# mappe2_SOK_2009

### What we will be looking at:

Data is taken from [principlesofeconometrics](https://www.principlesofeconometrics.com/poe5/data/def/mroz.d

We will be looking at these variables from the data set, seeing if anyone stands out:

hfathereduc husband's father's education level

hmothereduc husband's mothers's education level

hhours Husband's hours worked in 1975

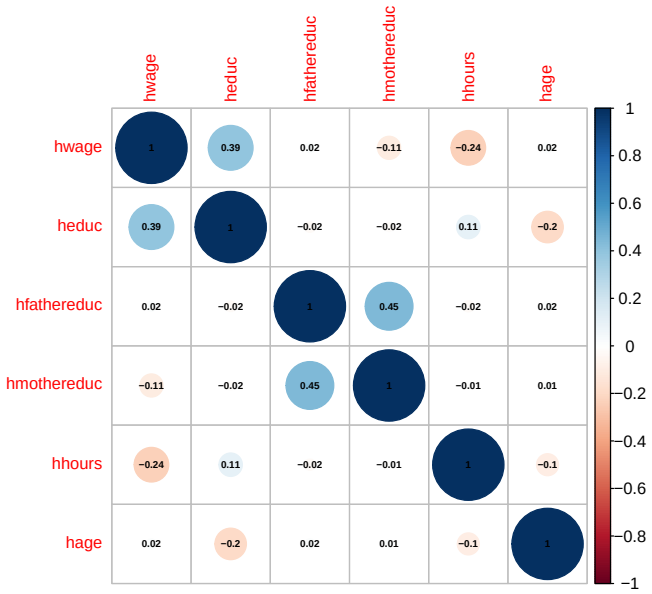
hage Husband's age

heduc Husband's educational attainment, in years

hwage Husband's wage, in 1975 dollars

The goal with these variable will determine the husbands wage (hwage), and see what influences it. We will first be looking at the thing I will assume has the largest influence, education (heduc). And that is a pretty straightforward assumption that people with higher education normally seeks higher wages, and we will be using a simple linear regression on that with the heduc as the target variable, and heduc as independent. After that we will be doing a multiple linear regression with the other variables.

But first lets create a correlation matrix to see what we can expect to see in the model. As we can see in the graph under at the first column, we see that education is the only one that has a moderate correlation to the wage. While husbands mothers education and hours worked has a weak negative correlation, and lastly husbands fathers education and husbands age as close to no correlation, so we can just as well take those out.




So based on the correlation matrix we have a few tings we expect to see if we graph the different variables up against the husbands wages. First the only one we expect to have a positive correlation is the education, and as we can see in the first graph top left, there is an upward trend as education increases so does wages. While for hours worked and mothers education we expect a weak negative correlation, and as we can see as hours goes up, wages goes down. And with mothers education we can see the same, but it is extremely messy and not really that useful, as we expected with a correlation of -0.11.
Lastly I added in the husbands age, and as we can see it is more or less flat and is not a strong indicater of wages.


## Linear regression:

To create a simple linear regression line we will be using:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where

$y$ = target variable

$\beta_0$ = y-intersept

$\beta_1$ = slop relationship

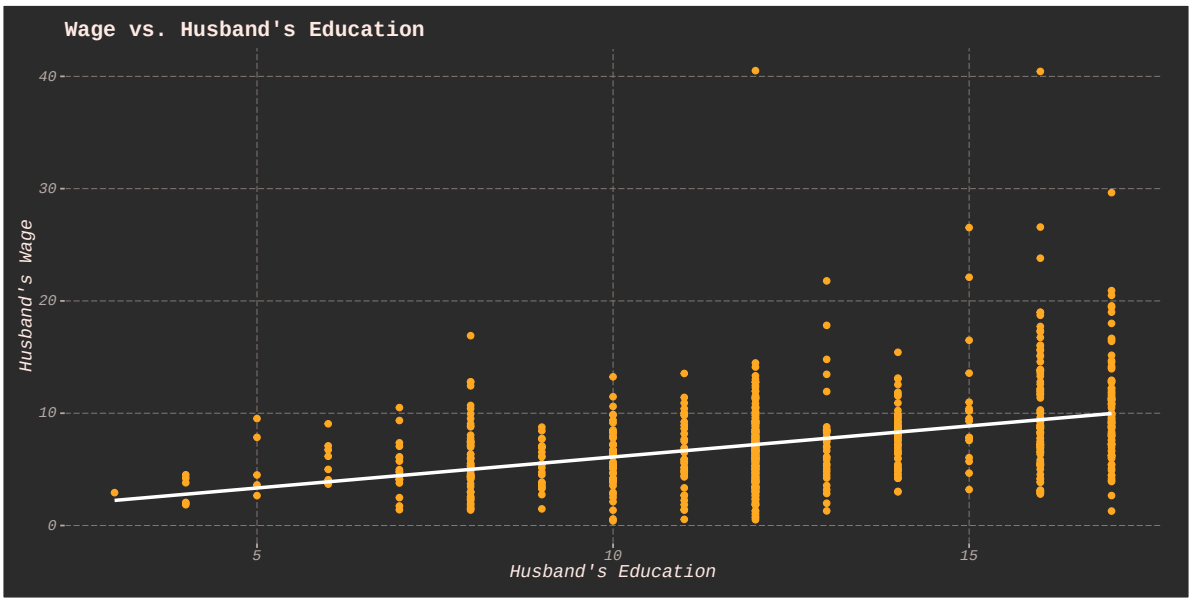
$x$ = value for feature

$\epsilon$ = Error rate, distance from expected target posision

Or since we are using code, we will be using this:

```
model_1 <- lm(hwage ~ heduc, data = df_mroz_f)
```

And then based on that we will create a liner regression line through the data. As we can see this is the same scatter plot graph we saw earlier called "Wage vs. Husband's Education", just that here we have a least squares linear regression going through. And this clearly shows the upward trend we have talked about earlier, as education increases, wages goes up.




We can also look more in depth on what the regression model give us.

```
summary(model_1)
```

```
Call:
lm(formula = hwage ~ heduc, data = df_mroz_f)

Residuals:
   Min     1Q Median     3Q    Max
-8.688 -2.317 -0.431  1.531 33.298

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57798    0.60343   0.958    0.338
heduc        0.55272    0.04696  11.771   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.89 on 751 degrees of freedom
Multiple R-squared:  0.1558,    Adjusted R-squared:  0.1546
F-statistic: 138.6 on 1 and 751 DF,  p-value: < 2.2e-16
```

One of the first things we can see is here:

```
heduc        0.55272    0.04696  11.771   <2e-16 ***
```

The three asterisks (***) indicate that the variable heduc is statistically significant in predicting the husband's wage. This is further supported by the extremely low p-value (<2e-16), suggesting that the relationship between education and wage is not due to random chance.

Another important statistic is the R-squared value:

```
Multiple R-squared:  0.1558
Adjusted R-squared:  0.1546
```

An R-squared value of 0.1558 means that approximately 15.58% of the variability in the husband's wage can be explained by his level of education according to our model. And when adjusted its still quite close, on 15.46%. While this indicates that education is a factor in determining wages, it also suggests that there are other variables not included in this simple model that contribute to the variation in wages.

Lastly we can look at the F-statistic:

```
F-statistic: 138.6 on 1 and 751 DF,  p-value: < 2.2e-16
```

And here we can see the F-statistic value of 138.6, quite large, especially for a simple regression with only one predictor. This high value indicates that the model variance is significantly greater than the error variance. In other words, the model explains a substantial amount of the total variance in the dependent variable (hwage)

The "on 1 and 751 DF" refers to the degrees of freedom used for the F-test. The first number (1) is the number of independent variables in the model (just heduc in this case), and the second number (751) is the degrees of freedom of the residuals (the number of observations minus the number of estimated coefficients).

We can also check if the p-value is less then an alpha variable, normally 0.05. So by doing that we get 2.2e-16 $<$ 0.05, which is significant less then 0.05, and we can be certain or model is better than just taking the mean value.

### Multiple Linear regression:

Now we will move into multiple linear regression, and similar to the simple regression formula, this one is almost the same, just more $\beta_n x_n$ values. So the old formula was:

Simple Regression:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Multiple Linear Regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

The problem we will have with multiple regression is to visualize it, as each new variable is like adding a new dimension to the graph. Graphic the 1 dimensional data points we had with a 2 dimensional graph worked fine, and we can do a multiple regression line with two variables, stretching out the line we had into a plain, and graphing it 3 dimensional. But anything more then that can not easily be graphed, so to work with higher dimensions we normally does something called partial regression modeling, where we visualy show each independent variable separate while holding the other independent variables constant. But we can still do the calculations for a full multiple linear regression to get the data we are looking for.

So because of that we will simplify things and only add one new variable in and as we saw in the correlation matrix. There was really just two variables that was moderate to weakly correlated with wages, and that was education and hours worked, and we have already looked at education, so now we will add in hours worked as well.

So we will be using:

hhours Husband's hours worked in 1975

heduc Husband's educational attainment, in years

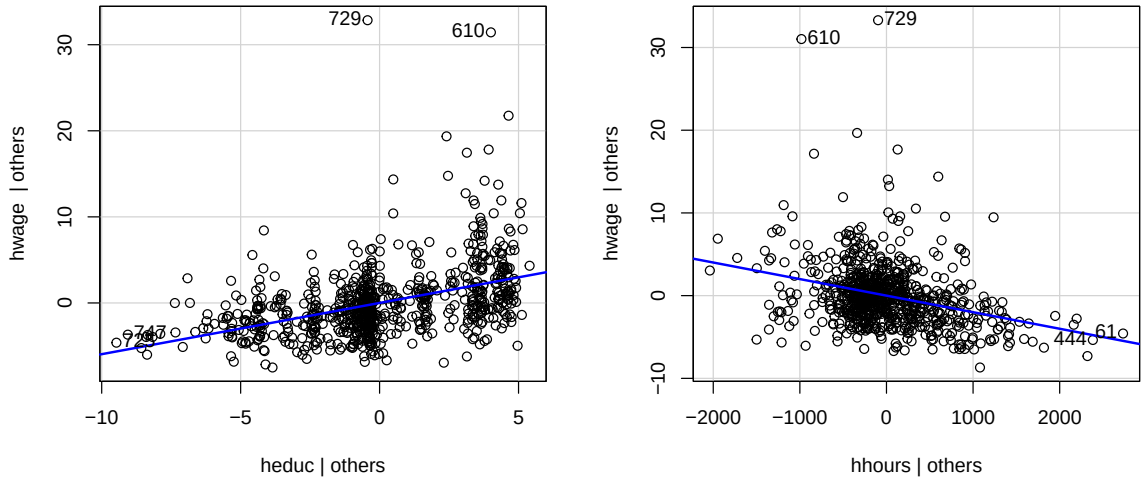
hwage Husband's wage, in 1975 dollars

With this we will see how education and hours affect wages.

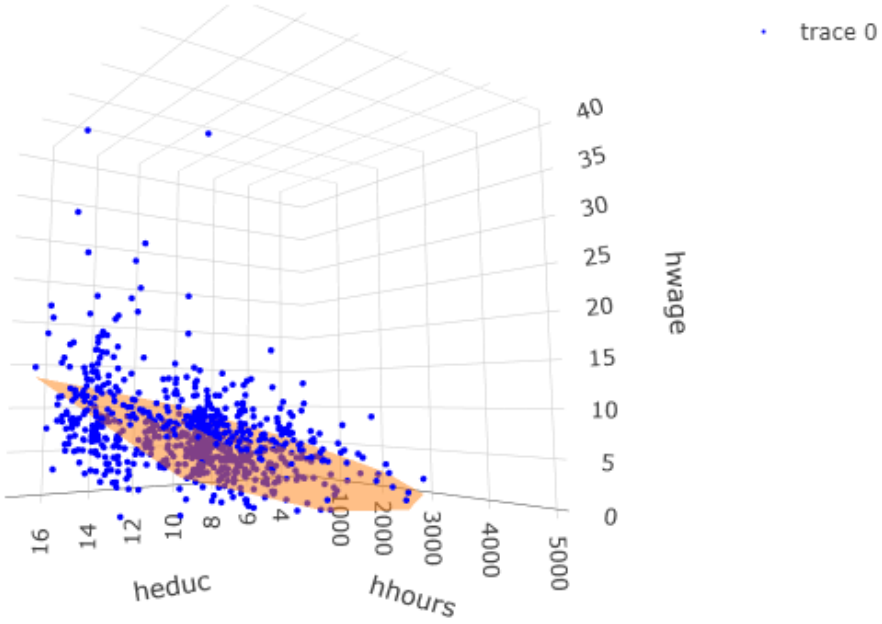
Similar to simple liner regression, here we will use this code for multiple:

```
model_2 <- lm(hwage ~ heduc+hhours, data = df_mroz_f)

#summary(model_2)
```




This is what a partial regression modeling would look like. So as we expected, education (heduc) goes up similar to when we did the simple linear regression. Hours worked data as we saw earlier has a negative correlation, so a downward trend. But for fun since we only have two variables we can also look at a 3d graph.



Here in the 3d graph we can see how the plane tilts downwards the closer we get to 0 hours worked and low education, while the best place seems to be high education with many hours.

Lets go over the data we got from this regression.

```
summary(model_2)
```

```
Call:
lm(formula = hwage ~ heduc + hhours, data = df_mroz_f)

Residuals:
   Min     1Q Median     3Q    Max
-8.318 -2.185 -0.308  1.462 33.105

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5856955  0.7345307   6.243 7.18e-10 ***
heduc        0.5952885  0.0450111  13.225  < 2e-16 ***
hhours      -0.0020022  0.0002283  -8.770  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.707 on 750 degrees of freedom
Multiple R-squared:  0.2343,    Adjusted R-squared:  0.2322
F-statistic: 114.7 on 2 and 750 DF,  p-value: < 2.2e-16
```

The first we see is that there are now multiple values in the coefficient area:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5856955  0.7345307   6.243 7.18e-10 ***
heduc        0.5952885  0.0450111  13.225  < 2e-16 ***
hhours      -0.0020022  0.0002283  -8.770  < 2e-16 ***
```

Here we can see that all of them has three stars (***), and as we said that shows significance. We can see that the P-value is still under 0.05, and we can assume that the model has benefited by adding in the new data.

To double check that, lets see how much of the wage can be explained by this model by looking at the R-squared

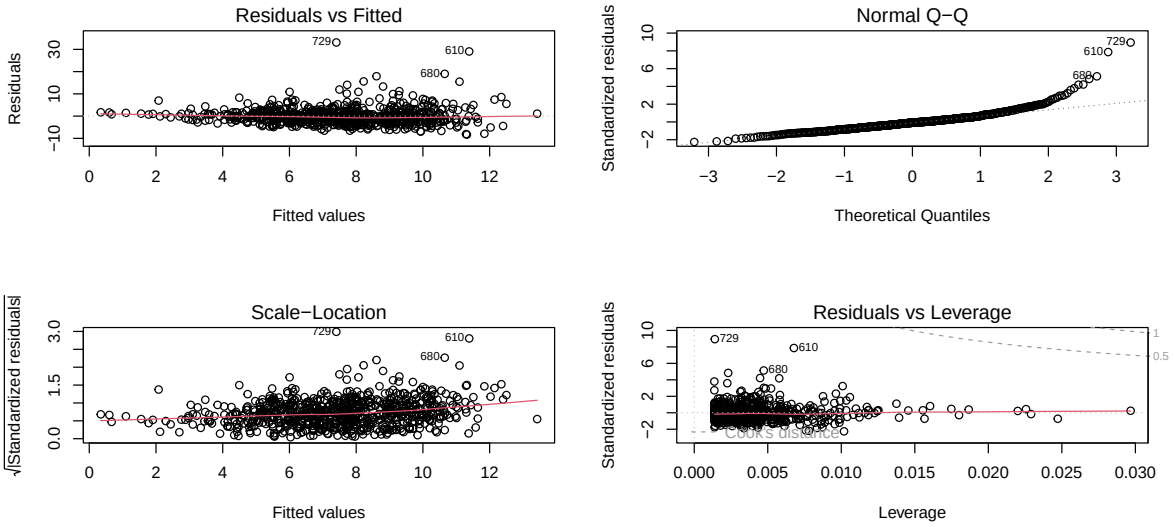
```
Multiple R-squared:  0.2343,    Adjusted R-squared:  0.2322
```

As we see the model went from 15.58% and 15.46% when adjusted, to 23.43% and 23.22% when adjusted. So the model has defiantly become better at describing the wage.

As well we should check the multicollinearity, normally more relevant when we have multiple varietals to check, but it can't hurt to check it.

```
   heduc   hhours
1.011767 1.011767
```

And here we can see that both are close to one so multicollinearity should not a problem in this model.




We can also look at these graphs to check for things like homogeneous and heterogeneous in the data. And just by looking at the first graph "Residual vs Fitted" we can see that the data is more spread out the farther from 0 we are, which is not unexpected since we found out that our model only explain around 23% of the wage.

As well in the Normal Q-Q graph we can see two things, we can see the distribution and the extreme values. For distribution it seems fine most data is between 2 units of 0, and this will most lightly create a expected bell curve. Tho when it comes to extreme values we can see that most of the data follows the expected line, until we hit 2 units. After that our models would have a hard time to accurately predict the wages based on education and hours worked,

arguably not surprising considering highly educated and a hard worker will naturally end up at the extreme. Based on the graph we would under predict.

When we look at the third graph "Scale-Location" we can more or less confirm what we observed in the two previous graphs.

And lastly we can look at the Residuals vs Leverage graph, and see if we have any data point with a large amount of leverage. And here in this data it does not seem like we have any data point worth removing, as non is crossing over the cook's line.

So in the end our model is not optimal, and could benefit from a few extra values for optimization. But it is still better then taking the mean based on the data, as it does explain roughly 23% of the wage for the husband.