Devoir 2 - Solutions

Patrick Boily

2023-02-08

Préliminaires

Nous importons l'ensemble `Autos.xlsx` se retrouvant sur Brightspace: ave prédicteur `VKM.q` (X, distance quotidienne moyenne, en km) et réponse `CC.q` (Y, consommation de carburant quotidienne moyenne, en L).

```
library(tidyverse)  # pour avoir acces a select() et |>
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Autos <- readxl::read_excel("Autos.xlsx") |> select(VKM.q,CC.q)
str(Autos)
```

```
## tibble [996 x 2] (S3: tbl_df/tbl/data.frame)
##  $ VKM.q: num [1:996] 330 264 251 235 230 230 215 208 203 196 ...
##  $ CC.q : num [1:996] 49 33 44 22 38 31 28 19 31 19 ...
```

```
x = Autos$VKM.q
y = Autos$CC.q
```

Q11

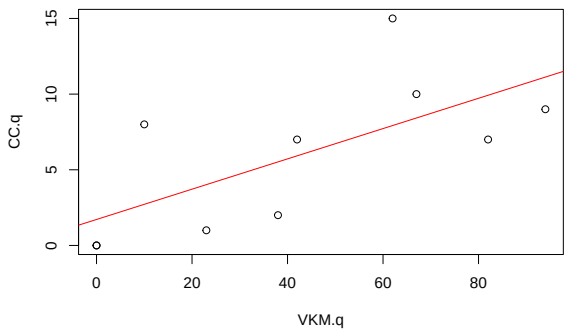
À l'aide de R, prélevez au hasard n paires d'observations dans l'ensemble de données. Déterminez la droite de meilleure ajustement au sens des moindres carrés L_n et calculez son coefficient de détermination R_n^2. Répétez pour $n = 10, 50, 100, 500$ et pour toutes les observations. Y a-t-il quelquechose d'intéressant à signaler? Si oui, comment cela s'explique-t-il?

Solution: on peut se servir du code suivant (vous aurez des réponses spécifiques différentes en fonction des échantillons aléatoires prélevés).

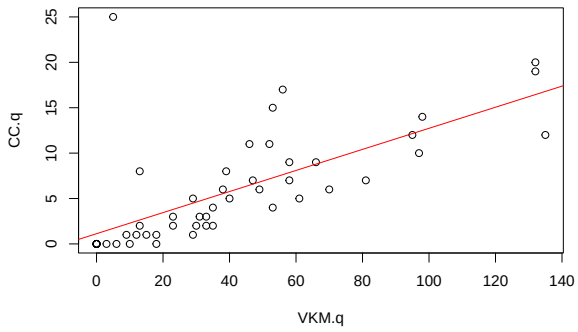
```
R2n = function(n){
  ind_n = sample(nrow(Autos), n)
  data_n = Autos[ind_n,]
  mod_n = lm(CC.q ~ VKM.q, data=data_n)
  plot(data_n)
  abline(mod_n, col="red")
  R2n = summary(mod_n)$r.squared
  return(R2n)
}
```

Pour les valeurs n suggérées, voici une série de réalisations:

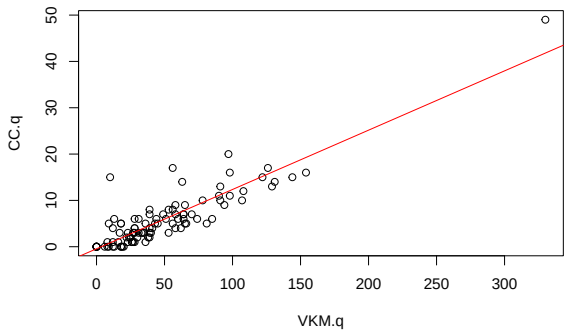
```
set.seed(1)
R2.10  = R2n(10)
```



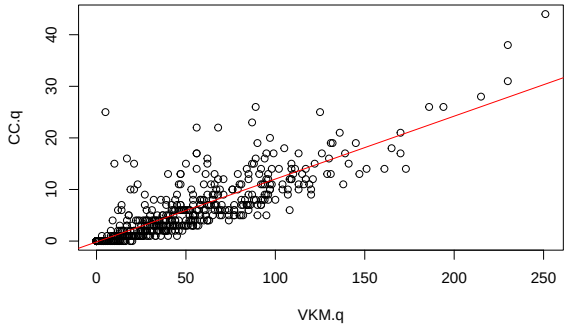

```
R2.50  = R2n(50)
```



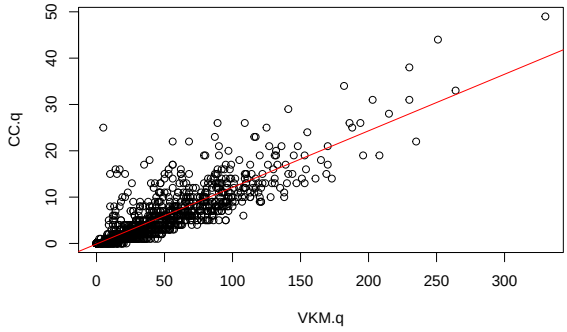
```
R2.100 = R2n(100)
```




```
R2.500 = R2n(500)
```




```
R2.All = R2n(nrow(Autos))
```

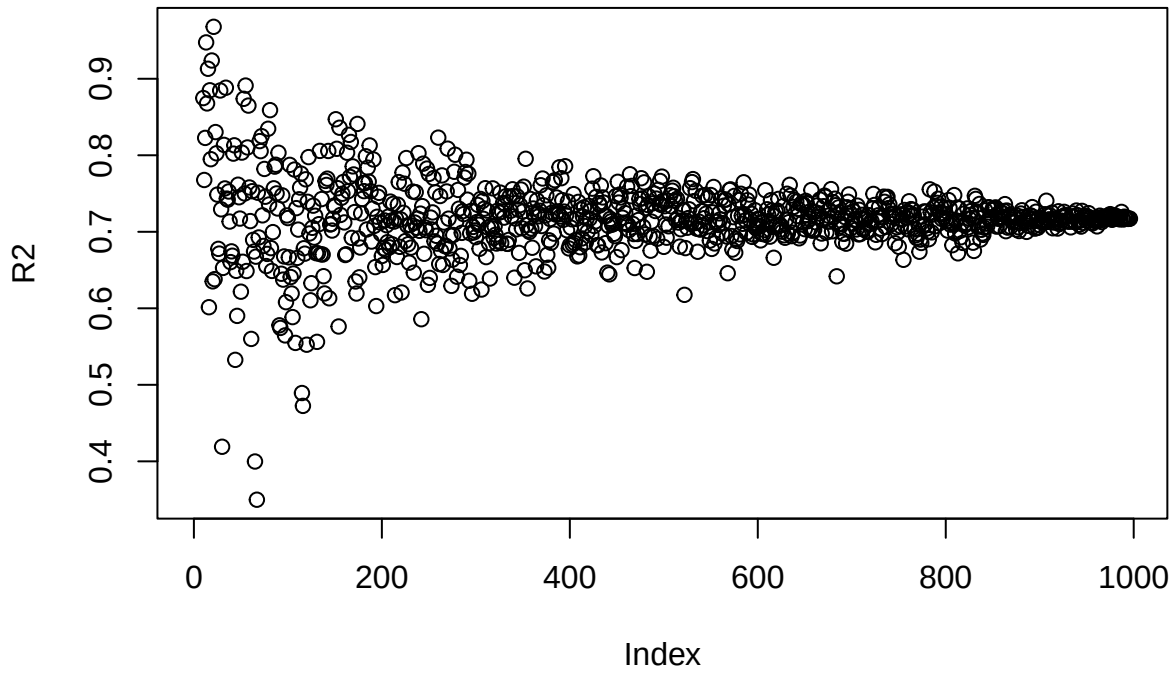



```
c(R2.10, R2.50, R2.100, R2.500, R2.All)
```

```
## [1] 0.4570214 0.4703746 0.7977084 0.6951693 0.7171661
```

Dans cette version, il n'y a pas de schéma évident. Peut-etre qu'en s'y prenant différemment, en augmentant le nombre d'observations un par un?

```
R2n = function(n){
  ind_n = sample(nrow(Autos), n)
  R2n = (cor(Autos[ind_n,])[1,2])^2
  return(R2n)
}

R2 = c()
for(j in 10:nrow(Autos)){
  R2[j]=R2n(j)
}

plot(R2)
```



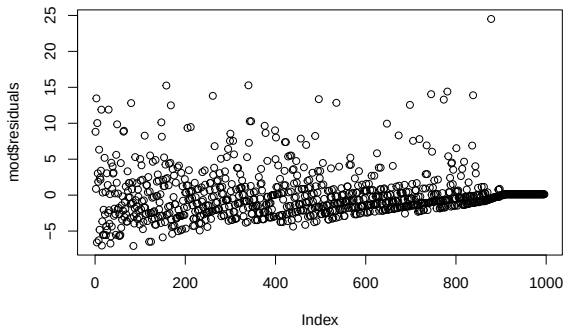
Qu'en dites-vous?

Q12

À l'aide de R, tracez le diagramme des résidus correspondant à la droite de meilleur ajustement lorsque l'on utilise toutes les observations de l'ensemble. Visuellement, les hypothèses du modèle de RLS sur les termes d'erreurs semblent-elles être satisfaites? Donnez une approximation visuelle de σ^2. Calculez ensuite l'estimateur $\hat{\sigma}^2$. Comparez.

Solution: on peut le faire directement comme suit.

```
mod = lm(CC.q ~ VKM.q, data = Autos)
plot(mod$residuals)
```




```
hist(mod$residuals)
```

Il y a clairment une structure dans les résidus – les hypothèses du modèle de RLS sur les termes d'erreurs ne sont vraisemblablement pas satisfaites. Les infractions sont cependant légères: la moyenne semble tout de même près de zéro, et l'écart-type est à peu près de 5. Vérifions si c'est bien le cas.

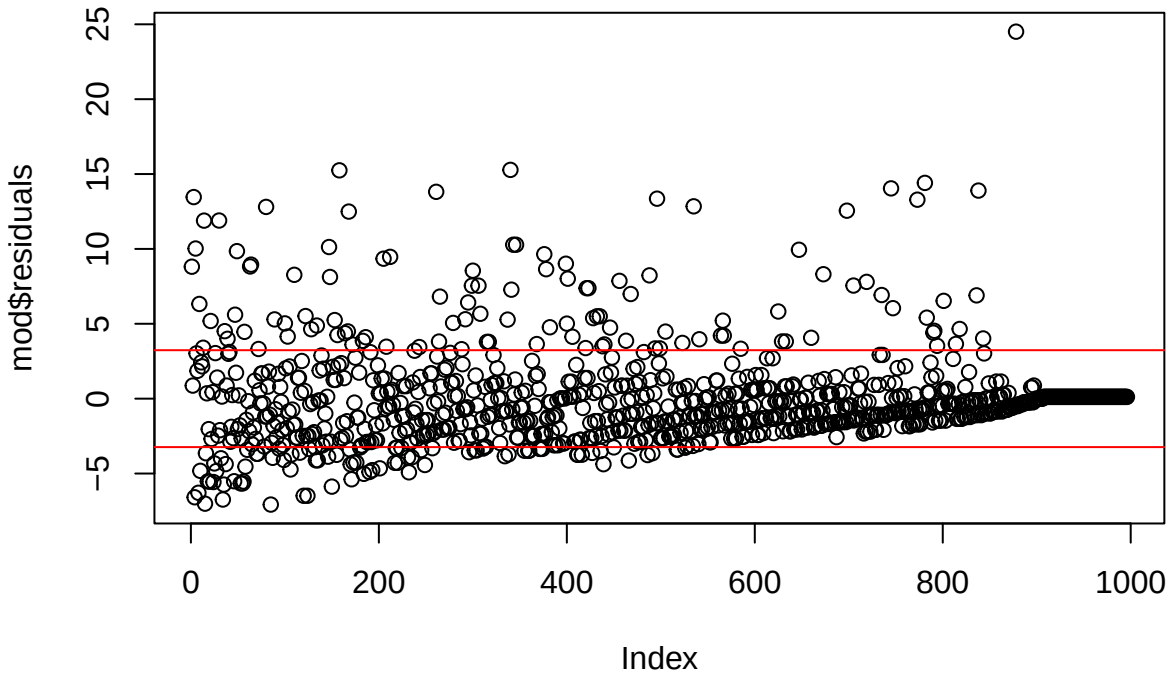
```
(e.bar = mean(mod$residuals))
```

```
## [1] 2.215574e-15
```

```
(s.e= sd(mod$residuals))
```

```
## [1] 3.235502
```

```
plot(mod$residuals)
abline(a=e.bar+s.e, b=0, col="red")
abline(a=e.bar-s.e, b=0, col="red")
```



Pas mal... mais faisons tout de même attention!

Q13

En utilisant R, calculez les valeurs de $\frac{\text{SST}}{\widehat{\sigma^2}}, \frac{\text{SSE}}{\widehat{\sigma^2}}, \frac{\text{SSR}}{\widehat{\sigma^2}}$. Semble-t-il possible que $\frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1), \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2), \frac{\text{SSR}}{\sigma^2} \sim \chi^2(1)$? Déterminez l'intervalle de confiance de la pente de la droite de régression à environ 95% et à environ 99%.

Solution: à ce stade, on peut calculer les sommes de carrés comme bon nous semble (on peut aussi tout simplement se servir des calculs des exercices précédents).

```
n = nrow(Autos)
(SST = sum(y*y)-n*(mean(y))^2)
```

```
## [1] 36827.72
```

```
(SSR = (sum(x*y)-n*mean(x)*mean(y))^2/(sum(x*x)-n*(mean(x))^2))
```

```
## [1] 26411.59
```

```
(SSE = SST - SSR)
```

```
## [1] 10416.13
```

On ne connait pas la valeur exacte de σ^2, alors on utilise l'approximation non-biasée MSE:

```
(sigma.2.hat = SSE/(n-2))
```

```
## [1] 10.479
```

Les quantités recherchées sont alors:

```
(SST/sigma.2.hat)
```

```
## [1] 3514.43
```

```
(SSR/sigma.2.hat)
```

```
## [1] 2520.43
```

```
(SSE/sigma.2.hat)
```

```
## [1] 994
```

Engendrons plusieurs valeurs provenant des lois $\chi^2(n-1) = \chi^2(995)$, $\chi^2(n-2) = \chi^2(994)$, et $\chi^2(1)$.

```
m = 50000
chi.2.995 = rchisq(m,df=n-1)
chi.2.1 = rchisq(m,df=1)
chi.2.994 = rchisq(m,df=n-2)
```

Dans quelle proportion est-ce que les valeurs échantillonnées sont plus élevées que les statistiques observées?

```
sum(chi.2.995>SST/sigma.2.hat)/m
```

```
## [1] 0
```

```
sum(chi.2.1>SSR/sigma.2.hat)/m
```

```
## [1] 0
```

```
sum(chi.2.994>SSE/sigma.2.hat)/m
```

```
## [1] 0.49372
```

C'est seulement dans le cas SSE/$sigma^2$ que la valeur observée est raisonable (ce qui implique que β_1 n'est sans doute pas nul - pourquoi?).

On peut calculer l'intervalle de confiance pour la pente β_1 directement:

```
b1 = ((sum(x*y)-n*mean(x)*mean(y))/(sum(x*x)-n*(mean(x))^2))
s.b1 = sqrt(sigma.2.hat/(sum(x*x)-n*(mean(x))^2))
```

Lorsque $\alpha = 0.05$, nous avons:

```
alpha=0.05
c(b1-qt(1-alpha/2,n-2)*s.b1, b1+qt(1-alpha/2,n-2)*s.b1)
```

```
## [1] 0.1173671 0.1269156
```

Lorsque $\alpha = 0.01$, nous avons:

```
alpha=0.01
c(b1-qt(1-alpha/2,n-2)*s.b1, b1+qt(1-alpha/2,n-2)*s.b1)
```

```
## [1] 0.1158625 0.1284201
```

Q14

Avant même de faire les calculs avec R, croyez-vous qu'on devrait être en mesure de déterminer si l'intervalle de confiance de l'ordonnée à l'origine de la droite de régression est plus petit ou plus grand que l'intervalle correspondant pour la pente? Si oui, pourquoi cela serait-il le cas? Déterminez l'intervalle de confiance de l'ordonnée à l'origine de la droite de régression à environ 95% et à environ 99%.

Solution: nous avons

$$s^2\{b_1\} = \frac{\text{MSE}}{S_{xx}} \quad \text{and} \quad s^2\{b_0\} = \frac{\text{MSE}}{S_{xx}} \cdot \overline{X}^2 + \frac{\text{MSE}}{n} = \frac{\text{MSE}}{S_{xx}} \left(\overline{X}^2 + \frac{S_{xx}}{n} \right).$$

Mais nous savons que $\overline{X}^2 \gg 1$, d'où on s'attend à ce que $s^2\{b_0\} \gg s^2\{b_1\}$, et donc à ce que $s\{b_0\} \gg s\{b_1\}$.

En effet,

```
n = nrow(Autos)
b0 = mean(y)-b1*mean(x)
s.b0 = sqrt(sigma.2.hat*(1/n+(mean(x))^2/(sum(x*x)-n*(mean(x))^2)))
```

Lorsque $\alpha = 0.05$, nous avons:

```
alpha=0.05
c(b0-qt(1-alpha/2,n-2)*s.b0, b0+qt(1-alpha/2,n-2)*s.b0)
```

```
## [1] -0.4247139  0.1879373
```

Lorsque $\alpha = 0.01$, nous avons:

```
alpha=0.01
c(b0-qt(1-alpha/2,n-2)*s.b0, b0+qt(1-alpha/2,n-2)*s.b0)
```

```
## [1] -0.5212517  0.2844751
```

Q15

En vous servant de l'ajustement des questions précédentes:

a) Testez pour $H_0 : \beta_0 = 0$ vs.@ $H_1 : \beta_0 > 0$.

b) Testez pour $H_0 : \beta_1 = 10$ vs.@ $H_1 : \beta_1 \neq 10$.

c) Testez pour $H_0 : \beta_1 = 0$ vs.@ $H_1 : \beta_1 \neq 0$.

Justifiez et expliquez vos réponses.

Solution: nous savons que

```
n = nrow(Autos)
b1 = ((sum(x*y)-n*mean(x)*mean(y))/(sum(x*x)-n*(mean(x))^2))
s.b1 = sqrt(sigma.2.hat/(sum(x*x)-n*(mean(x))^2))
b0 = mean(y)-b1*mean(x)
s.b0 = sqrt(sigma.2.hat*(1/n+(mean(x))^2/(sum(x*x)-n*(mean(x))^2)))
```

On ne spécifie pas le niveau de confiance dans la question, alors on utilise $\alpha = 0.05$. Nous aurons besoin des valeurs critiques de la loi T de Student avec $\nu = 996 - 2 = 994$ degrés de liberté, aux niveaux de confiance $1 - \alpha = 0.95$ et $1 - \alpha/2 = 0.975$:

```
(t.0975 = qt(0.975,n-2))
```

```
## [1] 1.962353
```

```
(t.095 = qt(0.95,n-2))
```

```
## [1] 1.646388
```

a) On effectue un test unilatéral à droite : la statistique observée est

```
(t.star = (b0 - 0)/s.b0)
```

```
## [1] -0.7584077
```

```
t.star > t.095
```

```
## [1] FALSE
```

Puisque la statisque n'est pas plus grande que la valeur critique, on ne peut pas rejeter $H_0 : \beta_0 = 0$ lorsque $\alpha = 0.05$.

b) On effectue un test bilatéral : la statistique observée est

```
(t.star = abs((b1 - 10)/s.b1))
```

```
## [1] 4060.107
```

```
t.star > t.0975
```

```
## [1] TRUE
```

Puisque la statisque est plus grande que la valeur critique, nous rejetons $H_0 : \beta_1 = 10$ lorsque $\alpha = 0.05$.

c) On effectue un test bilatéral : la statistique observée est

```
(t.star = abs((b1 - 0)/s.b1))
```

```
## [1] 50.20388
```

```
t.star > t.0975
```

```
## [1] TRUE
```

Puisque la statisque est plus grande que la valeur critique, nous rejetons $H_0 : \beta_1 = 0$ lorsque $\alpha = 0.05$.

Q16

a) À l'aide des formules démontrées en classe, calculez la covariance $\sigma\{b_0, b_1\}$.

b) Sélectionnez au hasard un échantillon de 50 paires d'observations dans `Autos.xlsx` (avec ou sans remise, c'est au choix). Calculez les paramètres de régression $(b_0^{(1)}, b_1^{(1)})$ correspondant. Répétez la procédure 300 fois, afin de produire 300 paires $(b_0^{(j)}, b_1^{(j)})$. Placez toutes les paires dans un diagramme de dispersion.

c) Commentez des résultats. Sont-ils compatibles avec ce que vous avez obtenu en a)?

Solution:

a) Nous avons $s\{b_0, b_1\} \approx -\overline{X}s^2\{b_1\}$.

```
(cov.b0.b1 = -mean(x)*(s.b1)^2)
```

```
## [1] -0.0002862827
```

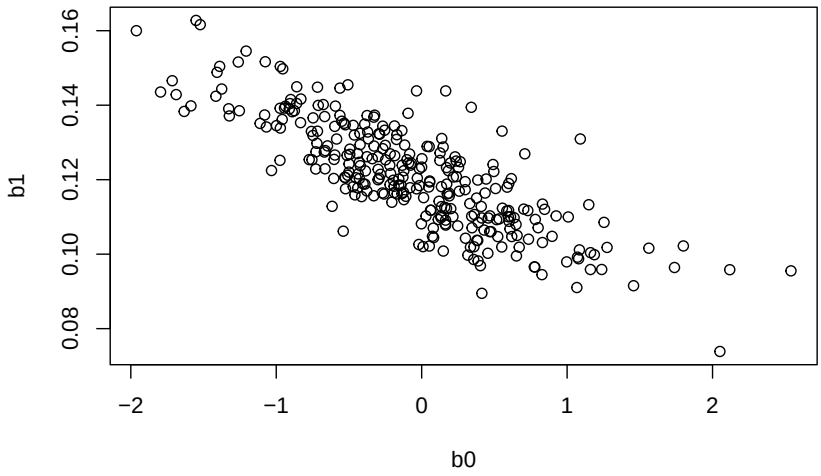
b) on choisit un échantillon aléatoire de 50 paires d'observations dans `Autos.xlsx`, sans remise:

```
set.seed(0)
n=50
ind_n = sample(nrow(Autos), n)
data_n = Autos[ind_n,]
mod_n = lm(CC.q ~ VKM.q, data=data_n)
mod_n$coefficients
```

```
## (Intercept)       VKM.q
##   1.5619088   0.1016047
```

Répétons cette procédure à $m = 300$ reprises :

```
m=300
set.seed(0)
b0 = c()
b1 = c()

for(j in 1:m){
  ind_n = sample(nrow(Autos), n)
  data_n = Autos[ind_n,]
  mod_n = lm(CC.q ~ VKM.q, data=data_n)
  b0[j] = mod_n$coefficients[[1]]
  b1[j] = mod_n$coefficients[[2]]
}

plot(b0,b1)
```



c) On peut calculer la covariance des coefficients:

```
cov(b0,b1)
```

```
## [1] -0.008188861
```

Très près de ce qui a été calculé en a). En passant, la corrélation devient:

```
cor(b0,b1)
```

```
## [1] -0.8078804
```

Q17

Déterminez l'intervalle de confiance de la réponse moyenne $\mathrm{E}\{Y\}$ à un niveau de confiance de 95% lorsque le prédicteur est $X = X^*$. Quel est l'intervalle spécifique lorsque $X^* = 27$? Calculez la moyenne des réponses $\{Y^*\}$ lorsque $X^* = 27$? Cette moyenne se retrouve-t-elle dans l'intervalle de confiance? Répétez l'exercice pour $X^* = 5$. Testez $H_0 : \mathrm{E}\{Y^* \mid X^* = 5\} = 0$ vs. $H_1 : \mathrm{E}\{Y^* \mid X^* = 5\} > 0$ à un niveau de confiance de 95%.

Solution: selon la formule,

$$\mathrm{IC}(E\{Y^* \mid X = X^*\}; 1-\alpha) = b_0 + b_1 X^* \pm t(1-\alpha/2; n-2)\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right)};$$

Si $X^* = 27$, cela devient:

```
n=nrow(Autos)
alpha=0.05
X.star = 27
mod = lm(CC.q ~ VKM.q, data=Autos)
b0 = mod$coefficients[[1]]
b1 = mod$coefficients[[2]]
x=Autos$VKM.q
y=Autos$CC.q
X.barre=mean(x)
SST = sum(y*y)-n*(mean(y))^2
SSR = (sum(x*y)-n*mean(x)*mean(y))^2/(sum(x*x)-n*(mean(x))^2)
SSE = SST - SSR
MSE = SSE/(n-2)

c(b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))),
  b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))))
```

```
## [1] 2.953772 3.405084
```

Allons chercher les observations de l'ensemble de données lorsque $X^* = 27$.

```
petit.ensemble.27 = Autos[Autos$VKM.q == X.star,]
nrow(petit.ensemble.27)
```

```
## [1] 19
```

```
mean(petit.ensemble.27$CC.q)
```

```
## [1] 3.368421
```

La moyenne se retrouve bien dans l'intervalle de confiance.

On répète le tout pour $X^* = 5$:

```
X.star = 5
c(b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))),
  b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))))
```

```
## [1] 0.2035608 0.7810760
```

```
petit.ensemble.5 = Autos[Autos$VKM.q == X.star,]
nrow(petit.ensemble.5)
```

```
## [1] 6
```

```
mean(petit.ensemble.5$CC.q)
```

```
## [1] 4.166667
```

La moyenne des $\{Y^* \mid X^* = 5\}$ ne se retrouve pas dans l'intervalle de confiance.

On test pour $H_0 : \mathrm{E}\{Y^* \mid X^* = 5\} = 0$ vs. $H_1 : \mathrm{E}\{Y^* \mid X^* = 5\} > 0$ (un test unilatéral) à un niveau de confiance de 95% en calculant la statisque observée:

```
(T.star = ((b0+b1*5)-(0))/sqrt(MSE*(1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))))
```

```
## [1] 3.345722
```

La valeur critique de la loi T de Student à $996 - 2$ degrés de liberté est:

```
(t.crit = qt(1-alpha,n-2))
```

```
## [1] 1.646388
```

Puisque $T^* > t(0.95, 964)$, on rejete l'hypothèse nulle H_0 en faveur de l'aternative $H_1 : E\{Y^* \mid X^* = 5\} > 0$ lorsque $\alpha = 0.05$.

Q18

Déterminez l'intervalle de prédiction d'une nouvelle réponse Y_p^* à un niveau de confiance de 95% lorsque le prédicteur est $X = X^*$. Quel est l'intervalle spécifique lorsque $X^* = 27$? Quel pourcentage des réponses Y_p^* se retrouvent dans l'intervalle de préduction d'une nouvelle réponse lorsque $X^* = 27$? Répétez l'exercice pour $X^* = 5$. Est ce que les résultats sont compatibles avec la notion d'intervalle de prédiction? L'observation $(5, 25)$ est-elle probable (à un niveau de confiance de 95%)?

Solution: on répète la procédure de la Q18, en remplaçant $s\{Y^*\}$ par $s\{\text{pred}^*\}$.

Si $X^* = 27$, nous avons:

```
n=nrow(Autos)
alpha=0.05
X.star = 27
mod = lm(CC.q ~ VKM.q, data=Autos)
b0 = mod$coefficients[[1]]
b1 = mod$coefficients[[2]]
x=Autos$VKM.q
y=Autos$CC.q
X.barre=mean(x)
SST = sum(y*y)-n*(mean(y))^2
SSR = (sum(x*y)-n*mean(x)*mean(y))^2/(sum(x*x)-n*(mean(x))^2)
SSE = SST - SSR
MSE = SSE/(n-2)

c(b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))),
  b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/(sum(x*x)-n*(mean(x))^2))))
```

```
## [1] -3.176970  9.535825
```

Allons chercher les observations de l'ensemble de données lorsque $X^* = 27$.

```
petit.ensemble.27 = Autos[Autos$VKM.q == X.star,]
nrow(petit.ensemble.27)
```

```
## [1] 19
```

```
Sxx = sum(x*x)-n*(mean(x))^2
mean(petit.ensemble.27$CC.q > b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/Sxx)) &
       petit.ensemble.27$CC.q < b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/Sxx)))
```

```
## [1] 1
```

Toutes les observations se retrouvent dans l'intervalle de prédiction lorsque $X^* = 27$ et $\alpha = 0.05$.

On répète le tout pour $X^* = 5$:

```
X.star = 5
c(b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/Sxx)),
  b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1/n+(X.star-X.barre)^2/Sxx)))
```

```
## [1] 0.2035608 0.7810760
```

```
petit.ensemble.5 = Autos[Autos$VKM.q == X.star,]
nrow(petit.ensemble.5)
```

```
## [1] 6
```

```
mean(petit.ensemble.5$CC.q > b0+b1*X.star-qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/Sxx)) &
       petit.ensemble.5$CC.q < b0+b1*X.star+qt(1-alpha/2,n-2)*sqrt(MSE*(1+1/n+(X.star-X.barre)^2/Sxx)))
```

```
## [1] 0.8333333
```

Seulement 5 des $\{Y^* \mid X^* = 5\}$ se retrouvent dans l'intervalle de prédiction à environ 95% (c'est acceptable puisqu'il n'y a que 6 observations dans l'ensemble de données pour lesquelles $X^* = 5$.

Mais $Y_p^* = 25$ ne se retrouve pas dans l'IP lorsque $X^* = 5$ et $\alpha = 0.05$; l'observation $(5, 25)$ n'est ainsi que peu probable (même si elle se trouve dans l'ensemble de données).

Q19

a) Effectuez une estimation conjointe des paramètres β_0 et β_1 à environ 95%. Comparez avec les résultats de la question 16.

b) Calculez la bande de confiance de Working-Hostelling pour la réponse moyenne lorsque $X = X^*$, à un niveau conjoint de confiance d'environ 95%. Superposez la droite d'ajustement et la bande en question sur le nuage de points.

c) Calculez la bande de confiance de Scheffé pour la prédiction de $g = 20$ nouvelles réponses Y_k^* pour $X = X_k^*$, $k = 1, \ldots, 20$, à un niveau conjoint de confiance d'environ 95%. Superposez la droite d'ajustement et la bande en question sur le nuage de points.

Solution:

a) On utilise la procédure de Bonferroni avec $g = 2$. Les valeurs requises ont été calculées lors des questions précédentes:

```
g=2
alpha=0.05
n
```

```
## [1] 996
```

```
b0
```

```
## [1] -0.1183883
```

```
s.b0
```

```
## [1] 0.1561011
```

```
b1
```

```
## [1] 0.1221413
```

```
s.b1
```

```
## [1] 0.002432906
```

À un niveau de confiance de $\alpha = 0.05$, les intervalles de confiance simultanés sont:

```
c(b0-qt(1-(alpha/g)/2,n-2)*s.b0,b0+qt(1-(alpha/g)/2,n-2)*s.b0)
```

```
## [1] -0.4688047  0.2320281
```

```
c(b1-qt(1-(alpha/g)/2,n-2)*s.b1,b1+qt(1-(alpha/g)/2,n-2)*s.b1)
```

```
## [1] 0.1166799 0.1276027
```

On remarque qu'ils sont tous deux plus large que les intervalles de confiance individuels.

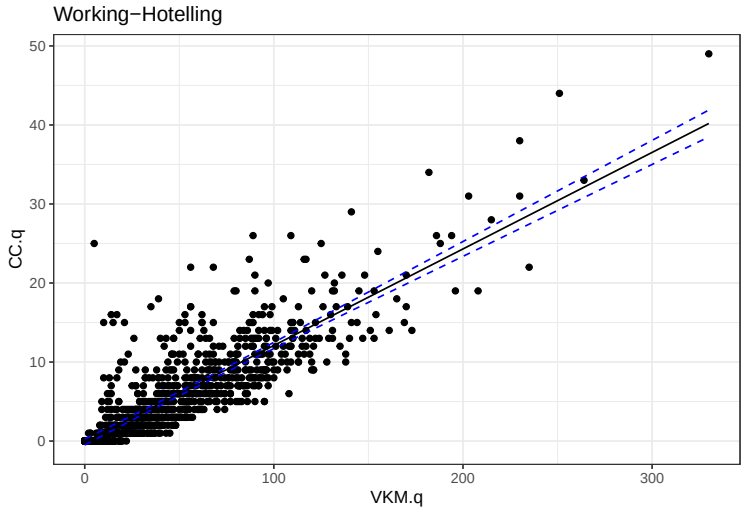
b) On pourrait utiliser le code suivant pour trouver le coefficient de Working-Hotelling:

```
y <- Autos$CC.q
x <- Autos$VKM.q
n <- length(y)

fit <- lm(y ~ x)
fit <- fit$fitted.values

se <- sqrt(sum((y - fit)^2) / (n - 2)) *
    sqrt(1 / n + (x - mean(x))^2 / sum((x - mean(x))^2))

(W <- sqrt(2 * qf(p = 0.95, df1 = 2, df2 = n - 2)))
```

```
## [1] 2.45144
```

Le nuage de point et la bande de confiance se tracent comme suit:

```
wh.upper <- fit + W * se
wh.lower <- fit - W * se

library(ggplot2)
ggplot(Autos, aes(x=VKM.q, y=CC.q)) +
  geom_point() +
  geom_line(aes(y=fit, x=VKM.q)) +
  geom_line(aes(x=VKM.q, y=wh.upper), colour='blue', linetype='dashed') +
  geom_line(aes(x=VKM.q, y=wh.lower), colour='blue', linetype='dashed') +
  labs(title='Working-Hotelling') + theme_bw()
```



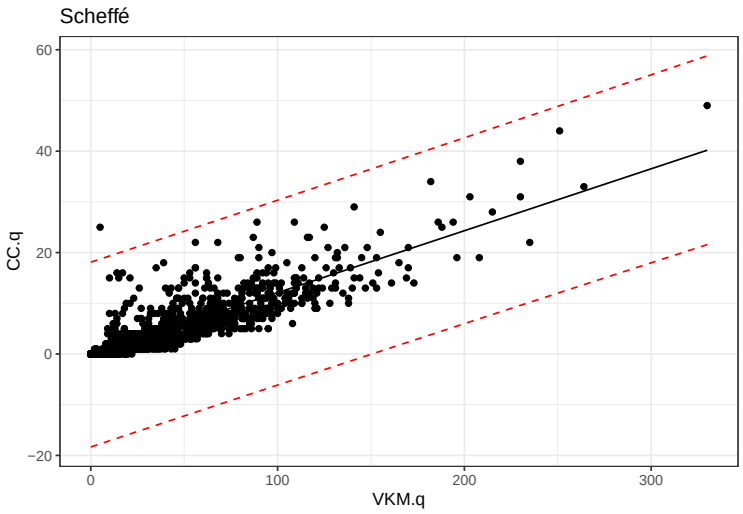
c) On pourrait utiliser le code suivant pour trouver le coefficient de Scheffé pour $g = 20$ nouvelles prédictions:

```
g=20
s.pred <- sqrt(sum((y - fit)^2) / (n - 2)) *
    sqrt(1 + 1 / n + (x - mean(x))^2 / sum((x - mean(x))^2))
(S <- sqrt(g * qf(p = 0.95, df1 = g, df2 = n - 2)))
```

```
## [1] 5.623381
```

Le nuage de point et la bande de confiance se tracent comme suit:

```
scheffe.upper <- fit + S * s.pred
scheffe.lower <- fit - S * s.pred

library(ggplot2)
ggplot(Autos, aes(x=VKM.q, y=CC.q)) +
  geom_point() +
  geom_line(aes(y=fit, x=VKM.q)) +
  geom_line(aes(x=VKM.q, y=scheffe.upper), colour='red', linetype='dashed') +
  geom_line(aes(x=VKM.q, y=scheffe.lower), colour='red', linetype='dashed') +
  labs(title='Scheffé') + theme_bw()
```



Q20

Effectuez une analyse de la variance afin de déterminer si la régression est significative ou non.

Solution: nous avons calculé les sommes de carrés au préalable.

```
SST
```

```
## [1] 36827.72
```

```
SSR
```

```
## [1] 26411.59
```

```
SSE
```

```
## [1] 10416.13
```

La statistique observée dans l'ANOVA est ainsi:

```
(F.star = (SSR/1)/(SSE/(n-2)))
```

```
## [1] 2520.43
```

Lorsque $\alpha = 0.05$, la valeur critique est

```
F.crit = qf(0.95,1,n-2)
```

Puisque $F^* > F(0.95; 1, 994)$, on rejette $H_0 : \beta_1 = 0$ en faveur de $H_1 : \beta_1 \neq 0$.