Homework-3

108048110

2022-10-31

Problem 1.

This data was drawn as a sample from the Current Population Survey in 1988.

- *wage*: weekly wages in dollars.
- *educ*: Years of education.
- *exper*: Years of experience.
- *race*: 1=black; 0=white (other races are dropped)
- *smsa*: 1=living in SMS area; 0=not
- *ne*: 1=living in the North East
- *mw*: 1= living in the Midwest
- *so*: 1=living in the South
- *pt*: 1=working part time; 0=not

```
summary(data1)
```

```
##       wage              educ          exper          race        
##  Min.   :   50.05   Min.   : 0.00   Min.   :-4.0   Min.   :0.00000  
##  1st Qu.:  308.64   1st Qu.:12.00   1st Qu.: 8.0   1st Qu.:0.00000  
##  Median :  522.32   Median :12.00   Median :16.0   Median :0.00000  
##  Mean   :  603.73   Mean   :13.07   Mean   :18.2   Mean   :0.07928  
##  3rd Qu.:  783.48   3rd Qu.:15.00   3rd Qu.:27.0   3rd Qu.:0.00000  
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0   Max.   :1.00000  
##       smsa              ne                mw                so        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :0.7435   Mean   :0.2288   Mean   :0.2438   Mean   :0.3111  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##        we               pt          
##  Min.   :0.0000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.00000  
##  Median :0.0000   Median :0.00000  
##  Mean   :0.2163   Mean   :0.08965  
##  3rd Qu.:0.0000   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1.00000
```

- There is a negative min in the predictor variable *exper*, which is not reasonable, so I look into the negative observations.

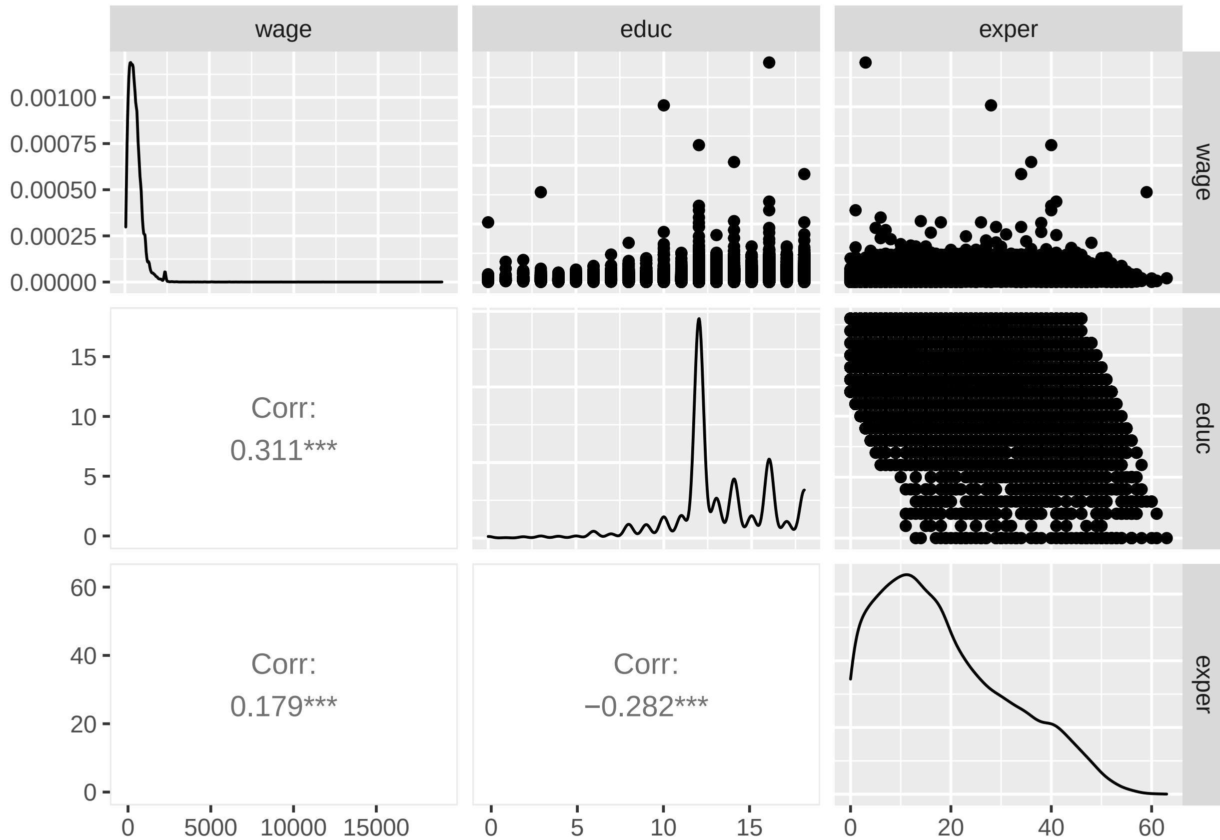
```
## [1] 438
```

- There are 438 record that should be resurveyed, and the negative values range from -4 to -1. Since there is no way to conducted the questionnaire again, I would remove these rows.

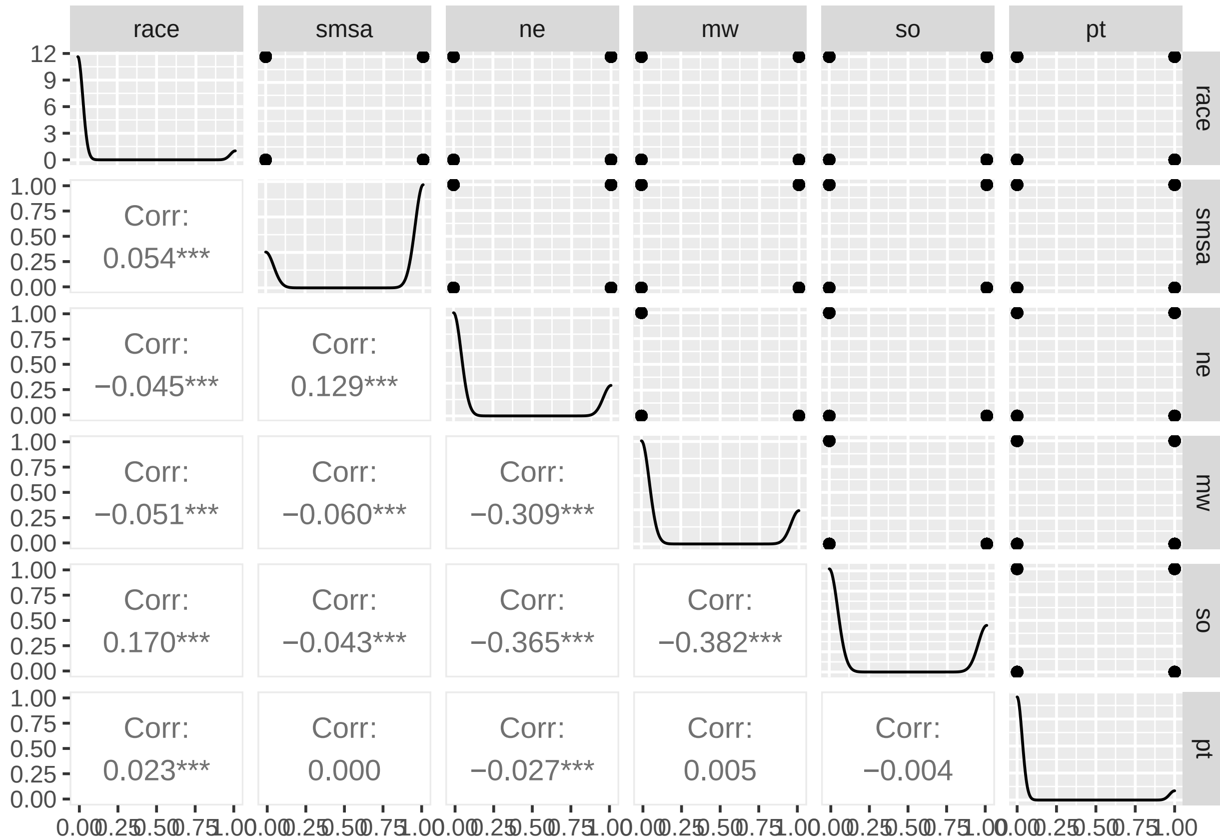
```
data1 = data1[data1$exper>=0,]
attach(data1)
```

```
summary(data1)
```

```
##       wage               educ           exper           race
##  Min.   :   50.05   Min.   : 0.00   Min.   : 0.0   Min.   :0.00000
##  1st Qu.:  313.41   1st Qu.:12.00   1st Qu.: 8.0   1st Qu.:0.00000
##  Median :  522.32   Median :12.00   Median :16.0   Median :0.00000
##  Mean   :  609.73   Mean   :13.05   Mean   :18.5   Mean   :0.07966
##  3rd Qu.:  790.36   3rd Qu.:15.00   3rd Qu.:27.0   3rd Qu.:0.00000
##  Max.   :18777.20   Max.   :18.00   Max.   :63.0   Max.   :1.00000
##       smsa              ne              mw               so
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.7423   Mean   :0.228   Mean   :0.2438   Mean   :0.3113
##  3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we               pt
##  Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000
##  Mean   :0.2169   Mean   :0.08226
##  3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000
```



wage
educ
exper
0.00100
0.00075
0.00050
0.00025
0.00000
15
10
5
0
60
40
20
0
Corr:
0.311***
Corr:
0.179***
Corr:
−0.282***
0
5000
10000
15000
0
5
10
15
0
20
40
60



- $race, smsa, ne, mw, s0, pt$ are qualitative variables, $smsa, mw, s0$ has little correlation associated with pt; while mw is the only variable that seems not to be significance to $wage$.
- $wage, educ, exper$ are quantitative variables, $wage$ and $educ$ appear to be positively correlated, the trend can also been deduced by observing the scatter plot; while $wage$ and $exper$ seem to be negatively correlated, but it's relatively unclear when looking at the scatter plot. It is worth noting that $exper$ and $educ$ seems to have a normally distributed variance.
- None of the correlation values between variables are bigger than 0.5.
- As we can observe from the scatter plots, the distribution of $wage, exper, race, ne, mw, so, we, pt$ are right skewed; yet the distribution of $educ, smsa$ are left skewed.

a. Fit a model with $wage$ as response and $educ, exper$ as predictors. Report `test statistics` and `p-values` for the following tests.

$$Model\ 1: \ wage = \beta_0 + \beta_1 * educ + \beta_2 * exper + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-375.32154	13.3002561	-28.21912	0
educ	61.11897	0.8849420	69.06551	0
exper	10.14211	0.1989626	50.97494	0

	x
value	2924.922
numdf	2.000
dendf	27714.000

sigma	r squared
411.7249	0.17429

$$RSS_{\Omega} = 411.725$$
$$dim(\Omega) = 3\ ;\ df(\Omega) = 27714$$

- $R^2 = 0.1743$, indicating that only 17.43% of the *wage* variation is interpreted by the model, there may be important explanatory variables that have not been included.
- All two variables seemed to have significant fitting results, which is consistent with the EDA graphical observation and correlation coefficient results.

i. Neither *educ* nor *exper* have predictive value for *wage*.

$$\beta_1 = \beta_2 = 0$$
$$True\ model : wage = \beta_0 + \beta_1 * educ + \beta_2 * exper + \epsilon$$
$$Fitted\ model : wage = \beta_0 + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	609.7314	2.721477	224.0443	0

sigma	r squared
453.0829	0

$$RSS_{\omega} = 453.083$$
$$dim(\omega) = 1\ ;\ df(\omega) = 27716$$

```
## Analysis of Variance Table
##
## Model 1: wage ~ 1
## Model 2: wage ~ educ + exper
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1  27716 5689655646
## 2  27714 4698005414  2 991650232 2924.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Chi-squared test statistics is 5308.1,F test statistics is 2924.9. The corresponding p-value is very small, less than 0.05, so we reject the null hypothesis that *educ*, *exper* should not be removed from the full model at the same time.

ii. ***educ*** **has no predictive value for** ***wage*** **when** ***exper*** **is included in the model.**

$$H_0: \ \beta_1 = 0$$

$$True\ model: wage = \beta_0 + \beta_1 * educ + \beta_2 * exper + \epsilon$$

$$Fitted\ model: \ wage = \beta_0 + \beta_2 * exper + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	493.642655	4.6686916	105.73469	0
exper	6.273532	0.2066895	30.35245	0

	x
value	921.2711
numdf	1.0000
dendf	27715.0000

sigma	r squared
445.7432	0.0321715

$$RSS_\omega = 445.7432$$

$$dim(\omega) = 2\ ;\ df(\omega) = 27715$$

- $R^2 = 0.03217$, this model has less interpretive ability comparing to the previous model.

```
## Likelihood ratio test
##
## Model 1: wage ~ exper
## Model 2: wage ~ educ + exper
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -208394
## 2   4 -206193  1 4401.8  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ exper
## Model 2: wage ~ educ + exper
##   Res.Df        RSS Df Sum of Sq    F    Pr(>F)
## 1  27715 5506611028
## 2  27714 4698005414  1 808605614 4770 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Chi-squared test statistics is 4401.8 and F test statistics is 4770. The corresponding p-value is still less than 0.05, so we reject the null hypothesis, which indicates that when *exper* is included in the model, one should not remove *educ* from the model.

iii. ***educ*** **has no predictive value for** ***wage*** **when** ***exper*** **is `not` included in the model.**

$$\beta_1 = 0$$

$$True\ model : wage = \beta_0 + \beta_1 * educ + \epsilon$$

$$Fitted\ model : wage = \beta_0 + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-21.96346	11.8709524	-1.850185	0.0642975
educ	48.41937	0.8880493	54.523288	0.0000000

	x
value	2972.789
numdf	1.000
dendf	27715.000

sigma	r squared
430.5863	0.096872

```
## Likelihood ratio test
##
## Model 1: wage ~ 1
## Model 2: wage ~ educ
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -208847
## 2   3 -207435  1 2824.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ 1
## Model 2: wage ~ educ
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1  27716 5689655646
## 2  27715 5138487047  1 551168599 2972.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Chi-squared test statistics is 2824.1 and F test statistics is 2972.8. The corresponding p-value is less than 0.05, so we reject the null hypothesis, which indicates that whether *exper* is included in the true model or not, one should not remove *educ* from the fitted model, and that *educ* may be an important variable for predicting wages.

b. For the model of question a, give the predicted effect of 1 additional year of experience.

$$Fitted\ model:\ wage = \beta_0 + \beta_1 * educ + \beta_2 * (exper + 1) + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-375.32154	13.3002561	-28.21912	0
educ	61.11897	0.8849420	69.06551	0
exper	10.14211	0.1989626	50.97494	0

	x
value	2924.922
numdf	2.000
dendf	27714.000

sigma	r squared
411.7249	0.17429

- Looks the same as the model of question a. The two models basically provide the same prediction results regardless of the offset value given to the variable *exper*.

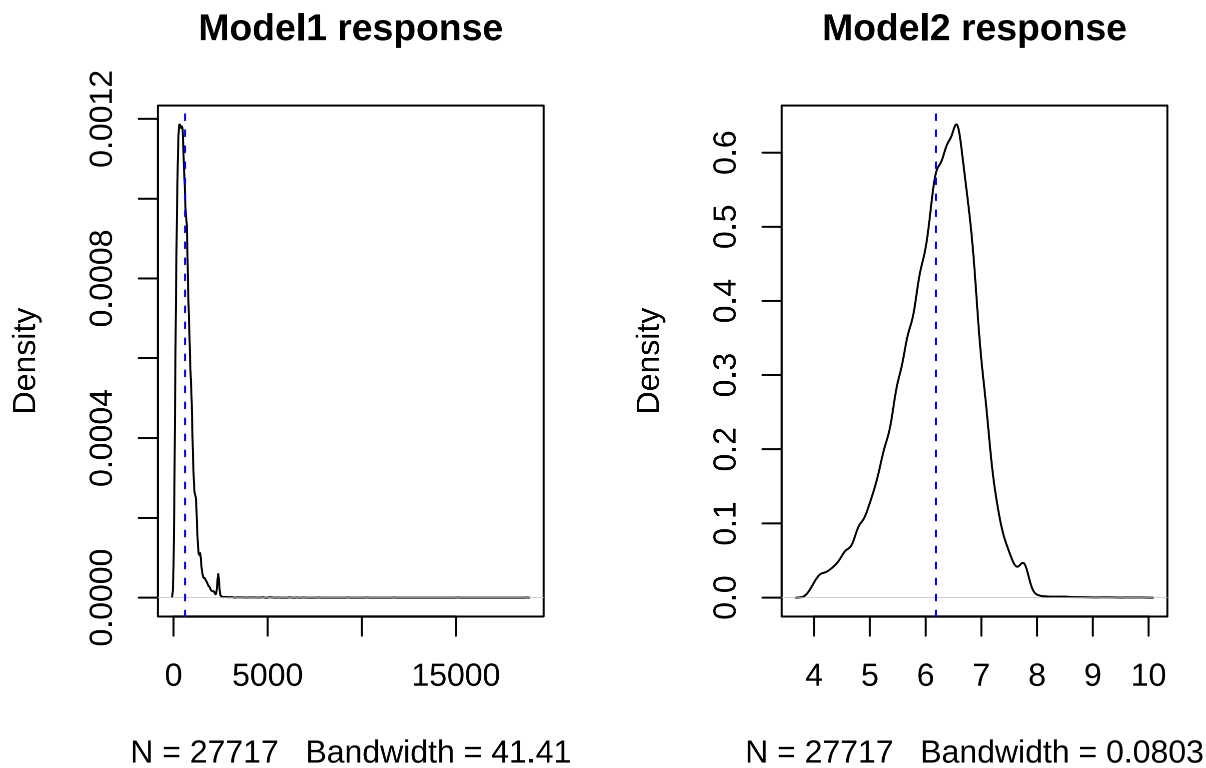
c. Fit a model with the log of weekly wages as the response and years of education and experience as predictors.

$$Model\ 2: log(wages) = \beta_0 + \beta_1 * educ + \beta_2 * exper + \epsilon$$

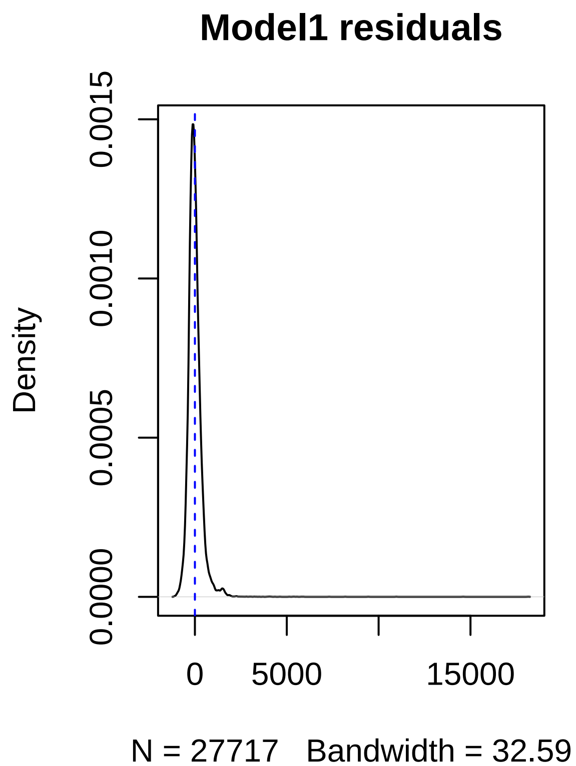
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.5247948	0.0202176	223.80482	0
educ	0.1016610	0.0013452	75.57357	0
exper	0.0181656	0.0003024	60.06332	0

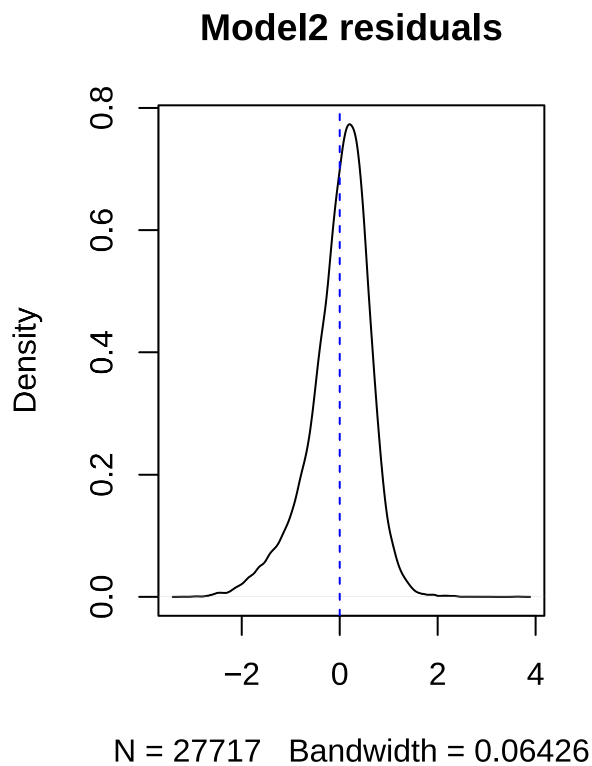
	x
value	3672.665
numdf	2.000
dendf	27714.000

sigma	r squared
0.6258592	0.2095114



- It's like standardizing the residuals.





***i.* Can you use an F-test to compare `Model 2` to `Model 1`? Do the F-test or Explain why not.**

	Estimate	Std..Error	t.value	Pr...t..
Model 1 educ	61.1189720	0.8849420	69.06551	0
Model 1 exper	10.1421092	0.1989626	50.97494	0
Model 2 educ	0.1016610	0.0013452	75.57357	0
Model 2 exper	0.0181656	0.0003024	60.06332	0

	value	numdf	dendf
Model 1	2924.922	2	27714
Model 2	3672.665	2	27714

	sigma	r squared
Model 1	411.7249165	0.1742900
Model 2	0.6258592	0.2095114

$$Model\ 1 : r^2 = 0.1742\ ;\ \ Model\ 2 : r^2 = 0.2095$$

- No, I don't. Hypothesis testings provide conjectures to respond to the question, "Which of the model spaces is more adequate in describing the data?" , we use F-test to compare two competing regression models in their ability to "explain" the variance in the predictors.
- But taking log on the response variable does not make a model simpler. Furthermore, you can't compare to model predicting different things.
- As we can observe from the general form of F statistic below,

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df(\omega) - df(\Omega))}{RSS_\Omega/df(\Omega)} = \frac{(RSS_\omega - RSS_\Omega)/(p-q)\sigma^2}{RSS_\Omega/(n-p)\sigma^2}$$

  , since both models have the same number of parameters, the denominator would be zero, and the calculated value could then not be defined.

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

a_model	c_model
411.72492	0.6258592
0.17429	0.2095114

- Since if I calculated the test statistics as $\frac{RSS_\omega}{RSS_\Omega}$, the value is just the sum of squares of ***`Model 2`*** divided by the sum of squares of ***`Model 1`*** and I assumed that the model with the lower value for the SS will fit the data better because this number represent the total distance the model is from the true data points and this was minimized during the regression procedure.
- Based on the sum of squares and the testing results, I expect the result to indicate that ***`Model 2`*** is statistically better than ***`Model 1`***.

d. For the model of question c, give the predicted effect of 1 additional year of experience.

$$Model\ 3 : log(wage) = \beta_0 + \beta_1 * educ + \beta_2 * (exper + 1) + \epsilon$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.5247948	0.0202176	223.80482	0
educ	0.1016610	0.0013452	75.57357	0
exper	0.0181656	0.0003024	60.06332	0

	x
value	3672.665
numdf	2.000
dendf	27714.000

sigma	r squared
0.6258592	0.2095114

- Contrasted to ***`Model 2`***, ***`Model 3`*** giving one additional year of experience to the parameter of the model does not change the predicted effect.

e. For the model of question c, test $\beta_1 = 0.1$

$$Full\ model : log(wage) = \beta_0 + \beta_1 * educ + \beta_2 * exper + \epsilon$$
$$H_0 : \beta_1 = 0.1\ ;\ H_1 : \beta_1 \neq 0.1$$
$$Fitted\ model : log(wage) = \beta_0 + 0.1 * educ + \beta_2 * exper + \epsilon$$

	exper	RSS	R squared
Full	10.1421092	411.7249165	0.1742900
Fitted	0.0180605	0.6258651	0.2053067

```
lrtest(lm1ii, model2)
```

```
## Likelihood ratio test
##
## Model 1: log(wage) ~ offset(0.1 * educ) + exper
## Model 2: log(wage) ~ educ + exper
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   3 -26339
## 2   4 -26338  1 1.5247     0.2169
```

```
anova(lm1ii, model2)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ offset(0.1 * educ) + exper
## Model 2: log(wage) ~ educ + exper
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1  27715 10856
## 2  27714 10856  1   0.59718 1.5246 0.2169
```

- From both likelihood ratio test and anova, I found out that the test statistics are all above the critical values, and the corresponding p-values are far above 0.05, which indicate that I should not reject the null hypothesis, that is, the parameter associated with *educ* is fairly resonable.

f. Extract every 1000th row from the dataset and refit the model of question c.

```
newdata <- data1[1000*(1:28), ]
```

- Since we have accessed indicies that is above the size of the data, I remove the last row.

***i.* Which fit has the higher R^2? Would a reduced data always have a higher or lower value than the full data?**

	sigma	r squared
Full data	0.6258592	0.2095114

	sigma	r squared
Reduced data	0.7856563	0.0738386

- The reduced data has a higher value of R^2.
- No, a reduced data would not always result in a higher R^2 value than the full data, it largely depends on how you sampled your data. If the reduced data is randomly sampled from the full data, the data for each sample would be different, on the other hand, R^2 is a measure of regression model performance, which represents the proportion of variance in response variable $wage$ that can be explained from predictors $educ, exper$; therefore, every time one reduced data from full data by sampling, it give one different model matricies to interpret the response variable, accordingly, the full model's R^2 is conducted simply by taking average on all of these sampled interpretation results (RSS_ω).
- In conclusion, the reduced-data's R^2 would varied along the full-data's R^2, not necessarily be higher or lower than 0.2095.

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

	Estimate	Std. Error	t value	Pr(>|t|)
Full intercept	4.5247948	0.0202176	223.8048157	0.0000000
Full educ	0.1016610	0.0013452	75.5735707	0.0000000
Full exper	0.0181656	0.0003024	60.0633242	0.0000000
Reduced intercept	6.2401667	1.0942652	5.7026092	0.0000071
Reduced educ	-0.0313647	0.0783957	-0.4000819	0.6926350
Reduced exper	0.0145880	0.0118877	1.2271557	0.2316660

- It is clear that only both of which predictors $educ\ andexper$ are not significant in the reduced-data model.
- Compare to the full data version model, where every predictors are significant to the response ($log(wage)$), the reduced data version of regression model suggested that $educ, exper$ are not significant to the response, that is to say, we can not use the estimated values that are obtained from this reduced version of data to infer something about the full data (population).
- In my opinion, I think there are two reasons why the reduced data generated different results comparing to the full model. First, the sample size of the reduced data is too small. There are 28155 rows of observations in the full data while in the *newdata* there are only 28 observations, (and that do not even capture 1% of the full data) which is way too small to represent the original dataset. Secondly, the reduced data is not randomly sampled from the full data, it should be generated using simple random sampling in order to be representative enough of the full data. Thus, the result generated from the reduced data may be biased, and it might not be a good sample to reach any conclusion about the full data.

```
newdata = data1[sample(nrow(data1), size=nrow(data1)*0.01),]
model4 = lm(log(wage)~educ+exper, data = newdata)
smodel4 = summary(model4)
smodel4
```

```
##
## Call:
```

```
## lm(formula = log(wage) ~ educ + exper, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9611 -0.3859  0.0445  0.4562  1.3008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.349595   0.193480  22.481  < 2e-16 ***
## educ        0.108391   0.012458   8.700 3.13e-16 ***
## exper       0.021487   0.002835   7.578 5.43e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5888 on 274 degrees of freedom
## Multiple R-squared:  0.2707, Adjusted R-squared:  0.2653
## F-statistic: 50.84 on 2 and 274 DF,  p-value: < 2.2e-16
```

- This *newdata* is now randomly sampled from the full data and have a size of (281, 10). As we can observe from the model's summary above, *educ*, *exper* are both significant again in the *newdata*.

Problem 2

- A study of infant mortality.
- Response: Baby's birth wright
- Predictors: Age of the mother, whether the birth was out of wedlock, whether the mother smoked or took drugs during pregnancy, the amount of medical attention the mother had, the mother's income...
- $R^2 = 0.092$
- Predictors was all significant at 0.01 significance level.

Explain the significance of the study.

- Significant at 1% means that every predictors' p-values are less than 0.01. And the lower the significance level (10% > 5% > 1%), the more conservative the statistical analysis and the more the data must diverge from the null hypothesis to be significant.
- A good R^2 value signifies that the model eplains a good proportion of the variability in the response variable; while a low R^2 value indicates that the model still have a great deal of unexplained variance.
- Correspondingly, the statistical significance indicates hat changes in the predictors correlate with shifts in the response variable.
- As a result, low p-value tells that one can be reasonably sure that the predictors do have an effect on the dependent variable. And **interpreting a regression coefficient that is statistically significant does not change based on the R^2 value.**

Words for the obstetrician and possible reasons.

- So, from the previous lectures we know that R^2 isn't the best measure to use when determining model's predictions are sufficiently large enough. Humans are hard to predict, it's okay to have a low R^2 value, the possible reasons why you had obtained such a low R^2 value may result from the noisiness nature of the predicted variable.

- Yet, the statistically significant between variables tells us that the knowing variables provide information about the response variable. Since you used many variables to fit the regression model, it would be easier to assess precision (rather than R^2 value) using prediction intervals, where a single new observation is likely to fall given values of the predictors that you had specified.

- As for what you can do about that low R^2 value, my suggestion is to add more predictors to your model, just keep in mind that for every study area there is an inherent amount of inexplicable variability, so certainly, you can force your regression model to fo past this issue and reach a high R^2 value but it comes at the cost of misleading regression coefficients and p-values.

- High variability around the regression line produces a lower R^2 value, and a low R^2 value may indicates that current predictors do not account for much of the variance in birth weight (underfit), and the predictors ending up with low p-values are due to the fact that regardless of other variables that may have an effect on birth weight, the mother's age, whether or not a mother took drugs, etc. babies' birth weights do tend to be affected by these variables.

- Therefore, to recapitulate, there is a statistically significant effect of current predictors on birth weight, but not enough predictors to conduct an accurate prediction.