Introductory Econometrics

Solutions Problem Set 1: Descriptive Statistics and Simple Linear Regression

Brief Solutions

The solution is aimed to help you understand where to find the numbers. There is no need for you to copy and paste the whole table when it comes to the project report. Only the result is good enough.

EX1

1 enf18-enf6

2

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
enf3		4503	0.1154786	0.3345734	0	2.0000000
enf6		4503	0.2744837	0.5657047	0	4.0000000
enf18		4503	0.9089496	1.0513968	0	6.0000000
enf6_18		4503	0.6344659	0.9021829	0	6.0000000
age	agd	4503	40.4394848	8.5917741	25.0000000	55.0000000
educ	educ	4503	18.2191872	2.6984810	6.0000000	26.0000000
dip	dip	4503	16.5689540	4.9017975	6.0000000	26.0000000
ltsal	ltsal	4503	2.3743936	0.3115555	1.7272209	3.4812400

s				
s	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2179	48.39	2179	48.39
2	2324	51.61	4503	100.00

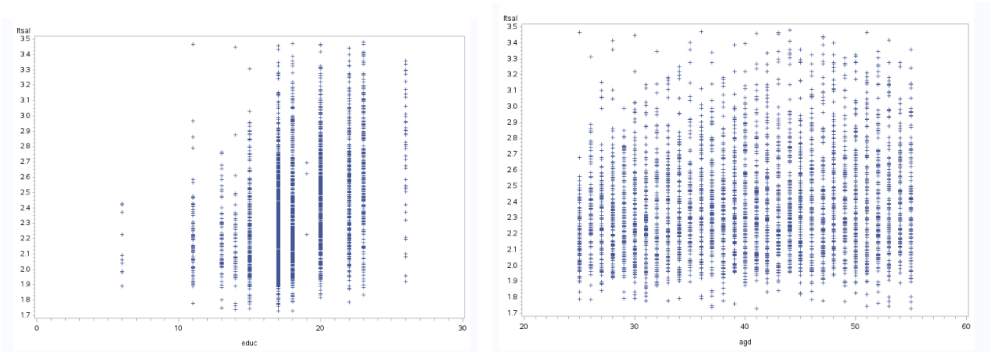
tech				
tech	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1879	41.73	1879	41.73
1	2624	58.27	4503	100.00

3

Pearson Correlation Coefficients, N = 4503 Prob > |r| under H0: Rho=0	
	ltsal
educ educ	0.36788 <.0001
age agd	0.12740 <.0001

Correlation between ltsal and educ is significant at 1%. Correlation between ltsal and age is significant at 1%. Reason: p-value is smaller than 0.01% (thus 1%).

4



The correlations are too weak to see any clear relationship in this graph.

EX2

1

Parameter Estimates

Variable	Label	DF	Parameter Estimate	St	t Value	Pr > |t|
Intercept	Intercept	1	1.60054	0.0	4.31	<.0001
educ	educ	1	0.04247	0.0	6.54	<.0001

$\widehat{\beta}_1$ $\widehat{\beta}_2$

Interpretation of $\widehat{\beta_2}$: one additional year of education is associated with a wage increase of 4,2%.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	1	59.14245	59.14245		<.0001
Error	4501	377.85228	0.08395		
Corrected Total	4502	436.99474			

$\widehat{\sigma^2}$

2

Analysis of Variance

Source	DF	Sum of Squares		F Value	Pr > F
Model	1	59.14245	59.	704.51	<.0001
Error	4501	377.85228	0.0		
Corrected Total	4502	436.99474			

ESS RSS TSS

TSS=RSS+ESS

3

Root MSE	0.28974	**R-Square**	0.1353
Dependent Mean	2.37439	**Adj R-Sq**	0.1351
Coeff Var	12.20264		

R2

Interpretation:

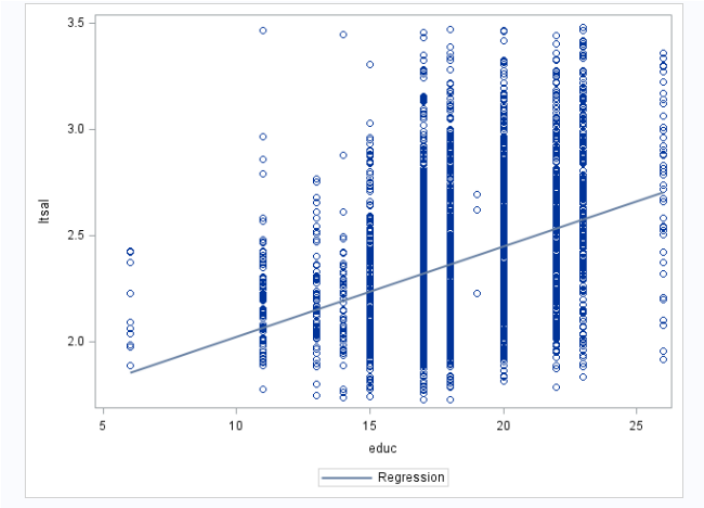
It measures the part of the log-wage variation that is explained by the variation in the level of education: here, 13,5% of the variation in the log hourly wage is explained by the model, i.e., by education.

$R^2 = 0.1353$

$\rho = 0.36788$

$0.36788^2 = 0.1353$

4



Review of concepts

1 Pearson's Empirical correlation

1.1 Coefficient

Theoretical correlation $\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$ is estimated by the Pearson's empirical correlation coefficient $r_{xy} = \frac{\widehat{Cov}(x,y)}{\widehat{\sigma_x}\widehat{\sigma_y}}$

with $\widehat{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ and $\widehat{\sigma_x}^2 = \widehat{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

The empirical correlation coefficient belongs to $[-1;1]$.

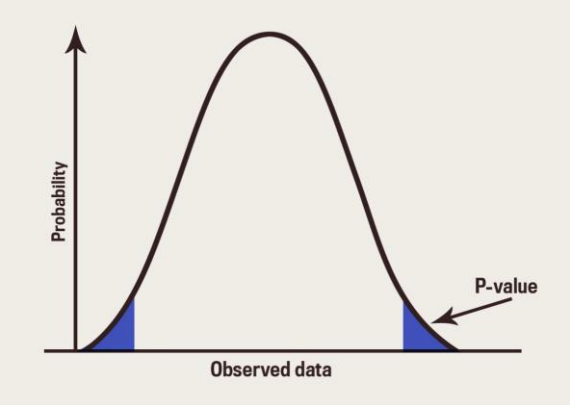
The closer the correlation coefficient is to the extreme values -1 and 1, the stronger is the correlation between the variables. A correlation equal to 0 means that the variables are not correlated (which does not mean that they are independent!)

1.2 Significance

$H_0 : \rho = 0$ (Absence of a linear relationship between the two variables)

Under H_0, $\widehat{\rho} \sim t(n-2)$ Student distribution with (n-2) degrees of freedom (df) (draw a graph to demonstrate the decision rule)

Decision rule: If the p-value $\leq \alpha$: you reject H_0, if not, you do not reject H_0



Blue region: rejected

p-value small: reject

2 Simple Linear Model

2.1 coefficient (beta's)

Interpretation: one additional amount of increase in x is associated with

1) For level y-variable: a $\widehat{\beta}$ amount of increase/decrease in y

2) For ln(y): a $100\,\widehat{\beta}\,\%$ percentage change (or growth rate) in y

2.2 significance

p-value small/ t-value big, significant

2.3 TSS=RSS+ESS

Analysis of the variance

	Sum of Squares	ddl	Average Sum of Squares	Fisher
regression	ESS	k	$\frac{ESS}{k}$	$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-(k+1)}}$
residual	RSS	$n-(k+1)$	$\frac{RSSR}{n-(k+1)} = \widehat{\sigma}^2$	
total	TSS	$n-1$		

n=number of observations=4503

k=number of explanatory variables (except the constant)=1

2.4 R^2

Meaning: fitness of the model, bigger better.

$$R^2 = \frac{ESS}{TSS}$$

The closer the explained variance is to the total variance, the better is the adjustment of the cloud of points by the OLS regression line.

Interpretation: It measures the part of y variation that is explained by the variation in x.

Improve R^2: 1) add more variables 2) add non linear terms.