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 - $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z \leqslant 2.326) = 99\%$.

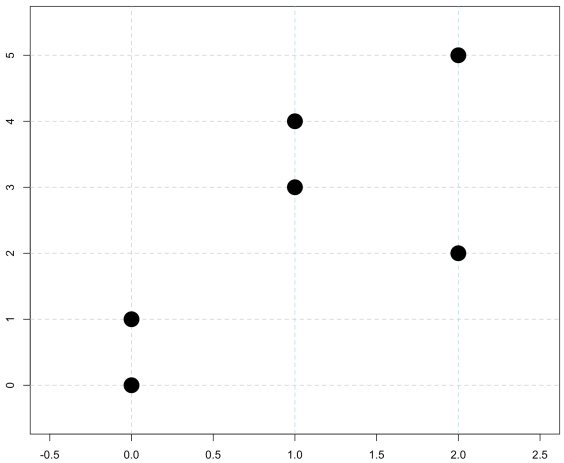
	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Student (50)	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
Student (30)	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
Student (20)	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.849
Student (15)	0.691	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 :



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 (associé à x) ?

A) -0.02

B) $+0.02$

C) 0

D) $+0.25$

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

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

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)

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

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

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

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

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

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 $s^2 = 5$

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]

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

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

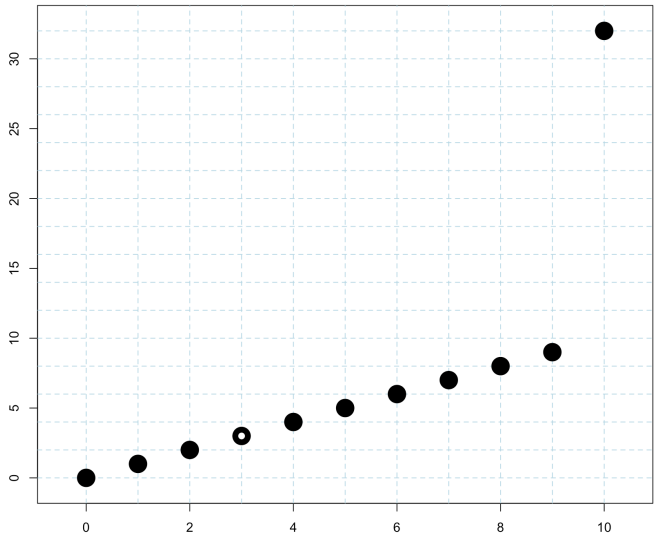
B) $\widehat{b} \sum_{i=1}^{n} x_i = \widehat{a} \sum_{i=1}^{n} x_i$

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

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

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: XXXXXXXXXXX
```

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

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

12 Les résidus sont dits homoscédastiques si

A) les résidus sont robustes

B) les résidus sont Gaussien

C) les résidus sont de variance constante

D) les résidus sont positifs

E) les résidus sont non-nuls

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.

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

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 $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}$

A) 0.0001

B) 0.0108

C) 0.1039

D) 0.1662

E) 0.4077

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

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

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%

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

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

(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)

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,

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

```
> 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
```

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

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) est 20% plus élevé

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

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

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

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}

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

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

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

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

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%