Introductory Econometrics

Solutions Problem Set 2: Multiple Linear Regression

Brief Solutions

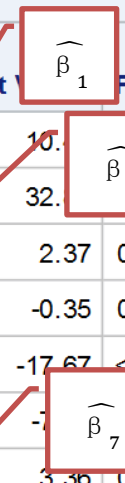
The solution is aimed to help you understand where to find the numbers. There is no need for you to copy and paste the whole table when it comes to the project report. Only the result is good enough.

EX1

1

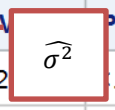
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.05884	0.10107	10.	01
educ	educ	1	0.05304	0.00161	32.	01
age	agd	1	0.01196	0.00505	2.37	0.0178
age2		1	-0.00002223	0.00006311	-0.35	0.7246
female		1	-0.14342	0.00812	-17.67	< 0001
tech	tech	1	-0.06066	0.00836	-7	001
enf18		1	0.01412	0.00421	3.36	0.0008



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	113.20063	18.86677	2	.0001
Error	4496	323.79411	0.07202		
Corrected Total	4502	436.99474			



$\widehat{\beta}_1 = 1.05$; $\widehat{\beta}_2 = 0.05$; $\widehat{\beta}_3 = 0.01$; $\widehat{\beta}_4 = -0.00002$; $\widehat{\beta}_5 = -0.14$; $\widehat{\beta}_6 = -0.06$; $\widehat{\beta}_7 = 0.014$

Interpretation:

Age: An older individual has a higher wage rate but this advantage is decreasing with age (as $\beta_4 < 0$), all else being equal .

Gender: Being a woman (female=1) leads to a 14% smaller wage rate all else being equal.

Tech: Individuals with a technical education have a smaller wage rate than others (by 6%) ceteris paribus.

2

Analysis of Variance					
Source	DF	Sum of Squares	Square	F Value	Pr > F
Model	6	113.20063	18.86	61.97	<.0001
Error	4496	323.79411	0.07		
Corrected Total	4502	436.99474			

ESS

RSS

TSS

	Sum of squares	df	Sum of mean squares
regression	$ESS = 113.2$	$k = 6$	$\frac{ESS}{k} = 18.86$
residuals	$RSS = 323.79$	$n-(k+1) = 4496$	$\frac{RSS}{n-(k+1)} = \widehat{\sigma}^2 = 0.072$
total	$TSS = 436.99$	$n-1 = 4502$	

TSS=RSS+ESS

3

Root MSE	0.26836	R-Square	0.2590
Dependent Mean	2.37439	Adj R-Sq	0.2581
Coeff Var	11.30235		

R^2

R^2_{adj}

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	113.20063	18.86677	261.97	<.0001
Error	4496	323.79411	0.07202		
Corrected Total	4502	436.99474			

F-test statistics

$$R^2 = \frac{\sum (y_i - \overline{y})^2}{\sum (\widehat{y}_i - \overline{y})^2} = \frac{ESS}{TSS} = 0,2590 \quad R^2_{adj} = \frac{(n-1)R^2 - (k)}{n-(k+1)} = 0,2581 \quad F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-(k+1)}} = 261.97$$

Interpretation:

25.90% of the variation in the log hourly wage rate is explained by model M6, i.e., by education, age, age square, gender, technical diploma and the number of children aged less than 18 years old.

4

We check $\sum \widehat{u}_i = 0$ (proof in Tutorials) using PROC UNIVARIATE (that gives more information than PROC MEANS)

Moments			
N	4503	Sum Weights	4503
Mean	0	Sum Observations	0
Std Deviation	0.2681833	Variance	0.07192228
Skewness	0.57114921	Kurtosis	1.07331893
Uncorrected SS	323.794105	Corrected SS	323.794105
Coeff Variation	.	Std Error Mean	0.00399651

5

$$\begin{aligned}\widehat{ltsal}_i^{(1)} &= \widehat{\beta}_1 + \widehat{\beta}_2 * 21 + \widehat{\beta}_3 * 30 + \widehat{\beta}_4 * 900 + \widehat{\beta}_5 * 1 + \widehat{\beta}_6 * 0 + \widehat{\beta}_7 * 1 \\ &= 1.05 + 0.05 * 21 + 0.01 * 30 - 0.00002 * 900 - 0.14 + 0.014 \\ &= 2.256\end{aligned}$$

$$\begin{aligned}\widehat{ltsal}_i^{(2)} &= \widehat{\beta}_1 + \widehat{\beta}_2 * 18 + \widehat{\beta}_3 * 40 + \widehat{\beta}_4 * 1600 + \widehat{\beta}_5 * 1 + \widehat{\beta}_6 * 0 + \widehat{\beta}_7 * 0 \\ &= 1.05 + 0.05 * 18 + 0.01 * 40 - 0.00002 * 1600 - 0.14 \\ &= 2.178\end{aligned}$$

$$\widehat{ltsal}_i^{(2)} - \widehat{ltsal}_i^{(1)} = 2.178 - 2.256 = -0.078$$

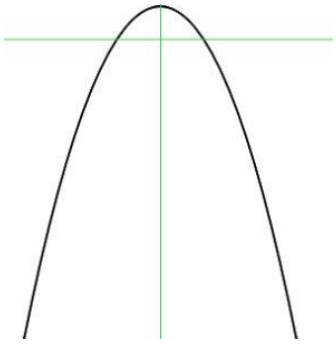
Review of concepts

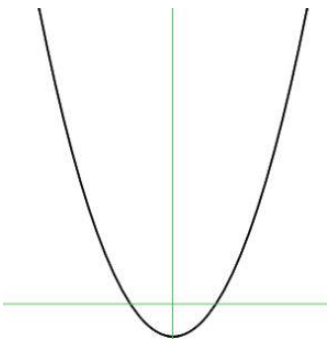
Multiple Linear Regression

1 Coefficient (beta's)

Interpretation:

1) Dummy: 0 for category A and 1 for category B. y is $\widehat{\beta}$ more (less if -$\widehat{\beta}$) for category B than category A. [e.g. female]
2) Non-linear: first increase then decrease/first decrease then increase. [e.g. age]





2 R^2 R^2_{adj} F-test statistics

R^2	R^2_{adj}	F-test statistics
$R^2 = \frac{ESS}{TSS}$	$R^2_{adj} = \frac{(n-1)R^2 - (k)}{n-(k+1)}$ $R^2_{adj} = 1 - \frac{RSS/(n-(k+1))}{TSS/(n-1)}$	$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-(k+1)}}$
fitness of the model	fitness of the model	Total significance of the model
It measures the part of y variation that is explained by the variation in x.	It measures the part of y variation that is explained by the variation in x, adjusted for the number of explanatory variables. It is useful to compare models having a different number of explanatory variables.	This statistic is the test statistic associated with testing the hypothesis H0: All the parameters are equal to zero except the intercept.
bigger better	bigger better	bigger better

Remarks:

1 H0:$\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

2 R^2_{adj} is independent of the number of explanatory variables. It is adjusted for the number of degrees of freedom.

3 Relationship between the R^2 and the Fisher statistic:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$F = \frac{ESS/(k)}{RSS/(n-(k+1))} = \frac{n-(k+1)}{k}\frac{ESS}{RSS} = \frac{n-(k+1)}{k}\frac{ESS/TSS}{RSS/TSS} = \frac{n-(k+1)}{k}\frac{R^2}{1-R^2}$$

3 Predicted values & Residuals

Predicted values: $\widehat{Y} = X\widehat{\beta}$

Residuals: $\widehat{u}_i = y_i - \widehat{y}_i$ = difference between the observed values of the variable and the predicted values using the coefficient estimates of the model.

$\sum \widehat{u}_i = 0$