PS2 R Solutions

Matthew Alampay Davis

October 11, 2021

Question 1

First, let's replicate the table

```
cancer <- data.frame(country = c("Switzerland", "Finland", "Great Britain",
    "Canada", "Denmark"), cigarettes = c(530, 1115, 1145, 510,
    380), deaths = c(250, 350, 465, 150, 165))
cancer  # print the table
```

```
##         country cigarettes deaths
## 1   Switzerland        530    250
## 2       Finland       1115    350
## 3 Great Britain       1145    465
## 4        Canada        510    150
## 5       Denmark        380    165
```

1i)

```
mean(cancer$cigarettes)
```

```
## [1] 736
```

The sample mean for the number of cigarettes consumed per capita in 1930 is 736

```
mean(cancer$deaths)
```

```
## [1] 276
```

The sample mean for the number of lung cancer deaths per million people in 1950 is 276

1ii)

```
sd(cancer$cigarettes)
```

```
## [1] 364.4071
```

The sample standard deviation for the number of cigarettes consumed per capita in 1930 is 364

```r
sd(cancer$deaths)
```

```
## [1] 132.3537
```

The sample standard deviation for the number of lung cancer deaths per million people in 1950 is 132

1iii)

```r
cor(cancer$cigarettes, cancer$deaths)
```

```
## [1] 0.9262529
```

The sample correlation between the two variables is 0.926

1iv)

```r
cancer.model <- lm_robust(deaths ~ cigarettes, data = cancer,
    se_type = "stata")
summary(cancer.model)
```

```
##
## Call:
## lm_robust(formula = deaths ~ cigarettes, data = cancer, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper DF
## (Intercept)  28.3966    54.6853  0.5193  0.63945 -145.6365 202.4296  3
## cigarettes    0.3364     0.0795  4.2314  0.02415    0.0834   0.5894  3
##
## Multiple R-squared:  0.8579 ,    Adjusted R-squared:  0.8106
## F-statistic:  17.9 on 1 and 3 DF,  p-value: 0.02415
```

The estimated $\hat{\beta}_1$ is 0.336

1v)

From the same output, we can see that the estimated$\hat{\beta}_0$ is 28.4

1vi)

```r
cancer$predictions <- cancer.model$fitted.values  # add a column called 'predictions' to the cancer dat
cancer$predictions  # print the predictions
```

```
## [1] 206.6979 403.5023 413.5948 199.9696 156.2353
```

1vii)

```r
cancer$residuals <- cancer$deaths - cancer$predictions
cancer$residuals
```

```
## [1]  43.302050 -53.502316  51.405153 -49.969595   8.764708
```

One thing to note here is that lm_robust models don't allow you to directly pull residuals. Instead, we just create them as the difference between the true values and the fitted/predicted values. Alternatively, we could create a non-robust lm model object and use the $ symbol to pull them:

```r
cancer.model.nonrobust <- lm(deaths ~ cigarettes, data = cancer)
cancer.model.nonrobust$residuals
```

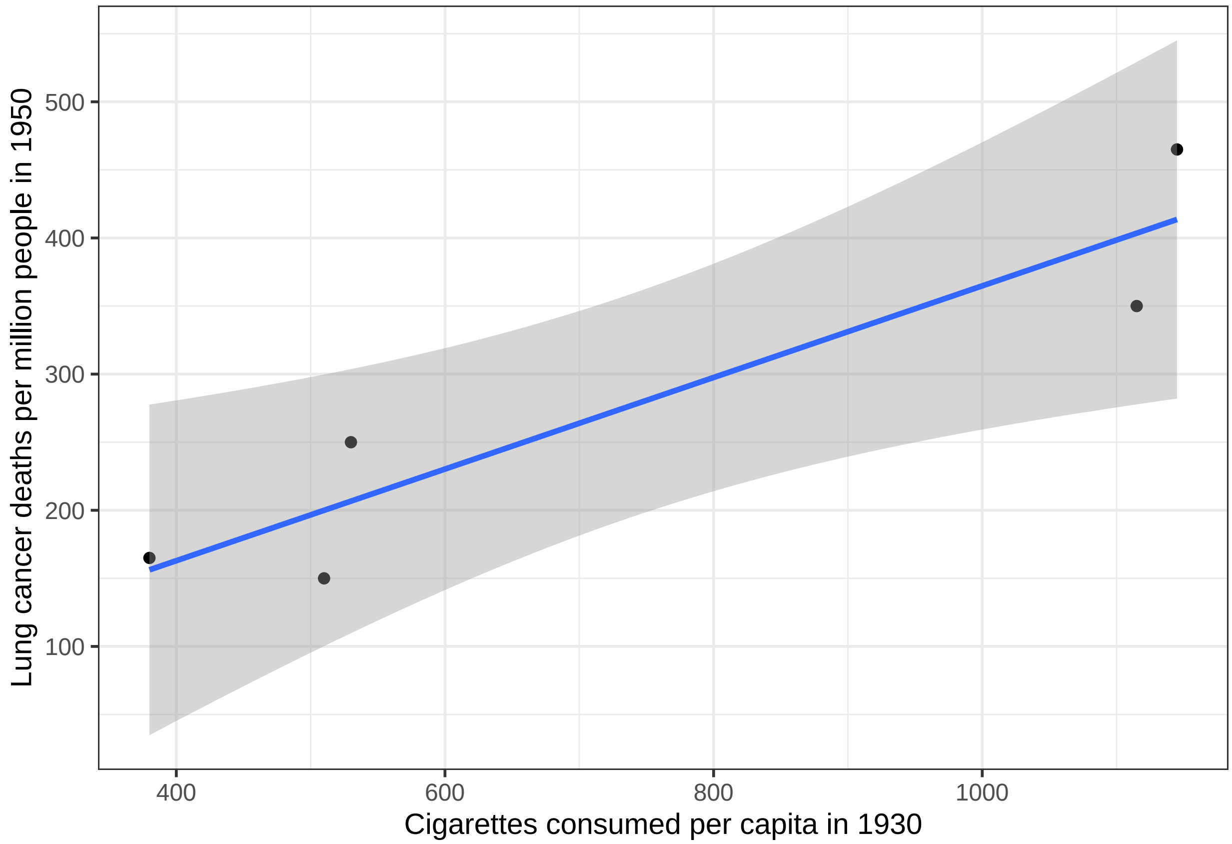
```
##          1          2          3          4          5
##  43.302050 -53.502316  51.405153 -49.969595   8.764708
```

They are exactly the same since residuals are invariant to the type of standard error used

Question 2

```r
cancer.plot <- ggplot(cancer, aes(x = cigarettes, y = deaths)) +
  theme_bw() +
  geom_point() +
  geom_smooth(method = 'lm') + # Can add an 'se = FALSE' argument to remove the confidence interval
  xlab('Cigarettes consumed per capita in 1930') + ylab('Lung cancer deaths per million people in 1950')
cancer.plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The estimated intercept is 28.4, which suggests that the linear model would predict that a country that consumes zero cigarettes in 1930 would have 28.4 lung cancer deaths per million people in 1950.

The slope of the regression is 0.336, which suggests that an increase by one cigarette concsumed per capita in 1930 is associated with an increase in the death rate of 0.336 lung cancer deaths per million in 1950.

Question 3

3(a)

```r
wage <- read.dta13("WAGE.dta")
head(wage)
```

```
##   wage hours  IQ KWW educ exper tenure age married black south urban sibs
## 1  769    40  93  35   12    11      2  31       1     0     0     1    1
## 2  808    50 119  41   18    11     16  37       1     0     0     1    1
## 3  825    40 108  46   14    11      9  33       1     0     0     1    1
## 4  650    40  96  32   12    13      7  32       1     0     0     1    4
## 5  562    40  74  27   11    14      5  34       1     0     0     1   10
## 6 1400    40 116  43   16    14      2  35       1     1     0     1    1
##   brthord meduc feduc    lwage
## 1       2     8     8 6.645091
## 2      NA    14    14 6.694562
```

```
## 3        2    14    14 6.715384
## 4        3    12    12 6.476973
## 5        6     6    11 6.331502
## 6        2     8    NA 7.244227
```

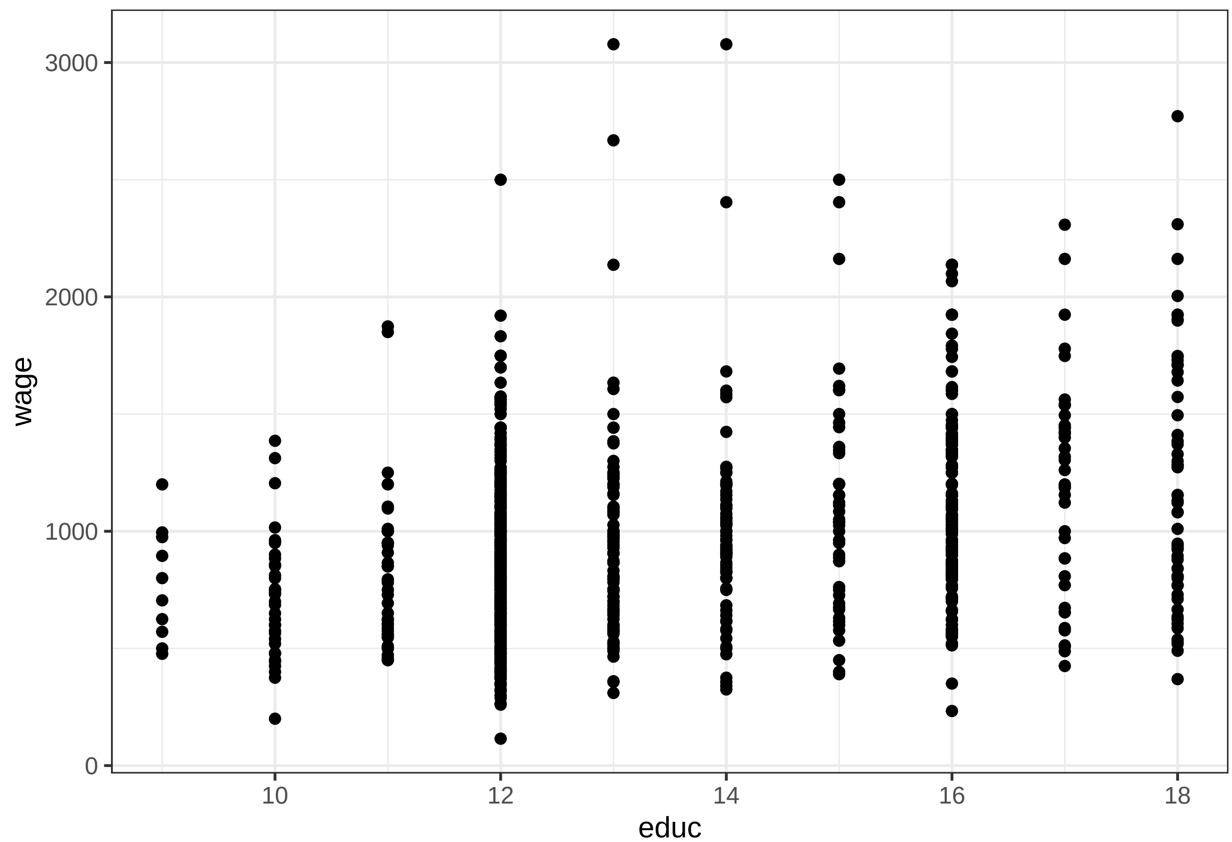
```
summary(wage)
```

```
##       wage            hours             IQ              KWW
##  Min.   : 115.0   Min.   :20.00   Min.   : 50.0   Min.   :12.00
##  1st Qu.: 669.0   1st Qu.:40.00   1st Qu.: 92.0   1st Qu.:31.00
##  Median : 905.0   Median :40.00   Median :102.0   Median :37.00
##  Mean   : 957.9   Mean   :43.93   Mean   :101.3   Mean   :35.74
##  3rd Qu.:1160.0   3rd Qu.:48.00   3rd Qu.:112.0   3rd Qu.:41.00
##  Max.   :3078.0   Max.   :80.00   Max.   :145.0   Max.   :56.00
##
##       educ           exper           tenure             age
##  Min.   : 9.00   Min.   : 1.00   Min.   : 0.000   Min.   :28.00
##  1st Qu.:12.00   1st Qu.: 8.00   1st Qu.: 3.000   1st Qu.:30.00
##  Median :12.00   Median :11.00   Median : 7.000   Median :33.00
##  Mean   :13.47   Mean   :11.56   Mean   : 7.234   Mean   :33.08
##  3rd Qu.:16.00   3rd Qu.:15.00   3rd Qu.:11.000   3rd Qu.:36.00
##  Max.   :18.00   Max.   :23.00   Max.   :22.000   Max.   :38.00
##
##     married          black            south             urban
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.893   Mean   :0.1283   Mean   :0.3412   Mean   :0.7176
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##       sibs           brthord           meduc            feduc
##  Min.   : 0.000   Min.   : 1.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 8.00   1st Qu.: 8.00
##  Median : 2.000   Median : 2.000   Median :12.00   Median :10.00
##  Mean   : 2.941   Mean   : 2.277   Mean   :10.68   Mean   :10.22
##  3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:12.00   3rd Qu.:12.00
##  Max.   :14.000   Max.   :10.000   Max.   :18.00   Max.   :18.00
##                   NA's   :83       NA's   :78      NA's   :194
##      lwage
##  Min.   :4.745
##  1st Qu.:6.506
##  Median :6.808
##  Mean   :6.779
##  3rd Qu.:7.056
##  Max.   :8.032
##
```

The mean value of wage is 957.95 units, its SD = 404.36, with min value of 115 and max value of 3078. Do the same for the other model variables.

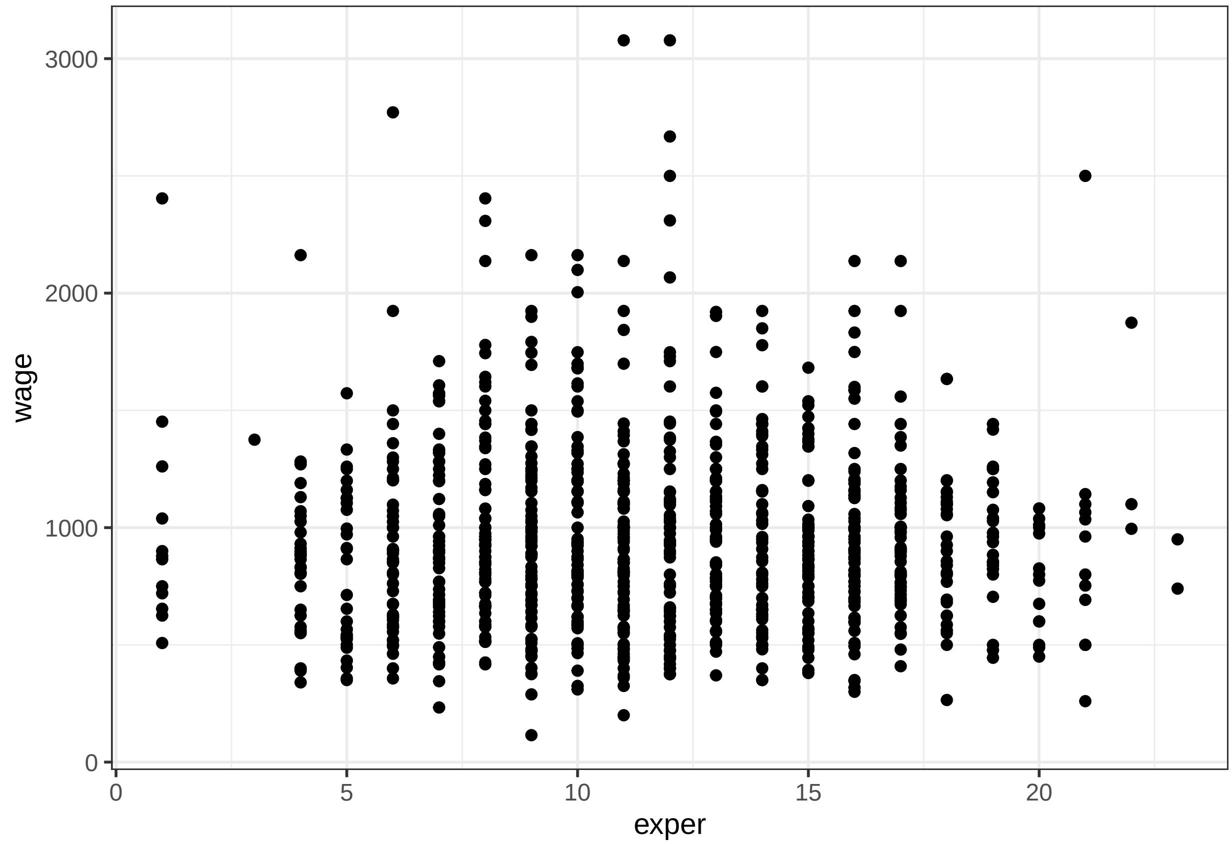
3(b)

```r
wage.plot.1 <- ggplot(wage, aes(x = educ, y = wage)) + theme_bw() +
    geom_point()

wage.plot.2 <- ggplot(wage, aes(x = exper, y = wage)) + theme_bw() +
    geom_point()

wage.plot.3 <- ggplot(wage, aes(x = tenure, y = wage)) + theme_bw() +
    geom_point()

wage.plot.1
```



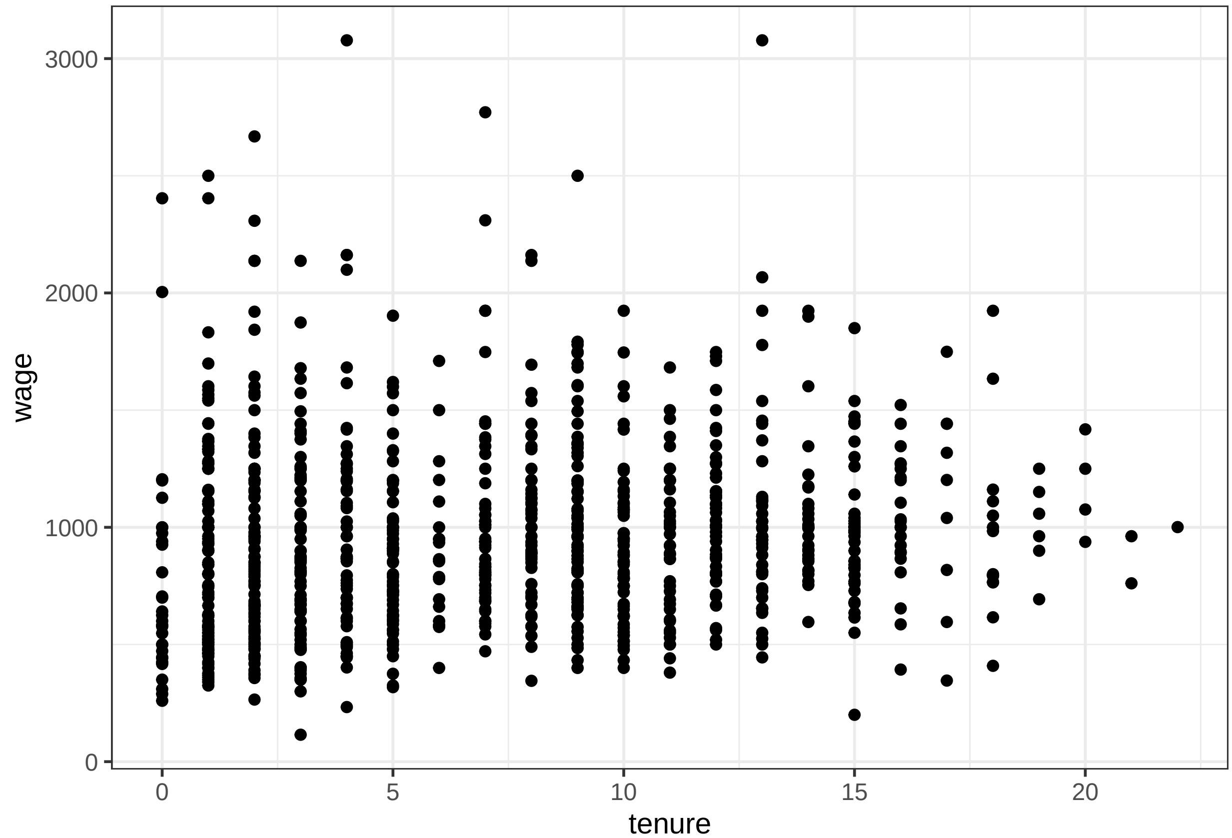
It looks like that as the number of years of education increases, wage tends to increase. It seems that they are possitively correlated, though it is not clear.

```r
wage.plot.2
```



It looks like that as the number of years of experience increases, wages tend to increase first then start to decrease after a certain threshold level of years of experience. It seems that they are not linearly correlated.

```
wage.plot.3
```



There seems to be no meaningful or clear positive or negative relationship between these two variables by visual inspection of this plot. It seems that they may have a slightly positive correlation.

3(c)

```
wage.model1 <- lm_robust(wage ~ educ, wage, se_type = "stata")
wage.model2 <- lm_robust(wage ~ exper, wage, se_type = "stata")
wage.model3 <- lm_robust(wage ~ tenure, wage, se_type = "stata")
summary(wage.model1)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   146.95     80.270   1.831 6.746e-02   -10.58    304.5 933
## educ           60.21      6.157   9.780 1.432e-21    48.13     72.3 933
##
## Multiple R-squared:  0.107 , Adjusted R-squared:  0.106
## F-statistic: 95.65 on 1 and 933 DF,  p-value: < 2.2e-16
```

In this regression, the slope coefficient is statistically significant as $t = 9.780$ but the intercept is not. This implies that as the number of years of education increases by one unit, earnings tend to increase by 60.21 units. About 10.6% of the variation in wages is explained by our explanatory variable. The 95% confidence interval for the slope is (48.13, 72.3). This interval doesn't contain zero so we can reject a null of zero slope. This is also the same for the intercept term as the confidence interval for the intercept doesn't contain zero in it and we reject a null of zero intercept.

Matt: note that our t statistic is smaller and our confidence interval is wider than appears in the 'official' Stata solutions, which did not include robust standard errors. They'd be identical if it had.

```
summary(wage.model2)
```

```
##
## Call:
## lm_robust(formula = wage ~ exper, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error  t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept) 955.6049     37.043 25.79716 3.351e-111  882.908 1028.302 933
## exper         0.2024      2.881  0.07026  9.440e-01   -5.451    5.856 933
##
## Multiple R-squared:  4.795e-06 , Adjusted R-squared:  -0.001067
## F-statistic: 0.004936 on 1 and 933 DF,  p-value: 0.944
```

In this regression, the slope coefficient is not statistically significant as $t = 0.07$ and the intercept is statistically significant with $t = 25.8$. Only 1% of variation in wage is explained by our explanatory variable, suggesting experience doesn't explain much of the variation in wages. The 95% confidence interval for the slope is (-5.45, 5.86), which contains zero and thus we cannot reject the null hypothesis of a zero slope coefficient. However, the confidence interval on the intercept coefficient does allow us to reject a null hypothesis of a zero intercept.

```
summary(wage.model3)
```

```
##
## Call:
## lm_robust(formula = wage ~ tenure, data = wage, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   884.02     24.446  36.162 1.071e-179  836.040   931.99 933
## tenure         10.22      2.507   4.076  4.963e-05    5.299    15.14 933
##
## Multiple R-squared:  0.01645 ,   Adjusted R-squared:  0.0154
## F-statistic: 16.62 on 1 and 933 DF,  p-value: 4.963e-05
```

In this regression, the slope coefficient is statistically significant at $t = 4.076$ and the intercept coefficient is also statistically significant. The implication is that an additional year working with the company is associated with an increase in wages of 10.2 units. About 17% of variation in wages is explained by our explanatory variable. The 95% confidence interval for the slope is (5.30, 15.1). This interval doesn't contain zero so we can easily reject a null of zero slope. The same is true for the intercept term.

3(d)

```r
confint(wage.model1, level = 0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) -60.23198 354.13686
## educ         44.32251  76.10606
```

```r
confint(wage.model2, level = 0.99)
```

```
##                   0.5 %      99.5 %
## (Intercept) 859.992853 1051.217025
## exper        -7.233639    7.638445
```

```r
confint(wage.model3, level = 0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) 820.91786 947.11320
## tenure        3.74868  16.69026
```