Homework 2

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

Question 1

Here is code to import and clean the cps data, with the additional filtering line added in.

```r
D <- read.csv("../data/cps-econ-4261.csv") %>%
  filter(YEAR>=2014,YEAR<=2018,AGE<=40) %>% #<- see, right here!
  mutate(EARNWEEK = na_if(EARNWEEK,9999.99),
         UHRSWORKT = na_if(na_if(na_if(UHRSWORKT,999),997),0),
         HOURWAGE = na_if(HOURWAGE,999.99)) %>%
  mutate(Wage = case_when(PAIDHOUR==1 ~ EARNWEEK/UHRSWORKT,PAIDHOUR==2 ~ HOURWAGE)) %>%
  mutate(kids = NCHILD>0,female = SEX==2) %>%
  filter(!is.na(Wage))
```

Question 2

Estimate the model:

$$\log(W_n) = \beta_0 + \beta_1 F_n$$

where W_n is the wage and F_n a dummy variable that is equal to one if the individual is female.

```r
lm(log(Wage) ~ female,D) %>%
  summary()
```

```

Call:
lm(formula = log(Wage) ~ female, data = D)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1332 -0.4317 -0.0634  0.3624  3.6552 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.856502   0.004771  598.72   <2e-16 ***
femaleTRUE  -0.119104   0.006862  -17.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5857 on 29174 degrees of freedom
Multiple R-squared:  0.01022,	Adjusted R-squared:  0.01019 
F-statistic: 301.3 on 1 and 29174 DF,  p-value: < 2.2e-16
```

Question 3

Now calculate the difference between the sample mean of log wages for women and the sample mean for men. What do you notice? Explain why.

```r
d <- D %>%
  group_by(SEX) %>%
  summarize(meanwage = mean(log(Wage)))

gap <- d$meanwage[2] - d$meanwage[1]
print(gap)
```

```
[1] -0.1191041
```

The difference is exactly the same as the point estimate for β_1. This is because the OLS estimator (as we discussed in recitation) must give that $\hat{\beta}_0$ is the sample mean of log wages for all men, and $\hat{\beta}_0 + \hat{\beta}_1$ gives the sample mean of log wages for all women, implying that $\hat{\beta}_1$ must be the difference in sample means.

Question 4

Write down a linear model that allows for wage gaps to be different by the individual's fertility status.

$$\mathbb{E}[\log(W_n)|F_n, kids_n] = \beta_0 + \beta_1 F_n + \beta_2 kid_n + \beta_3 kids_n F_n$$

where $kids_n$ is a dummy that indicates children in the household. β_3 represents the difference in age gaps by fertility status.

Question 5

Suppose that the null hypothesis is that wage gaps are the same for each. Write this null hypothesis in terms of the parameters of your model.

This is equivalent to $\beta_3 = 0$.

Question 6

Test the null hypothesis against a two-sided alternative. Make your test size 5%. Recall that we used the variable `NCHILD` to impute fertility status.

We use the variable `kids` that we constructed already:

```r
lm(log(Wage) ~ female*kids,D) %>%
  summary()
```

```

Call:
lm(formula = log(Wage) ~ female * kids, data = D)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2899 -0.3896 -0.0565  0.3633  3.5451 
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.764539   0.005932 466.010   <2e-16 ***
femaleTRUE          -0.076768   0.008912  -8.614   <2e-16 ***
kidsTRUE             0.248696   0.009756  25.493   <2e-16 ***
femaleTRUE:kidsTRUE -0.141810   0.013800 -10.276   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5782 on 29172 degrees of freedom
Multiple R-squared:  0.03567,   Adjusted R-squared:  0.03557 
F-statistic: 359.7 on 3 and 29172 DF,  p-value: < 2.2e-16
```

We estimate the the wage gap is 14 percentage points larger among individuals with children. The last column gives p-values for the two-sided hypothesis test that $\beta_3 = 0$ and strongly rejects the null. It certainly does at 5% size. Equivalently, the test statistic is much large than the 5% critical value of 1.96.

Question 7

Re-write the model to allow for:

(1) A linear trend for all wages with age;
(2) A linear trend for wage gaps with age; AND
(3) A linear trend for the the *difference* in wage gaps by fertility status.

Use this model to test the null hypothesis that the difference in wage gaps by fertility status does not change with age. Use a two-sided alternative with size 10%.

There are several models that satisfy these properties. Here are two examples.

Example 1

The model is:

$$\mathbb{E}[\log(Wage)|Age_n, kids_n, F_n] = \beta_0+\beta_1 F_n+\beta_2 kid_n+\beta_3 kids_n F_n+\beta_4 Age_n+\beta_5 Age_n F_n+\beta_6 Age_n F_n kids_n$$

Which can be estimated using:

```r
D %>%
  mutate(female = as.integer(female)) %>% #<- convert female from a factor variable to an in
  lm(log(Wage) ~ female*kids + AGE + female:AGE + female:kids:AGE,data=.) %>%
  summary()
```

```
Call:
lm(formula = log(Wage) ~ female * kids + AGE + female:AGE + female:kids:AGE,
    data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9132 -0.3202 -0.0350  0.3160  3.3358

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.6709800  0.0224386  74.469  < 2e-16 ***
female              -0.1876136  0.0364279  -5.150 2.62e-07 ***
kidsTRUE             0.0164775  0.0101212   1.628 0.103531
AGE                  0.0395450  0.0007869  50.255  < 2e-16 ***
female:kidsTRUE      0.1721093  0.0526579   3.268 0.001083 **
female:AGE           0.0050601  0.0013027   3.884 0.000103 ***
female:kidsTRUE:AGE -0.0104521  0.0016573  -6.307 2.89e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5337 on 29169 degrees of freedom
Multiple R-squared:  0.1785,    Adjusted R-squared:  0.1783
F-statistic:  1056 on 6 and 29169 DF,  p-value: < 2.2e-16
```

Under the null hypothesis, $\beta_6 = 0$. The point estimate of β_6 suggests that the difference in the wage gaps grows with age, and the two-sided p-value strongly rejects the null. Since the p-value is less than 0.05, we reject the null at 95% significance (as well as at much higher levels of significance).

Example 2:

The model is:

$$\mathbb{E}[\log(Wage)|Age_n, kids_n, F_n] = \beta_0+\beta_1 F_n+\beta_2 kid_n+\beta_3 kids_n F_n+\beta_4 Age_n+\beta_5 Age_n F_n+\beta_6 Age_n kids_n+\beta_6 Age_n F_n$$

Notice that compared to example 1, this model has an additional lower order interaction $(Age_n \times kids_n)$. This makes β_7 now the coefficient that represents how the difference in wage gaps grows with age.

We can estimate it as:

```r
lm(log(Wage) ~ female*kids*AGE,D) %>%
 summary()
```

```
Call:
lm(formula = log(Wage) ~ female * kids * AGE, data = D)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9352 -0.3196 -0.0352  0.3154  3.3358 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.6466563  0.0261550  62.958  < 2e-16 ***
femaleTRUE              -0.1632899  0.0388268  -4.206 2.61e-05 ***
kidsTRUE                 0.1180983  0.0570545   2.070  0.03847 *  
AGE                      0.0404246  0.0009249  43.709  < 2e-16 ***
femaleTRUE:kidsTRUE      0.0704884  0.0769768   0.916  0.35983    
femaleTRUE:AGE           0.0041806  0.0013904   3.007  0.00264 ** 
kidsTRUE:AGE            -0.0031852  0.0017599  -1.810  0.07033 .  
femaleTRUE:kidsTRUE:AGE -0.0072669  0.0024174  -3.006  0.00265 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5336 on 29168 degrees of freedom
Multiple R-squared:  0.1786,	Adjusted R-squared:  0.1784 
F-statistic: 905.8 on 7 and 29168 DF,  p-value: < 2.2e-16
```

And similarly strongly reject the null that $\beta_7 = 0$.