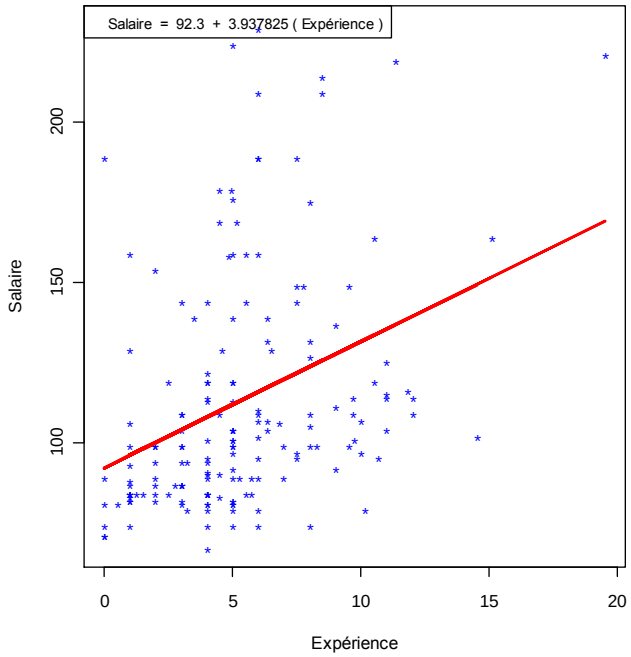
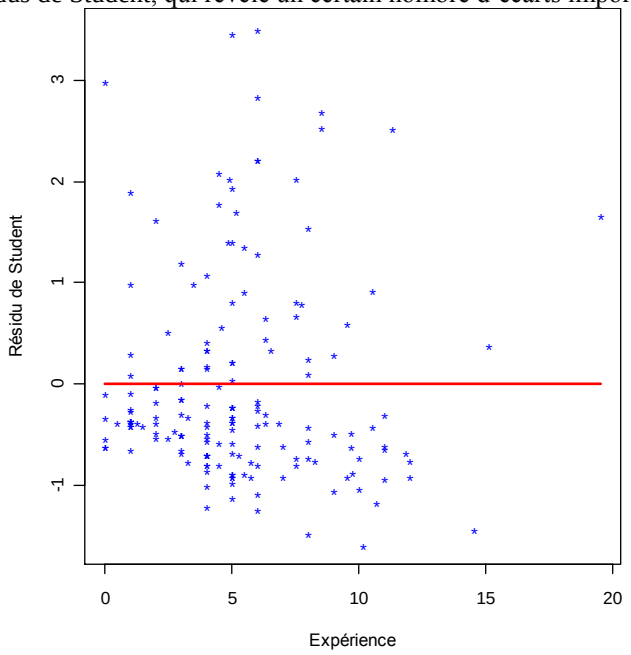
MAT7381 Esquisse de solution — Numéro 5.9

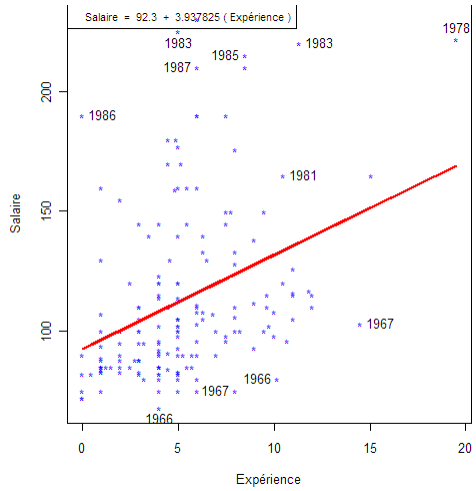
a)



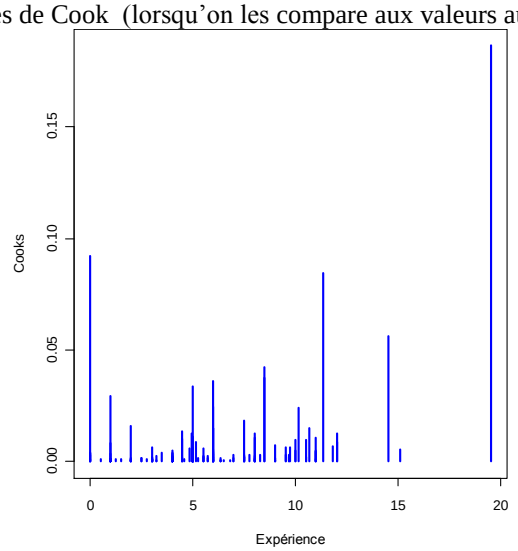
Il est évident que l'expérience préalable a joué dans la détermination du salaire à l'entrée. Bien que la dépendance est visiblement très faible, avec plusieurs données très éloignées de la droite, rien ne permet de proposer un modèle autre que linéaire.
La droite de régression est
Salaire à l'entrée = 92,295 + 3,938(Expérience)
Voici le graphique des résidus de Student, qui révèle un certain nombre d'écarts importants :



Le graphique suivant identifie certaines des données très éloignées. Il montre clairement que les valeurs très supérieures aux valeurs prédites sont expliquées par le fait qu'elles correspondent à des engagements relativement récents (par rapport à 1991).



Ces écarts, cependant, ne cause pas trop d'inquiétude étant donné qu'aucune donnée n'est particulièrement influente, selon le graphique des statistiques de Cook (lorsqu'on les compare aux valeurs au point critique $F_{2\ ;171-2\ ;0,05}$= 3,05.:



Le coefficient de détermination, R^2 = 0,1288, confirme la faiblesse de la dépendance, alors que la valeur p de 0,00000143 affirme que la dépendance, aussi faible soit-elle, est significative : elle existe vraiment dans la population et ne peut être attribuée uniquement au hasard.

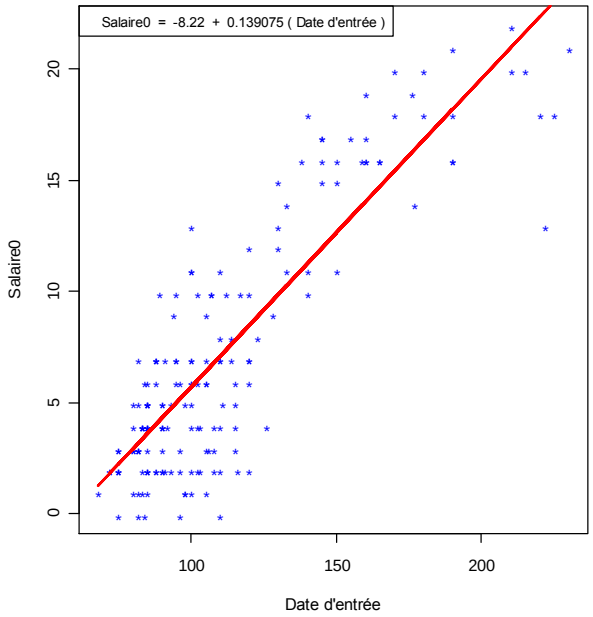
```
> a<-lm(Salaire0~Expérience)
> a
Coefficients:
(Intercept)   Expérience
     92.295        3.938

> summary(a)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.2953     4.8654  18.970  < 2e-16 ***
Expérience    3.9378     0.7877   4.999 1.43e-06 ***
```

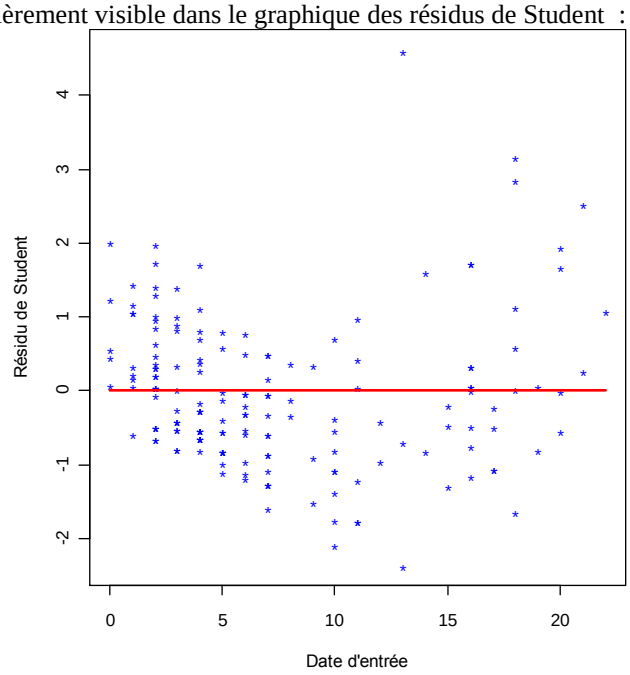
```
Residual standard error: 33.63 on 169 degrees of freedom
Multiple R-squared: 0.1288,     Adjusted R-squared: 0.1237
F-statistic: 24.99 on 1 and 169 DF,  p-value: 1.431e-06
```

b-(i)

On voit effectivement dans le nuage de points une certaine tendance non linéaire : au milieu, un concentration de points au dessus de la droite au milieu, et au-dessous de la droite aux extrémités gauche et droite :



Cette tendance est particulièrement visible dans le graphique des résidus de Student :

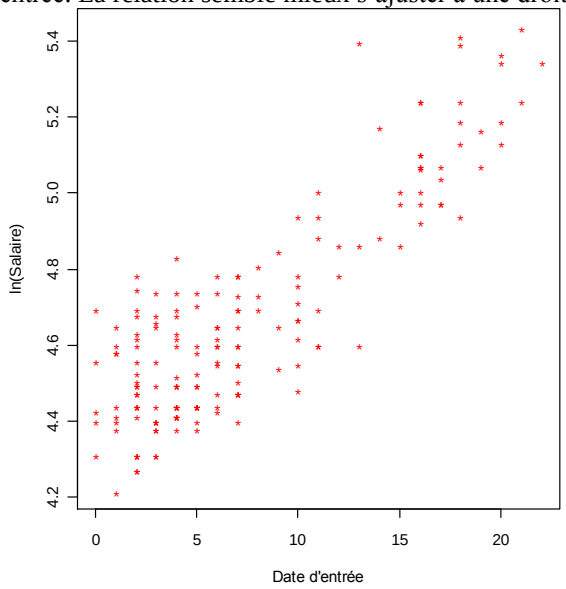



```
> a<-lm(y~x)
> summary(a)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.0708     2.2934   31.86   <2e-16 ***
x             5.3265     0.2424   21.98   <2e-16 ***

Residual standard error: 18.34 on 169 degrees of freedom
Multiple R-squared: 0.7408,     Adjusted R-squared: 0.7392
F-statistic:   483 on 1 and 169 DF,  p-value: < 2.2e-16
```

b-(ii)

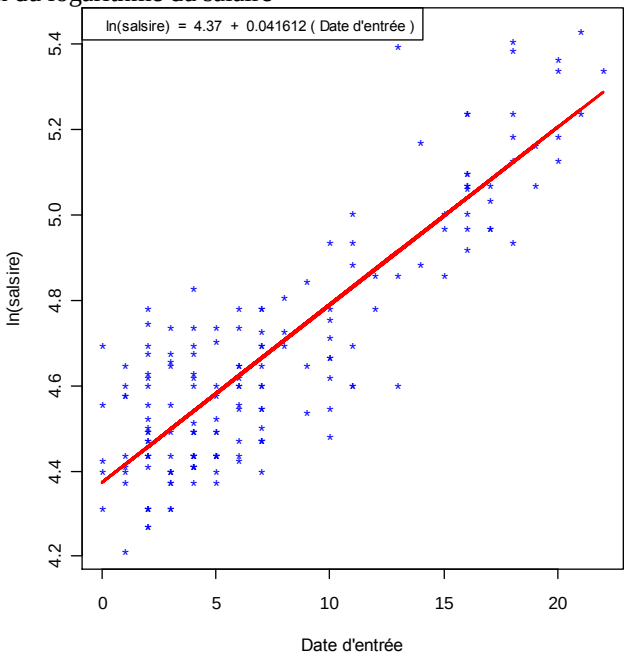
Voici le nuage de points représentant la relation entre le logarithme du salaire (à l'entrée) et la date d'entrée. La relation semble mieux s'ajuster à une droite.



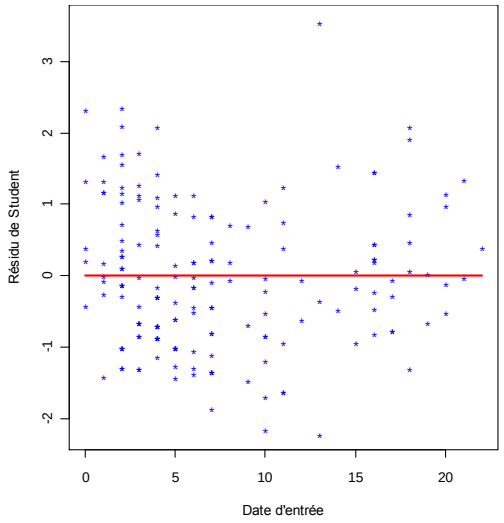
b-(iii)

> b<-lm(ly~x)

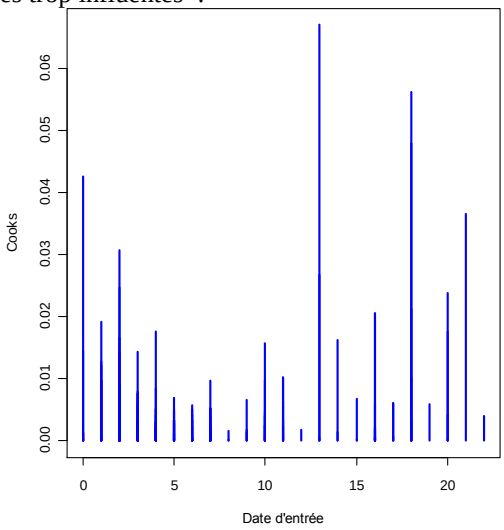
Voici la droite de régression du logarithme du salaire



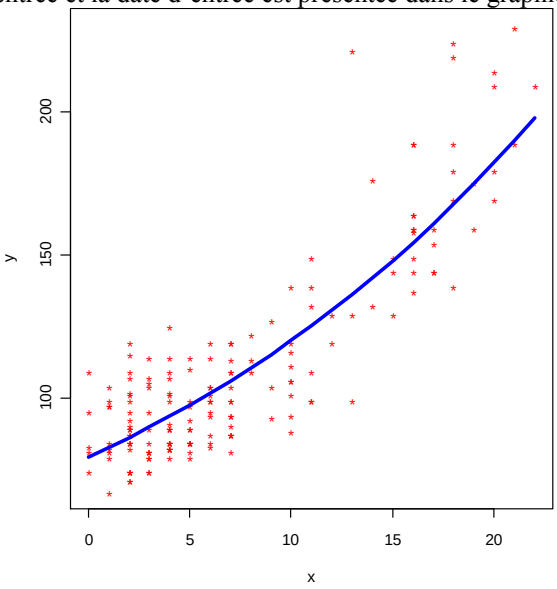
Les résidus de Student suggère qu'une certaine courbure se maintient dans les données, mais la droite semble quand même un ajustement adéquat.



Il ne semble pas y avoir de données trop influentes :



La relation entre le *salaire* à l'entrée et la date d'entrée est présentée dans le graphique suivant :



```
> summary(b)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.373165   0.017804  245.63   <2e-16 ***
x           0.041612   0.001882   22.11   <2e-16 ***

Residual standard error: 0.1424 on 169 degrees of freedom
Multiple R-squared: 0.7432,     Adjusted R-squared: 0.7417
F-statistic: 489.1 on 1 and 169 DF,  p-value: < 2.2e-16
```

b- (iv)

L'intervalle de confiance pour le logarithme du salaire est

```
> predict(a,data.frame(x=(1978-1965)),interval="confidence")
       fit      lwr      upr
1 4.914122 4.884427 4.943818
```

L'intervalle de confiance pour le salaire est

```
> exp(predict(a,data.frame(x=(1978-1965)),interval="confidence"))
       fit      lwr      upr
1 136.1997 132.2147 140.3049
```

b-(v)

Le tableau ci-dessous permet d'obtenir l'estimation de β et l'estimation de son écart-type :

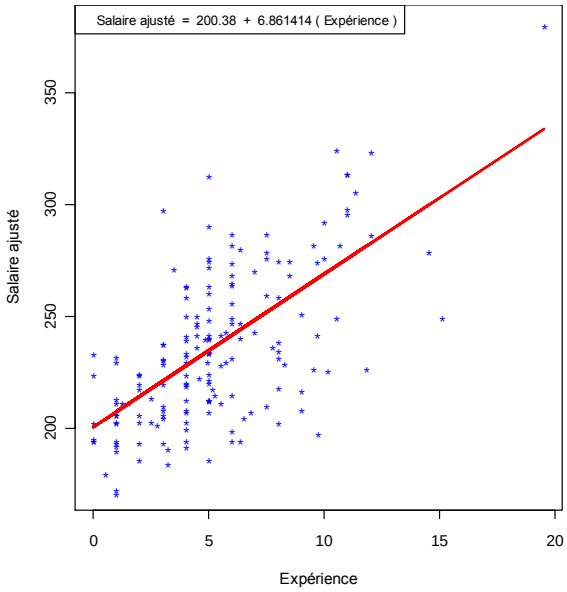
$\hat{\beta}$ = 0.0416121; $\hat{\sigma}_{\hat{\beta}}$ = 0.001881597.

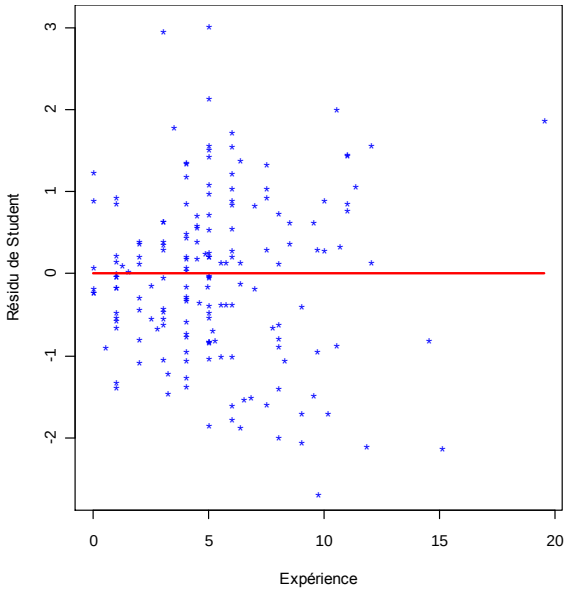
L'estimation de e^{β} est $e^{\hat{\beta}}$ = 1,04249 et l'estimation de l'écart-type de $e^{\hat{\beta}}$ est $e^{\hat{\beta}}\hat{\sigma}_{\hat{\beta}}$ = 0.001961546

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.373165   0.017804  245.63   <2e-16 ***
x           0.041612   0.001882   22.11   <2e-16 ***
> betahat
         x
0.04161211
> sigmab
[1] 0.001881597
> exp(2*b)*sb^2
[1] 3.847664e-06
> exp(b)*sb
[1] 0.001961546
```

c)

La relation semble en effet être plus forte :






```
> thetahat<-exp(betahat)-1
> thetahat
          x
0.04249003
> salaj<-Salaire0*(1+thetahat)^(26-x)

> baj<-lm(salaj~Expérience)
> baj

> summary(baj)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 200.3754     3.8534   52.00   <2e-16 ***
Expérience    6.8614     0.6239   11.00   <2e-16 ***
Residual standard error: 26.64 on 169 degrees of freedom
Multiple R-squared: 0.4172,     Adjusted R-squared: 0.4137
F-statistic:   121 on 1 and 169 DF,  p-value: < 2.2e-16
```

d)

Le salaire en 1991 est élevé pour ceux qui ont été engagés il y a longtemps. Leur salaire à l'entrée était donc peu élevé. Ce qui explique le coefficient de corrélation négatif. Lorsque les salaires sont ajustés, on observe un coefficient de corrélation positif, faible mais significatif.

```
> cor(Salaire0,Salaire91)
[1] -0.6005449
> cor(salaj,Salaire91)
[1] 0.3674278
> summary(lm(Salaire91~salaj))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 224.2512    36.8333   6.088 7.44e-09 ***
salaj         0.7919     0.1542   5.136 7.67e-07 ***
Residual standard error: 69.93 on 169 degrees of freedom
Multiple R-squared: 0.135,      Adjusted R-squared: 0.1299
F-statistic: 26.38 on 1 and 169 DF,  p-value: 7.672e-07
```