STT5100 - Hiver 2020 - Examen Intra (OLS)

Arthur Charpentier

Examen **A**

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une page d'aide mémoire. L'examen dure 3 heures, mais toute sortie avant midi est autorisée, et sera définitive.

Dans les feuilles qui suivent, il y a 30 questions relatives au cours sur les modèles linéaires. Pour chaque question (sauf deux), cinq réponses sont proposées. Une seule est valide, et vous ne devez en retenir qu'une,

- vous gagnez 1 point par bonne réponse
- vous ne perdez pas de points pour une mauvaise réponse
- vous ne gagnez pas de point pour plusieurs réponses

Aucune justification n'est demandée. Une question repose sur un graphique qu'il faudra tracer sur la feuille de réponses (au dos). Votre note finale est le total des points (sur 30). Il y a une 31ème question, bonus. Une prédiction parfaite (sur 30) donnera un point bonus qui s'ajoutera à la note.

La page de réponses est au dos de celle que vous lisez présentement : merci de décrocher ladite feuille et de ne rendre que cette dernière, après avoir indiqué votre code permanent en haut à gauche.

Merci de cocher le carré en bleu ou en noir. En cas d'erreur, vous pouvez cocher une autre case en rouge. Seule cette dernière sera alors retenue.

En plus des 15 pages du présent document, vous devez avoir une annexe de sortie R de 24 pages.

Le surveillant ne répondra à aucune question durant l'épreuve : en cas de soucis sur une question (interprétation possiblement fausse, typo, etc), vous pouvez mettre un court commentaire sur la feuille de réponses.

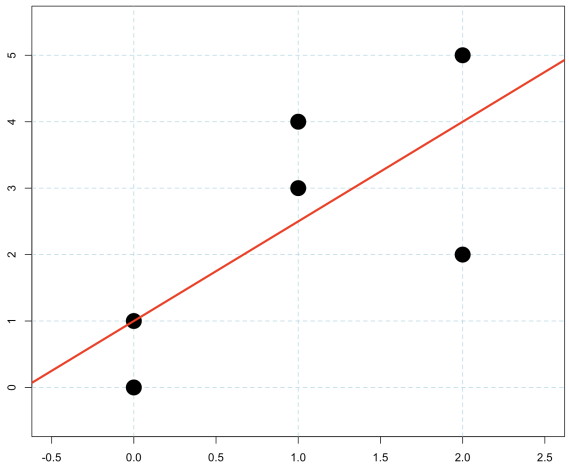
Formulaire : Quantiles de lois usuelles. Exemple pour une loi normale - $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z \leqslant 2.326) = 99\%$.

	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Student (50)	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
Student (30)	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
Student (20)	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.849
Student (15)	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
Student (10)	0.700	1.372	1.812	2.228	2.764	3.169	4.143	4.587
Student (9)	0.703	1.383	1.833	2.262	2.821	3.250		
Student (8)	0.706	1.397	1.860	2.306	2.896	3.355		
Student (7)	0.711	1.415	1.895	2.365	2.998	3.499		
Student (6)	0.718	1.440	1.943	2.447	3.143	3.707		
Student (5)	0.727	1.476	2.015	2.571	3.365	4.032		
Student (4)	0.741	1.533	2.132	2.776	3.747	4.604		
Student (3)	0.765	1.638	2.353	3.182	4.541	5.841		

Code permanent : _____ Sujet : A

question 1	☐ A	☐ B	■ C	☐ D	☐ E
question 2	☐ A	■ B	☐ C	☐ D	☐ E
question 3	☐ A	☐ B	■C	☐ D	☐ E
question 4	☐ A	☐ B	☐ C	■ D	☐ E
question 5	☐ A	☐ B	☐ C	■ D	☐ E
question 6	■ A	☐ B	☐ C	☐ D	☐ E
question 7	☐ A	☐ B	■ C	☐ D	☐ E
question 8	☐ A	☐ B	☐ C	☐ D	■ E
question 9	☐ A	☐ B	☐ C	■ D	☐ E
question 10	☐ A	■ B	☐ C	☐ D	☐ E
question 11	☐ A	☐ B	☐ C	■ D	☐ E
question 12	☐ A	☐ B	■C	☐ D	☐ E
question 13	figure à droite				
question 14	☐ A	☐ B	☐ C	☐ D	■ E
question 15	☐ A	■ B	☐ C	☐ D	☐ E
question 16	■ A	☐ B	☐ C	☐ D	☐ E
question 17	☐ A	☐ B	■C	☐ D	☐ E
question 18	■ A	☐ B	☐ C	☐ D	☐ E
question 19	☐ A	☐ B	☐ C	■ D	☐ E
question 20	☐ A	☐ B	☐ C	■ D	☐ E
question 21	☐ A	☐ B	☐ C	■ D	☐ E
question 22	☐ A	☐ B	☐ C	■ D	☐ E
question 23	■ A	☐ B	☐ C	☐ D	☐ E
question 24	■ A	☐ B	☐ C	☐ D	☐ E
question 25	☐ A	■ B	☐ C	☐ D	☐ E
question 26	■ A	☐ B	☐ C	☐ D	☐ E
question 27	☐ A	■ B	☐ C	☐ D	☐ E
question 28	☐ A	■ B	☐ C	☐ D	☐ E
question 29	☐ A	☐ B	☐ C	☐ D	■ E
question 30	☐ A	■ B	☐ C	☐ D	☐ E
question 31	*Combien de bonnes réponses pensez vous avoir ?*				

question 13 :



_____

Les sujets A et B posaient les mêmes questions (partiellement) dans un ordre différent, et pour la question 13 ici (9 pour le sujet B) les données étaient symmétriques, en x ($x_i \to 2 - x_i$). Je ne referais pas ce cas dans la correction qui suit.

1 On a simulé les données suivantes :

$x_i = i$ pour $i = 1, \cdots, 50$, les ε_i sont tirés suivant des lois $\mathcal{N}(0,1)$, et $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ avec $\beta_0 = 2$ et $\beta_1 = 1$.

On obtient la sortie suivante

```
> x = 1:50
> epsilon = rnorm(50)
> y = 2+x+epsilon
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.250      0.281    7.99 2.26e-10 ***
x              0.980      0.014    69.9  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Que vaut le biais de $\widehat{\beta}_1$ obtenu par moindres carrés ?

A) -0.02

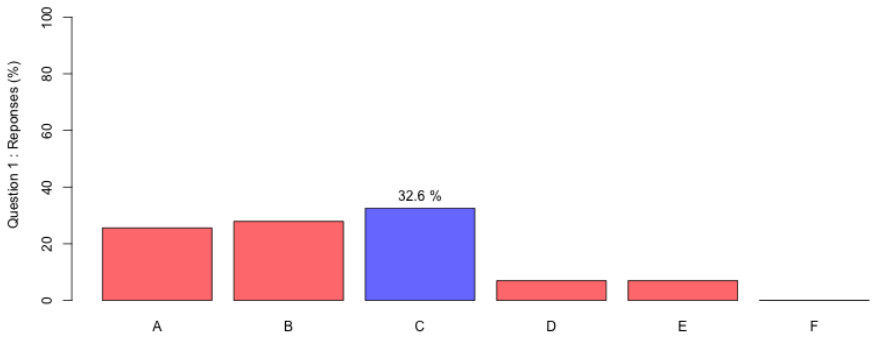
B) $+0.02$

C) 0

D) $+0.25$

E) On n'a pas assez d'information pour répondre

Je renvoie au cours (STT5100, mais aussi probableement *stats 1*), mais le *biais* n'est pas quelque chose que l'on observe sur les données, c'est une propriété théorique d'un estimateur. On a ici les hypothèses H1 (plus d'observations que de variables et identifiabilité du modèle - ce qui arrive toujours avec une seule variable explicative qui n'est pas catégorielle) et H2 (bruit centré, de variance constante et indépendent) donc – cf cours – *l'estimateur $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ obtenu par moindres carrés est sans biais* – au sens où $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Donc a fortiori, le biais de $\widehat{\beta}_1$ est nul ! Donc C est la bonne réponse.



2 De manière générale, on considère le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où on suppose ε_i centré, de variance constante, et indépendants les uns des autres. On propose plusieurs estimateurs pour β_1,

$$\widehat{\beta}_1^{(1)} = \frac{\overline{y}}{\overline{x}}, \widehat{\beta}_1^{(2)} = \frac{y_2 - y_1}{x_2 - x_1} \text{ et } \widehat{\beta}_1^{(3)} = \frac{\max\{y_i\} - \min\{y_i\}}{\max\{x_i\} - \min\{x_i\}}.$$

A) $\widehat{\beta}_1^{(1)}$ est un estimateur sans biais de β_1

B) $\widehat{\beta}_1^{(2)}$ est un estimateur sans biais de β_1

C) $\widehat{\beta}_1^{(3)}$ est un estimateur sans biais de β_1

D) les trois sont des estimateurs sans biais de β_1

E) aucun n'est un estimateur sans biais de β_1

Faisons simple : montront que A) est fausse (ce qui exclue A et D) et que B) est correcte (ce qui exclura E... et C car il faut une seule réponse).

On nous dit que $\widehat{\beta}_1^{(2)} = \dfrac{y_2 - y_1}{x_2 - x_1}$ et il faut calculer son biais, donc on va voir $\widehat{\beta}_1^{(2)}$ comme une variable aléatoire (et remplace y_i par Y_i), i.e.

$$\mathbb{E}(\widehat{\beta}_1^{(2)}) = \mathbb{E}\left(\frac{Y_2 - Y_1}{x_2 - x_1}\right) = \mathbb{E}\left(\frac{(\beta_0 + \beta_1 x_2 + \varepsilon_2) - (\beta_0 + \beta_1 x_1 + \varepsilon_1)}{x_2 - x_1}\right)$$

bref

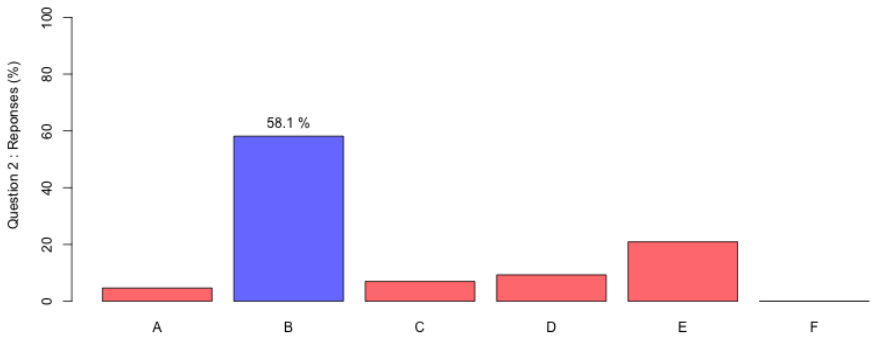
$$\mathbb{E}(\widehat{\beta}_1^{(2)}) = \beta_1 \frac{x_2 - x_1}{x_2 - x_1} + \underbrace{\mathbb{E}\left(\frac{\varepsilon_2 - \varepsilon_1}{x_2 - x_1}\right)}_{=0} = \beta_1$$

donc oui : $\widehat{\beta}_1^{(2)}$ est un estimateur sans biais de β_1.

Pour le premier estimateur,

$$\mathbb{E}(\widehat{\beta}_1^{(1)}) = \mathbb{E}\left(\frac{\overline{Y}}{\overline{x}}\right) = \mathbb{E}\left(\frac{\beta_0 + \beta_1 \overline{x} + \overline{\varepsilon}}{\overline{x}}\right) = \frac{\beta_0}{\overline{x}} + \beta_1 + 0 \neq \beta_1$$

donc non : $\widehat{\beta}_1^{(1)}$ n'est pas un estimateur sans biais de β_1



3 Toujours sur le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où on suppose ε_i centré, de variance constante, et indépendants les uns des autres, on estime les coefficients par moindres carrés. Quelles affirmations parmi les suivantes sont justes ?

i) la somme des carrés des résidus estimés est toujours nulle

ii) la somme des carrés des résidus estimés est nulle à condition que $\overline{y} = 0$

iii) si $R^2 = 0$, $\widehat{\beta}_1 = 0$ (et la droite de régression est horizontale)

iv) la droite de régression $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ passe par le point $(\overline{x}, \overline{y})$ à condition que ce point soit un point de l'échantillon

A) (i) seulement

B) (ii) seulement

C) (i) et (iii)

D) (ii) et (iii)

E) (i) et (iv)

(i) oui, on l'a vu plusieurs fois en cours, c'est simplement la condition du premier ordre lorsqu'on différencie la somme des carrés des erreurs par rapport à β_0, i.e. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$, ce qui se traduit aussi par $\sum_{i=1}^{n} \widehat{\varepsilon}_i$.

A fortiori (ii) est fausse.

(iii) si $\widehat{\beta}_1 = 0$ alors $R^2 = 0$ (on reverra ce point dans la question 16 - la seconde sur les sorties de régression des annexes). Mais là, on nous demande la réciproque. Si $R^2 = 0$ alors $\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = 0$. La seule possibilité pour qu'une somme de termes positifs soit nulle est que *tous* les termes sont nuls. Donc pour tout i, $\widehat{y}_i - \overline{y}$, autrement dit, le modèle estimé est tout simplement un modèle constant, prenant la valeur $\overline{y}$, soit $\widehat{\beta}_0 = \overline{y}$ et $\widehat{\beta}_1 = 0$. Donc oui, (iii) est juste.

(iv) La droite de régression passe *toujours* par $(\overline{x}, \overline{y})$ - au risque de me répéter, c'est une conséquence directe de la condition du premier ordre $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$. En effet

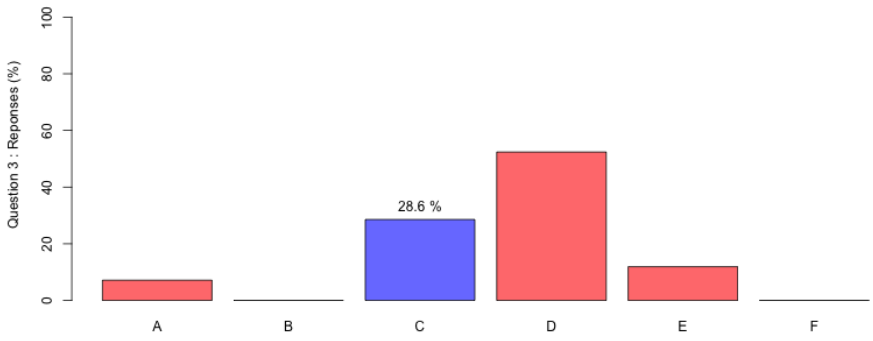
$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} \widehat{\beta}_0 + \widehat{\beta}_1 x_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i$$

soit, en divisant par n,

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\overline{y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\overline{x}}$$

ce qui signifie que la droite de régression passe par $(\overline{x}, \overline{y})$. Donc (iv) est fausse.

Si on conclue, (i) et (iii) sont valides (et ce sont les seules), donc la bonne réponse est C.



4 On a estimé un modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, sur un premier échantillon. On a obtenu

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 50, \sum_{i=1}^{n} x_i = 10, \sum_{i=1}^{n} x_i^2 = 100, \widehat{\beta}_0 = 2 \text{ et } \widehat{\beta}_1 = 1$$

Sur un autre échantillon de même taille, on a obtenu

$$\sum_{i=1}^{n} \tilde{\varepsilon}_i^2 = 80, \sum_{i=1}^{n} x_i = 10, \sum_{i=1}^{n} x_i^2 = 100, \tilde{\beta}_0 = 2 \text{ et } \tilde{\beta}_1 = 1$$

Que peut-on dire sur les statistiques de test t pour nos différents estimateurs (estimés par moindres carrés)

A) $t_{\hat{\beta}_0} \leqslant t_{\tilde{\beta}_0}$ et $t_{\hat{\beta}_1} \leqslant t_{\tilde{\beta}_1}$

B) $t_{\hat{\beta}_0} \leqslant t_{\tilde{\beta}_0}$ et $t_{\hat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$

C) $t_{\hat{\beta}_0} \geqslant t_{\tilde{\beta}_0}$ et $t_{\hat{\beta}_1} \leqslant t_{\tilde{\beta}_1}$

D) $t_{\hat{\beta}_0} \geqslant t_{\tilde{\beta}_0}$ et $t_{\hat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$

E)$t_{\hat{\beta}_0} = t_{\tilde{\beta}_0}$ et $t_{\hat{\beta}_1} = t_{\tilde{\beta}_1}$

Comme les estimateurs sont identiques ($\widehat{\beta}_1 = \tilde{\beta}_1$ et $\widehat{\beta}_0 = \tilde{\beta}_0$), comparer les statistiques de test revient à comparer les variances : comme les statistiques sont positives (elles sont toujours du même signe que les estimateurs, qui sont ici positifs), une statistique plus grande correspond à une variance plus petite.

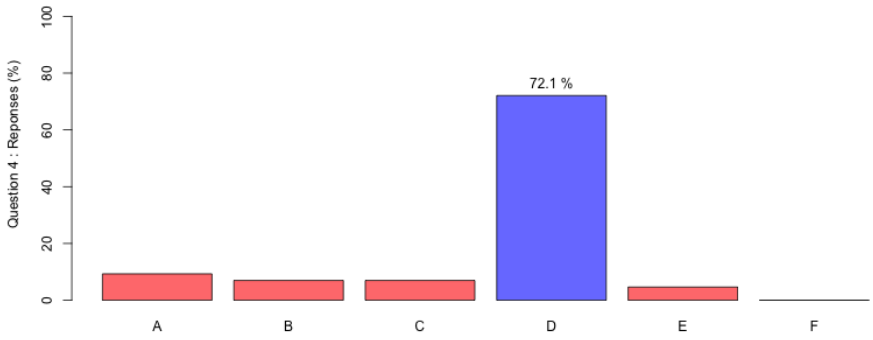
Si on utilise une écriture matricielle, on sait que la variance de nos estimateurs s'écrit $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Or comme la somme des X_i et la somme des X_i^2 est la même, le terme de droite $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ ne change pas. Aussi, ce qu'il faut comparer, ce sont juste les estimations de variance des résidus, σ^2. Comme les résidus sont centrés, et que n est la même, on compare la somme des carrés des résidus. Qui est plus grande pour le second modèle que pour le premier. Aussim

$$\text{Var}(\tilde{\beta}_0) > \text{Var}(\hat{\beta}_0) \text{ et } \text{Var}(\tilde{\beta}_1) > \text{Var}(\hat{\beta}_1)$$

de telle sorte que

$$t_{\hat{\beta}_0} \geqslant t_{\tilde{\beta}_0} \text{ et } t_{\hat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$$

qui est la réponse D.



5 On a estimé une régression simple, $y = \beta_0 + \beta_1 x + \varepsilon$, avec n observations. On note $\sigma^2 = \text{Var}(\varepsilon)$ et s^2 la variance de la variable x. Parmi les 5 cas suivant, quel cas correspond au cas où la variance de $\widehat{\beta}_1$ est la plus faible

A) $\sigma^2 = 4$, $n = 12$ et $s^2 = 5$

B) $\sigma^2 = 3$, $n = 11$ et $s^2 = 6$

C) $\sigma^2 = 4$, $n = 11$ et $s^2 = 5$

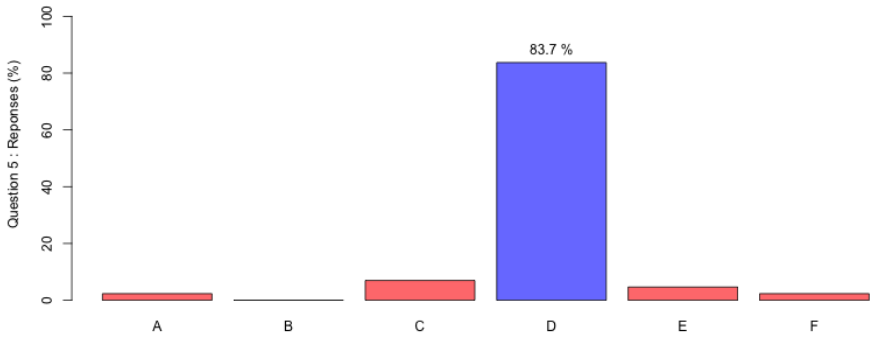
D) $\sigma^2 = 2$, $n = 13$ et $s^2 = 6$

E) $\sigma^2 = 3$, $n = 13$ et $s^2 = 5$

Pour rappels la variance de $\widehat{\beta}_1$ est

$$\text{Var}(\widehat{\beta}_1) = \frac{\text{Var}(\varepsilon)}{n\text{Var}(X)}$$

(je renvoie au cours). On peut calculer cette valeur dans les 5 cas, mais on peut aussi noter que cette variance sera minimal si σ^2 est faible, et si n et S^2 est grand. Ce qui correspond au cas D (pas besoin de faire ici tous les calculs).



6 Toujours dans un modèle de régression simple, $y = \beta_0 + \beta_1 x + \varepsilon$, on utilise 6 observations. L'intervalle de confiance à 99% pour β_1 est $[0, 2]$. Quel serait l'intervalle de confiance à 95% pour β_1 ?

A) [0.232, 1.768]

B) [0.397, 1.603]

C) [0.537, 1.463]

D) [0.216, 1.784]

E) [0.362, 1.638]

Commençons par le faire (trop) rapidement. L'intervalle de confiance à 99% pour β_1 est

$$\left[\widehat{\beta}_1 \pm 2.575829\sqrt{\text{Var}(\widehat{\beta}_1)}\right]$$

car $\mathbb{P}(Z < -2.576) = 0.5\%$, si $Z \sim \mathcal{N}(0,1)$, i.e. $\mathbb{P}(Z \in [-2.576; 2.576]) = 99\%$, compte tenu des valeurs indiquées sur la page de garde. Aussi, comme l'intervalle de confiance est ici centré sur 1 on en déduit que $\widehat{\beta}_1 = 1$,

et que $\sqrt{\text{Var}(\widehat{\beta}_1)} = 1/2.575829$. Maintenant, on sait aussi que l'intervalle de confiance à 95% pour β_1 est

$$\left[\widehat{\beta}_1 \pm 1.959964\sqrt{\text{Var}(\widehat{\beta}_1)}\right]$$

donc, par substitution,

$$\left[1 \pm \frac{1.959964}{2.575829}\right] = [1 \pm 0.760906] = [0.239094; 1.760906]$$

qui ressemble à la réponse A.

Le raisonnement précédant est juste, sauf qu'on a seulement 6 observations, donc on doit utiliser une loi de Student à 6-2=4 degrés de liberté. Auss l'intervalle de confiance à 99% pour β_1 est

$$\left[\widehat{\beta}_1 \pm 4.604095\sqrt{\text{Var}(\widehat{\beta}_1)}\right]$$

car $\mathbb{P}(Z < -4.604095) = 0.5\%$, si $Z \sim \mathcal{N}(0,1)$, i.e. $\mathbb{P}(Z \in [-4.604095; 4.604095]) = 99\%$, compte tenu des valeurs indiquées sur la page de garde (le tableau indiquait 4.604). Aussi, comme l'intervalle de confiance est ici centré sur 1 on en déduit que $\widehat{\beta}_1 = 1$, et que $\sqrt{\text{Var}(\widehat{\beta}_1)} = 1/4.604095$. Maintenant, on sait aussi que l'intervalle de confiance à 95% pour β_1 est

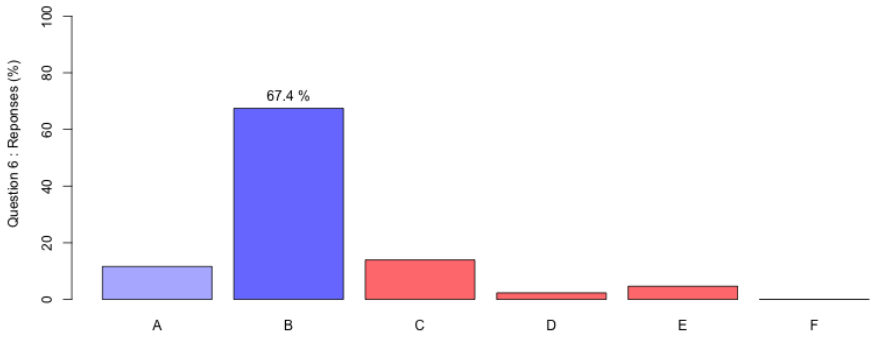
$$\left[\widehat{\beta}_1 \pm 2.776445\sqrt{\text{Var}(\widehat{\beta}_1)}\right]$$

(en prenant le quantile à 97.5% de la loi normale) donc, par substitution,

$$\left[1 \pm \frac{2.776445}{4.604095}\right] = [1 \pm 0.6030382] = [0.3969618; 1.6030382]$$

qui ressemble à la réponse B. La bonne réponse était donc B.

Cet exercice était un vieil exercice de la SOA, et comme je l'ai dit en cours, je ne suis pas un fan des régressions avec 6 observations, donc je donne un point pour la réponse A aussi



7 On ajuste un modèle $y = \beta_0 + \beta_1 x + \varepsilon$ sur $n = 100$ observations, où x est une variable prenant les valeurs 0 ou 1. Dans 40% des cas, x_i a pris la valeur 1. On nous dit que

$$\widehat{\beta}_1 = 1.4 \text{ et } \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = 920.$$

Que vaut la statistique du test de Student associé au test de significativité $H_0 : \beta_1 = 0$?

A) 1.15

B) 1.78

C) 2.26

D) 2.46

E) 3.51

L'estimateur (classique) de σ^2 est

$$\widehat{\sigma}^2 = \frac{1}{100-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{920}{98} \sim 9.2$$

Pour avoir l'écart type de notre estimateur $\widehat{\beta}_1$, il nous manque le terme

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

Comme ici x_i prend les valeurs 0 ou 1, $x_i^2 = x_i$. Donc la relation précédante se simplifie,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i - \overline{x}^2 = n\overline{x} - n\overline{x}^2 = 40 - 16 = 24$$

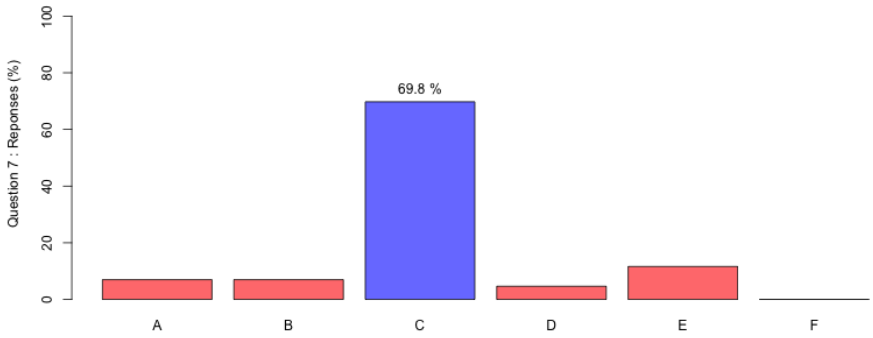
(qui est juste la variance d'une loi binomiale, $n\overline{X}(1-\overline{x})$). Aussi, l'écart type de notre estimateur $\widehat{\beta}_1$ s'écrit

$$\sqrt{\mathsf{Var}(\widehat{\beta}_1)} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \sqrt{\frac{9.2}{24}} = 0.6191.$$

Aussi, la statistique de test est

$$t = \frac{\widehat{\beta}_1}{\sqrt{\mathsf{Var}(\widehat{\beta}_1)}} = \frac{1.4}{0.6191} = 2.2612$$

qui correspond au cas C.



8 On dispose d'un jeu de données $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. On considère deux modèles

$$y_i = bx_i + u_i \text{ et } x_i = ay_i + v_i$$

avec les conditions usuelles (en particulier H2). On consière les estimateurs par moindres carrés de a et b

A) $\widehat{a} \cdot \widehat{b} = 1$

B) $\widehat{b} \sum_{i=1}^{n} x_i = \widehat{a} \sum_{i=1}^{n} x_i$

C) $\widehat{b} \sum_{i=1}^{n} x_i = \widehat{a} \sum_{i=1}^{n} y_i$

D) $\widehat{b} \sum_{i=1}^{n} y_i^2 = \widehat{a} \sum_{i=1}^{n} x_i^2$

E) $\widehat{b} \sum_{i=1}^{n} x_i^2 = \widehat{a} \sum_{i=1}^{n} y_i^2$

Pour la première équation, l'estimateur par moindres carrés de b est

$$\widehat{b} = \text{argmin} \left\{ \sum_{i=1}^{n} (y_i - bx_i)^2 \right\},$$

dont la condition du premier ordre est

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - bx_i)^2 \Big|_{b=\widehat{b}} = \sum_{i=1}^{n} \frac{\partial (y_i - bx_i)^2}{\partial b} \Big|_{b=\widehat{b}} = \sum_{i=1}^{n} 2x_i \big(y_i - \widehat{b} x_i\big) = 0$$

donc

$$\widehat{b} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

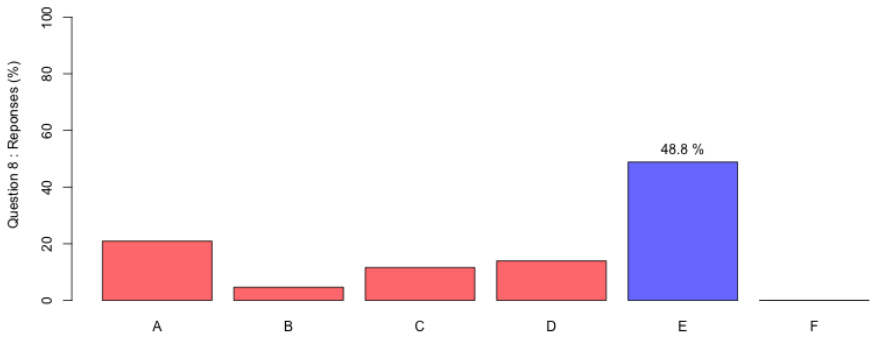
et on va s'arrêter là. Parce que si on regarde le second modèle, et qu'on cherche l'estimateur par moindres carrés de a, on va aboutir à l'équation

$$\widehat{a} \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i x_i,$$

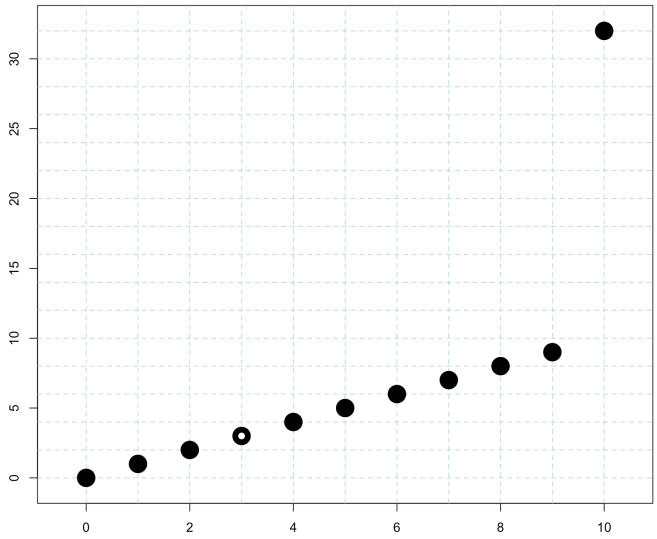
et donc, en égalisant, on obtient

$$\widehat{b} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i = \widehat{a} \sum_{i=1}^{n} y_i^2,$$

qui est l'expression E.



9 Sur la Figure suivante, avec 11 observations (avec $x_i = (i-1)$ pour $i = 1, 2, \ldots, 11$), on nous dit que $\widehat{\varepsilon}_4 = 0$. Quelle proportion de résidus estimés $\widehat{\varepsilon}_i$ sont strictement positifs



A) 7 sur 11

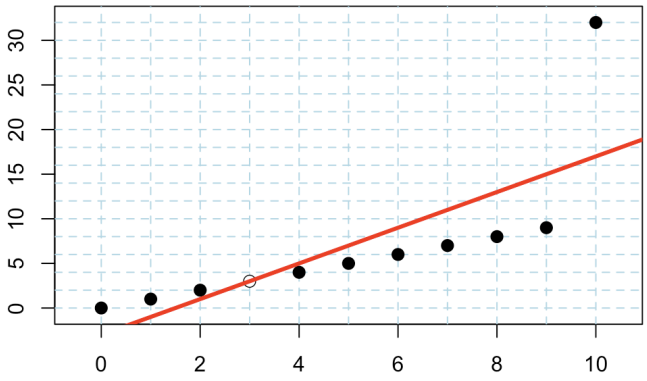
B) 6 sur 11

C) 5 sur 11

D) 4 sur 11

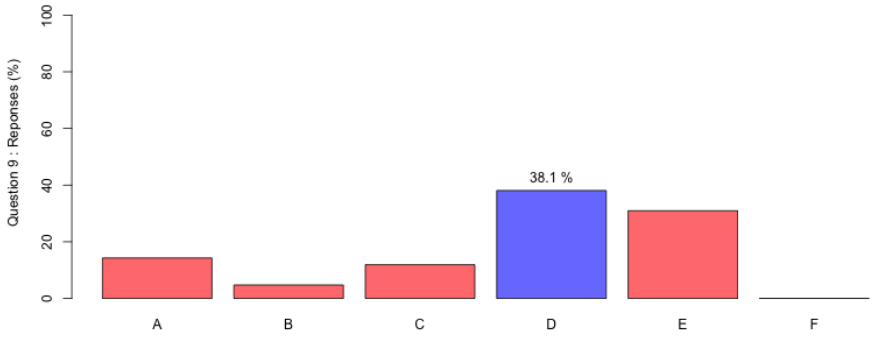
E) on ne peut pas savoir

Bon, on peut le faire avec les mains... On nous dit qu'on passe par le quatrième point, $(3, 3)$. Tout à droite (en $x = 10$) on peut légitimement penser que $\widehat{y}$ est entre 10 (c'est ce qu'on aurait eu si les 11 points étaient alignés) et 32 (qui est la valeur observée) : ca serait irréaliste d'être sous 10, et au dessus de 32. Bref, on a forcément une droite comme sur le dessin suivant



Si on compte, ca fait 3+1 résidus estimés strictement positifs (3 tout à gauche, et 1 à droite).

Mais on peut le faire plus rigoureusement, en notant que $\overline{x} = 5$ et $\overline{y} = 7$, donc on peut calculer proprement l'équation de la droite, qui passe par $(3, 3)$ et $(5, 7)$: $y = -3 + 2x$. Si on trace cette droite, on a exactement la droite de la figure ci-dessus, et on retrouve le fait que 4 résidus estimés sont strictement positifs.



Les deux prochaines questions (10 et 11) portent sur les sorties suivantes. On considère dans un premier temps la régression double

```
> summary(lm(y~x1+x2, data=df))


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.06328    0.30678   3.466  0.00121 **
x1           0.92449    0.04714  19.613  < 2e-16 ***
x2          -0.98824    0.08822 -11.202 2.46e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3963 on 43 degrees of freedom
Multiple R-squared:  0.9704,Adjusted R-squared:  0.969
F-statistic: 704.3 on 2 and 43 DF,  p-value: < 2.2e-16
```

Dans un second temps, on décide de centrer et réduire les variables explicatives

```
> (m <- sapply(df, mean))
         y         x1         x2
-0.4643975  1.4143586  2.8689534
> (s <- sapply(df, sd))
        y        x1        x2
2.2508261 1.6639146 0.8890444
> df$x1tilde = (df$x1 - m[2])/s[2]
> df$x2tilde = (df$x2 - m[3])/s[3]
> summary(lm(y~x1tilde+x2tilde,data=df))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.46440    0.05843  -7.948 5.62e-10 ***
x1tilde     XXXXXXXX    0.07843  XXXXXXXXXXXXXXXXXX
x2tilde     XXXXXXXX    0.07843  XXXXXXXXXXXXXXXXXX
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: XXXXXX on 43 degrees of freedom
Multiple R-squared:  XXXXXX,Adjusted R-squared:  XXXXX
F-statistic: XXXXXX on 2 and 43 DF,  p-value: XXXXXXXXXXXX
```

On a ici, avec le premier modèle

$$y_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_{1,i} + \widehat{\alpha}_2 x_{2,i} + \widehat{\varepsilon}_i$$

que l'on peut réécrire

$$y_i = \underbrace{\widehat{\alpha}_0 + \widehat{\alpha}_1 \overline{x}_1 + \widehat{\alpha}_2 \overline{x}_2}_{\widehat{\beta_0}} + \underbrace{s_1 \widehat{\alpha}_1}_{\widehat{\beta_1}} \underbrace{\frac{x_{1,i} - \overline{x}_1}{s_1}}_{\widetilde{x}_{1,i}} + \underbrace{s_2 \widehat{\alpha}_2}_{\widehat{\beta_2}} \underbrace{\frac{x_{2,i} - \overline{x}_2}{s_2}}_{\widetilde{x}_{2,i}} + \widehat{\varepsilon}_i$$

autrement dit, si on régresse sur les variables centrées, on a

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 \widetilde{x}_{1,i} + \widehat{\beta}_2 \widetilde{x}_{2,i} + \widehat{\varepsilon}_i$$

10 Que vaut $\widehat{\sigma}^2$ dans le second modèle ?

A) 0

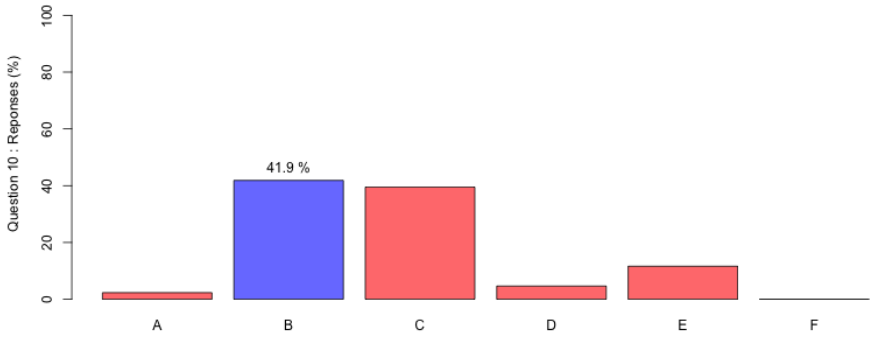
B) 0.157

C) 0.396

D) 0.970

E) On n'a pas assez d'information pour savoir

Des calculs faits en intro, notons que les deux modèles donnt *exactement* les mêmes résidus, et donc les mêmes prévisions... ce sont les mêmes modèles ! Les modèles sont identiques quand on transforme linéairement les variables. Mais ici, comme on a les mêmes résidus, ils sont la même variance... donc $\widehat{\sigma}^2$ est identique dans les deux modèles, i.e. 0.3963^2 d'après la première sortie, ce qui correspond à la réponse B.



11 Que vaut $\widehat{\beta}_1$ dans le second modèle ?

A) 0.490

B) 0.924

C) 1.414

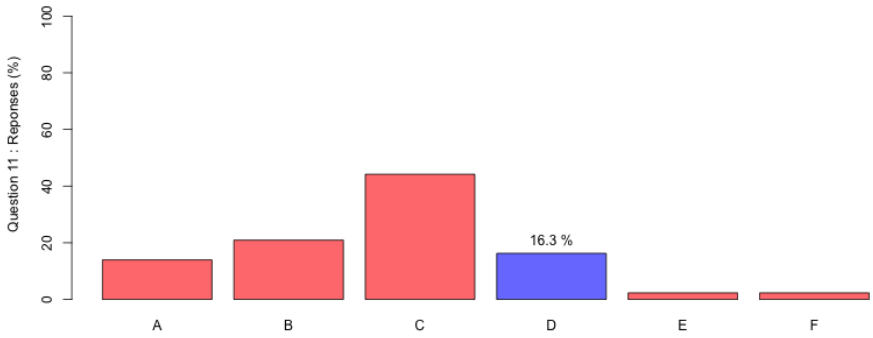
D) 1.538

E) 2.338

On a les mêmes modèles, et en particulier, comme on se contente de tranformer les variables une à une, les estimateurs ont les mêmes significativité ! Autrement dit, les statistiques de Student sont identiques, pour $\widehat{\beta}_1$ et $\widehat{\alpha}_1$ (et $\widehat{\beta}_2$ et $\widehat{\alpha}_2$ aussi). De là, on peut en déduire facilement la valeur de $\widehat{\beta}_1$:

$$\frac{\widehat{\alpha}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\alpha}_1]}} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = 19.613 \text{ et donc } \widehat{\beta}_1 = 19.61\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}$$

soit ici $19.61 \cdot 0.07843 = 1.538$ qui et la réponse D.



12 Les résidus sont dits homoscédastiques si

A) les résidus sont robustes

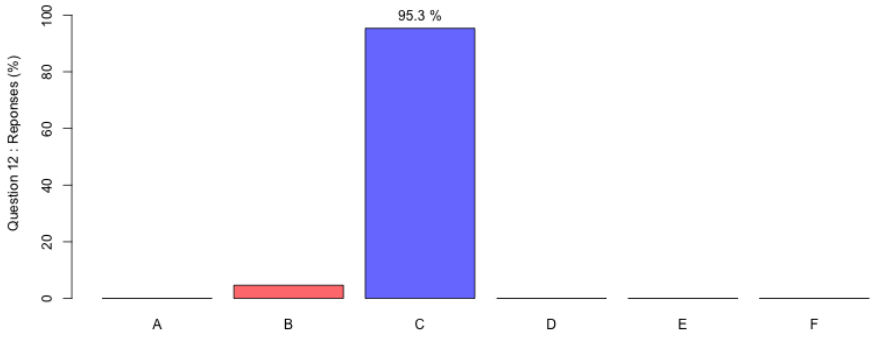
B) les résidus sont Gaussiens

C) les résidus sont de variance constante

D) les résidus sont positifs

E) les résidus sont non-nuls

C'est du cours, réponse C.



13 Sur la Figure de la page 2, tracez très exactement la droite de régression, sachant qu'elle passe par (au moins) un des points.

On a ici une information : la droite de régression passe par un des points. Via le cours, on a une seconde information : la droite estimée par moindres carrés passe par $(\overline{x}, \overline{y}$ (c'est la condition du premier ordre obtenue en dérivant par rapport à β_0). Ici, calculer les deux moyennes était facile

$$\boldsymbol{x} = \{0, 0, 1, 1, 2, 2\} \text{ donc } \overline{x} = 1$$

$$\boldsymbol{x} = \{0, 1, 2, 3, 4, 5\} \text{ (dans le désorrdre) donc } \overline{x} = \frac{5}{2} = 2.5$$

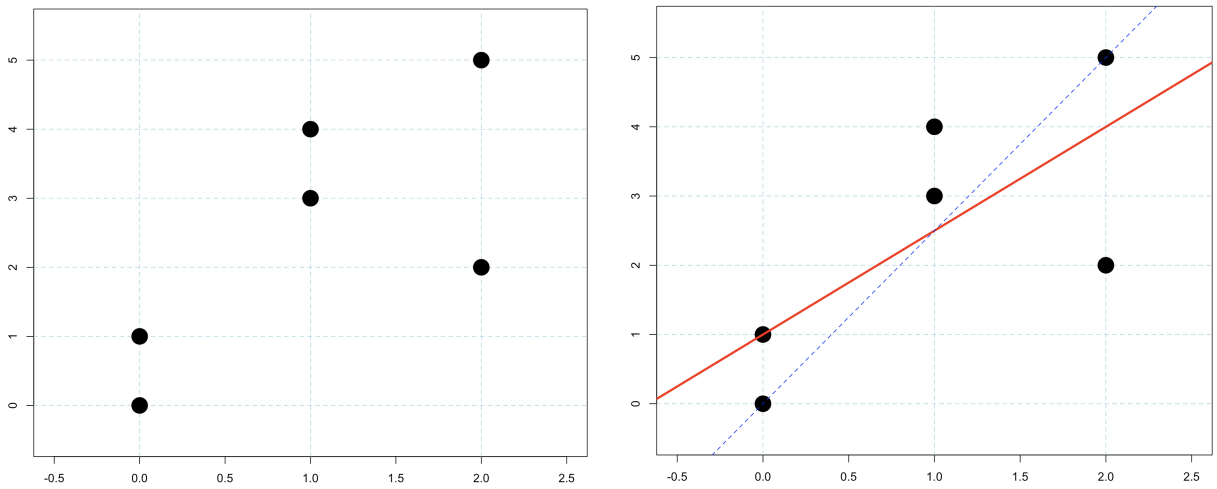
Comme la droite passe par $(1; 2.5)$, il est exclus qu'elle passe par les deux points au centre. On a alors 3 possibilités

(i) passer par le point en bas à droite, $(2, 2)$: ce cas est exclus, les résidus sont trop importants, et la pente est négatif, ce qui n'a pas trop de sens... (mais les laisses les sceptiques faire les calculs);

(ii) passer par le point en haut à gauche, $(0, 1)$;

(iii) passer par les deux autres points, $(0, 0)$ et $(2, 5)$ (car ces points sont alignés avec $(1; 2.5)$);

Ces deux derniers cas sont visualisables ci-dessous



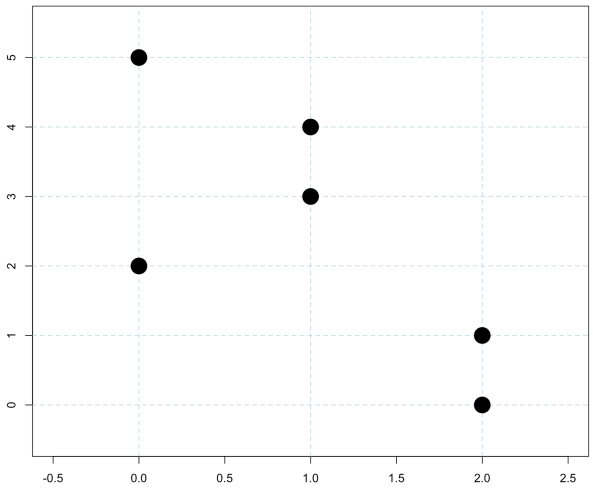
La première chose qu'on peut noter est que, quelle que soit la droite retenue, les résidus aux centres sont identiques. La seconde chose qu'on peut noter est que, quelle que soit la droite retenue, les carrés des résidus à gauche sont identiques, avec soit 0 et $(+1)^2$, soit 0 et $(-1)^2$. Donc le meilleur modèle sera celui qui a les résidus dont la somme des carrés sera la plus faible...

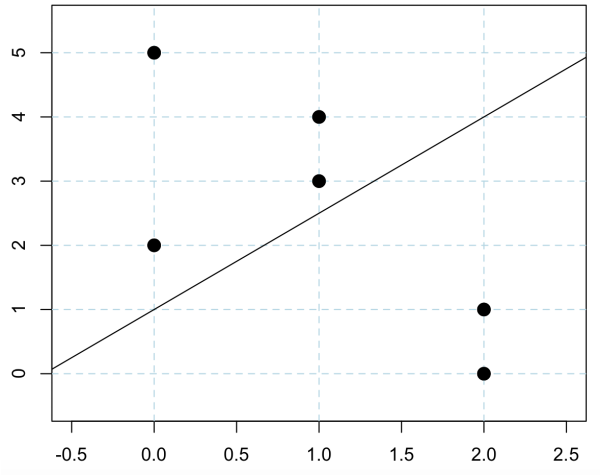
Pour les deux points, si on retient la courbe *rouge*, la prévision (en $x = 2$) sera $\widehat{y} = 4$. Donc les résidus sont respectivement $+1$ et -2. Si on retient la courbe *bleue*, la prévision (en $x = 2$) sera $\widehat{y} = 5$. Donc les résidus sont respectivement 0 et -3. Or

$$\underbrace{(+1)^2 + (-2)^2 = 5}_{\text{rouge}} < \underbrace{(0)^2 + (-3)^2 = 9}_{\text{bleue}}$$

donc la courbe rouge est celle qui correspond à la plus petite somme des carrés des résidus, on va donc retenir ce modèle. L'estimation de la courbe de régression par moindres carrés donne le modèle rouge.

Sinon, je ne pas en parler, mais le sujet B demandait de travailler sur le jeu de données suivant (qui est exactement le même, avec $2 - x_i$ au lieu de x_i – ce qui correspond à une symmétrie par rapport à la droite verticale $x = 1$, qui est en fait $x = \overline{x}$). Au lieu d'avoir la droite $y = 1 + 3x/2$, on a la droite de régression $y = 4 - 3x/2$, qui est décroissante... Je rajoute ce paragraphe car une des copies que j'ai reçu présentait le dessin de droite. Au delà du fait que visuellement ça n'a pas de sens si on comprends ce qu'est la régression linéaire, je rappelle ici que les sujets sont différents, donc c'est assez stupide de copier sur la copie de son voisin / sa voisine...





(oui, la régression de droite n'a pas de sens)

14 Considérons un modèle $y = \beta_0 + \beta_1 x + \varepsilon$. Sur 32 observations, on a obtenu la régression suivante

```
> summary(lm(y~x, data = database))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0862     0.1396   7.782 1.11e-08 ***
x             1.0364     0.2477   4.185 0.000229 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.789 on 30 degrees of freedom
Multiple R-squared:  0.3686,Adjusted R-squared:  0.3475
F-statistic: 17.51 on 1 and 30 DF,  p-value: 0.0002293
```

Sur un autre jeu de donnée, on a obtenu un autre jeu d'estimateurs $(\widehat{\beta}_0, \widehat{\beta}_1)$. Parmi les propositions suivante, quelle paire vous semble la moins vraisemblable ?

A) $(\widehat{\beta}_0, \widehat{\beta}_1) = (0.80, 1.54)$

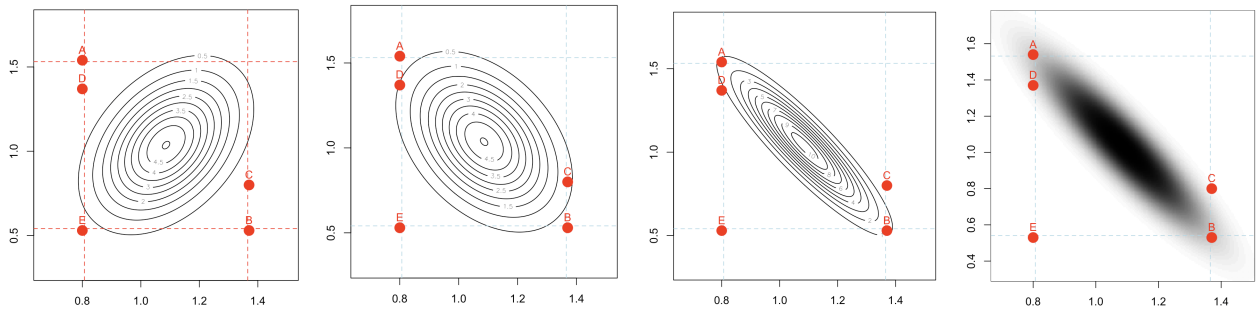
B) $(\widehat{\beta}_0, \widehat{\beta}_1) = (1.37, 0.53)$

C) $(\widehat{\beta}_0, \widehat{\beta}_1) = (1.37, 0.80)$

D) $(\widehat{\beta}_0, \widehat{\beta}_1) = (0.80, 1.37)$

E) $(\widehat{\beta}_0, \widehat{\beta}_1) = (0.80, 0.53)$

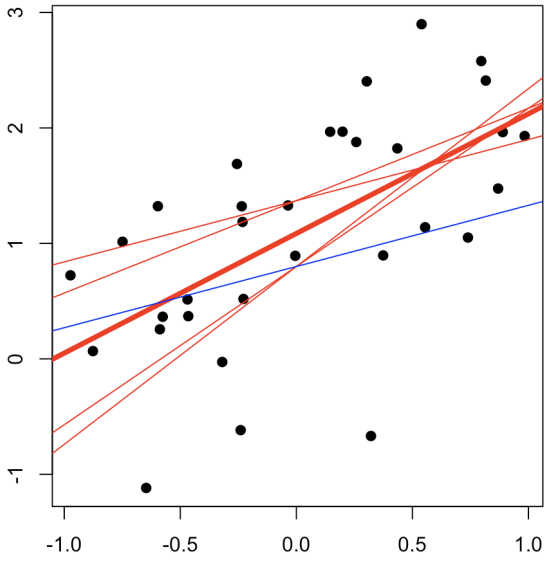
Faisons des petits dessins pour expliquer ce qui se passe ici : on supposant les résidus Gaussiens, $\widehat{\boldsymbol{\beta}}$ suit un vecteur Gaussien, dont les courbes de niveau sont des ellipses, comme sur les dessins ci-dessous

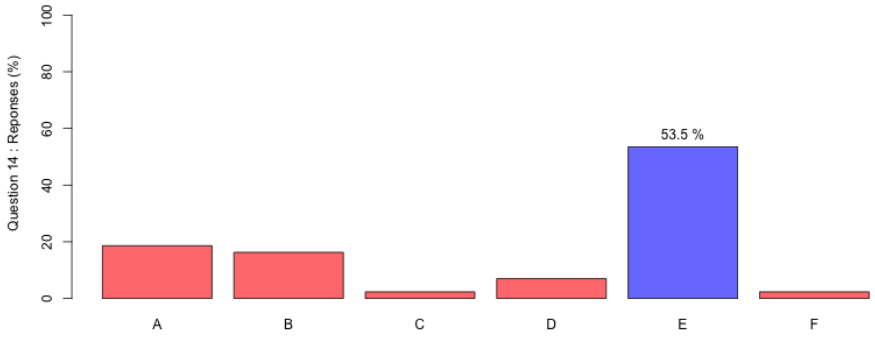


avec les courbes d'iso-densité à gauche, et une forme de visualisation de la densité à droite : plus on et sombre, plus on a de chances d'avoir des observations. Tous les points sont, marginalement, au bord des intervalles de confiance, les fameux $[\widehat{\beta}_j \pm 2\widehat{\text{Var}}[\widehat{\beta}_j]]$, donc ça ne va pas trop servir...

Le point important ici, comme on l'a vu en cours, c'est que dans la régression simple, $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont très fortement négativement corrélés...! On n'a pas le dessin de gauche, loin de là ! On est plutôt comme sur les dessins de droite... Autrement dit, le point qu'on a le moins de chances de voir est E !

On peut d'ailleurs vérifier ce que ca signifie: les données avaient été simulées ici pour obtenir la sortir R. Le nuage de points, la régression initiale (en trait épais rouge), les 4 premières (en rouge plus fin) et la dernière (en bleu) sont sur la figure ci-dessous





Pour répondre aux questions de 15 à 30, vous devez vous aider de l'annexe (correspondant à 24 pages de sorties R).

Sauf mention contraire, les tests seront analysés avec un seuil de significativité $\alpha = 5\%$.

Les questions 15 et 16 portent sur la sortie suivante :

On fait la régression de y (logPrix) sur une constante, $y_i = \beta_0 + \varepsilon_i$, et on obtient la sortie suivante

```
> summary(lm(logPrix~1, data=database))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4077 on 1424 degrees of freedom
```

15 Donnez une valeur approchée de $\sqrt{\widehat{\text{Var}}(\widehat{\beta}_0)}$

A) 0.0001

B) 0.0108

C) 0.1039

D) 0.1662

E) 0.4077

Pour commencer, un petit exercice de statistique (du niveau de stat-1) : on suppose ici que $y_i \sim \mathcal{N}(\beta_0, \sigma^2)$. L'estimateur par moindres carrés est simplement la moyenne empirique, $\widehat{\beta}_0 = \overline{y}$. Et on sait (je renvoie aux cours précédants) que quand on a des observations indépendantes

$$\text{Var}(\overline{Y}) = \text{Var}\left[\frac{1}{n}\sum i = 1^n Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{\sigma^2}{n}$$

Aussi,

$$\mathsf{Var}(\widehat{\beta}_0) = \mathsf{Var}\big(\overline{Y}\big) = \frac{\sigma^2}{n}$$

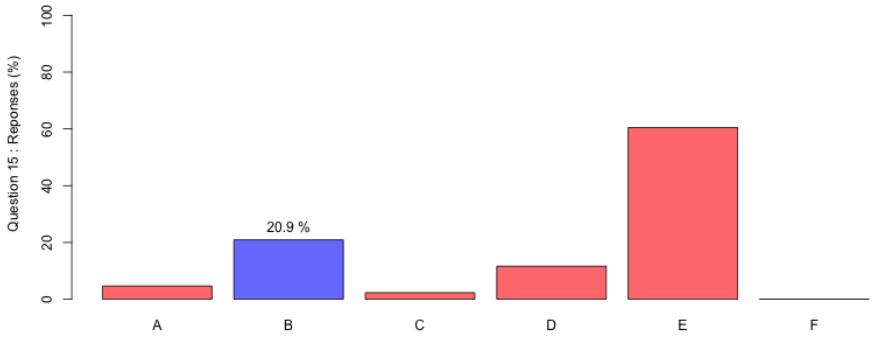
ou encore

$$\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta}_0)} = \frac{\widehat{\sigma}}{\sqrt{n}}.$$

A-t-on les information pour répondre ? Oui : `1424 degrees of freedom` signifie que $n = 1424 + 1 = 1425$ et `Residual standard error: 0.4077` signifie que $\widehat{\sigma} = 0.4077$. Aussi

$$\sqrt{\widehat{\mathsf{Var}}(\widehat{\beta}_0)} = \frac{\widehat{\sigma}}{\sqrt{n}} = \frac{0.4077}{\sqrt{1425}} = 0.0108$$

qui est la réponse B.



Il y a ici beaucoup de réponses E. Pour info, cette quantité est l'estimateur de σ i.e. $\sqrt{\widehat{\mathsf{Var}}(\epsilon)}$ mais aussi $\sqrt{\widehat{\mathsf{Var}}(y)}$ (puisqu'on régresse juste sur la constante.)

16 Que vaut le R^2 ce ce modèle ?

A) 0

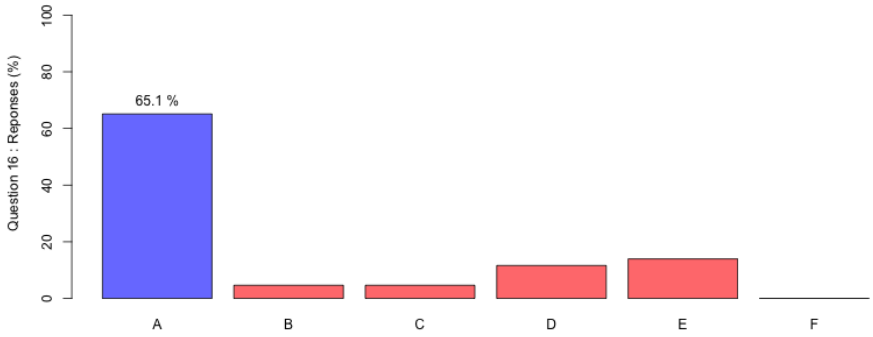
B) 0.1662

C) 0.4077

D) 1

E) on n'a pas assez d'éléments pour répondre à cette question

Revenons à la définition du R^2, qui vaut simplement $1 - \mathsf{Var}[\widehat{\varepsilon}]/\mathsf{Var}[y]$. Or ici $Y = \beta_0 + \varepsilon$ et donc $\mathsf{Var}[\widehat{\varepsilon}] = \mathsf{Var}[y]$. Autrement dit, quand on régresse sur une constante seulement, on a un R^2 nul ! le 'modèle' n'explique ici pas grand chose (et se contente de prédire la valeur moyenne). C'est donc la réponse A.



Les questions 17, 18 et 19 portent sur la sortie suivante :

On fait la régression de y (logPrix) sur la variable indiquant la présence (ou pas) d'une piscine,

$$y_i = \beta_0 + \beta_1 \mathbf{1}(\text{piscine}_i) + \varepsilon_i,$$

et on obtient la sortie suivante

```
> database$I_piscine = (database$Piscine_Surface>0)
> summary(lm(logPrix~I_piscine, data=database))

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)   XXXXXXXX    0.01077 XXXXXXXX XXXXXXXX XXX
I_piscineTRUE XXXXXXXX    0.16595 XXXXXXXX XXXXXXXX XXX
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4056 on 1423 degrees of freedom
Multiple R-squared:  0.01088,Adjusted R-squared:  0.01019
F-statistic: 15.66 on 1 and 1423 DF,  p-value: 7.965e-05
```

17 Donnez la valeur de $\widehat{\beta}_1$:

A) 12.01913

B) 12.67577

C) 0.65664

D) -0.65664

E) 12.0219

Il faut reprendre ici la partie du cours sur les modèles avec des variables explicatives catégorielles. Pour rappel, comme on a ici deux classes, et que l'on a une variable indicatrice qui prend la valeur 1 s'il y a une piscine,

$$\widehat{y}_i = \begin{cases} \widehat{\beta}_0 & \text{si } i \text{ n'a pas de piscine} \\ \widehat{\beta}_0 + \widehat{\beta}_1 & \text{si } i \text{ a une piscine} \end{cases}$$

et comme on l'a vu, quand on utilise la méthode des moindres carrés,

$$\widehat{y}_i = \frac{1}{n_{\text{pas de piscine}}} \sum_{j:\text{pas de piscine}} y_j = 12.01913, \text{ si } i \text{ n'a pas de piscine}$$

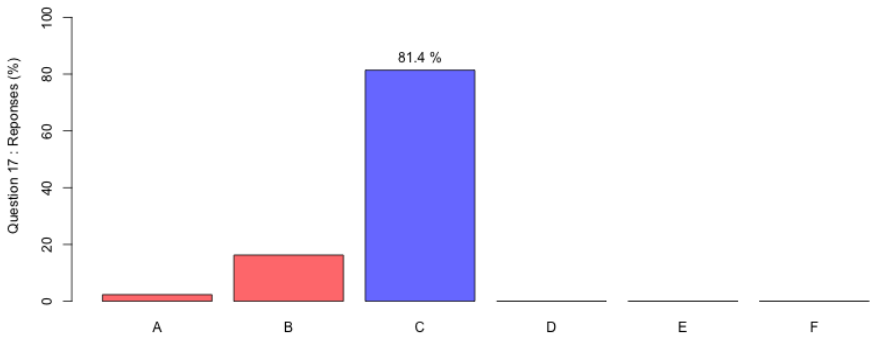
(où $n_{\text{pas de piscine}}$ est le nombre de personnes qui n'ont pas de piscine) et

$$\widehat{y}_i = \frac{1}{n_{\text{piscine}}} \sum_{j:\text{piscine}} y_j = 12.67577, \text{ si } i \text{ a une piscine}$$

en utilisant les annexes (page 5). Aussi, on a le système suivant

$$\begin{cases} \widehat{\beta}_0 = 12.01913 \\ \widehat{\beta}_0 + \widehat{\beta}_1 = 12.67577 \end{cases}$$

et donc $\widehat{\beta}_1 = 12.67577 - 12.01913 = 0.65664$ qui est la réponse C.



18 Que vaut la p-value du test de Student associé à l'hypothèse $H_0 : \beta_1 = 0$ (contre $H_1 : \beta_1 \neq 0$) :

A) moins de 0.01%

B) entre 0.01% et 0.1%

C) entre 0.1% et 1%

D) entre 1% et 10%

E) plus de 10%

On vient de voir que $\widehat{\beta}_1 = 0.65664$. La statistique du test est ici

$$t = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\text{Var}}[\widehat{\beta}_1]}} = \frac{0.65664}{0.16595} = 3.956854.$$

Pour rappel, comme on a plus de mille observations, sous $H_0 : \beta_1$, T doit suivre une loi Gaussienne. La p-value est ici

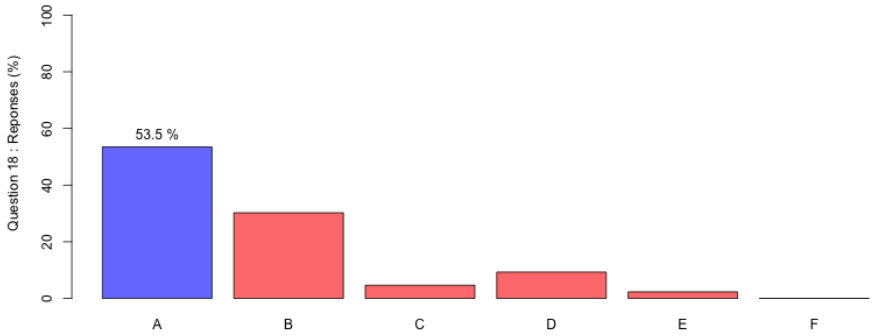
$$p = \mathbb{P}[|T| > |t|] = 2\mathbb{P}[T > |t|] = 2(1 - \mathbb{P}[T < |t|])$$

Si on regarde dans la table donnée page 1, on sait que $\mathbb{P}[T < 3.29] = 99.95\%$ donc $1 - \mathbb{P}[T < |t|] < 0.05\%$ et donc $p < 0.1\%$. On se retrouve à hésiter entre A et B...

En fait, on avait des éléments de réponse dans les 24 pages d'annexe... par exemple en haut de la page 20

```
## FoundationB          2.490e-02  1.341e-02   1.856 0.063631 .
## FoundationC          6.621e-02  1.636e-02   4.046 5.49e-05 ***
```

on voit qu'avait une statistique de test proche de 4, on a une p-value est probablement un peu plus grande que `5.49e-05`, soit 0.005%, ce qui correspond à la réponse A. (En fait, la valeur exacte était `7.97e-05` soit 0.008%)



19 On estime le modèle suivant par moindres carrés

$$y_i = \gamma_0 + \gamma_1 \mathbf{1}(\text{pas de piscine}_i) + \eta_i,$$

et on a les affirmations suivantes

(i) $\widehat{\eta}_i = \widehat{\varepsilon}_i$ pour tout $i = 1, \cdots, n$

(ii) $\displaystyle\sum_{i=1}^{n} \widehat{\eta}_i = \sum_{i=1}^{n} \widehat{\varepsilon}_i$

(iii) $\displaystyle\sum_{i=1}^{n} \widehat{\eta}_i^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$

(iv) $\widehat{\beta}_0 + \widehat{\beta}_1 = \widehat{\gamma}_0$

(v) $\widehat{\beta}_0 = \widehat{\gamma}_0 + \widehat{\gamma}_1$

Lesquelles (ou laquelle) sont justes ?

A) (ii) seulement

B) (ii) et (iii)

C) (iv) et (v)

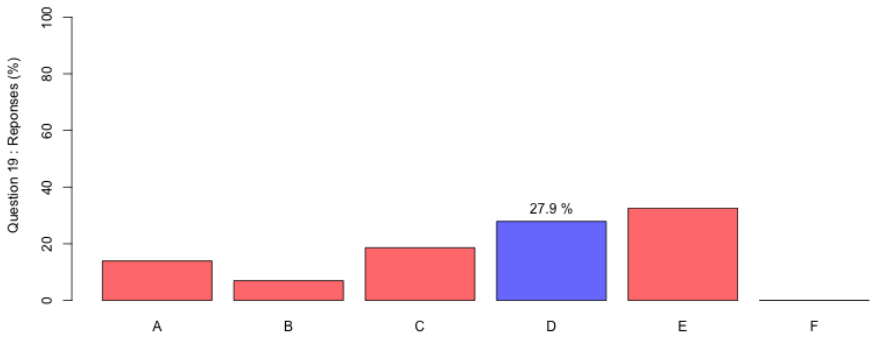
D) les cinq affirmations sont justes

E) ni (A), ni (B), ni (C), ni (D)

C'est toujours la même histoire... quand on a une variable factorielle, on a un soucis d'identifiabilité. Les deux modèles sont rigoursesement équivalent ! (on l'a vu en cours). En particulier $\widehat{\eta}_i = \widehat{\varepsilon}_i$ pour toutes les observations ! donc (i) est vraie. Et *a fortiori* (ii) et (iii) aussi. Maintenant, si on regarde les deux prévisions qu'on peut faire on a, respectivement

$$\widehat{y}_i = \begin{cases} \widehat{\beta}_0 = \widehat{\gamma}_0 + \widehat{\gamma}_1 & \text{si } i \text{ n'a pas de piscine} \\ \widehat{\beta}_0 + \widehat{\beta}_1 = \widehat{\gamma}_0 & \text{si } i \text{ a une piscine} \end{cases}$$

où on retrouve respectivement (v) et (iv). Autrement dit, les cinq affirmations sont juste, qui était la réponse D.



Les questions 20 et 21 portent sur les sorties suivantes :

$$y_i = \beta_0 + \beta_1 \mathsf{Construction\_Annee}_i + \varepsilon_i$$

```
> summary(reg_Construction_Annee)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.8312288  0.5648195  -6.783 1.72e-11 ***
Construction_Annee  0.0080416  0.0002865  28.071  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3272 on 1423 degrees of freedom
Multiple R-squared:  0.3564,Adjusted R-squared:  0.3559
F-statistic:   788 on 1 and 1423 DF,  p-value: < 2.2e-16
```

De plus, une régression a été faite sur une sous-base, en excluant les maisons construites avant 1920,

$$y_i = \beta_0^{\mathsf{new}} + \beta_1^{\mathsf{new}} \mathsf{Construction\_Annee}_i + \varepsilon_i^{\mathsf{new}}$$

```
> new = with(database, (Construction_Annee>=1920))
> reg_Construction_Annee_new = lm(logPrix ~ Construction_Annee, data = database, subset = new)
> summary(reg_Construction_Annee_new)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -7.4056035  0.6442509  -11.49   <2e-16 ***
Construction_Annee  0.0098428  0.0003261   30.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3136 on 1342 degrees of freedom
Multiple R-squared:  0.4043,Adjusted R-squared:  0.4039
F-statistic: 910.9 on 1 and 1342 DF,  p-value: < 2.2e-16
```

Juste une remarque, car quelqu'un s'est senti obligé de laisser un commentaire sur sa copie : on peut faire une prévision pour 1900 même si les données utilisées ne commence qu'en 1920 ! l'avantage du modèle linéaire est qu'il donne une estimation pour toutes les valeurs x (ce qui ne serait pas forcément le cas avec des modèles non-linéaires, genre des splines). C'est ce qu'on utilise pour faire de la prévision, si une des composantes est le temps ! Si je veux conclure, je dirais juste qu'a priori, ce modèle sera a priori moins bon en 1900, car on n'a aucune donnée de cette époque, mais on peut utiliser le modèle malgré tout !

20 Par rapport à la régression reg_Construction_Annee, on veut faire un test $H_0 : \beta_1 = \beta_1^{\text{new}}$,

A) on accepte H_0 si le seuil de significativité est de 10%

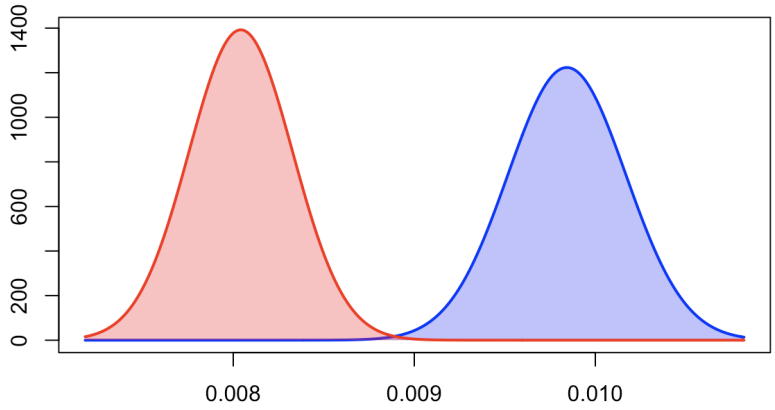
B) on accepte H_0 si le seuil de significativité est de 5%

C) on accepte H_0 si le seuil de significativité est de 1%

D) on rejette H_0 si le seuil de significativité est de 1%

E) on ne peut pas savoir

La question est un peu tordue, et on n'a pas ce test dans les sortie. Néanmoins, on a des propriétés de normalité de nos estimateurs. Aussi, rappelons que les estimateurs sont Gaussiens, et plus précisément, les densités de $\widehat{beta}_1^{\text{new}}$ et $\widehat{\beta}_1$ sont visualisées sur la figure ci-dessous



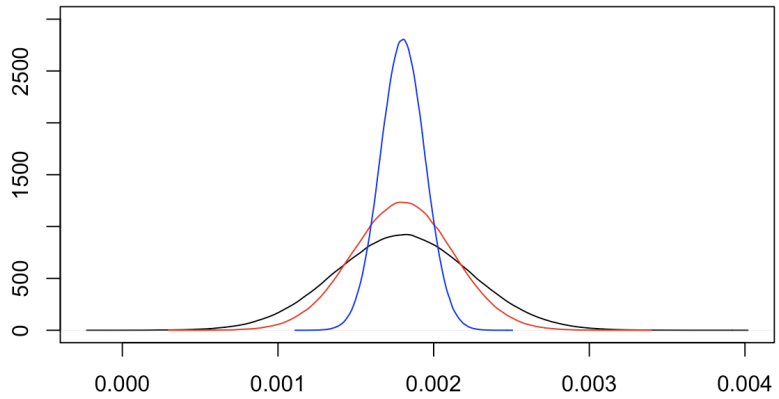
Reprenons le cours de statistique, avec la différence entre deux échantillons $\{z_{A,1}, \dots, z_{A,n}\}$ et $\{z_{B,1}, \dots, z_{B,n}\}$. Pour tester s'ils ont la même moyenne, la statistique de test est $t = \overline{z}_A - \overline{z}_B$ qui soit suivre, sous l'hypothèses que les moyennes sont identiques, une loi normale centrée, de variance la somme des variances, si les échantillons sont indépendants. Ici, on devrait

$$t = \frac{\widehat{\beta}_1^{\text{new}} - \widehat{\beta}_1}{\sqrt{s_1^2 + s_1^{2\text{new}}}} \sim \mathcal{N}(0,1)$$

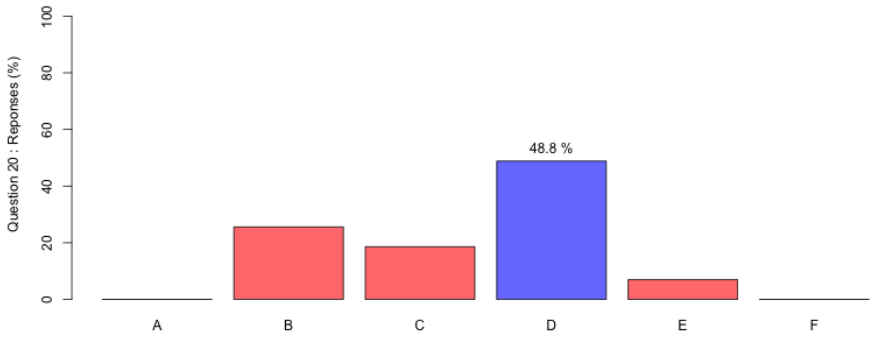
si les deux coefficients sont égaux. Or ici

$$t = \frac{0.0018012}{0.0004340777} = 4.149488$$

On rejette alors ici l'hypothèse d'égalité des deux avec un seuil α de l'ordre de 1%, ce qui est la réponse D.



En fait, les échantillons ne sont pas ici indépendants, mais au contraire très fortement dépendant: on a encore plus de chance d'avoir des valeurs différentes. Sur la figure ci-dessous, on a la distribution de la différence entre les deux moyennes, en fonction de la corrélation : nulle en noir, un peu corrélée (positivement) en rouge, et très corrélée en bleu... la variance est alors beaucoup plus faible, ce qui augementera encore la valeur de T, et diminera la p-value... Bref, on doit rejeter H_0 ici sans trop d'hésitation...



21 On souhaite comparer ces deux modèles, pour une maison construite en 1900 (le prix de la maison est $\exp[y]$). Le prix donné par le premier modèle, par rapport au prix donné par le modèle 'new'

A) est sensiblement identique

B) est 5% plus élevé

C) est 10% plus élevé

D) est 15% plus élevé

E) aucune des réponses proposées

On a ici une maison construite en 1900. Avec le premier modèle,

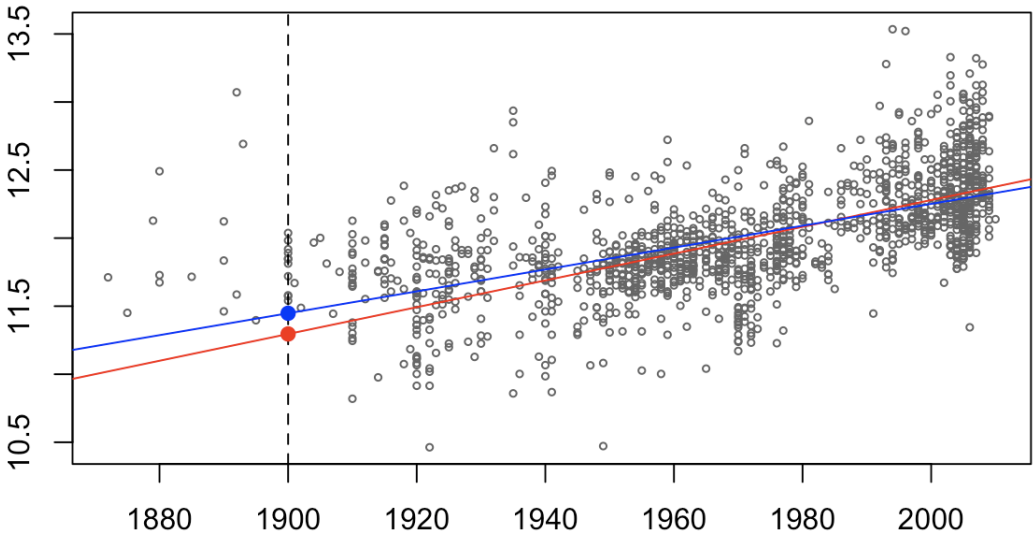
$$\widehat{y} = -3.8312288 + 0.0080416 \cdot 1900 = 11.44781$$

alors qu'avec le second

$$\widehat{y} = -7.4056035 + 0.0098428 \cdot 1900 = 11.29572$$

ce qui donne une différence de 0.15. Aussi, en revenant au prix initial $\exp[y]$, la différence (multiplicative) est $\exp[0.15] = 1.161834$, soit 16.18% moins cher.

On peut visualiser cette différence, en logarithme, sur la figure suivante, avec en *bleu* la régression sur toutes les observations, et en *rouge* la régression sur les données ultérieures à 1920. On observe la hausse de 0.15.



En réalité, si on veut on prévision sans biais du prix réel de la maison, on devrait prendre $\exp[\widehat{y} + \widehat{\sigma}^2/2]$, ce qui donne respectivement

$$\exp\left(11.44781 + \frac{0.3272^2}{2}\right) = 98848.13$$

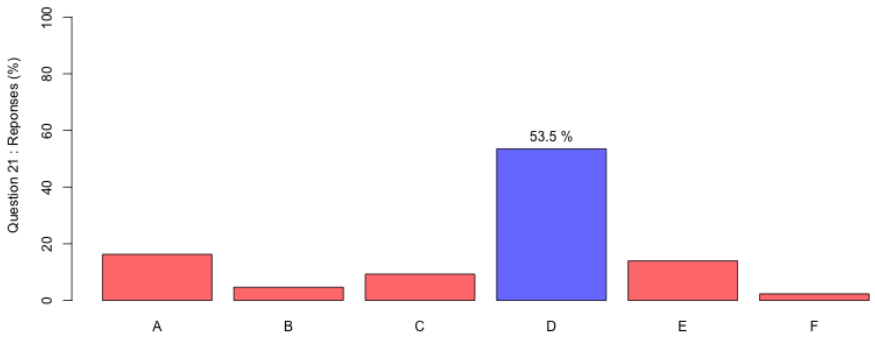
pour l'ancien prix, et, pour le nouveau

$$\exp\left(11.29572 + \frac{0.3136^2}{2}\right) = 84532.6$$

ce qui donne un prix moins cher, relativement de

$$\frac{98848.13 - 84532.6}{84532.6} = 16.93492\%$$

On va arrondir à 15% plus cher...



Les questions 22 et 23 portent sur la sortie suivante, où un modèle assez simple a été envisagé

```
> reg_total_3 = lm(logPrix ~ Surface_Lot + Surface_RdC + Construction_Annee + Pieces,
                   data = database)
> summary(reg_total_3)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Surface_Lot         7.285e-06  1.103e-06   6.606 5.57e-11 ***
Surface_RdC         3.618e-04  1.924e-05  18.805  < 2e-16 ***
Construction_Annee  6.196e-03  2.135e-04  29.015  < 2e-16 ***
Pieces              7.852e-02  4.140e-03  18.965  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2318 on 1420 degrees of freedom
Multiple R-squared:  XXXXX, Adjusted R-squared:  XXXXXXX
F-statistic: 746.6 on 4 and 1420 DF,  p-value: < 2.2e-16

> predict(reg_total_3, newdata= data.frame(Surface_Lot = 8396,
+         Surface_RdC = 847, Construction_Annee = 2003,
+         Pieces = 9))

12.29062
```

22 Que vaut le R^2 de ce modèle ?

A) 0.32

B) 0.43

C) 0.56

D) 0.67

E) 0.94

Il suffit de se souvenir de ce qu'est le R^2 :

$$R^2 = 1 - \frac{\sum \widehat{\varepsilon}_i^2}{\sum (y)i - \overline{y})^2} = 1 - \frac{(n-p)\widehat{\sigma}^2}{(n-1)\mathsf{Var}[y]}$$

Pour la variance de y, elle est dans la page 2 des annexes

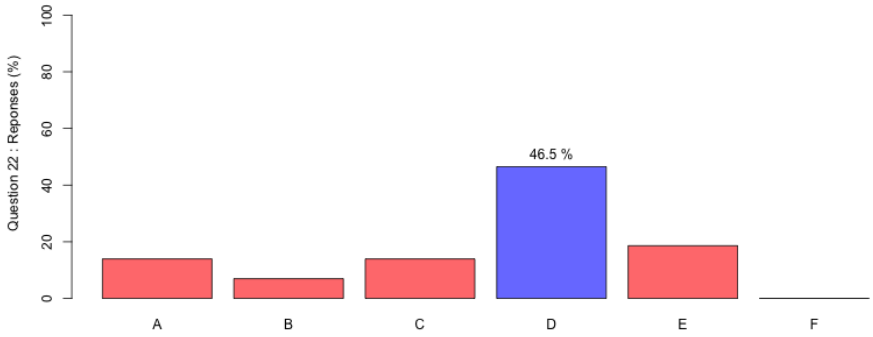
```
var(database$logPrix)

## [1] 0.1662255
```

i.e. 0.1662255. Et $\widehat{\sigma}^2$ est donné dans la sortie, avec 0.2318^2. Donc ici

$$R^2 = 1 - \frac{(1425-5)\cdot 0.2318^2}{(1425-1)\cdot 0.1662255} = 0.6776649$$

ce qui correspond à la réponse D.



23 Pour la première observation, on a une prévision de 12.29062. Que vaut $\widehat{\beta}_0$ associé à ce modèle ?

A) -1.194

B) -0.594

C) 0.135

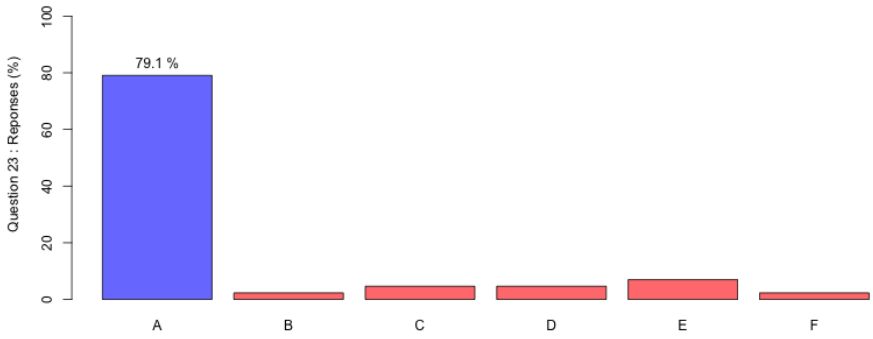
D) 0.594

E) 1.194

Pour la prévision, on nous donne toutes les informations,

$$\widehat{y} = \widehat{\beta}_0 + 0.000007285 \cdot 8396 + 0.0003618 \cdot 847 + 0.006196 \cdot 2003 + 0.07852 \cdot 9 = \widehat{\beta}_0 + 13.48488 = 12.29062$$

donc $\widehat{\beta}_0 = 12.29062 - 13.48488 = -1.19426$ qui correspond à la réponse A.



24 On souhaite expliquer le prix juste par l'année de construction. En utilisant reg_Construction_Annee_2, en quelle année de construction a-t-on les logements les moins chers :

A) 1925

B) 1930

C) 1935

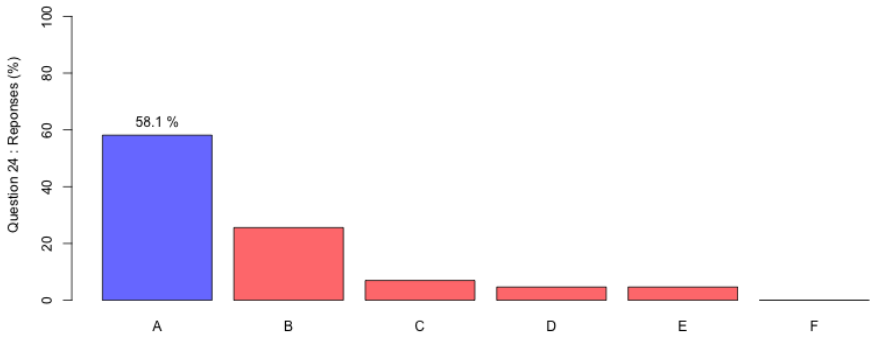
D) 1940

E) 1945

On l'avait fait en cours ! Page 8, on a un parabole, d'équation $y = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$, et on sait que le minimum est atteint en... $-\widehat{\beta}_1/2\widehat{\beta}_2$ (c'est la condition du première ordre). Aussi, ici,

$$\frac{-\widehat{\beta}_1}{2\widehat{\beta}_2} = \frac{0.4135}{2 \cdot 0.0001074} = 1925$$

ce qui correspond à la réponse A. Mais ca se voyait sur le dessin, non ?



25 La variable Fondation comprenait initialement 6 modalités, { Slab, BrkTil, Stone, CBlock, Wood, PConc }. Quel regroupement de modalité vous semble pertinent ?

A) { BrkTil, Stone}

B) { Slab, BrkTil, Stone }

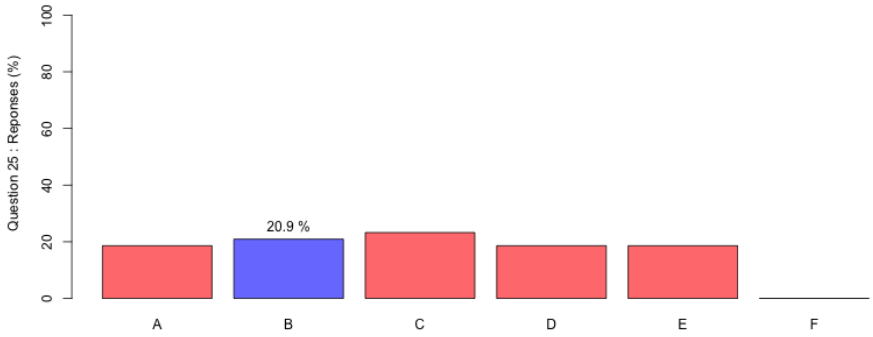
C) { BrkTil, Stone, CBlock }

D) { Slab, BrkTil, Stone, CBlock }

E) { CBlock, Wood}

La réponse se trouve page 15. On va créer 3 modalités, et la première est celle qui regroupe { Slab, BrkTil, Stone }. La bonne réponse est la réponse B.

Pour être plus précis, la réponse A n'est pas fausse, mais le but de regroupe des modalités est de regrouper *toutes* celles qui ne sont pas significativement différentes. Dans la première sortie, on teste $H_0 : \beta_{BrkTil} = \beta_{Stone} = 0$, qui est un teste multiple. On accepte ici cette hypothèse (on a un p-value supérieure à 71%), qui correspond à C : ces deux modalités ne sont pas significativement différentes de la modalité de référence. A forciori $H_0 : \beta_{BrkTil} = \beta_{Stone}$ donc oui, A est juste, mais le test nous dit bien plus. Le regroupement le plus pertinent était celui avec les trois modalités !



26 La variable Chauff_Qualite comprenait initialement 4 modalités. En regroupant les modalités non-significativement différentes, combien la variable devrait avoir au final de modalité

A) 4

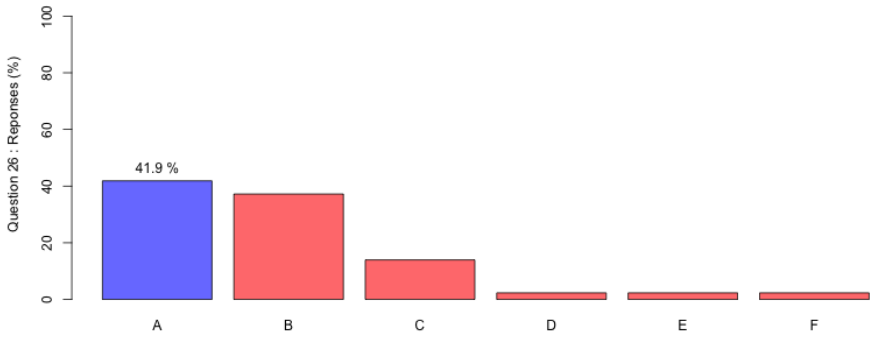
B) 3

C) 2

D) 1

E) 0

Toujours 4. Les sorties laissent entendre que tous les regroupements tentés ont échoué.



27 Que vaut le résidu $\widehat{\varepsilon}_{373}$ (associé à l'observation 373) pour le modèle reg_total_2 ?

A) -5.870

B) -0.785

C) -0.342

D) 0.785

E) 5.870

Toutes les informations sont page 23 du document

```
##     Vente_Annee Vente_Type Vente_Condition  logPrix I_piscine
## 373        2009        WD           Family 11.32055     FALSE
```

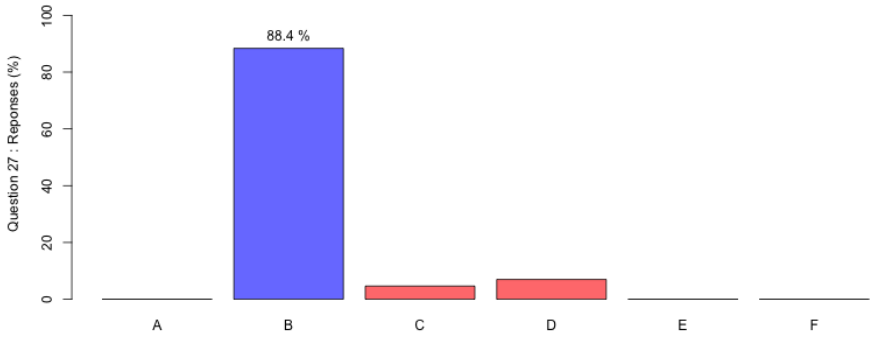
```
predict(reg_total_2,newdata=database["373",])
```

```
##      373
## 12.10604
```

On nous dit que le vrai log-prix est $y_{373} = 11.32055$ et que la prévision est $\widehat{y}_{373} = 12.10604$, donc

$$\widehat{\varepsilon}_{373} = y_{373} - \widehat{y}_{373} = 11.32055 - 12.10604 = -0.78549$$

(ce qui se retrouve d'ailleurs sur la figure de la page 21). donc on retient la réponse B.



28 Par rapport à la prédiction donnée par le modèle reg_total_2, le vrai prix ($\exp[y]$ et non pas un logarithme du prix) pour l'observation 373 était

A) 78% moins cher

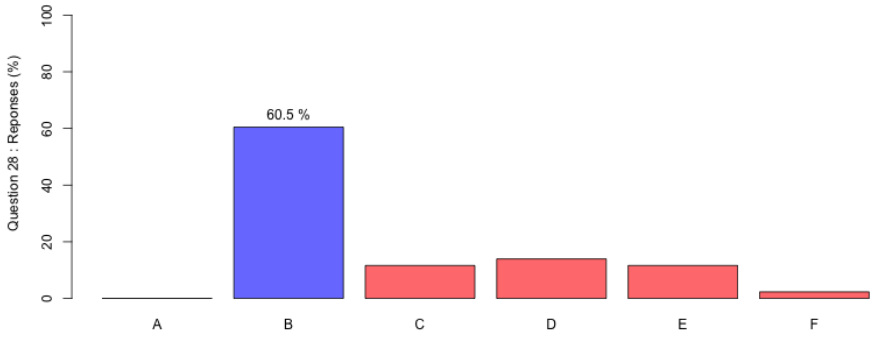
B) 54% moins cher

C) 29% moins cher

D) 5% moins cher

E) 12% plus cher

Avec une différence absolue de -0.78549, l'impact sur le prix sera de $\exp[-0.78549] = 0.4558963$ soit 54.41% moins cher. C'est la réponse B.



29 Que vaut le résidu Studentisé $\widehat{e}_{2072}$ (associé à l'observation 2072) pour le modèle reg_total_2 ?

A) -0.34

B) -0.93

C) -1.25

D) -2.53

E) -2.93

Je renvoie au dernier cours... on avait noté que $\text{Var}[\widehat{\varepsilon}] = \sigma^2(\mathbb{I} - \boldsymbol{H})$, où $\boldsymbol{H}$ est juste la matrice de projection orthogonale sur l'espace engendré par les combinaisons linéaires des variables de $\boldsymbol{X}$. Bref, on définie les résidus Studentisés par

$$\widehat{e}_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-\boldsymbol{H}_{i,i}}}$$

Or dans les annexes, on a

```
lm.influence(reg_total_2)$hat[which(lm.influence(reg_total_2)$hat>.15)]
```

```
##      2072
## 0.2530952
```

autrement dit, la seule observation pour laquelle $\boldsymbol{H}_{i,i} > 15\%$ est justement celle pour $i = 2072$, et précisément $\boldsymbol{H}_{i,i} = 0.253$. Juste après, on a les information pour $\widehat{\varepsilon}_i$ (comme pour la question 27)

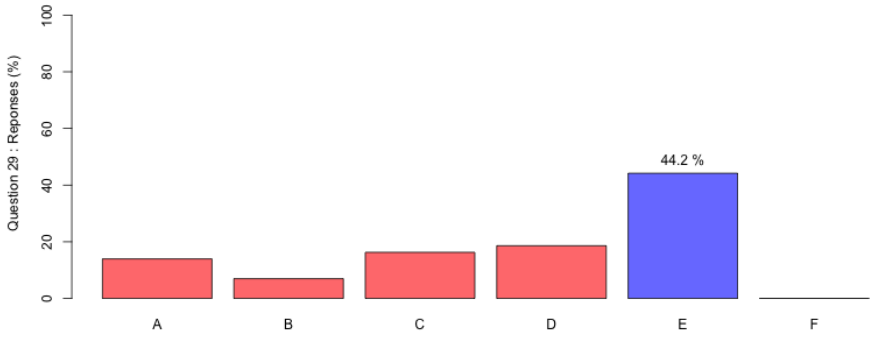
```
##      Vente_Annee Vente_Type Vente_Condition  logPrix I_piscine
## 2072        2007        WD           Normal 12.61818     FALSE
predict(reg_total_2,newdata=database["2072",])
```

```
##     2072
## 12.95709
```

donc ici $\widehat{\varepsilon}_i = -0.33891$. Quant à l'écart-type des résidus, il est dans la sortie de la régression, $\widehat{\sigma} = 0.1338$. Et donc

$$\widehat{e}_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-\boldsymbol{H}_{i,i}}} = \frac{-0.33891}{0.1338 \cdot \sqrt{1-0.253}} = -2.930677$$

qui est la réponse D.



30 On lit l'information suivante dans un journal "*L'indice de qualité* Int_Qualite *impacte significativement la valeur de votre maison : gagner 1 point augmente la valeur de votre de maison de 25%*". Si on compare des maisons comparables (même surfaces, même année de construction, même nombre de pièces, même qualités extérieure et de cuisine, même fondation), quel est le vrai impact d'un point de l'indice de qualité Int_Qualite sur le prix ($\exp[y]$) ?

A) -1%

B) +1%

C) +5%

D) +12%

E) +25%

bonne réponse) +8%

Toutes mes excuses, j'ai fait une faute de frappe, la bonne réponse est 8.8% (et pas 0.88% comme je l'avais tapé initialement...) Toutes mes excuses pour la typo, qui rendait la réponse fausse. J'ai donné un point pour les réponses C et D qui étaient les plus proches...

Revenons un peu en arrière. La première affirmation se comprend si on regarde la sortie page 4

```
summary(reg_Int_Qualite)
```

```
##
## Call:
## lm(formula = logPrix ~ Int_Qualite, data = database)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.547252   0.026776  393.90   <2e-16 ***
## Int_Qualite  0.242737   0.004295   56.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2264 on 1423 degrees of freedom
## Multiple R-squared:  0.6917, Adjusted R-squared:  0.6915
## F-statistic:  3193 on 1 and 1423 DF,  p-value: < 2.2e-16
```

On a ici $\widehat{y}_i = 10.54 + 0.24 x_i$, donc 1 points de plus sur x_i aura un impact *additif* sur $\widehat{y}_i$ de 0.24. Et donc un point de plus pour x_i aura un impact *multiplicatif* sur $\exp[\widehat{y}_i]$ de $\exp[0.24] \approx 1.27$, ce qui correspond à une hausse de $+27\%$. Donc la première partie de l'affirmation est légitime.

Si on regarde des maisons comparables, pour des facteurs $x_1, \dots, x_k$, cela signifie qu'on veut mesure l'impact de Int_Qualite sur y dans la régression de y sur $\{x_1, \dots, x_k, \text{Int\_Qualite}\}$. C'est ce que fait la régression reg_total_2 (cela dit, on peut aussi regarder reg_total_2 et l'interprétation sera du même ordre...)

```
## Surface_Etage       2.824e-04  1.511e-05  18.693  < 2e-16 ***
## Int_Qualite         8.451e-02  4.634e-03  18.237  < 2e-16 ***
## Construction_Annee  3.473e-03  2.101e-04  16.526  < 2e-16 ***
```

et cette fois, le coefficient est de l'ordre de 0.085. Autrement dit, une hausse d'un point sur la qualité de la maison aura un impact multiplicatif de $\exp[0.085] \approx 1.088$ sur $\exp[\widehat{y}]$, ce qui correspond à une hausse de $+8.8\%$, soit un peu moins de 10%. Je donne un point pour ceux qui ont répondu 5% et 12%...

(le but de l'exerice était de comprendre qu'oublier des variables importantes donne un estimateur potentiellement biaisé : si on compare des maisons identiques, un point ne vaut pas $+25\%$ mais "seulement" $+9\%$)