Assignment 7

```
library(car)
library(here)
library(dplyr)
library(purrr)
library(ggplot2)
```

- **age**: Age of household head.
- **educ**: Years of education for the household head.
- **earnings**: Annual earnings for the household head in 1996.
- **hours**: Annual hours worked for the household head in 1996.
- **married**: Marriage indicator for household head.
  - 1 = married
  - 0 = not married

```
data <-
    haven::read_dta(here("data", "psid97.dta")) %>%
    filter(across(everything(), ~ !is.na(.x)))
```

Question 0 - Creating new variables

- **hourly_wage**: Estimated average hourly wage found by dividing **earnings** by **hours**.
- **employed**: Indicates if head of household had recorded work or not in 1996 (**hours** is greater than 0).
  - 1 = worked
  - 0 = no recorded work
- **labor_exprnc**: Years of *prior* labor market experience (**age** - **educ** - 6).
  - **Data issue**: One individual had negative years of prior labor market experience. They were left as is however it may make sense to recode them to be 0.
  - I also clarified the labor experience variable is counting **prior** years of labor market experience because we aren't taking into account the current year.

```
dataNew <-
    data %>%
    mutate(hourly_wage = if_else(hours != 0, earnings / hours, 0),
           employed = if_else(hours > 0, 1, 0),
           labor_exprnc = age - educ - 6)
```

```
map(dataNew, summary)
```

```
## $age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   34.00   42.50   44.77   53.00   95.00
##
## $educ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   13.00   13.31   16.00   17.00
##
## $earnings
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    7280   26000   32965   45001  700021
##
## $hours
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1568    2070    1880    2480    5307
##
## $married
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.8237  1.0000  1.0000
##
## $hourly_wage
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.922  12.400  15.748  20.545 250.000
##
## $employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.8582  1.0000  1.0000
##
## $labor_exprnc
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00   14.00   23.00   25.46   33.00   83.00
```

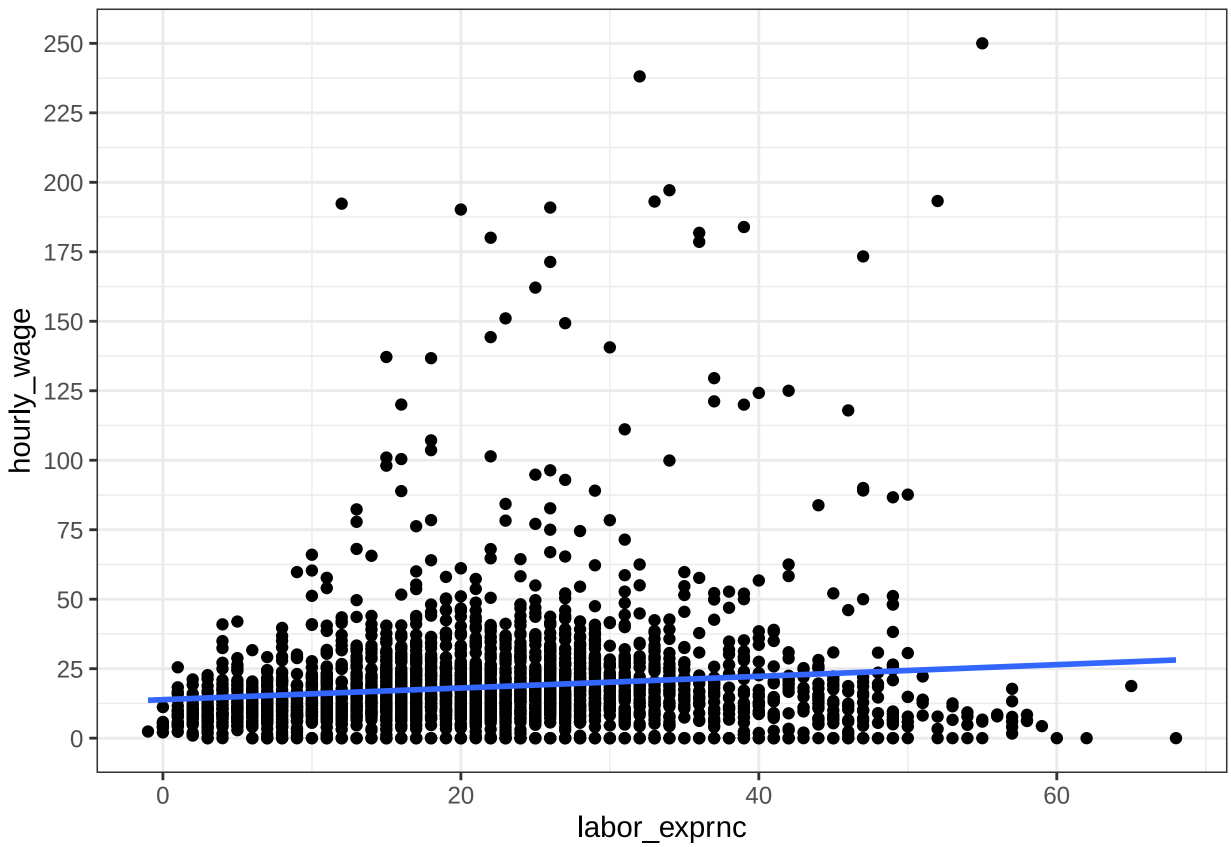
Question 1

Limit analysis only to households where the head is working

```r
dataWorking <- dataNew %>% filter(employed == 1)
```

Parts A and B

```r
ggplot(dataWorking, aes(x = labor_exprnc, y = hourly_wage)) +
    geom_point() +
    geom_smooth(method = "lm", se = F, formula = y ~ x) +
    scale_y_continuous(breaks = seq(0, 250, 25)) +
    theme_bw()
```



Correlation Coefficient: 0.1212883

The sample correlation is quite weak. Both the *eyeball test* as well as the numerically calculated correlation indicate there isn't much of an association between the two variables. This is counterintuitive because one would think the more labor market experience a person had the higher their wages would be. There are a vast number of explanations which might explain why the two aren't related:

- the relationship could be mediated by demographics (race, gender, age).
- it might be the case one needs *relevant* work experience thus not all years of work experience are created equal.
- etc.

Part C

```
olsLaborE <- lm(hourly_wage ~ labor_exprnc, data = dataWorking)
summary(olsLaborE)
```

```
##
## Call:
## lm(formula = hourly_wage ~ labor_exprnc, data = dataWorking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.128  -8.986  -3.548   4.229 224.602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.84929    0.78090  17.735  < 2e-16 ***
## labor_exprnc  0.20998    0.03202   6.557 6.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.01 on 2880 degrees of freedom
## Multiple R-squared:  0.01471,    Adjusted R-squared:  0.01437
## F-statistic:    43 on 1 and 2880 DF,  p-value: 6.465e-11
```

```
anova(olsLaborE)
```

```
## Analysis of Variance Table
##
## Response: hourly_wage
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## labor_exprnc    1   17214 17213.8      43 6.465e-11 ***
## Residuals    2880 1152928   400.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ssrrLaborE <- anova(olsLaborE)["Sum Sq"]["Residual",]
```

The estimate for β_1 is approximately 0.20998. It is significantly different from 0 at 0.05 significance level. This means in this sample for each year of work experience accumulated you would, on average, earn about 21 cents more per hour.

Part D

The F-statistic for the model is 43 (with 1 and 2880 degrees of freedom). The critical value one should use to reject the null hypothesis at $\alpha = 0.05$ level of significance would be 3.8446897. Thus we can reject the null hypothesis.

Part E

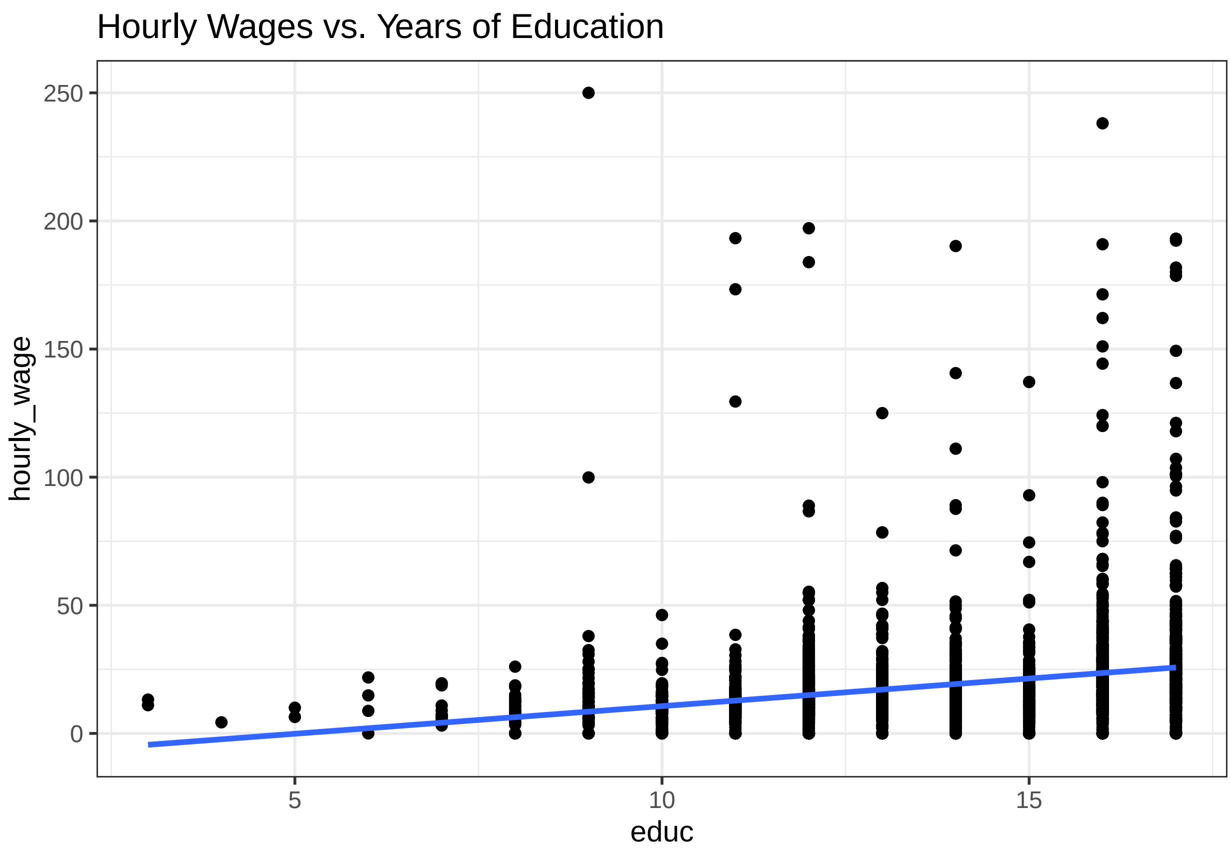
t-statistic: 6.5574254
t-statistic (squared): 42.999828
F-statistic: 42.999828

The F-statistic is equal to the t-statistic squared.

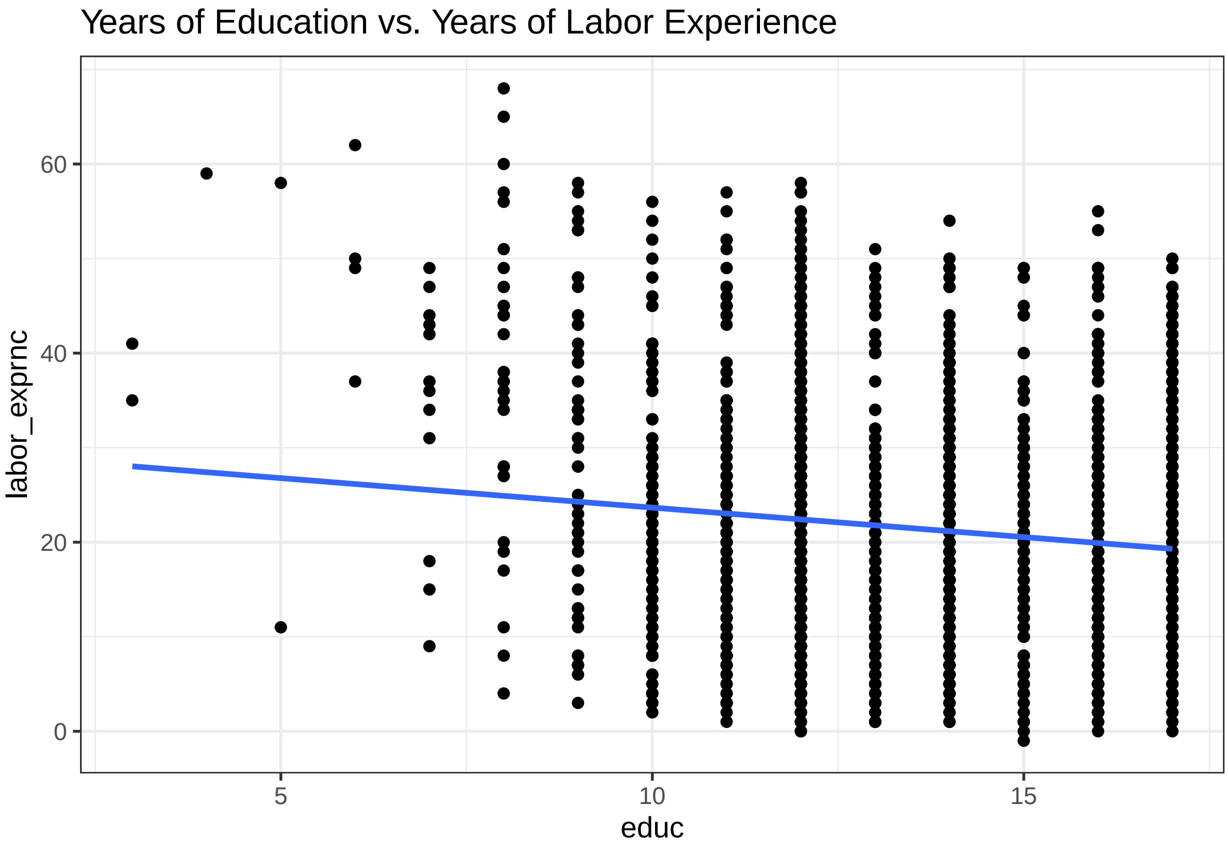
Question 2

Part A

```
ggplot(dataWorking, aes(x = educ, y = hourly_wage)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x, se = F) +
    labs(title = "Hourly Wages vs. Years of Education") +
    theme_bw()
```



```r
ggplot(dataWorking, aes(x = educ, y = labor_exprnc)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x, se = F) +
    labs(title = "Years of Education vs. Years of Labor Experience") +
    theme_bw()
```



Correlation Coefficient Between Education and Wages: 0.2513287
Correlation Coefficient Between Education and Work Experience: -0.1258688

The sample correlation being weakly negative for years of education and years of labor market experience makes sense I think. You would expect individuals with more years of education, all else being equal, would have less years in the labor market. There are many other variables though which come to bear on this relationship (not everything else is equal) hence the relationship isn't stronger.

Part B

```
olsEducLaborE <-lm(hourly_wage ~ labor_exprnc + educ, data = dataWorking)
summary(olsEducLaborE)
```

```
##
## Call:
## lm(formula = hourly_wage ~ labor_exprnc + educ, data = dataWorking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.720  -7.665  -2.027   3.879 233.253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -18.95945    2.30054  -8.241 2.56e-16 ***
## labor_exprnc   0.26901    0.03108   8.656  < 2e-16 ***
## educ           2.32344    0.15397  15.090  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.26 on 2879 degrees of freedom
## Multiple R-squared:  0.08693,    Adjusted R-squared:  0.08629
## F-statistic:   137 on 2 and 2879 DF,  p-value: < 2.2e-16
```

```
anova(olsEducLaborE)
```

```
## Analysis of Variance Table
##
## Response: hourly_wage
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## labor_exprnc    1   17214   17214  46.385 1.178e-11 ***
## educ            1   84504   84504 227.707 < 2.2e-16 ***
## Residuals    2879 1068424     371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ssruEducLaborE <- anova(olsEducLaborE)["Sum Sq"]["Residual",]
```

The estimate for β_1 is approximately 0.26901. It is significantly different from 0 at the $\alpha = 0.05$ level of significance. In the first regression, education was implicitly in the error term. Because it is negatively correlated with labor market experience, it was slightly downvoting the coefficient for labor market experience. I.e. the effect of labor market experience was being slightly confounded by education. Once the education variable was made explicit, its negative effect on labor market experience went away and the coefficient increased reflecting a more accurate estimate. This means I do believe the estimate for β_1 is a better estimate in regression 2 versus regression 1.

Part C

$\alpha = 0.05$
degrees of freedom are 2879
$H_o : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
t-statistic of rejection: 1.6453831
t-statistic for education from regression 2: 15.089968

We can therefore reject the null hypothesis and state the coefficient on education is statistically significantly different from 0.

```
fstatEduc <- ((ssrrLaborE - ssruEducLaborE) / 1) / (ssruEducLaborE / (nrow(dataWorking) - 3))
```

$\alpha = 0.05$
$H_o : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
Degrees of freedom for the numerator: 2
Degrees of freedom for the denominator: 2879
Sum of Squared Residuals from Restricted Model (Model #1): 1.1529284×10^6
Sum of Squared Residuals from Unrestricted Model (Model #2): 1.0684241×10^6
F-statistic of rejection: 2.9988516
F-statistic: 227.7071333
t-statistic squared: 227.7071333

We can therefore reject the null hypothesis and state that adding in the education variable statistically significantly increases the amount of variation explained in the dependent variable by the regression model.

The F-test and t-test agree with each other, and the F-statistic is the square of the t-statistic.

Part D

$\alpha = 0.05$
$H_o : \beta_1 = \beta_2 = 0$
$H_a : \beta_1 \neq 0, \beta_2 \neq 0$
Degrees of freedom for the numerator: 2
Degrees of freedom for the denominator: 2879
F-statistic of rejection: 2.9988516
F-statistic: 137.0459056

We can therefore reject the null hypothesis and state that at least one of the estimators explains a statistically significant amount of the variation in the dependent variable.

Part E

```r
linearComboHypo <- linearHypothesis(olsEducLaborE, "labor_exprnc + educ = 2")
```

$\alpha = 0.05$
$H_o : \beta_1 + \beta_2 = 2$
$H_o : \beta_1 + \beta_2 \neq 2$
Degrees of freedom for the numerator: 2
Degrees of freedom for the denominator: 2879
F-statistic of rejection: 2.9988516
F-statistic: 13.5635331

We can reject the null hypothesis and state that at the $\alpha = 0.05$ confidence level we are confident the linear combination of the coefficients does not equal 2.