HW 3 - Linear model

ID：111024517　　　　Name：鄭家豪

due on 11/03

1.

讀取資料:

```r
dat1 <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/uswagesall.txt",
                   header = TRUE)
```

(a) Model a： wages $= \beta_0 + \beta_1(educ) + \beta_2(exper) + \epsilon$

```r
a_fit <- lm(wage ~ educ + exper,data= dat1)
summary(a_fit)
```

```
Call:
lm(formula = wage ~ educ + exper, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1136.1  -220.8   -48.3   154.5 18156.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -385.0834    13.2428  -29.08   <2e-16 ***
educ          60.8964     0.8828   68.98   <2e-16 ***
exper         10.6057     0.1957   54.19   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411.5 on 28152 degrees of freedom
Multiple R-squared:  0.1768,	Adjusted R-squared:  0.1768 
F-statistic:  3024 on 2 and 28152 DF,  p-value: < 2.2e-16
```

i. 這裡我們用 F-statistic 的值來對其假設 H_{0i}：$\beta_1 = \beta_2 = 0$ vs. H_{1i} :at least one β_k does not equal to zero 做檢定，可以發現 the calculated F-statistic 3024 is larger than the critical value $F_{(0.95,2,28152)} = 2.996051$ and the provided p-value is smaller than 0.05，因此在顯著水準為 0.05 下拒絕 H_{0i}。

ii. $H_{0ii}: \beta_1 = 0$ vs. $H_{1ii}: \beta_1$ and β_2 does not equal 0

```
aii_nullfit <- lm(wage ~exper,data = dat1)
anova(aii_nullfit,a_fit)
```

```
Analysis of Variance Table

Model 1: wage ~ exper
Model 2: wage ~ educ + exper
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1  28153 5572962645
2  28152 4767264752  1 805697893 4757.9 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

以 ANOVA 的結果來看，p-value 極小表示 H_{1ii} 比較顯著。

因此在顯著水準為 0.05 下拒絕 H_{0ii}。

iii. $H_{0iii}: wages = \beta_0 + \epsilon$ vs. $H_{1iii}: wages = \beta_0 + \beta_1(educ) + \epsilon$

```
aiii_nullfit <- lm(wage ~ 1,data=dat1)
aiii_fit <- lm(wage ~ educ,data = dat1)
anova(aiii_nullfit,aiii_fit)
```

```
Analysis of Variance Table

Model 1: wage ~ 1
Model 2: wage ~ educ
  Res.Df        RSS Df Sum of Sq    F    Pr(>F)
1  28154 5791424164
2  28153 5264467695  1 526956469 2818 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

以 ANOVA 的結果來看，p-value 極小表示 H_{1iii} 比較顯著。

因此在顯著水準為 0.05 下拒絕 H_{0iii}。

(b) The effect of 1 additional year of experience to this model is $\beta_2 = \dfrac{\partial(wage)}{\partial(exper)}$.
So,the predict effect of 1 additional year of experience to this model is $\hat{\beta}_2 = 10.6057$.

(c) Model c : $\log(wages) = \beta_0 + \beta_1(educ) + \beta_2(exper) + \epsilon$

```
c_fit <- lm(I(log(wage))~educ + exper,data= dat1)
```

(i) 這裡 F-test 的檢定統計量為: $\dfrac{(RSS_c - RSS_a)/(df_a - df_c)}{RSS_a/(n - df_a)}$, where RSS_i is the residual sum of square in model i. 由於對 wage 取 log 後，與原本 wage 的尺度不一致，在計算 RSS 時會與 question a 的 RSS 不一致，因此不能使用 F-test 來比較兩個 response 尺度不一樣的模型。

(ii)

```r
summary(c_fit)
```

```
Call:
lm(formula = I(log(wage)) ~ educ + exper, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2412 -0.3308  0.0888  0.4211  3.7032 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.4887875  0.0204402  219.60   <2e-16 ***
educ        0.1013404  0.0013627   74.37   <2e-16 ***
exper       0.0196442  0.0003021   65.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6352 on 28152 degrees of freedom
Multiple R-squared:  0.2128,	Adjusted R-squared:  0.2128 
F-statistic:  3806 on 2 and 28152 DF,  p-value: < 2.2e-16
```

這裡我們可以發現解釋變數皆顯著，與 a 一致。但 $R^2 = 0.2128 > 0.1768407$: R^2 of model a，代表 wage 能被這些解釋變數解釋的比例，model c 略勝一籌。另外，model a 的 fitted value，會有在負數值，檢驗如下:

```r
length(fitted(a_fit)[fitted(a_fit)<0])
```

```
## [1] 83
```

代表 fitted value 有 83 個是負數，這不應屬於 wage 變數的定值範圍。另外，model c 因為 expoential function 的特性，可保證每組新資料預測的 wage 恆為正。因此 model c is better fitting than model a。

(d) The effect of 1 additional year of experience to model c is $\beta_2 = \dfrac{\partial\ ln(wage)}{\partial(exper)}$.

So,the predict effect of 1 additional year of experience to model c is $\hat{\beta}_2 = 0.0196442$.

(e)

```r
e_fit <- lm(I(log(wage))~ offset(0.1*educ)+exper,data= dat1)
anova(e_fit,c_fit)
```

```
Analysis of Variance Table

Model 1: I(log(wage)) ~ offset(0.1 * educ) + exper
Model 2: I(log(wage)) ~ educ + exper
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1  28153 11358
2  28152 11358  1   0.39034 0.9675 0.3253
```

P-value= $0.3253 > 0.05$，故無足夠證據說明 $H_{0e}: \beta_1 = 0.1$ 不會成立，因此在顯著水準 0.05 下不拒絕 H_{0e}。

(f)

i. Model f : $\log(\text{wages}) = \beta_0 + \beta_1(educ) + \beta_2(exper) + \epsilon$ based on reduced data.

```
newdata <- dat1[1000*(1:28),]
f_fit <- lm(I(log(wage))~educ + exper,data= newdata)
summary(c_fit)$r.squared - summary(f_fit)$r.squared
```

```
## [1] -0.07679701
```

由於 R^2 of model c - R^2 of model f < 0 ，因此 model of this reduced data version ，在這組數據上是有較高的 R^2。

這裡 model f 的 R^2 比 model c 大，可能是因為減少後的數據能被解釋的變異比例比原始的還多，所以減少後的數據不一定總會有比原本數據高或低的 R^2。

ii.

```
summary(f_fit)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.43154 -0.27358  0.05187  0.40237  0.91710

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.736873   0.510565   9.278 1.42e-09 ***
educ        0.113308   0.035595   3.183  0.00387 **
exper       0.004255   0.008418   0.505  0.61765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6134 on 25 degrees of freedom
Multiple R-squared:  0.2896,	Adjusted R-squared:  0.2328
F-statistic: 5.096 on 2 and 25 DF,  p-value: 0.01392
```

educ 最為顯著，exper 的 p-value=0.61765 > 0.05 ，並不顯著。
利用 $T_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{(X^TX)^{-1}_{ii}}\hat{\sigma}}$ ，來檢驗 $\beta_i = 0$ 是否足夠拒絕。

```
beta_2_hat <-c(c_fit$coefficients[3],f_fit$coefficients[3])
sigma_hat <-c(summary(c_fit)$sigma,summary(f_fit)$sigma)
dataXform_c <- as.matrix(cbind(dat1$educ,dat1$exper))
dataXform_f <- as.matrix(cbind(newdata$educ,newdata$exper))
root_c <- sqrt(solve(t(dataXform_c)%*% dataXform_c)[2,2])
root_f <- sqrt(solve(t(dataXform_f)%*% dataXform_f)[2,2])
comparison <- data.frame("beta2 hat"=beta_2_hat,
                         "sigma hat"=sigma_hat,
                         "root(Gram)22"=c(root_c,root_f),
                         row.names = c("model c","model f"))
knitr::kable(comparison)
```

	beta2.hat	sigma.hat	root.Gram.22
model c	0.0196442	0.6351662	0.0004066
model f	0.0042549	0.6134343	0.0118778

以上列表為計算 T-statistic 所需的量值 (beta2.hat $= \hat{\beta}_2$,sigma.hat $= \hat{\sigma}$,root.Gram.22 $= \sqrt{(X^TX)^{-1}_{22}}$)，可以發現 model c 與 model f 的 $\hat{\sigma}$ 差不多，model c 的 $\hat{\beta}_2$ 大約是 model f 的 4.6 倍，但 $\sqrt{(\text{Gram matrix})^{-1}_{22}} = \sqrt{(X^TX)^{-1}_{22}}$ 的值，兩個模型差很多，model c 的此值太小會使得 T-statistic 過大，導致顯著。相似地，model f 的此值約為 0.012，影響 T-statistic 的幅度沒有比 model c 還要強烈。
因此，相較於 model c，expr 在 model f 不顯著的主要原因是在於 $\sqrt{(\text{Gram matrix})^{-1}_{22}}$ 不夠小。

2.

因為 $se(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}_{ii}}\hat{\sigma}$. 故 n 越大代表會讓 $se(\hat{\beta}_i)$ 變得很小以致 $\hat{\beta}_i$ 越接近真實的 β_i。還有 R^2 很小代表此數據被這模型解釋的變異比例非常少，可能因為能”主要”解釋 birth weight 的變數並不在所假設的模型上，所以在這情況下，會發生每一個解釋變數會在顯著水準 0.01 下顯著，但不足以解釋 birth weight 的情況。