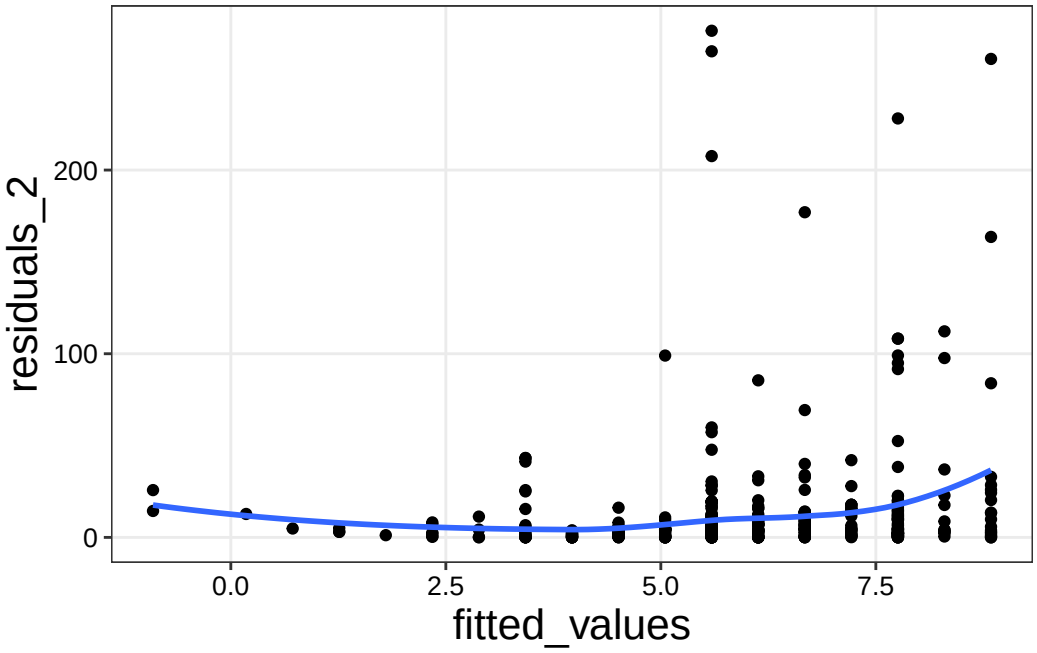
Attention car la non-normalité des résidus signifie quand même que tous les éléments générateurs des données n'ont pas été pris en compte ce qui peut être dérangeant notamment pour faire de la prévision puisque l'on sous-estime l'apparition de valeurs extrêmes.

### Tester l'homogénéité de la variance

Comme pour la normalité, on va commencer par une analyse graphique. Pour cela on va représenter les valeurs prédites du modèle en fonction des résidus au carré et observer la relation entre les deux.

```
tibble(
  residuals_2 = reg1$residuals**2,
  fitted_values = reg1$fitted.values
) |>
  ggplot(aes(x = fitted_values, y = residuals_2)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

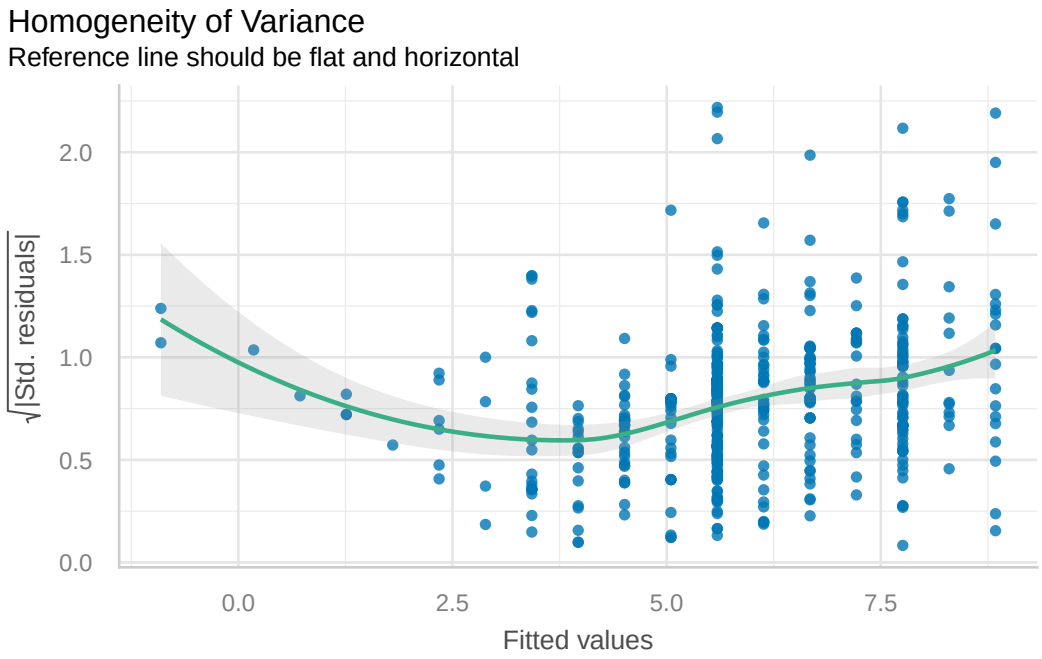
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




Il semblerait que la variance des résidus ne soit pas constante pour les différents niveaux de valeurs prédites ce qui est une suspicion de présence d'hétéroscédasticité parmi nos résidus.

La librairie `performance` permet également de réaliser ce genre de graphiques.

```
performance::check_heteroscedasticity(reg1) |>
  plot()
```




Pour tester la présence d'hétéroscédasticité, on peut utiliser le test de Breusch-Pagan (utilisé par la fonction `check_heteroscedasticity` mais qui encore une fois ne permet pas d'extraire les résultats) et le test de White qui tests tous les deux l'hypothèse nulle d'homoscédasticité des résidus.

```
# Breusch-Pagan test
performance::check_heteroscedasticity(reg1)
```

```
Warning: Heteroscedasticity (non-constant error variance) detected (p <
.001).
```

```
# Breusch-Pagan test
lmtest::bptest(reg1)
```

```

    studentized Breusch-Pagan test

data:  reg1
BP = 15.306, df = 1, p-value = 0.00009144
```

```
# White test
whitestrap::white_test(reg1)
```

```
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 23.24
P-value: 0.000009
```

Les tests de Breusch-Pagan et de White indiquent avec un risque faible que les résidus ne sont pas homoscédastiques. Ils sont caractérisés par une variance non constante Nos estimateurs ne sont donc plus de variance minimale ce qui réduit notre capacité à inférer l'information.

### Indépendance des variables explicatives

Afin que les estimateurs soient sans biais il est nécessaire qu'il n'existe aucun lien statistique entre les résidus et ces variables (si une variable est corrélée avec les résidus alors seul le coefficient de cette variable sera biaisé). Cela signifie qu'il faut s'assurer que notre modèle inclue toutes les variables pouvant expliquer $Y_i$ et étant corrélées avec $X_{ki}$. Si une telle variable a été omise alors le coefficient de la variable explicative sera biaisée puisque l'on ne peut pas savoir si l'évolution observée de la variable $X_{ki}$ est dû à sa propre évolution ou à celle de la variable omise.

On peut utiliser certaines techniques pour tester si nos variables sont exogènes (test de Hausman avec les variables instrumentales) mais il s'agit surtout d'une histoire d'argumentation sur le choix des variables inclues dans le modèle ou non.

Dans notre cas on peut légitimement penser que notre coefficient associé à *educ* n'est pas exogène. En effet, la durée de scolarisation peut être liée au sexe de la personne qui lui même peut influer sur le salaire. La durée d'éducation peut aussi être liée au talent de l'individu qui peut lui-même influer sur le salaire…

## Régression robuste à l'hétéroscédasticité

Nous venons de montrer que nos résidus ne sont pas caractérisés par une variance constante rendant ainsi la variance des estimateurs MCO non-minimale.

La méthode la plus pratique consiste à corriger les écarts-types de l'hétéroscédasticité et à utiliser ce que l'on appelle des "erreurs standards robustes".

```
car::hccm(reg1) # version raffinée de White
```

```
            (Intercept)         educ
(Intercept)   0.5378539 -0.044577403
educ         -0.0445774  0.003826987
```

```
car::hccm(reg1, type = "hc0") # version standard de White
```

```
            (Intercept)         educ
(Intercept)  0.52431924 -0.043490427
educ        -0.04349043  0.003738473
```

On peut obtenir une table de régression avec la fonction `coeftest()`.

```
lmtest::coeftest(reg1) # table standard non-corrigée
```

```
t test of coefficients:

             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.904852   0.684968  -1.321               0.1871
educ         0.541359   0.053248  10.167 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmtest::coeftest(reg1, vcov=hccm) # Version corrigée avec la correction
raffinée de White
```

```
t test of coefficients:

             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.904852   0.733385 -1.2338              0.2178
educ         0.541359   0.061863  8.7510 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmtest::coeftest(reg1, vcov=hccm(reg1, type = "hc0")) # Version corrigée avec
la version standard de White
```

```
t test of coefficients:

             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.904852   0.724099 -1.2496               0.212
educ         0.541359   0.061143  8.8540 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On peut voir que les résultats sur les écarts-types et p.values ne changent pas énormément ce qui semble indiquer que l'hétéroscédasticité n'est pas un trop grand problème. Notre coefficient de l'éducation est toujours significativement différent de 0 avec un risque très faible.

### Changement de forme fonctionnelle

La première régression nous indiquait que le salaire moyen horaire d'un individu augmentait de 0.541 dollars en moyenne lorsque son niveau d'étude augmentait de 1. On peut se demander s'il ne serait pas plus pertinent de réfléchir l'augmentation du salaire en pourcentage plutôt qu'en unité monétaire. D'une part cela rend probablement la compréhension plus facile et d'autre part cela permet d'avoir un effet non-linéaire de l'augmentation monétaire, mais un effet constant en pourcentage.

Le modèle devient donc :

$$log(\text{wage}_i) = \alpha + \beta \times \text{educ}_i + u_i$$

Pour intégrer une telle non-linéarité, on peut modifier notre jeu de données et intégrer une nouvelle variable `log_wage = log(wage)`. Ou alors on peut intégrer `log(wage)` directement dans notre formule.

```
reg2 <- lm(log(wage) ~ educ, data = wage1)

reg2_summary <- summary(reg2)

(lmtest::coeftest(reg2, vcov = hccm))
```

```
t test of coefficients:

             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 0.5837727  0.0994161   5.872        0.000000007651 ***
educ        0.0827444  0.0078291  10.569 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dans un tel cas, nous sommes dans un modèle Log-Niveau. L'interprétation de $\beta \approx 0.083$ est la suivante : une année d'étude supplémentaire augmente en moyenne le salaire horaire de l'individu de 8.3%. Ce coefficient est significativement différent de 0 même en prenant en compte une possible hétéroscédasticité.

Ici passer `educ` sous forme logarithme n'aurait aucun intérêt. Augmenter son nombre d'années d'études de 1% ne fait aucun sens. On raisonne en termes d'années supplémentaires.

### Régression multiple

Comme nous avons pu le mentionner, il est fort probable que

1. Le coefficient de l'éducation soit biaisé à cause de variables omises.
2. D'autres variables soient intéressantes dans une analyse s'intéressant aux déterminants du salaire.

Il pourrait être pertinent de se dire que le salaire dépend certes de son niveau d'éducation, mais aussi de son expérience professionnelle `exper` et de son nombre d'années avec son employeur actuel `tenure`.

Le modèle devient donc :

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{tenure}_i + u_i$$

```
reg3 <- lm(log(wage) ~ educ + exper + tenure, data = wage1)

(reg3_summary <- summary(reg3))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05802 -0.29645 -0.03265  0.28788  1.42809

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 0.284360   0.104190   2.729              0.00656 **
educ        0.092029   0.007330  12.555 < 0.0000000000000002 ***
exper       0.004121   0.001723   2.391              0.01714 *
tenure      0.022067   0.003094   7.133     0.00000000000329 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4409 on 522 degrees of freedom
Multiple R-squared:  0.316, Adjusted R-squared:  0.3121
F-statistic: 80.39 on 3 and 522 DF,  p-value: < 0.00000000000000022
```

Ces résultats nous indiquent que lorsque le nombre d'années d'études d'un individu augmente toutes choses égales par ailleurs (à expérience et tenure constante), alors le salaire moyen horaire augmente de 9.2% environ. Chaque année d'étude augmente en moyenne de 9.2% le salaire horaire d'un individu. Ce coefficient est significativement différent de 0 à un seuil inférieur à 1%.

Une année d'expérience professionnelle supplémentaire augmente en moyenne le salaire horaire de 0.41% toute chose égale par ailleurs. Ce coefficient est significativement différent de 0 à un seuil de 5% mais pas de 1%.

Une année supplémentaire avec le même employeur augmente en moyenne le salaire horaire de 2.2% tous les autres facteurs maintenus constants. Ce coefficient est significativement différent de 0 à un seuil de 1%.

Ce modèle permet d'expliquer 31.6% de la variance de la variable endogène. Ce modèle a statistiquement un pouvoir explicatif à un seuil de 1%.

### Ajout d'un terme au carré

Théoriquement il est possible de penser qu'une année supplémentaire d'expérience fasse augmenter le salaire mais que passer un certain moment dans la carrière d'un individu cela ne soit plus vrai. On peut tester cela en ajoutant un terme au carré dans la régression qui devient :

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{tenure}_i + u_i$$

Comme pour les logarithme, on peut modifier notre jeu de données et inclure une variable `exper_2 = exper ** 2`. Mais on peut aussi l'inclure directement dans la formule de régression grâce à la fonction `I()`

```
reg4 <- lm(log(wage) ~ educ + exper + I(exper**2) + tenure, data = wage1)

(reg4_summary <- summary(reg4))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure,
    data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97087 -0.26809 -0.03463  0.27663  1.28678

Coefficients:
```

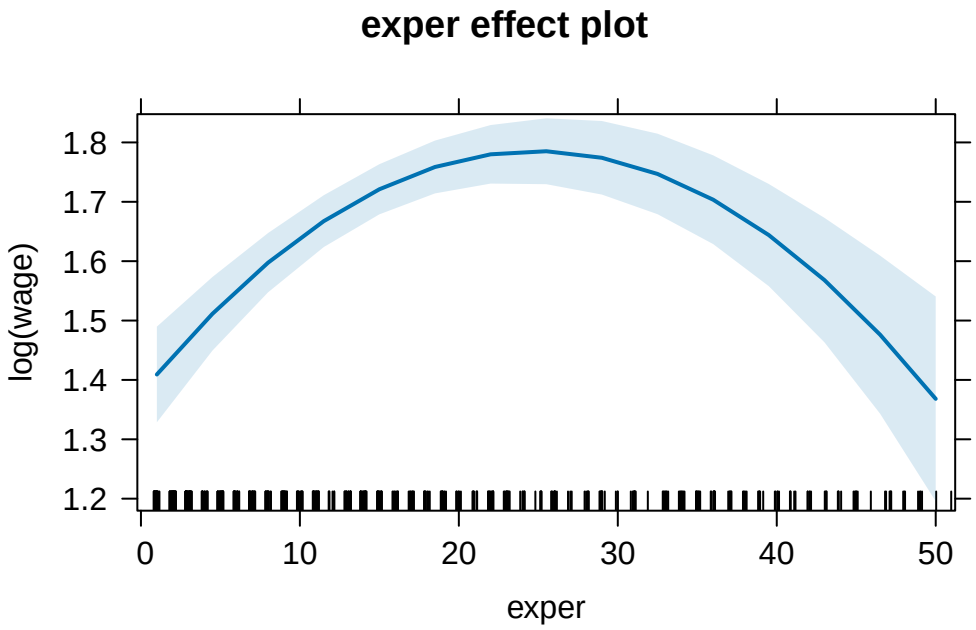
```
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.1983445  0.1019556   1.945                0.0523 .
educ         0.0853489  0.0071885  11.873 < 0.0000000000000002 ***
exper        0.0328542  0.0051135   6.425      0.0000000002979 ***
I(exper^2)  -0.0006606  0.0001111  -5.945      0.0000000050775 ***
tenure       0.0208413  0.0030037   6.938      0.0000000000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.427 on 521 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3545
F-statistic: 73.09 on 4 and 521 DF,  p-value: < 0.00000000000000022
```

Dans un tel modèle, l'interprétation des coefficients `educ` et `tenure` ne change pas. Par contre il n'est plus possible d'interpréter `exper` de la même manière puisqu'il y a désormais deux termes à prendre en compte. Il est possible de calculer les effets marginaux de manière manuelle. La fonction `effect()` du package `effect` permet de faire cela automatiquement. Basiquement, elle prend comme argument le nom de la variable dont on veut mesurer l'effet et le modèle de régression associé.

```
# Affciher l'effet non linéaire de la variable expérience
effects::effect("exper", reg4) |>
  print() |>
  plot()
```

```
 exper effect
exper
       1       10       30       40       50
1.409140 1.639426 1.768013 1.634120 1.368102
```




On peut voir que l'expérience a tout d'abord un effet positif sur le salaire puis un effet négatif en fin de carrière toutes choses égales par ailleurs. Attention la valeur sur l'axe des ordonnées ne peut pas être interpréter comme une valeur de coefficient. Par exemple, pour 10 années d'expérience $log(\text{wage}) = 1.639$ mais cela NE SIGNIFIE PAS qu'avoir 10 ans d'expérience fait augmenter le salaire de 163.9%. Pour avoir l'information de l'augmentation du salaire pour une année d'étude supplémentaire on utilise la formule :

$$\Delta \widehat{log(\text{wage})} \approx (\beta_2 + 2\beta_3 \text{exper}) \times \Delta \text{exper}$$

On peut calculer le point de retournement avec la formule :

$$x^* = \frac{-\hat{\beta}_3}{2\hat{\beta}_3}$$

```
coef_exper_reg4 <- reg4$coefficients[3]
coef_exper_2_reg4 <- reg4$coefficients[4]

retournement <- -coef_exper_reg4 / (2*coef_exper_2_reg4)

print(glue("Après {round(retournement)} années d'expérience, une année
d'expérience supplémentaire fait diminuer le salaire"))
```

```
Après 25 années d'expérience, une année d'expérience supplémentaire fait
diminuer le salaire
```

```r
exper_i <- 1
change <- 1
effet_1 <- (coef_exper_reg4 + 2 * coef_exper_2_reg4 * exper_i) * change

print(glue("Une personne avec {exper_i} ans d'expérience verra son salaire
varier de {round(effet_1,4)*100}% en augmentant son nombre d'années
d'expérience de {change}"))
```

```
Une personne avec 1 ans d'expérience verra son salaire varier de 3.15% en
augmentant son nombre d'années d'expérience de 1
```

```r
exper_i <- 30
change <- 1
effet_2 <- (coef_exper_reg4 + 2 * coef_exper_2_reg4 * exper_i) * change

print(glue("Une personne avec {exper_i} ans d'expérience verra son salaire
varier de {round(effet_2, 4)*100}% en augmentant son nombre d'années
d'expérience de {change}"))
```

```
Une personne avec 30 ans d'expérience verra son salaire varier de -0.68% en
augmentant son nombre d'années d'expérience de 1
```

Attention, la formule pour calculer le changement prédit n'est qu'une approximation. Elle convient bien pour des petits changements mais pas pour des grands. Dans ce cas on calcule la valeur prédite de $Y_i$ pour deux valeurs différentes de $X_i$ et on fait la différence entre les deux.

### Test de Fischer contraint

Pour tester la significativité d'une variable quadratique, on peut utiliser le test de Student classique sur la variable en niveau puis celle au carré. Cependant comme ces variables sont fortement corrélées, il est courant que ces variables ne soient pas individuellement significative. Pour tester la significativité, on peut utiliser le test de Fischer contraint avec l'hypothèse nulle que les coefficients associés à l'expérience et l'expérience au carrée sont égaux à 0.

```r
car::linearHypothesis(
  reg4,
  c("exper = 0", "I(exper^2) = 0")
)
```

```
Linear hypothesis test:
exper = 0
I(exper^2) = 0

Model 1: restricted model
Model 2: log(wage) ~ educ + exper + I(exper^2) + tenure

  Res.Df     RSS Df Sum of Sq      F         Pr(>F)
1    523 102.567
2    521  95.011  2    7.5561 20.717 0.000000002201 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
car::linearHypothesis(
  reg4,
  c("exper = 0", "I(exper^2) = 0"),
  vcov. = hccm
)
```

```
Linear hypothesis test:
exper = 0
I(exper^2) = 0

Model 1: restricted model
Model 2: log(wage) ~ educ + exper + I(exper^2) + tenure

Note: Coefficient covariance matrix supplied.

  Res.Df Df    F           Pr(>F)
1    523
2    521  2 22.6 0.0000000003869 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On peut voir que l'on peut rejeter l'hypothèse nulle de non significativité avec le test normal et le test robuste à l'hétéroscédasticité. Prise ensemble les variables de l'expérience et son terme au carré ont un effet significatif.

### Ajout d'une variable binaire

Il est commun de vouloir prendre en compte certaines caractéristiques des individus étudiés par le biais de variables binaires. On peut supposer par exemple que la relation liant le salaire à l'éducation est la même entre les hommes et les femmes à l'exception près que les hommes auront de manière systématique un salaire plus élevé que celui des femmes.

Pour tester cela on va introduire une variable binaire *female* qui va prendre la valeur 1 si l'individu est une femme.

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{tenure}_i + \delta_1 \text{female}_i + u_i$$

```
reg5 <- lm(log(wage) ~ educ + exper + I(exper^2) + tenure + female, data =
wage1)

(reg5_summary <- summary(reg5))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure +
    female, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83192 -0.26206 -0.01504  0.23972  1.14937

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.4146366  0.0994195   4.171  0.00003559829083431 ***
educ         0.0809586  0.0067841  11.934 < 0.0000000000000002 ***
exper        0.0328100  0.0048111   6.820  0.00000000002542656 ***
I(exper^2)  -0.0006480  0.0001046  -6.197  0.00000000116931519 ***
tenure       0.0162150  0.0028808   5.629  0.00000002974313522 ***
female      -0.2979067  0.0359802  -8.280  0.00000000000000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4018 on 520 degrees of freedom
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.4286
F-statistic: 79.77 on 5 and 520 DF,  p-value: < 0.00000000000000022
```

Une femme (female = 1) aura en moyenne toute chose égale par ailleurs (à caractéristiques égales à celle d'un homme) un salaire inférieur de 29% par rapport à un homme.

On peut également penser qu'il existe un écart permanent entre les individus catégorisés "blanc" et les autres. Pour cela, on intègre la variable *nonwhite* qui prend la valeur 1 si l'individu n'est pas blanc.

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{tenure}_i + \delta_1 \text{female}_i + \delta_2 \text{nonwhite}_i + u_i$$

```
reg6 <- lm(log(wage) ~ educ + exper + I(exper^2) + tenure + female +
nonwhite, data = wage1)

(reg6_summary <- summary(reg6))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure +
    female + nonwhite, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83436 -0.26015 -0.01746  0.23702  1.14824

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.4202759  0.1004005   4.186  0.00003334911554995 ***
educ         0.0807012  0.0068171  11.838 < 0.0000000000000002 ***
exper        0.0328548  0.0048161   6.822  0.00000000002511277 ***
I(exper^2)  -0.0006493  0.0001047  -6.202  0.00000000113854243 ***
tenure       0.0162263  0.0028832   5.628  0.00000002989768939 ***
female      -0.2981506  0.0360134  -8.279  0.00000000000000106 ***
nonwhite    -0.0243631  0.0580159  -0.420                0.675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4021 on 519 degrees of freedom
Multiple R-squared:  0.4343,    Adjusted R-squared:  0.4277
F-statistic:  66.4 on 6 and 519 DF,  p-value: < 0.00000000000000022
```

Il semblerait qu'il n'y ait pas de différence systématique entre les personnes blanches et les autres puisque la p.value du test de Student est largement supérieure à 5%.

### Terme d'interaction

Il existe un écart permanent entre le salaire des hommes et des femmes toutes choses égales par ailleurs. Mais on peut supposer que la différence entre les homme et les femmes se fait également sur le rendement de l'éducation. On peut penser qu'il existe des inégalités telles qu'une année d'éducation pour une femme est moins bien valorisée qu'une année d'éducation pour un homme. Pour tester cela on va intégrer une variable d'interaction entre la variable *educ* et la variable *female*.

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{tenure}_i + \delta_1 \text{female}_i + \delta_2 \text{educ}_i.\text{female}_i + u_i$$

```
reg7 <- lm(log(wage) ~ educ + exper + I(exper^2) + tenure + female +
educ*female, data = wage1)

# Similaire à
# reg7 <- lm(log(wage) ~ exper + I(exper^2) + educ*female, data = wage1)

(reg7_summary <- summary(reg7))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure +
    female + educ * female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8327 -0.2557 -0.0175  0.2447  1.1631

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.3945828  0.1192685   3.308                0.001 **
educ         0.0825245  0.0085128   9.694 < 0.0000000000000002 ***
exper        0.0327610  0.0048180   6.800      0.0000000000289 ***
I(exper^2)  -0.0006468  0.0001047  -6.175      0.0000000013380 ***
tenure       0.0162552  0.0028864   5.632      0.0000000292617 ***
female      -0.2478035  0.1681813  -1.473                0.141
educ:female -0.0039994  0.0131133  -0.305                0.760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4021 on 519 degrees of freedom
```

```
Multiple R-squared:  0.4342,    Adjusted R-squared:  0.4276
F-statistic: 66.37 on 6 and 519 DF,  p-value: < 0.00000000000000022
```

On remarque qu'il existe toujours toutes choses égales par ailleurs une différence moyenne de 24.7% de salaire entre les hommes et les femmes. Cependant ce coefficient n'est plus significatif. La variable d'interaction n'est quand à elle pas du tout significative individuellement. On remarque tout de même, qu'une année d'éducation supplémentaire pour une femme se traduit en moyenne par un rendement de l'éducation plus faible de 0.3% (même si on ne peut pas dire que ce résultat soit significativement de 0).

On peut interpréter ce type de modélisation comme le fait d'avoir un modèle pour les hommes (lorsque *female* égal 0) et un modèle pour les femmes (lorsque *female* égal 1). On peut donc tester si le modèle pour les hommes est le même modèle que pour les femmes en tester la significativité conjointe de la variable *female* avec le terme d'interaction.

```r
car::linearHypothesis(
  reg7,
  c("female = 0", "educ:female = 0")
)
```

```
Linear hypothesis test:
female = 0
educ:female = 0

Model 1: restricted model
Model 2: log(wage) ~ educ + exper + I(exper^2) + tenure + female + educ *
    female

  Res.Df    RSS Df Sum of Sq      F              Pr(>F)
1    521 95.011
2    519 83.929  2    11.082 34.264 0.00000000000001055 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prises conjointement les deux variables sont significatives au seuil de 5% et même en-dessous. Il semble donc que le modèle entre les hommes et les femmes diffère. La difficulté est de savoir s'il diffère uniquement au niveau de la constante (donc simplement un écart persistant) ou bien s'il diffère aussi au niveau de la relation. Au vu du faible niveau de significativité de la variable d'interaction et du fait que le $R^2$ ajusté diminue avec l'inclusion de cette variable, il semble que la différence se fasse uniquement au niveau de la constante.

### Ajout d'une variable qualitative à plusieurs modalités

Il est possible d'inclure dans le modèle une variable qualitative à plusieurs niveaux. La base de données *wage1* comprend les variables *northcen*, *south* et *west* qui indiquent si un individu vit dans la région nord-centrales des US, vit dans la région sud ou dans la région ouest. Il s'agit de trois variables binaires que l'on peut intégrer dans le modèle et que l'on interprétera par rapport aux habitants de la région est (donné par la valeur de la constante).

$$log(wage_i) = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{tenure}_i + \delta_1 \text{female}_i + \delta_2 \text{northcen} + \delta_3 \text{south} + \delta_4 \text{west} + u_i$$

Cependant, les variables qualitatives seront assez rarement pré-traitées comme ceci. il est plus courant d'avoir une variable contenant différentes chaînes de caractères indiquant les différentes modalités. Nous allons recréer une telle variable afin de montrer que cela s'utilise très facilement.

```
wage1_modif <-
  wage1 |>
  mutate(
    # Créer une variable région qui indique la région dans laquelle vit
    l'individu
    region =
      case_when(
        northcen == 1 ~ "north-central",
        south == 1 ~ "south",
        west == 1 ~ "west",
        .default = "east"
      ),
    # Mise sous type "factor". Pas obligatoire mais utile
    # Peut aussi utiliser la fonction : factor de base de R
    region = forcats::fct(region, levels = unique(region))
  ) |>
  select(wage, region, educ, exper, tenure, female)

head(wage1_modif)
```

```
  wage region educ exper tenure female
1 3.10   west   11     2      0      1
2 3.24   west   12    22      2      1
3 3.00   west   11     2      0      0
4 6.00   west    8    44     28      0
5 5.30   west   12     7      2      0
6 8.75   west   16     9      8      0
```

La variable région est désormais une variable de type “facteur” qui est le type utilisé pour les variables catégorielles. Il n’est pas obligé pour une régression de transformer la variable en type facteur mais cela a quelques avantages comme permettre d’éviter les typographies si on spécifie les différentes catégories possibles ou encore d’ordonner les catégories.

Dans la fonction `forcats::fct()` si l’argument `levels` n’est pas indiqué, alors la liste des catégories possibles sera définie comme étant l’ensemble unique des éléments dans la variable (ce qui ici ne change rien). La fonction `levels` permet d’obtenir les différentes catégories de la variable.

```
levels(wage1_modif$region)
```

```
[1] "west"          "east"          "south"         "north-central"
```

La variable de référence qui sera utilisée dans la régression correspond au premier niveau affiché. Ici ce sera *west*. Pour pouvoir comparer le résultat avec les variables dummy il faudrait que la variable de référence soit *east*.

```
# Définir l'ordre voulu pour les régions
region_order <- c("east", "west", "south", "north-central")

wage1_modif <-
  wage1_modif |>
  mutate(
    # On indique que "east" est la modalité de référence
    region = relevel(region, "east")
  )

levels(wage1_modif$region)
```

```
[1] "east"          "west"          "south"         "north-central"
```

Il n’est pas nécessaire pour l’estimation de créer les variables binaire. On peut directement donner la variable catégorielle.

```
reg8 <- lm(
  log(wage) ~ educ + exper + I(exper^2) + tenure + female + region,
            data = wage1_modif
  )

(reg8_summary <- summary(reg8))
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure +
    female + region, data = wage1_modif)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92320 -0.24657 -0.01521  0.24396  1.18278

Coefficients:
                     Estimate Std. Error t value             Pr(>|t|)
(Intercept)          0.4669302  0.1047449   4.458      0.0000101623059 ***
educ                 0.0795662  0.0067954  11.709 < 0.0000000000000002 ***
exper                0.0336184  0.0048290   6.962      0.0000000000102 ***
I(exper^2)          -0.0006615  0.0001049  -6.307      0.0000000006129 ***
tenure               0.0158252  0.0028709   5.512      0.0000000559773 ***
female              -0.3048205  0.0358330  -8.507 < 0.0000000000000002 ***
regionwest           0.0606929  0.0562190   1.080                0.2808
regionsouth         -0.0914848  0.0475479  -1.924                0.0549 .
regionnorth-central -0.0582118  0.0506611  -1.149                0.2511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3991 on 517 degrees of freedom
Multiple R-squared:  0.4449,    Adjusted R-squared:  0.4363
F-statistic: 51.79 on 8 and 517 DF,  p-value: < 0.00000000000000022
```

Ce qui donne le même résultat que :

```r
summary(
  lm(
    log(wage) ~ educ + exper + I(exper^2) + tenure + female + northcen +
    south + west,
             data = wage1
    )
)
```

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + tenure +
    female + northcen + south + west, data = wage1)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.92320 -0.24657 -0.01521  0.24396  1.18278

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.4669302  0.1047449   4.458      0.0000101623059 ***
educ         0.0795662  0.0067954  11.709 < 0.0000000000000002 ***
exper        0.0336184  0.0048290   6.962      0.0000000000102 ***
I(exper^2)  -0.0006615  0.0001049  -6.307      0.0000000006129 ***
tenure       0.0158252  0.0028709   5.512      0.0000000559773 ***
female      -0.3048205  0.0358330  -8.507 < 0.0000000000000002 ***
northcen    -0.0582118  0.0506611  -1.149                0.2511
south       -0.0914848  0.0475479  -1.924                0.0549 .
west         0.0606929  0.0562190   1.080                0.2808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3991 on 517 degrees of freedom
Multiple R-squared:  0.4449,    Adjusted R-squared:  0.4363
F-statistic: 51.79 on 8 and 517 DF,  p-value: < 0.00000000000000022
```

Seule la modalité *south* est statistiquement significative au seuil de 10%. Cela signifie que les individus vivant dans le sud des Etats-Unis ont en moyenne un salaire horaire inférieur de 9.1% toutes choses égales par ailleurs comparé aux individus vivant dans l'est des Etats-Unis. En revanche les individus vivant dans le nord ou l'ouest n'ont pas statistiquement un salaire horaire différent des individus vivant dans l'est.

Encore une fois, il peut être judicieux de tester si l'inclusion de toutes ces modalités a conjointement un effet significatif.

```
car::linearHypothesis(
  reg8,
  c("regionnorth-central = 0", "regionwest = 0", "regionsouth = 0")
  )
```

```
Linear hypothesis test:
regionnorth-central = 0
regionwest = 0
regionsouth = 0

Model 1: restricted model
Model 2: log(wage) ~ educ + exper + I(exper^2) + tenure + female + region
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    520 83.944
2    517 82.339  3    1.6056 3.3604 0.01862 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Toutes les modalités de lieux de vie sont conjointement significatives au seuil de 5%.

### Multicolinéarité

Une des conditions pour que le modèle soit identifiable est la non-colinéarité parfaite entre des variables du modèle. Il ne faut pas qu'une variable puisse être parfaitement expliquée par une combinaison linéaire d'autres variables. Le cas le plus classique se trouve lors de l'inclusion de variables qualitatives. Il faut dans ce cas retirer une modalité (ou la constante) qui deviendra la modalité de référence.

La multicolinéarité est un problème différent. On parle de multicolinéarité lorsque des variables sont fortement liées entre elles. Si des variables explicatives sont très fortement corrélées entre elles cela fait augmenter l'erreur-standard des estimateurs. En effet, les estimateurs étant des coefficients de corrélation partiels purgés des effets des autres variables, si des variables sont trop corrélées les unes aux autres cela brouille la précision de l'estimation.

On peut tester cela en regardant les facteurs d'inflation de la variance (VIF). Le VIF mesure à quel point la variance d'un paramètre augmente à cause de l'inclusion des autres variables. Pour tester la colinéarité des variables on peut utiliser la fonction `check_colinearity()` ou bien la fonction `vif` pour récupérer un tableau de résultat.

```
performance::check_collinearity(reg8)
```

```
# Check for Multicollinearity

Low Correlation

   Term  VIF     VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
   educ 1.17 [ 1.09,  1.33]     1.08      0.86     [0.75, 0.92]
 tenure 1.42 [ 1.29,  1.60]     1.19      0.71     [0.62, 0.78]
 female 1.06 [ 1.01,  1.29]     1.03      0.94     [0.78, 0.99]
 region 1.05 [ 1.01,  1.33]     1.01      0.95     [0.75, 0.99]

High Correlation

       Term   VIF     VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
      exper 14.16 [12.06, 16.65]     3.76      0.07     [0.06, 0.08]
 I(exper^2) 13.77 [11.73, 16.19]     3.71      0.07     [0.06, 0.09]
```

```r
vif(reg8)
```

```
                GVIF Df GVIF^(1/(2*Df))
educ        1.167169  1        1.080356
exper      14.159610  1        3.762926
I(exper^2) 13.766198  1        3.710283
tenure      1.418023  1        1.190808
female      1.058321  1        1.028747
region      1.047749  3        1.007804
```

On peut interpréter le coefficient VIF de la façon suivante : l'erreur standard du coefficient d'éducation serait $\sqrt{1.16} = 1.077$ fois plus élevée que l'erreur standard de ce même coefficient si aucune variable n'était incluse. Cela nous montre que l'erreur standard de l'expérience est $\sqrt{14.15} \approx 3.76$ fois plus élevée que si aucune autre variable n'était incluse. Cela tient au fait que l'expérience est très fortement corrélée avec l'expérience au carré.

généralement, on dira qu'un VIF supérieur 5 indique une colinéarité moyenne tandis qu'un VIF supérieur à 10 indique une colinéarité élevée. Dans ces cas là, plusieurs solutions sont possibles. Ou bien On supprime une des variables avec un VIF élevé pour voir si la colinéarité a disparu (ici on pourrait supprimer l'expérience au carré) :

```r
lm(
  log(wage) ~ educ + exper + tenure + female + region, data = wage1_modif
  ) |>
  performance::check_collinearity()
```

```
# Check for Multicollinearity

Low Correlation

   Term  VIF   VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
   educ 1.13 [1.06, 1.30]     1.06      0.88     [0.77, 0.94]
  exper 1.49 [1.35, 1.68]     1.22      0.67     [0.59, 0.74]
 tenure 1.41 [1.28, 1.60]     1.19      0.71     [0.63, 0.78]
 female 1.06 [1.01, 1.29]     1.03      0.94     [0.78, 0.99]
 region 1.03 [1.00, 1.70]     1.00      0.97     [0.59, 1.00]
```

Seulement parfois on ne peut/veut pas supprimer une variable de noter modèle puisque l'on estime que cette variable est importante. On estime par exemple que l'expérience a une relation quadratique qui est indispensable à prendre en compte. Dans ce cas on peut utiliser les tests de Fischer pour tester si conjointement les variables colinéaires ont un impact significatif sur le modèle.

### Prédiction

On souhaite prédire le salaire qu'aurait un homme "moyen" habitant dans l'est dans l'échantillon. Pour cela on va créer un dataframe qui contiendra les valeurs moyennes des prédicteurs puis on utilisera la fonction `predict` pour obtenir notre prédiction.

```
df_indiv_moyen <-
  wage1_modif |>
  summarize(
    across(
      .cols = c(educ, exper, tenure),
      ~ mean(.x)
    ),
    female = 0,
    region = "east"
  )

predict(reg8, df_indiv_moyen)
```

```
       1
1.927795
```

On peut voir qu'un homme "moyen" habitant dans l'est aurait un salaire en log prédit de 1.93 environ, soit un salaire horaire prédit de $exp(1.93) \approx 6.87$ dollars.

On peut aussi faire de multiples prédictions d'un coup. Par exemple on peut regarder le salaire prédit des individus caractéristiques à différents quantiles.

```
df_indiv_quantile <-
  tibble(
    educ = quantile(wage1$educ, probs = seq(0.1, 0.9, by = 0.1)),
    exper = quantile(wage1$exper, probs = seq(0.1, 0.9, by = 0.1)),
    tenure = quantile(wage1$tenure, probs = seq(0.1, 0.9, by = 0.1))
  )

# PAsser directement à l'exponentielle pour avoir une interprétation en
salaire horaire
exp(predict(reg4, df_indiv_quantile))
```

```
       1        2        3        4        5        6        7        8
2.799808 3.518482 4.137702 4.507997 4.890760 5.741605 6.710100 7.605957
       9
8.767651
```

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (Intercept) | 0.584*** | 0.284** | 0.198+ | 0.415*** |
| | (0.097) | (0.104) | (0.102) | (0.099) |
| educ | 0.083*** | 0.092*** | 0.085*** | 0.081*** |
| | (0.008) | (0.007) | (0.007) | (0.007) |
| exper | | 0.004* | 0.033*** | 0.033*** |
| | | (0.002) | (0.005) | (0.005) |
| tenure | | 0.022*** | 0.021*** | 0.016*** |
| | | (0.003) | (0.003) | (0.003) |
| I(exper^2) | | | −0.001*** | −0.001*** |
| | | | (0.000) | (0.000) |
| female | | | | −0.298*** |
| | | | | (0.036) |
| Num.Obs. | 526 | 526 | 526 | 526 |
| R2 Adj. | 0.184 | 0.312 | 0.355 | 0.429 |
| AIC | 2432.4 | 2344.8 | 2312.3 | 2249.1 |

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

### Tables de régression

Savoir présenter ses résultats est (presque) aussi important que d'obtenir les résultats. Il est évidemment hors de question de réécrire manuellement les résultats dans un tableau World ou une table LaTeX. De librairies telles que `stargazer` ou `modelsummary` existent afin de permettre de créer les tables voulues sous divers formats.

```
list_exported_models <-
  list(
    reg2, reg3, reg4, reg5
  )


modelsummary::modelsummary(
  list_exported_models,
  stars = TRUE,
  gof_map = c("nobs", "adj.r.squared", "aic")
  )
```