Econometrics FIN - 403: Problem Set 1

Due on Thursday October 21, 2021

Girolamo Vurro, Filippo Rondoni

Contents

Problem 1

Point a

We notice that the datataset presents many missing values N.A. . These were all located into the wages column (535), corresponding to unemployed people.Therefore we remove the observations with at least one missing value from the sample. We also noticed that the experience column contains negative value, that is a consequence of the definition of this parameter.

Point b

From the histogram we can notice that there is a larger number of observation for man with respect to woman. Most of the observation for the males are around 25 wph while form woman the value is lower. Here we summarize the statistics:

Table 1: Women's wages statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
WPH	342	17.876	9.640	5.482	12.100	21.490	87.787

Table 2: Men's wages statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
WPH	884	25.629	11.670	4.910	18.403	29.339	100.973

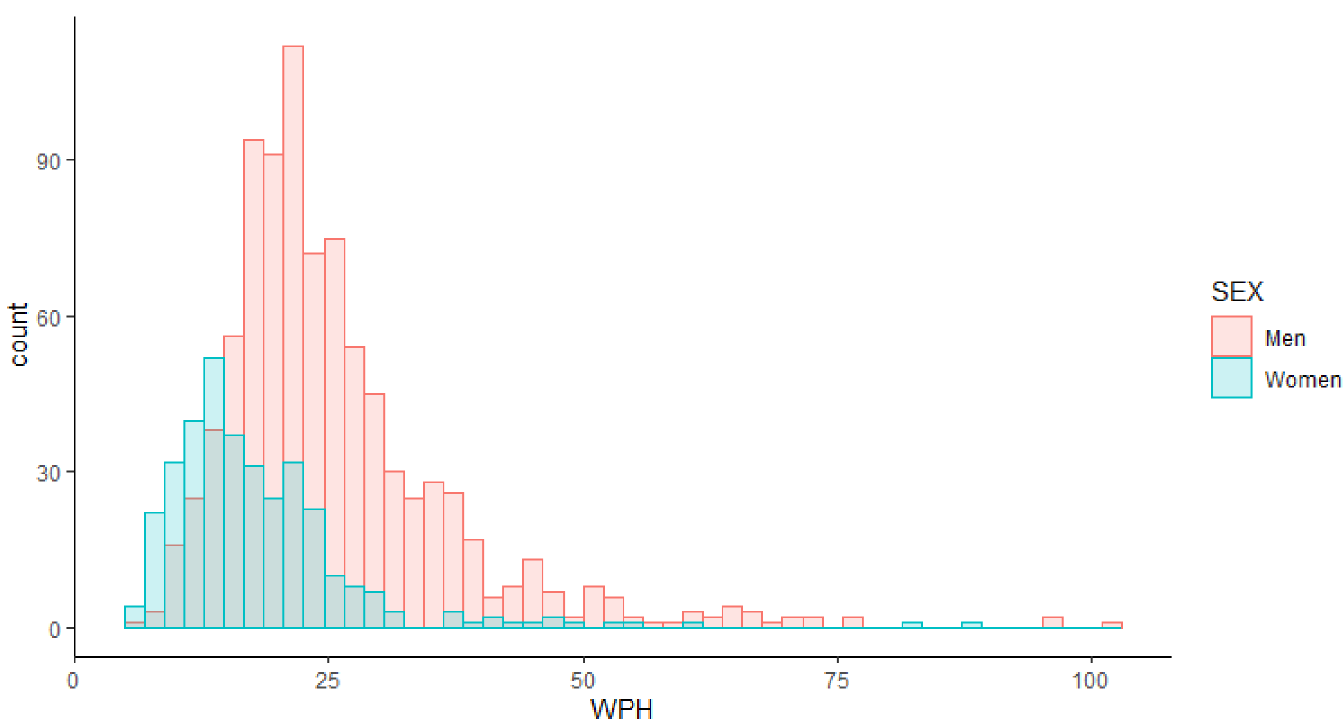

Figure 1: Histogram for men and women's wages

Problem 1 continued on next page. . .

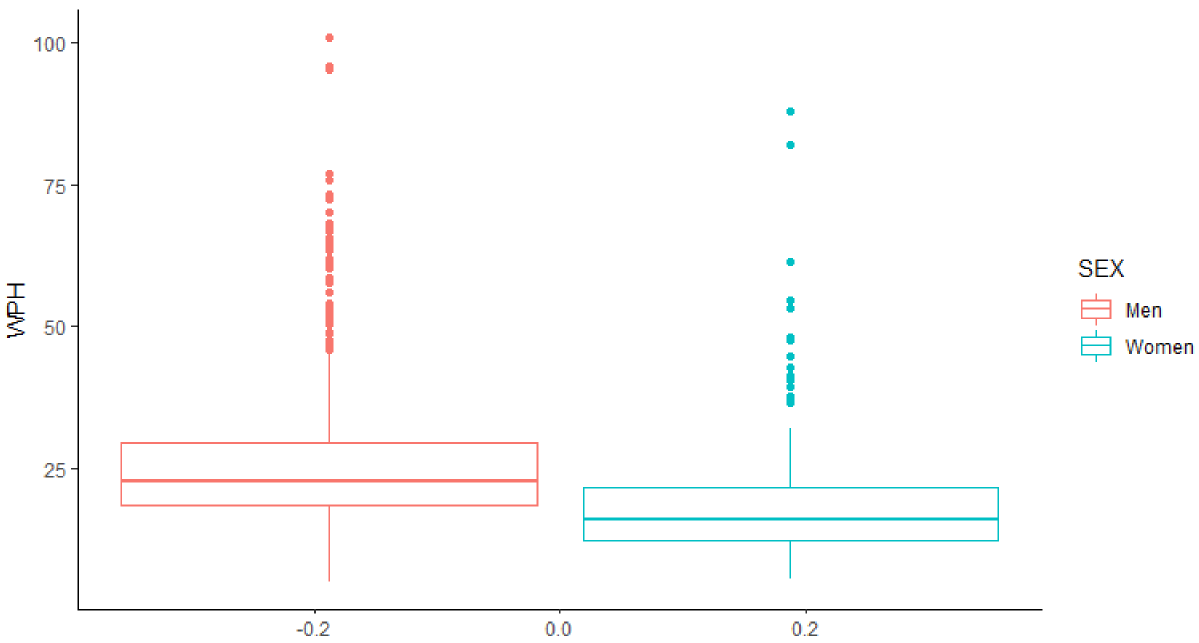

Figure 2: Boxplot for men and women's level of wph

Point c

β_2 is the *semi-elasticity*: as the independent variable EDU (which is in levels) increases by an additional unit, the expected wage increases by about β_2% in relative terms ($log(1+z) \approx z$ for z small). In our case the wage is expected to increase by about 7.7%. The other coefficients are indicated in the first column of Table (7).

Point d

EXP is statistically significant with a $p-value < 0.01$ as it can be seen from the table above. But we noticed that with growing experience your wage will increase more slowly.

Point e

$$\frac{\partial\, \mathbb{E}[WPH_i|EDU_i, EXP_i]}{\partial EDU_i} = \beta_3 + 2 * \beta_4 EXP_i \tag{1}$$

FOC

$$\beta_3 + 2 * \beta_4 EXP_i = 0$$
$$EXP_i = -\frac{\beta_3}{2 * \beta_4}$$

The second derivative is:

$$\frac{\partial^2\, \mathbb{E}[WPH_i|EDU_i, EXP_i]}{\partial^2 EDU_i} = 2 * \beta_4$$

Which will be strictly negative as $\hat{\beta}_4$ is negative.

Knowing $\hat{\beta}_3 = 0.036$ and $\hat{\beta}_4 = -0.001$ we can estimate our optimal level of experience:

$$EXP_i^* = -\frac{0.036}{2 * (0.001)} = 32.9$$

Point f

The predicted wage for MFE graduate with 17 years if education and 1 year of work experience:

$$E[WPH|EDU, EXP] = e^{log(\hat{W}PH)+0.5s^2} = 22.9 \text{ CHF per hour.}$$

Point g

i) We can observe that the variables *exp* and exp^2 have a high VIF value as expected as they are highly correlated.

Table 3: VIF

edu	exp	I(exp^2)
1.197	13.878	13.857

Deleting exp^2 from the regression, we observe two low VIF values.

Table 4: VIF

edu	exp
1.193	1.193

We checked that the R_k are high and therefore also the VIF relative to those variables will be high as the denominator of its formula is very small.

If wer regress exp^2 on the other independent variables we note that the $R_4^2 = 0.9278$ is very high and this will make the denominator of the formula utilized to calculate the VIF very small. Therefore we expect β_4 to have a high value of VIF. Same if we regress *exp* on the other independent variables, high $R_3^2 = 0.9279$ will have the same implication as the one above. So both the high VIF values for β_3 and β_4 are explained.

ii) We can visualize the positions of the potential outliers from the from figures (3) and (4). But for each variable using the function $|DFBETA| > 1$ we note that considering as critical value 1 there are no outliers.

iii) Then we compute the RESET test. That is usefull to inspect for potential functional form of misspecification. Since we get a p-value that is smaller than 0.05 we can reject the null- hypothesis, so it is likely that the functional form is misspecified or there are omitted variables. ($p - value = 0.0009234$)

Point h

Given that EXP is a linear combination of AGE and EDU, it turns that AGE does not add additional information and therefore the matrix of independent variables has $rank < K$. EXP contains the information of AGE causing collinearity. R omits the added variable AGE and makes the regression as the previous one.

Point i

We can see from table (7), in the regression (2) that the variable sex's p-value is really low from the summary in R. The coefficient of sex is statistically significant at the 99 % level, suggesting that there is indeed a difference in hourly wage by sex. Being a woman reduces the wages by approximately 28.5% the wage.

Point j

We apply the Chow test to verify the joint significance of the interaction terms. It is shown that they are jointly significant at the 95 % - and the 99%. Therefore we reject the null hypothesis (no structural break/ no group differences) as the $F = 52.522$ and the $p - value$ very close to 0.

continued on next page...

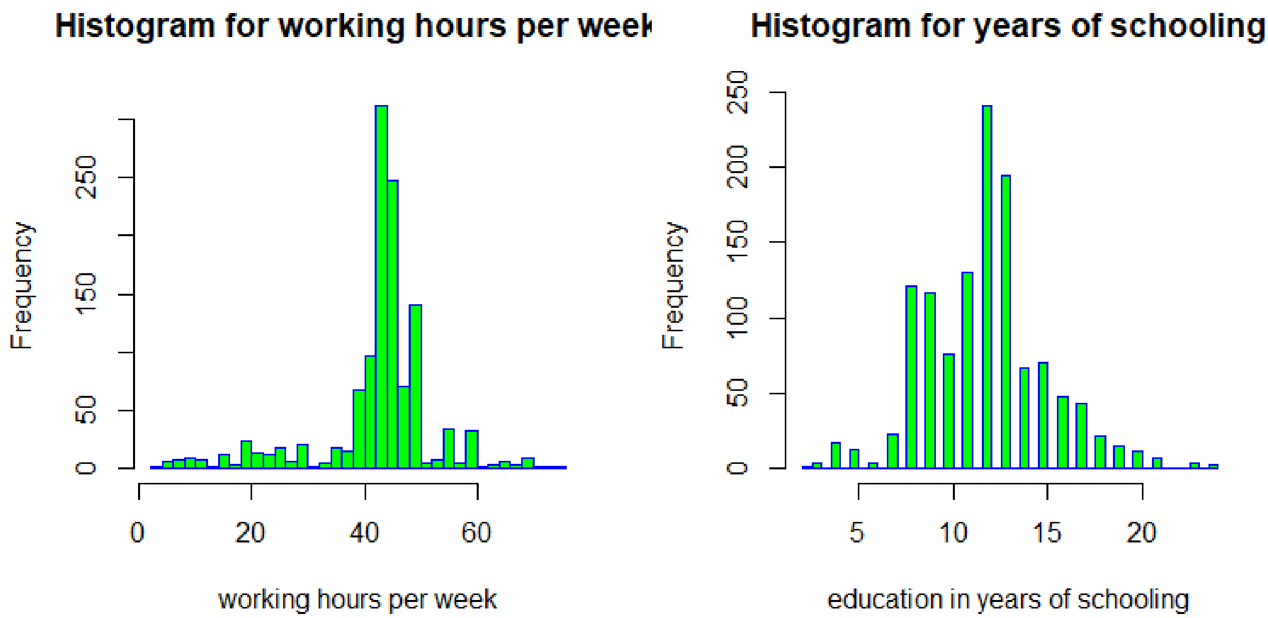

Figure 3: Histogram for the visualization of potential outliers in the independent variables

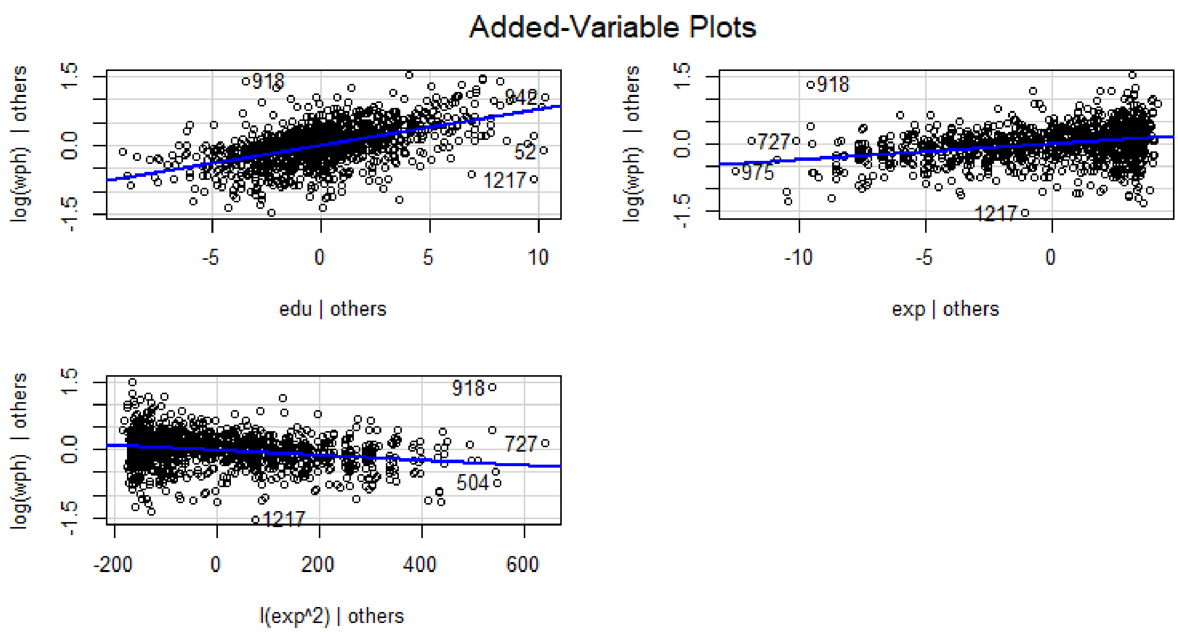

Figure 4: A-V plots

continued on next page...

Bonus question

We adopted the following procedure:

1. We can remove the iteration between edu and sex as it is not statistically relevant.
2. We added the variable hrs in the regression and we noticed that Adj. R^2, AIC and BIC better than the previous model
3. Adding the variables hrs and hi to the specification we obtain again an improvement for Adj. R^2, AIC and BIC if compared to the point 2.
4. Adding the iteration of kt and sex, as well as hrs, hi and kt we have a further improvement of Adj. R^2, AIC and BIC.

We obtain the following regression (Table 6) and AIC and BIC values:

Table 5:

AIC	BIC
774.197	835.535

continued on next page...

Table 6:

	Dependent variable:
	log(wph)
edu	0.066***
	(0.003)
exp	0.042***
	(0.004)
I(expˆ2)	−0.001***
	(0.0001)
sex	0.063
	(0.059)
I(exp ∗sex)	−0.018***
	(0.007)
I(expˆ2 ∗sex)	0.0002
	(0.0002)
hrs	−0.004***
	(0.001)
hi	0.004***
	(0.001)
kt	0.038***
	(0.011)
I(kt ∗sex)	−0.130***
	(0.024)
Constant	1.680***
	(0.112)
Observations	1,226
R^2	0.447
Adjusted R^2	0.442
Residual Std. Error	0.330 (df = 1215)
F Statistic	98.087*** (df = 10; 1215)
Note:	*p<0.1; **p<0.05; ***p<0.01

continued on next page. . .

Table 7: Regressions

	Dependent variable:		
	log(wph)		
	(1)	(2)	(3)
edu	0.077***	0.068***	0.067***
	(0.004)	(0.003)	(0.004)
exp	0.036***	0.035***	0.044***
	(0.003)	(0.003)	(0.003)
I(expˆ2)	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.0001)
sex		−0.285***	−0.005
		(0.022)	(0.123)
I(edu ∗sex)			0.004
			(0.008)
I(exp ∗sex)			−0.028***
			(0.006)
I(expˆ2 ∗sex)			0.0004***
			(0.0001)
Constant	1.713***	1.931***	1.842***
	(0.057)	(0.056)	(0.063)
Observations	1,226	1,226	1,226
R^2	0.322	0.401	0.422
Adjusted R^2	0.320	0.399	0.418
Residual Std. Error	0.364 (df = 1222)	0.343 (df = 1221)	0.337 (df = 1218)
F Statistic	193.316*** (df = 3; 1222)	204.380*** (df = 4; 1221)	126.834*** (df = 7; 1218)

Note: *p<0.1; **p<0.05; ***p<0.01