STOCKHOLM UNIVERSITY
Department of Statistics

Stockholms universitet

Michael Carlson
2017-11-24

Solutions to Plenary Exercises: Plenary Exercise 6

Basic Statistics for Economists, 15 ECTS, STE101

EXERCISE 1

Let's start by estimating the mean, variance and covariance of x and y.

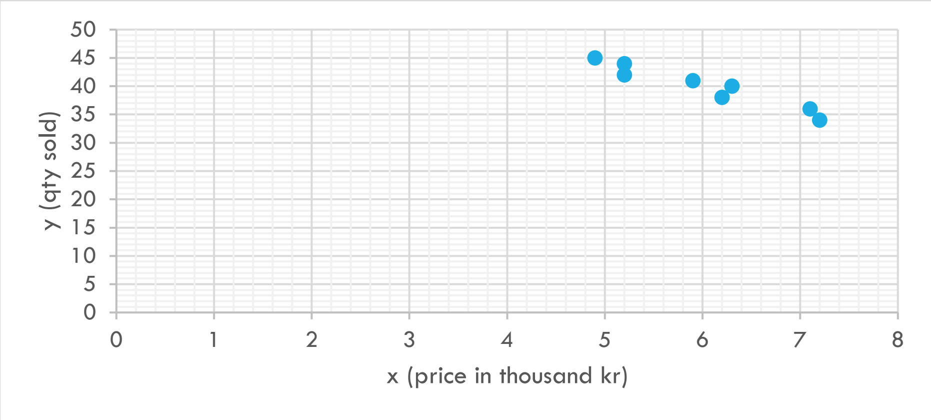
No.	x	x^2	y	y^2	xy
1	5.9	34.81	41	1681	241.9
2	6.2	38.44	38	1444	235.6
3	7.2	51.84	34	1156	244.8
4	6.3	39.69	40	1600	252.0
5	5.2	27.04	44	1936	228.8
6	7.1	50.41	36	1296	255.6
7	4.9	24.01	45	2025	220.5
8	5.2	27.01	42	1764	218.4
Σ	**48.0**	**293.28**	**320**	**12902**	**1897.6**

$$\bar{x} = \frac{48}{8} = 6 \qquad s_x^2 = \frac{293.28 - 8(6^2)}{7} = 0.754286 \qquad s_x = \sqrt{0.754286} = 0.868496$$

$$\bar{y} = \frac{320}{8} = 40 \qquad s_x^2 = \frac{12902 - 8(40^2)}{7} = 14.5714 \qquad s_x = \sqrt{0.754286} = 3.81725$$

$$s_{xy} = \frac{1897.6 - 8(6)(40)}{7} = -3.2$$

a. A suitable diagram is a scatter plot



Remember to put the explanatory variable (price) on the x axis, and the dependent variable on the y axis. What can be said about the relationship between price and quantity sold? Do you think it's appropriate to explain the relationship with the linear model $y = \beta_0 + \beta_1 x + \varepsilon$? Is the relationship positive or negative?

b. The parameter β_1 is the expected increase in y when x increases with one unit. It is the slope coefficient in the linear model $y = \beta_0 + \beta_1 x + \varepsilon$.

c.

$$b_1 = \frac{cov(x,y)}{s_x^2} = \frac{s_{xy}}{s_x^2} = -\frac{3.2}{0.754286} = \mathbf{-4.24242}$$
$$b_0 = \bar{y} - b_1\bar{x} = 40 - (-4.24242)(6) = \mathbf{65.4545}$$

An increase of 1000kr in price is estimated to decrease the quantity of items sold by an average of 4.242.

d. To calculate SSE, first we need to estimate the model $\hat{y} = 65.4545 - 4.24242x$, and the residuals $e = \hat{y} - y$.

No.	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{\hat{y}}$	$\boldsymbol{e}$	$\boldsymbol{e^2}$
1	5.9	41	40.424	0.576	0.332
2	6.2	38	39.151	-1.151	1.326
3	7.2	34	34.909	-0.909	0.826
4	6.3	40	38.727	1.273	1.62
5	5.2	44	43.394	0.606	0.367
6	7.1	36	35.333	0.667	0.444
7	4.9	45	44.667	0.333	0.111
8	5.2	42	43.394	-1.394	1.943
Σ	**48**	**320**		**0**	**6.97**

$$SSE = \sum_{i=1}^{n} e_i^2 = \mathbf{6.97}$$

e. The coefficient of determination can be calculated using the following formula:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{(n-1)S_y^2} = 1 - \frac{6.97}{(7)(14.5714)} \approx \mathbf{0.932}$$

93.2% of the variation in y, can be explained by the model.

f. Hypotheses:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 < 0$$
$$\alpha = 0.05$$

Test statistic:

$$t = \frac{b_1 - \beta_1^*}{s_{b_1}} \sim t_{n-K-1}$$

Decision rule:
We will reject the null hypothesis if the observed value of the test statistic is more extreme than the critical value. In other words, we reject H_0 if $t_{obs} < -t_{n-K-1;\alpha}$.

Observed value:

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{SSE/(n-K-1)}{(n-1)s_x^2} = \frac{6.97/6}{7(0.754286)} = 0.220013$$

$$s_{b_1} = \sqrt{0.220013} = 0.469$$

$$t_{obs} = -\frac{4.2424}{0.469} = \mathbf{-9.045}$$

The critical value:

$$-t_{n-K-1;\alpha} = -t_{6;0.05} = [table\ 3] = \mathbf{-1.94}$$

Conclusion:

We reject H_0 on the 5% significance level since $t_{obs} < -t_{n-K-1;\alpha}$. The estimated slope parameter is significant from zero.

g. We can calculate a 90% prediction interval with the following formula:

$$\hat{y}_{x_{n+1}=6} \pm t_{n-2;\frac{\alpha}{2}} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right)}$$

Where:

$$\hat{y}_{x_{n+1}=6} = b_0 + b_1 x_{n+1} = 65.4545 - 4.24242(6) = 40$$

$$t_{n-2;\frac{\alpha}{2}} = t_{6;0.05} = [table\ 3] = 1.943$$

$$s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right) = \frac{6.97}{6}\left(1 + \frac{1}{8} + \frac{(6-6)^2}{(7)0.754286}\right) = \frac{6.97}{6}\frac{9}{8} = 1.306875$$

Which gives:

$$40 \pm 1.943\sqrt{1.306875} = 40 \pm 2.22 \quad \Leftrightarrow \quad (37.88; 42.22)$$

h. The correlation coefficient r_{xy} can be calculated by taking the square root of the coefficient of determination R^2. *(However, the square root R^2 will only produce positive values, it will tell you about how strong the linear relationship between x and y is, but not whether the relationship is positive or negative).* Sinceour estimated slope coefficient b_1 was negative, we also know that the relationship between x and y is negative.

$$r_{xy} = -\sqrt{R^2} = -\sqrt{0.932} = -0.9652.$$

Nothing would happen to the correlation coefficient if we changed the scale of x, however the estimated regression coefficient would change.

$$b_1^{kr} = \frac{b_1^{1000\,kr}}{1000} = -0.00424$$

EXERCISE 2

a. A 95% confidence interval for β_1 can be calculated with the following formula:

$$b_1 \pm t_{n-2;\frac{\alpha}{2}} \times s_{b1}$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.27391304	1.40074452	0.90945424	0.38968736	-1.956209623	4.50403571
x1	0.33913043	0.08527819	3.97675473	0.00408018	0.142478582	0.53578229

From the model 1 regression output we get the coefficient estimate ($b_1 = 0.33913043$) and its standard error ($s_{b1} = 0.08527819$). From table 3 in the formula sheet we know that $t_{8;0.025} = 2.306$.

A 95% confidence interval for β_1 is then:

$$0.3391 \pm 2.306(0.08528) \Rightarrow \mathbf{0.3391 \pm 0.1967}$$

We can also get the confidence interval from the regression output (highlighted in green).

b. We do the same as above but using the regression output from model 2. (remember that we lose one degree of freedom when we include an additional variable in the model)

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	-0.868701467	0.951547725	-0.91294	0.3916343	-3.118754293	1.38135136
x_1	0.305672994	0.049442473	6.182397	0.000453	0.188760124	0.422585864
x_2	0.923425367	0.221113461	4.176251	0.0041566	0.400575115	1.446275618

$b_1 \pm t_{n-2;\frac{\alpha}{2}} \times s_{b1}$
$\Rightarrow\ 0.3057\ \pm t_{7;0.025}\,(0.049442473)$
$\Rightarrow 0.3057\ \pm 2.365\,(0.049442473)$
$\Rightarrow \mathbf{0.3057\ \pm 0.1168}$

c. We want to test whether at least one of the variables X_1 and X_2 can help explain the variation in Y. This is the same as testing whether either of the coefficients β_1 and β_2 are significant from zero.

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_A: At\ least\ one\ slope\ coefficient\ not\ equal\ to\ zero$$

Our test statistic is the F ratio, we can find the observed value and the p-value in the ANOVA table. If we find that the p-value is less than α=0.05 then we will reject the null hypothesis on the 5% significance level.

ANOVA	***df***	***SS***	***MS***	***F***	***Significance F***
Regression	2	21.60055651	10.80028	32.878367	0.00027624
Residual	7	2.299443486	0.328492		
Total	9	23.9			

Above we see that $F_{obs} = 32.88$ with a p-value of 0.000276. Thus, we reject the null hypothesis on the 5% significance level.

d. To test whether model 2 is better than model 1 in explaining the variation in Y, we want to test whether the β_2 coefficient is significant, given that X_1 is included in the model.

$$H_0: \beta_2 = 0 \mid X_1$$

$$H_0: \beta_2 \neq 0 \mid X_1$$

If we find that the p-value of the observed t statistic is less than $\alpha = 0.05$ we reject the null hypothesis. Alternatively, we can reject the null hypothesis on the 5% significance level if $t_{obs} > t_{7;0.025} = 2.365$.

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	-0.868701467	0.951547725	-0.91294	0.3916343	-3.118754293	1.38135136
x_1	0.305672994	0.049442473	6.182397	0.000453	0.188760124	0.422585864
x_2	0.923425367	0.221113461	4.176251	0.0041566	0.400575115	1.446275618

From the table above, we can see that the observed t stat is greater than the critical value ($4.176251 > 2.365 \Leftrightarrow t_{obs} > t_{7;0.025}$). In addition to this, the p-value is less than $\alpha = 0.05$. Thus, we reject the null hypothesis on the 5% significance level.

e. The adjusted coefficient of determination can be calculated with the following formula:

$$R_{adj}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{MSE}{s_y^2}$$

From the Excel table we know that the variance of y is $s_y^2 = 1.6296^2$ and from the ANOVA table we know that $MSE = 0.328492$. We then get that:

$$R_{adj}^2 = 1 - \frac{0.328492}{23.9/9} = 1 - 0.123698 = \mathbf{0.87632}$$

f. We get the expected value by substituting $X_1 = 15$ and $X_2 = 3$ into our estimated regression model.

$\hat{\mu}_{Y|X_1=15,X_2=3} = b_0 + b1(15) + b_2(3) = -0.8687 + 0.3057(15) + 0.9234\,(3) = \mathbf{6.487}$