

Case-based Social Statistics II

CSSS 322

Professor: Mark S. Handcock

Solutions to Homework 6

Due Friday, May 24, 2002

Problems to be handed in:

- 1) Suppose the regression model:

$$Y = \beta_0 + \beta_1 X$$

holds with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation of Y is made at $X = 5$.

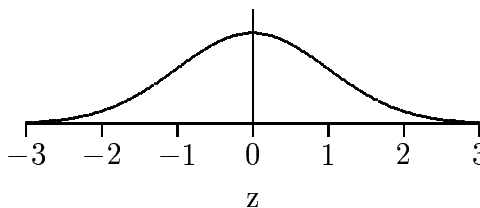
- a) Suppose the regression model holds without the normality assumption on the error term ϵ_i . This was assumption (iv) in the notes for Lecture 4.0. Can you state the exact probability that Y will fall between 195 and 205? Explain briefly.

Solution(3): The regression model is the one we have in class (*without the normality assumption on the error term ϵ_i*). Without this assumption we do not know the probability distribution of Y for each value of $X = x$. We do know that the mean of the distribution is $100 + 20x$ and the variance of the distribution is 25, but we do not know the shape of the distribution. Hence we can not calculate the exact probability. \boxtimes

- b) Suppose now that the regression model holds with the normality assumption on the error term ϵ_i . Can you now state the exact probability that Y will fall between 195 and 205? If so state it.

Solution(3): We now assume that the error term ϵ_i has a normal distribution. Hence the probability distribution of Y for each value of $X = x$ is also normal with the mean and variance given above. For $X = 5$ these values are 200 and 25 respectively. The exact probability is then

$$\begin{aligned} P(195 \leq Y \leq 205) &= P\left(\frac{195 - 200}{5} \leq \frac{Y - \mu}{\sigma} \leq \frac{205 - 200}{5}\right) \\ &= P(-1 \leq Z \leq 1) \\ &= 2 \times P(Z \leq 1) - 1 \\ &= 0.6826 \quad \text{from Table B.1, p. 1335} \end{aligned}$$



\boxtimes

- 2) A junk mail recycling company wants to promote its service. Suppose that four levels of direct mail advertisements were randomly assigned to four similar residential blocks in Greenwich village, resulting in the following people signing up for the companies service:

Advertisements (mailings/year)	Signups (signups per 100 residents/year)
1	70
2	70
4	80
5	100

- a) Calculate, by hand, the regression line of the signups against mailings

Solution(4): Let X be the advertisements in mailings/year and Y be the signups per 100 residents/year. Then

$$\bar{X} = 3 \text{ mailings}, \quad \bar{Y} = 80 \text{ signups per 100 residents/year}$$

$$\sum_{i=1}^4 x_i y_i = 1030, \quad \sum_{i=1}^4 x_i^2 = 46$$

So

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^4 x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^4 x_i^2 - n\bar{X}^2} \\ &= \frac{1030 - 4 \times 3 \times 80}{46 - 4 \times 3^2} \\ &= 7 \text{ signups per 100 residents/year per mailing} \end{aligned}$$

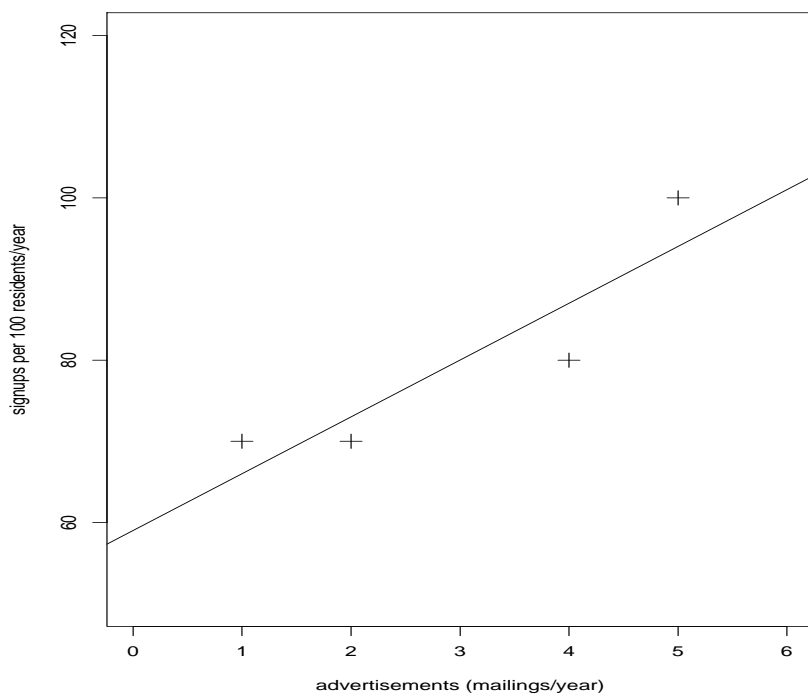
The intercept (or "constant") is,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 80 - 7 \times 3 = 59 \text{ signups per 100 residents/year}$$

⊗

- b) Graph the four points and the regression line. Check that the line fits the data reasonably well.

Solution(5): The plot is below:



⊗

c) Use the regression line to predict:

i) The signups if 3 mailings/year (mailings per year) were sent.

Solution(1): The fitted value at $x = 3$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 59 + 7 \times 3 = 80 \text{ signups per 100 residents/year}$$

⊗

ii) The signups if 4 mailings/year were sent.

Solution(1): The fitted value at $x = 4$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 59 + 7 \times 4 = 87 \text{ signups per 100 residents/year}$$

⊗

iii) The increase in signups for every 1 mailing/year increase in mailings.

Solution(1): This is just the slope of the regression line, $\hat{\beta}_1$, that is 7 signups per 100 residents/year.

⊗

Show these three on the graph.

- 3) A HeadStart type program aims to improve the job-prospects of participants. In seeking to determine just how influential the HeadStart training is, the review committee of a program has collected data over the previous year on the weekly income on participants, and the number of months they were trained within the program. The data are given below:

TRAINING	INCOME
3.0	50
5.0	250
7.0	700
6.0	450
6.5	600
8.0	1000
3.5	75
4.0	150
4.5	200
6.5	550
7.0	750
7.5	800
7.5	900
8.5	1100
7.0	600

The data are available on the course website under “Data”.

- a) Using the “Scatterplot” command in Datatools, plot the data. Does it appear that the length of training and income are linearly related?

Solution(1): Clearly a linear relationship is reasonable here.

✕

- b) Below is fit of the linear regression model of income based on training time.

LEAST SQUARES LINEAR REGRESSION OF INCOME Income (thousands per month)

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P	
CONSTANT	-644.951	72.8973	-8.85	0.0000	
TRAINING	195.074	11.5381	16.91	0.0000	
R-SQUARED	0.9565	RESIDUAL MEAN SQUARE (MSE)	5404.98		
ADJUSTED R-SQUARED	0.9532	STANDARD ERROR OF ESTIMATE	73.5186		
SOURCE	DF	SS	MS	F	P
REGRESSION	1	1.545E+06	1.545E+06	285.84	0.0000
RESIDUAL	13	70264.8	5404.98		
TOTAL	14	1.615E+06			

CASES INCLUDED 15 MISSING CASES 0

Fit the model in DataTools using the “Simple Linear Regression” command and hand in a copy of the output. You can use the above to make sure you are doing it right.

Solution(1): The output is:

Results:

Simple linear regression results:

Independent variable: training

Dependent variable: income

Sample size: 15

Correlation coefficient: 0.978

Estimate of sigma: 73.51859

Parameter	Estimate	Std. Err.	DF	Tstat	Pval
Intercept	-644.95074	72.89727	13	-8.847392	0
training	195.0739	11.538096	13	16.90694	0

c) Give an economic interpretation of the coefficient of training, $\hat{\beta}_1$, in the model.

Solution(1): The interpretation of the slope is an increase of \$195.07 in income for each month increase in training. Note that the units of income is in dollars and that of training is in months. \boxtimes

d) If the sign of the slope were negative, what would that say about the relationship between training and sales?

Solution(1): The interpretation of a negative slope is that income would decrease for each month increase in training would decrease. That is increased training decreases sales! \boxtimes

e) Give an economic interpretation of the intercept, $\hat{\beta}_0$, in the model. What does the value of the intercept tell us?

Solution(1): The interpretation of the intercept is the income if the training time was zero. Here the value is negative \$644.95. This is not reasonable and indicates that extrapolation is unreasonable for this model. \boxtimes

f) Calculate a 99% confidence interval for β_0 . Do you think that there will be no income if there was no training?

Solution(3): A 99% confidence interval for β_0 is

$$\begin{aligned}\hat{\beta}_0 &\pm t_{\alpha/2}(n-2) \times \widehat{SE}(\hat{\beta}_0) \\ &= -644.95 \pm t_{0.005}(13) \times 72.897 \\ &= -644.95 \pm 3.012 \times 72.897 \\ &= (-\$864.52, -\$425.38)\end{aligned}$$

\boxtimes

g) Calculate a 95% confidence interval for β_1 .

Solution(3): A 95% confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 &\pm t_{\alpha/2}(n-2) \times \widehat{SE}(\hat{\beta}_1) \\ &= 195.07 \pm t_{0.025}(13) \times 11.538 \\ &= 195.07 \pm 2.160 \times 11.538 \\ &= (170.15, 219.99) \text{ income dollar per training months}\end{aligned}$$

\boxtimes