Homework3

黃晨溦、廖偉傑、劉奕宏

2022-10-30

Problem 1

1.a.

Fit a model with weekly wages as the response and years of education and experience as predictors. Report the relevant test statistics and p-values for the following tests:

以下為本題所使用的變數與其變數解釋。

- **wage**: weekly wages in dollars
- **educ**: Years of education
- **exper**: Years of experience

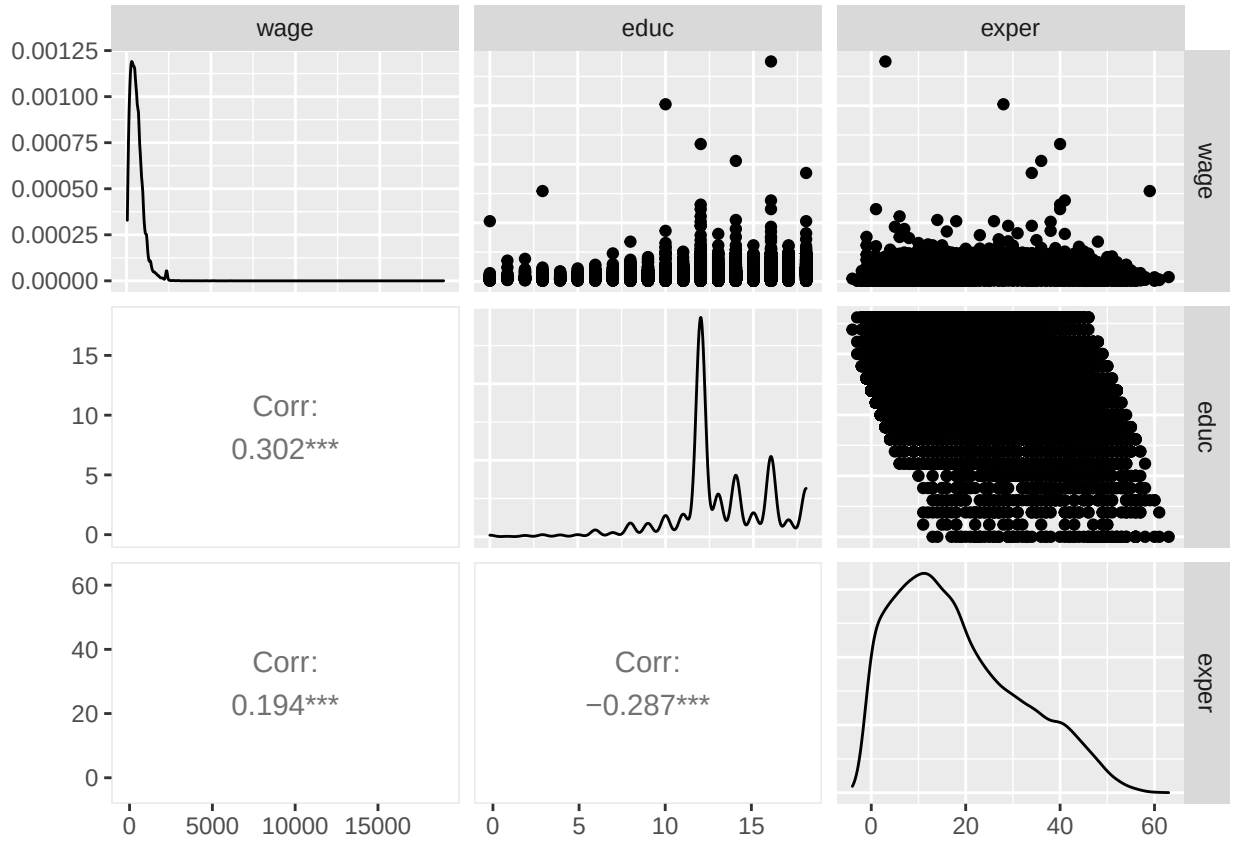
其中 **wage** 為反應變數 (response variable)，而 **educ** 和 **exper** 為解釋變數 (explanatory variable)。此資料共搜集 28155 筆樣本。

```
attach(data1)
summary(data1[,1:3])
```

```
##       wage              educ           exper
##  Min.   :   50.05   Min.   : 0.00   Min.   :-4.0
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0
##  Median :  522.32   Median :12.00   Median :16.0
##  Mean   :  603.73   Mean   :13.07   Mean   :18.2
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0
```

- 所有變數皆為 quantitative 或 continuous。

```
ggpairs(data1[,1:3],lower = list(continuous = "cor"),
        upper = list(continuous = "points"))
```



- **wage** 及 **exper** 的分布為右偏，**educ** 的分佈有左偏的趨勢，因 **wage** 為解釋變數，配適模型時考慮對此變數進行轉換。
- 3 個變數間的相關係數值皆不高。
  首先配適

$$\text{Model 1: } \mathbf{wage} = \beta_0 + \beta_1 \times \mathbf{educ} + \beta_2 \times \mathbf{exper} + \varepsilon,$$

  模型 1 配適結果如下、

```
lm1 <- lm(wage~educ+exper)
summary(lm1)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper)
##
## 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. That neither education or experience have predictive value for wages.

提出假設檢定為：

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs } H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

- H_0 所對應模型為 ω: $\textbf{wage} = \beta_0 + \varepsilon$
- H_a 所對應模型為 Ω: $\textbf{wage} = \beta_0 + \beta_1 \times \textbf{educ} + \beta_2 \times \textbf{exper} + \varepsilon$

建構檢定統計量為：

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} \overset{\text{Under } H_0 \text{ is true}}{\sim} F_{(df_\omega - df_\Omega), df_\Omega}$$

```
lm11 <- lm(wage~1)
anova(lm11,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ 1
## Model 2: wage ~ educ + exper
##   Res.Df        RSS Df  Sum of Sq    F    Pr(>F)
## 1  28154 5791424164
## 2  28152 4767264752  2 1024159412 3024 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

透過 ANOVA 表計算後可得 F 值為 3024，其自由度為 (2, 28152)，而其 $p - value < 2.2 \times 10^{-16} \ll 0.05$，檢定結果非常顯著，因此有足夠證據宣稱模型 Ω 為顯著。同樣的，我們可以從模型 1 的 summary output 得到假設檢定量 overall F-test=3024 with degree of freedom(2, 28152)，和其 $p - value < 2.2 \times 10^{-16}$。

ii. That education has no predictive value for wages when experience is also included in the model.

提出假設檢定為：

$$H_0 : \beta_1 = 0 | \ \beta_2 \text{ vs } H_a : \beta_1 \neq 0 | \ \beta_2.$$

- H_0 所對應模型為 ω: $\textbf{wage} = \beta_0 + \beta_2 \times \textbf{exper} + \varepsilon$
- H_a 所對應模型為 Ω: $\textbf{wage} = \beta_0 + \beta_1 \times \textbf{educ} + \beta_2 \times \textbf{exper} + \varepsilon$

其檢定統計量同 i. 為 F:

```
lm12 <- lm(wage~exper)
anova(lm12,lm1)
```

```
## 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
```

透過 ANOVA 表計算後可得檢定統計量 $F = 4757.9$，其自由度為 $(1, 28152)$，而其 $p - value < 2.2 \times 10^{-16} \ll 0.05$，檢定結果非常顯著，因此有足夠證據宣稱模型 Ω 為顯著。

另外，我們可以根據模型 1 之配適結果建構檢定統計量：

$$T = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \overset{\text{Under } H_0 \text{ is true}}{\sim} t_{n-p-1}$$

其中 n 為樣本數，p 為 predictor 個數。根據模型 1 之配適結果可得檢定統計量 $T = 68.98$，其自由度為 28152，而其 $p - value < 2.2 \times 10^{-16} \ll 0.05$，檢定結果非常顯著。這邊可以注意的是兩個檢定統計量之間的關係，$\sqrt{F} = 68.98 = T$，其關係可從 t 分配和 F 分配之關係式得 $t_\nu^2 = F_{1,\nu}$。

iii. That education has no predictive value for wages when experience is not included in the model.

提出假設檢定為：

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0.$$

- H_0 所對應模型為 ω: $\mathbf{wage} = \beta_0 + \varepsilon$
- H_a 所對應模型為 Ω: $\mathbf{wage} = \beta_0 + \beta_1 \times \mathbf{educ} + \varepsilon$

```
lm13 <- lm(wage~educ)
anova(lm11,lm13)
```

```
## 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
```

同樣使用檢定統計量 $F = 2818$，其自由度為 $(1, 28153)$，而其 $p - value < 2.2 \times 10^{-16} \ll 0.05$，檢定結果非常顯著，因此有足夠證據宣稱變數 **educ** 單獨對於 **wage** 有顯著的配適效果。

Note: 此檢定不同於 ii. 的地方為，ii. 的檢定是建立在給定 **exper** 的狀況下，因此 ii. 的 Ω model 有同時考慮 **educ** 和 **exper** 的效應，而 iii. 是考慮模型僅在加上 **educ** 的狀況下，對反應變數的配適效果，因此 iii. 的 Ω model 只有 **educ** 的效應。

同樣的，對於此假設檢定可從模型：

$$\mathbf{wage} = \beta_0 + \beta_1 \times \mathbf{educ} + \varepsilon,$$

配適結果得檢定統計量 T。

```
summary(lm13)$coef
```

```
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -12.82815 11.8969536 -1.078272 0.2809217
## educ         47.18097  0.8887803 53.085080 0.0000000
```

根據上面的 summary output 可得檢定統計量 $T = 53.085$，其自由度為 28153，而其 $p-value < 2.2\times10^{-16} \ll 0.05$，檢定結果同樣非常顯著。

1.b.

根據模型 1 的配適結果得到以下配適模型，

$$\widehat{\mathbf{wage}} = -385.0834 + 60.8964 \times \mathbf{educ} + 10.6057 \times \mathbf{exper},$$

根據模型 1 可推測，在固定其他變數的條件下，增加 1 單位 **exper** (1 additional year of experience) 時，預測 **wage** 平均將 增加$\hat{\beta}_1 \times 1 = 10.6057$單位。

1.c.

For a model with the log of weekly wages as the response and years of education and experience as predictors 根據題目配適

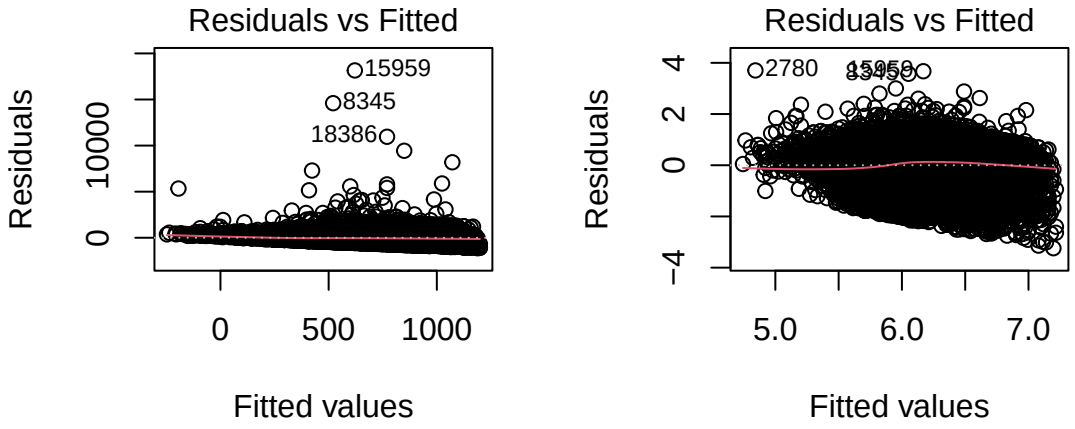
$$\text{Model 2: } \log(\mathbf{wage}) = \beta_{20} + \beta_{21} \times \mathbf{educ} + \beta_{22} \times \mathbf{exper} + \varepsilon,$$

模型 2 的配適結果為

```
lm2 <- lm(log(wage)~educ+exper)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper)
##
## 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
```

```
par(mfrow=c(1,2))
plot(lm1, which = 1)
plot(lm2, which = 1)
```



i. Can you use an F-test to compare this model to that in question a? Do it or explain why not.

雖然 a 小題模型 1 與 c 小題模型 2 放入模型的 predictors 變數一樣，但模型 2 的 response 經由 log 轉換後資料範圍乃至變異數數值都有所改變，因此和模型 1 之間沒有可比性。幾何上來說模型 1 和模型 2 之間並沒有在他們各自所展開的空間上有相互涵蓋的關係，所以不能用 F-test 直接比較。

ii. Is this a better fitting model than that in question a? Explain.

是否符合 Gauss-Markov conditions 是決定模型好壞的重要因素。在模型 1 裡 residuals 不均勻的分佈在-1300 到 20000 之間，且少數樣本擁有異常大的正值 residuals 讓模型 1 隱含 unequal variance 的問題。這裡模型 2 的 residuals 範圍相對均勻分佈在-4 到 4 之間有效解決此問題，因此，在這裡建議模型 2 會是較好的模型。並且在一開始的 EDA 裡，便有發現 wage 呈右偏的趨勢，log 轉換對於處理右偏分佈為慣用手法，所以在對於 wage 進行資料預處理的模型 2，會是相對好的模型。

注意在這裡模型 2 的 response 經由 log 轉換，單位不同，因此不太適合直接比較兩模型間的 RSS 或 R-squared 來決定哪個模型較好。若想要數值化地比較 response 經過轉換的各種模型，可以參考 Box-Cox transformation 的 criterion 值，關於 Box-Cox transformation 的使用會在之後的課程提到。

1.d.

根據模型 2 的配適結果得到以下配適模型，

$$\log(\widehat{\mathbf{wage}}) = -4.4888 + 0.1013 \times \mathbf{educ} + 0.0196 \times \mathbf{exper},$$

根據模型 2 可推測，在固定其他變數的條件下，增加 1 單位 **expr** (1 additional year of experience) 時，預測 **wage** 平均將增為 $\exp(\hat{\beta}_{22} \times 1) = \exp(0.0196) = 1.0198$ 倍 (增加 1.98%)。

1.e.

提出假設檢定為：

$$H_0 : \beta_{21} = 0.1| \ \beta_{22} \text{ vs } H_a : \beta_{21} \neq 0.1| \ \beta_{22}.$$

建構檢定統計量：

$$T = \frac{\hat{\beta}_{21} - \beta_{21}}{se(\hat{\beta}_{21})} \overset{\text{Under } H_0 \text{ is true}}{\sim} t_{n-p-1}$$

```
summary2<-summary(lm2)$coef
(summary2[2,1]-0.1)/summary2[2,2]
```

```
## [1] 0.9836363
```

```
pt((summary2[2,1]-0.1)/summary2[2,2],28152,lower.tail = F) *2
```

```
## [1] 0.3253028
```

根據上面的模型 2 的 summary output 可算得檢定統計量 $T = (\hat{\beta}_{21} - 0.1)/se(\hat{\beta}_{21}) = 0.9836$，其自由度為 28152，而其 $p-value = 0.3253 > 0.05$，檢定結果為不顯著，意即模型 2 中 **educ** 的係數為 0.1 可被接受。

1.f.

Extract every 1000th row from the dataset by "newdata <- fulldata[1000*(1:28),]", and refit the model of question c.

根據題目配適

$$\text{Model 3: } \log(wage) = \beta_{30} + \beta_{31} \times \textbf{educ} + \beta_{32} \times \textbf{exper} + \varepsilon,$$

模型 3 與模型 2 形式相同，差別在於模型 2 使用 fulldata，模型 3 使用 subdata(後面將以 subdata1 表示 fulldata[1000*(1:28),])。模型 3 結果為：

```
subdata1 <- data1[1000*(1:28), ]
lm3 <- lm(log(wage)~educ+exper, data = subdata1)
summary(lm3)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper, data = subdata1)
##
## 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
```

i. Which fit has the higher R-squared, this reduced data version or the full data version? Would a reduced data always have a higher (or lower) R-squared than the full data? Explain.

$R^2_{model3(subdata1)} = 0.2896$ 而 $R^2_{model2(fulldata)} = 0.2128$，以 subdata1 來說 reduced data version 有較高的 R^2。不過 reduced data 的 R^2 值在這裡易受影響，容易隨著選取不同的 subdata 而改變，並不一定會有較高的 R^2。當我們分別使用 subdata2 = fulldata[999 x (1:28),] 和 subdata3 = fulldata[1002 x (1:28),] 來配適模型 4 和 5 時，$R^2_{model4(subdata2)} = 0.4396$ 而 $R^2_{model5(subdata3)} = 0.0004124$。可以看到在不同選取資料的方式下，兩者的 R^2 值和原本的模型 2 相比，有異常大幅上升或降低。模型 4 和 5 的如下：

```
subdata2 <- data1[999*(1:28), ]
lm4 <- lm(log(wage)~educ+exper, data = subdata2);summary(lm4)$r.squared
```

```
## [1] 0.4395554
```

```
subdata3 <- data1[(1002*1:28), ]
lm5 <- lm(log(wage)~educ+exper, data = subdata3);summary(lm5)$r.squared
```

```
## [1] 0.0004123767
```

就 population 的概念來說，在 sub-data 的取樣下，不能保證可以抓到像在 full data 模型下反應變數和解釋變數之間的關係，而且對於反應變數本身的變異性也不能保證能由抽取出的樣本所代表，因此在配適 sub data 的模型時，計算 R-squared 的值不具有一定趨勢。

ii. Which predictors are statistically significant in this reduced data version? Compare this result to the significant predictors in the full data version and explain why the two results are different.

在模型 2 中，顯著項 **exper**，代表 Years of experience ，在模型 3 裡並不顯著。注意到從 28155 筆 reduce 至 28 筆，資料數相差約 1000 倍，此時會有兩個潛在問題：

(1) 樣本數暴力：樣本數大時即使 $\hat{\beta}_i$ 很接近 0，各變數 t-test 仍然容易因為估計精準度的上升，得到顯著的結果。樣本數小時，資料能提供的資訊不足以穩定估計 $beta$(精準度不足)，$se(\hat{\beta}_i)$ 大，對應的 $t_i = \hat{\beta}_i / se(\hat{\beta}_i)$ 值偏小，容易得到不顯著的 t-test 結果。

(2) subdata 的抽樣是否具代表性：(此部分相關結果與作圖集中顯示於此段回答之後)

(2-i) 在這 28 筆資料裡，**educ** 或 **exper** 對 log(**wage**) 的分佈是否有明顯缺失的資訊。然而，從 summary table 以及 **wage** 對 **educ**、**exper** 的 pointwise boxplot 觀察下，可以看到 subdata 提取到類似 full data 的 **educ** 和 **exper** 對於 **wage** 的趨勢，所以我們並沒有直接證據指出 fulldata 與 subdata 間 **educ** 或 **exper** 對 **wage** 的分佈不同，是造成 **exper** 不顯著的原因。

(2-ii) 這筆資料其實有許多其他類別型變數，但類別變數並沒有被包含在模型裡。當我們在不考慮這些類別因素來 reduced 資料時，連續變數與類別變數間分佈可能在取樣之後發生變化。背後牽扯的可能包含抽樣到特定族群，或是資訊被忽略。我們回過頭，對 fulldata 和 subdata 進行 EDA 比較，確認造成 **exper** 不顯著的原因。(因應

篇幅在類別變數上我們主要討論最有可能和 **exper** 相關的「是否為兼職」變數 **pt**，還有已知可能影響就業與工作處境的「種族類別」變數 **race**。)
從資料裡取出沒被模型涵蓋的類別變數（**pt**, **race**）對連續變數畫 density plot 分析時，我們可以發現 **exper** 與 **pt** 間的分佈在 fulldata 與 subdata1 的特徵明顯不同。在 fulldata 裡 $\mathbf{pt} = 1$ 的分佈呈明顯雙峰，且相較於 $\mathbf{pt} = 0$，$\mathbf{pt} = 1$ 能提供更多關於 **exper** 在數值較低與較高區域時的資訊。然而，到了 subdata1，前述 $\mathbf{pt} = 1$ 的特性消失，可能是造成 **exper** 不顯著的原因。此外，觀察到 **exper** 與 **race** 間的分佈在 fulldata 與 subdata1 趨勢是相反的，若種族議題確實存在於就業市場，subdata1 在 $\mathbf{race} = 1$（黑人族群）時有偏頗，可能會進而影響 **exper** 對 **wage** 配適的顯著性。

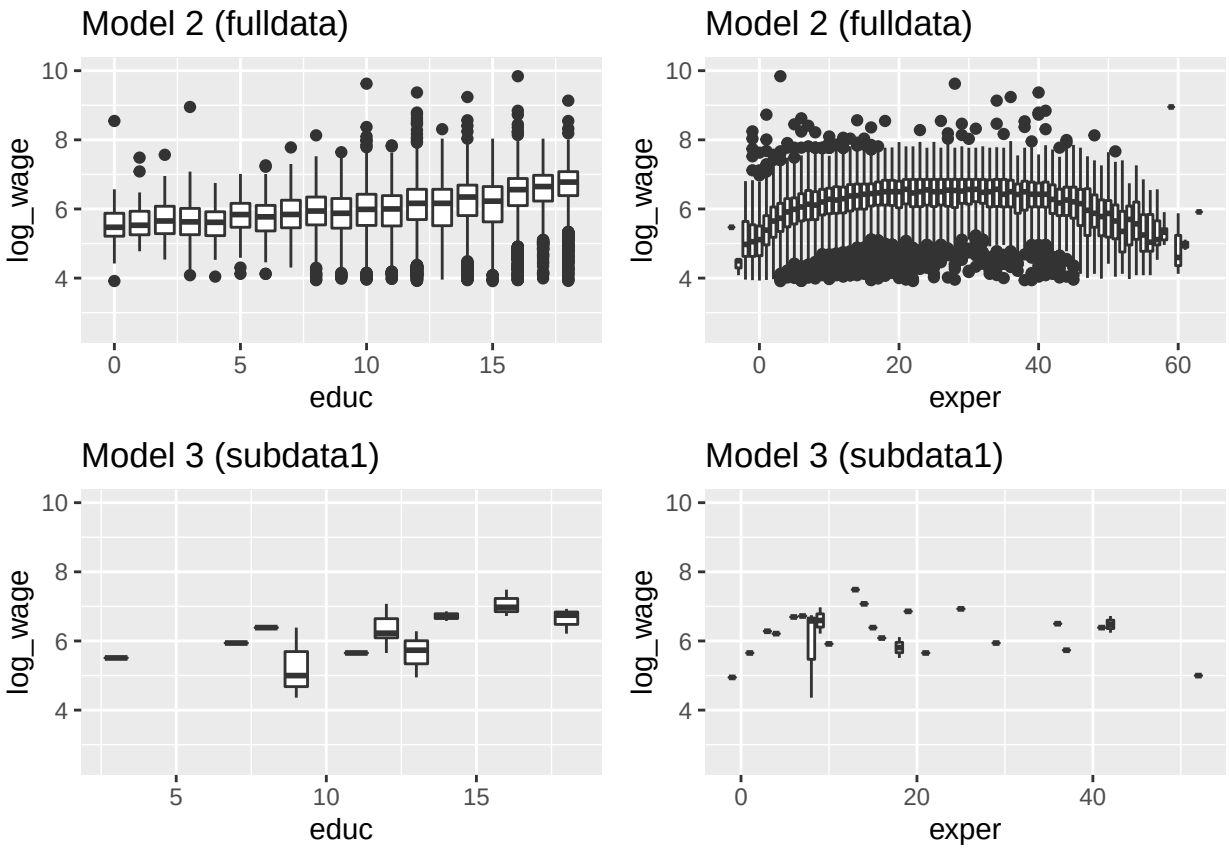
- **race**: 1 if Black, 0 if White
- **pt**: 1 if working part time, 0 if not

```
data1$race = factor(data1$race)
data1$smsa = factor(data1$smsa)
data1$pt = factor(data1$pt)
data1$log_wage = log(data1$wage)
subdata1 <- data1[1000*(1:28), ]
summary(data1[,c(1,2,11)]);summary(subdata1[,c(1,2,11)])
```

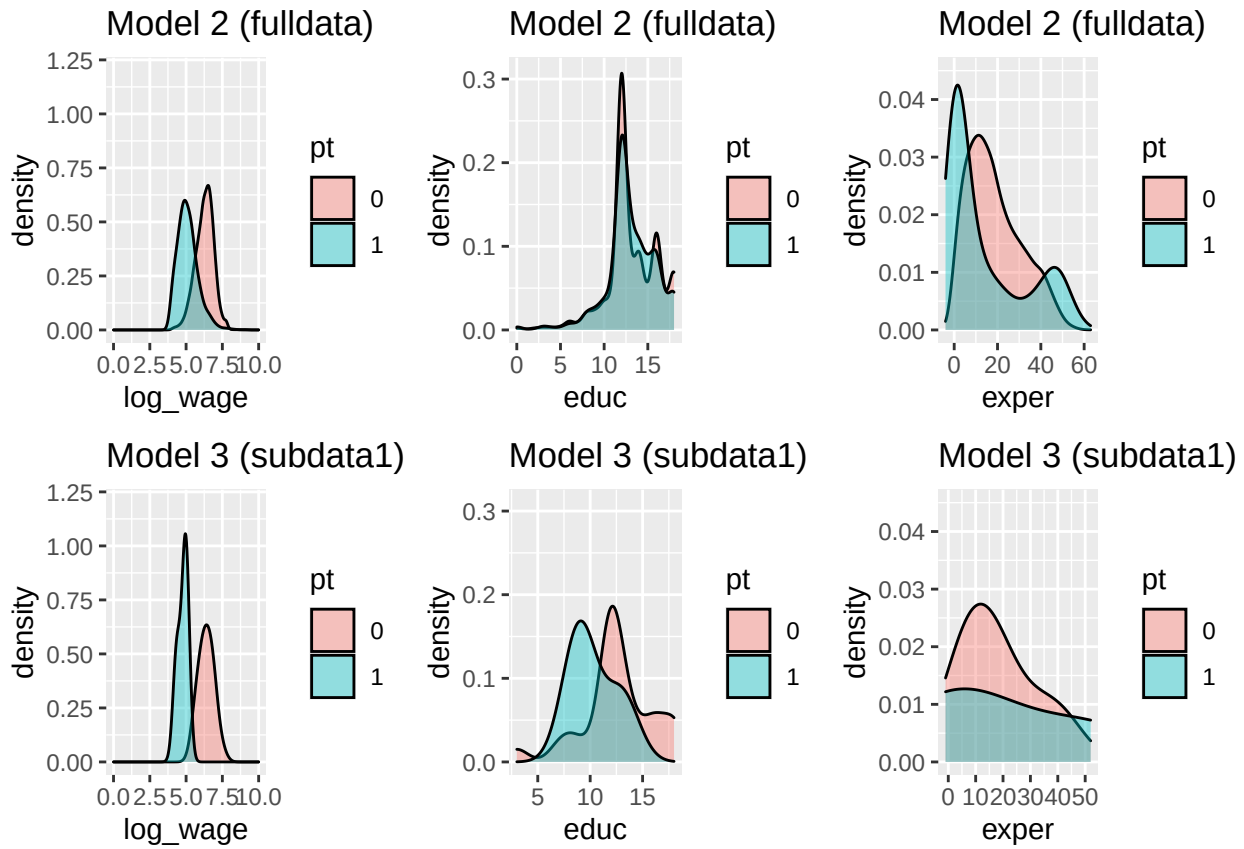
```
##       wage              educ          log_wage
##  Min.   :   50.05   Min.   : 0.00   Min.   :3.913
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.:5.732
##  Median :  522.32   Median :12.00   Median :6.258
##  Mean   :  603.73   Mean   :13.07   Mean   :6.171
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:6.664
##  Max.   :18777.20   Max.   :18.00   Max.   :9.840
```

```
##       wage             educ          log_wage
##  Min.   :  78.19   Min.   : 3.00   Min.   :4.359
##  1st Qu.: 354.94   1st Qu.:11.75   1st Qu.:5.869
##  Median : 523.00   Median :12.00   Median :6.259
##  Mean   : 609.55   Mean   :12.32   Mean   :6.210
##  3rd Qu.: 830.96   3rd Qu.:14.00   3rd Qu.:6.723
##  Max.   :1780.63   Max.   :18.00   Max.   :7.485
```

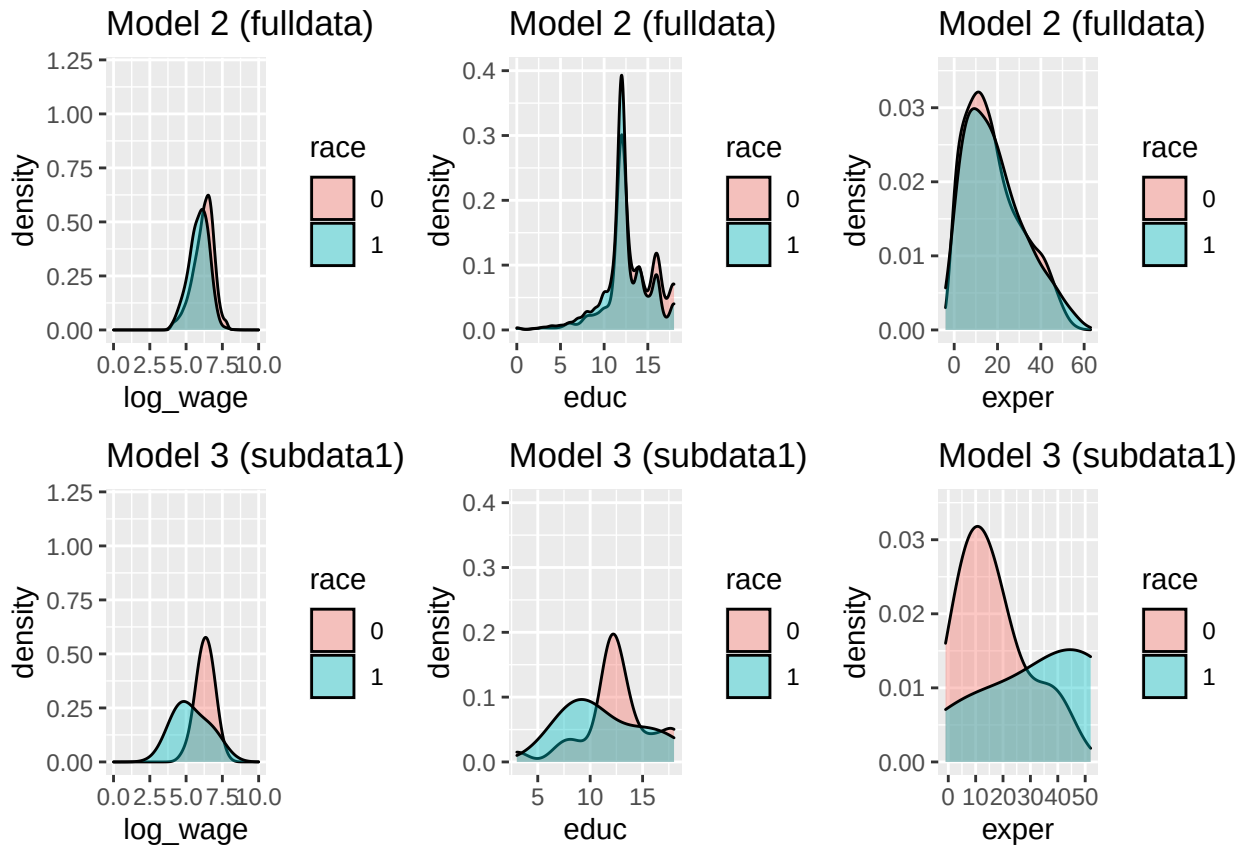
```
# If you would like to also do EDA toward living
ind <- which(data1[6:9] == 1, arr.ind = TRUE)
data1$living = factor(ifelse(rowSums(data1[6:9]) == 0, NA,
                      names(data1)[6:9][tapply(ind[, 2], ind[,1], FUN = max)]))

p1 <- ggplot(data1)+geom_boxplot(aes(educ, log_wage, group = educ)) + ylim(2.5, 10)+
  ggtitle("Model 2 (fulldata)")
p2 <- ggplot(data1)+geom_boxplot(aes(exper, log_wage, group = exper))+ylim(2.5, 10)+
  ggtitle("Model 2 (fulldata)")
p3 <- ggplot(subdata1)+geom_boxplot(aes(educ, log_wage, group = educ))+ylim(2.5, 10)+
  ggtitle("Model 3 (subdata1)")
p4 <- ggplot(subdata1)+geom_boxplot(aes(exper, log_wage, group = exper))+ylim(2.5, 10)+
  ggtitle("Model 3 (subdata1)")
gridExtra::grid.arrange(p1,p2,p3,p4, ncol = 2)
```




```
d1 <- ggplot(data1, aes(x=log_wage, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4)+xlim(0,10) + ggtitle("Model 2 (fulldata)") +
  ylim(0,1.2)
d2 <- ggplot(data1, aes(x=educ, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 2 (fulldata)") + ylim(0,0.31)
d3 <- ggplot(data1, aes(x=exper, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 2 (fulldata)") + ylim(0,0.045)
d4 <- ggplot(subdata1, aes(x=log_wage, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4) + xlim(0,10) + ggtitle("Model 3 (subdata1)") +
  ylim(0,1.2)
d5 <- ggplot(subdata1, aes(x=educ, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 3 (subdata1)") + ylim(0,0.31)
d6 <- ggplot(subdata1, aes(x=exper, group=pt, fill=pt)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 3 (subdata1)")+ ylim(0,0.045)
gridExtra::grid.arrange(d1,d2,d3,d4,d5,d6, ncol = 3)
```




```
d7 <- ggplot(data1, aes(x=log_wage, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4)+xlim(0,10) + ggtitle("Model 2 (fulldata)") +
  ylim(0,1.2)
d8 <- ggplot(data1, aes(x=educ, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 2 (fulldata)") + ylim(0,0.4)
d9 <- ggplot(data1, aes(x=exper, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 2 (fulldata)") + ylim(0,0.035)
d10 <- ggplot(subdata1, aes(x=log_wage, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4) + xlim(0,10) + ggtitle("Model 3 (subdata1)") +
  ylim(0,1.2)
d11 <- ggplot(subdata1, aes(x=educ, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 3 (subdata1)") + ylim(0,0.4)
d12 <- ggplot(subdata1, aes(x=exper, group=race, fill=race)) +
  geom_density(adjust=1.5, alpha=.4) + ggtitle("Model 3 (subdata1)") + ylim(0,0.035)
gridExtra::grid.arrange(d7,d8,d9,d10,d11,d12, ncol = 3)
```




```
detach(data1)
```

Problem 2

我們知道 $se(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}_{ii}}\hat{\sigma}$，所以當樣本數很大的時候就會導致 $se(\hat{\beta}_i)$ 變小，但也會進一步的使 $\hat{\beta}_i$ 更接近 β_i。當我們建立假設檢定為

$$H_0 : \beta_i = 0, \quad H_a : \beta_i \neq 0.$$

樣本數變大就會導致檢定統計量 t value $= \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$ 的分母變小，造成 t value 變大、p-value 變小讓係數檢定結果變的顯著。

而 R^2 很低代表了模型解釋的變異佔總體變異的比例很少，大部分的變異都沒有被這些變數解釋掉。因此，此題大致上會有兩種情況產生：

- 反應變數的變異很大，而且模型所包含的變數本身不是重要變數，所以並沒辦法有效的解釋掉反應變數的變異，從而導致模型的 R^2 很低。
- 儘管模型裡的變數皆為重要變數，但也有可能因為 measurement error 過大，大過模型的規律部分，所以就算解釋變數都是有效的，模型還是沒辦法完整的解釋反應變數的所有變異，從而導致模型的 R^2 很低。

所以我們可以跟這位婦產科醫生說明，基本上會產生這個結果是因為樣本數暴力再加上解釋變數不夠有效或是 measurement error 大過模型規律的部分而導致。

以下是模擬的 code：

```r
set.seed(100)
x1 = rnorm(10000000,0,1)
x2 = rnorm(10000000,0,0.01)
x3 = rnorm(10000000,0,0.1)
e = rnorm(10000000,0,1)
y = 3+0.01*x1+1*x2+0.1*x3+e
fit = lm(y~x1+x2+x3)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1163 -0.6750  0.0004  0.6745  5.2988
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.0000548  0.0003163 9485.75   <2e-16 ***
## x1          0.0098591  0.0003162   31.18   <2e-16 ***
## x2          1.0315346  0.0316358   32.61   <2e-16 ***
## x3          0.1080158  0.0031617   34.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 9999996 degrees of freedom
## Multiple R-squared:  0.0003204,  Adjusted R-squared:  0.0003201
## F-statistic:  1068 on 3 and 9999996 DF,  p-value: < 2.2e-16
```

從 summary 中我們就可以看到當樣本數非常大的時候 $(n = 10000000)$ 就有可能造成這種結果。