Lab Three: Bivariate Regression
Geog 301a

Introduction

Fitting a line to a bivariate data set is typically accomplished by developing a least-squares regression model. The rationale underlying the use of the traditional regression model is that it allows the researcher to develop an equation that minimizes the residual sum of squares between the fitted line and **all** of the residual values. Occasionally, the presence of extreme values, or outliers, detract from the predictive power of the traditional bivariate regression model. Consequently, traditional bivariate regression is not resistant to the effects of extreme values.

Computational Formulae

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\text{covariation}}{\text{variation in x}}$$

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \text{intercept with y axis when x} = 0$$

$$s^2 = \frac{\sum \hat{y}^2}{n} - \bar{y}^2 = \text{regression variance}$$

$$s_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \text{total variance}$$

$$r^2 = \frac{s^2}{s_y^2} = \text{coefficient of determination}$$

$$t = \frac{b - B}{s.e.b} \quad \text{where } B = \text{population parameter } H_o: B = 0$$

$$s.e.b = \frac{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Profit is sales minus expenses.

Instructions

The following data set contains raw information about the levels of agricultural capitalization, and the profitability of agriculture, for 18 of Alberta's 19 census tracts in 1986. Two extreme outliers are evident in the data, namely census tracts 8 and 9. As such, the appropriateness of examining the data set by the resistant line technique is clear. Treat profitability as the dependent variable, and capitalization as the independent variable.

Column One - the census tract number ID

Column Two - the total expenses associated with the agricultural sector for each census tract in millions of dollars

Column Three - the total sales for the agriculture sector in each census tract in millions of dollars

Column Four - the amount of capital committed to agriculture in the census tract in billions of dollars

ALBERTA AGRICULTURAL DATA BY CENSUS TRACT

ID	Expenses $millions	Sales $millions	Capital $billions	Profit
1	160.5	193.6	1.42	
2	518.6	583.3	2.88	
3	225.6	244.6	1.64	
4	130.3	154.6	1.16	
5	342.6	413.9	2.58	
6	343.2	375.1	2.75	
7	323.8	375.7	2.31	
8	353.3	399.1	0.45	
9	44.2	45.3	2.22	
10	401.3	473.2	2.95	
11	318.2	360.5	2.76	
12	93.2	100.4	0.79	
13	218.5	249	1.72	
14	24.5	23.3	0.22	
15	0	0	0	
16	19.9	19.1	0.22	
17	108.5	113	0.91	
18	26.3	25	0.22	
19	190.7	215.4	1.47	

Source: Statistics Canada (1987) Agriculture Alberta, vols 1&2. Minister of Supply and Services Canada, pp. 4.7-4.9.

To be calculated by hand or in a spreadsheet package:

1) Calculate profit levels in each of the given census tracts. Draw a scatterplot of profit as a function of capital.

2) Develop a function in the general form $y = ax + b$ that describes the relationship between profit and capital.

3) Calculate the following
   a) the slope
   c) the intercept
   d) $\hat{y}$
   e) the residuals
   f) coefficient of determination
   g) significance level of the b coefficient

4) On your scatterplot, draw in the line of best fit.

5) What does this analysis suggest about the economics of farming in Alberta? How could the predictive power of this model be improved?

6) Could your model be applied to Ontario?

7) Check your answers using SPSS.