

Homework 6 - handed out Friday, 9 Oct 2009

DUE DATE:

on campus Friday, 16 Oct 2009, in lecture (11 am)

or by e-mail to Chuanlong, dlong@iastate.edu, no later than noon.

off campus Monday, 5 Oct 2009, by 4 pm to Nicole Rembert, email: rembeall@iastate.edu or FAX: 515-294-4040 (please include cover page with Stat 500 / Nicole Rembert).

1. Regression using summary statistics – Suppose we consider the growth rate of the US economy (X) as a predictor of the proportion of the vote obtained by the Democratic candidate in presidential elections (Y). The following table gives the Democratic share of the two-party vote (ignoring third parties) and the growth rate (in %).

year	%Dem.	growth	year	%Dem.	growth	year	%Dem.	growth
1916	0.5168	6.38	1944	0.5377	6.88	1972	0.3821	5.05
1920	0.3612	-6.14	1948	0.5237	3.77	1976	0.5105	0.78
1924	0.4568	-2.16	1952	0.4460	-0.34	1980	0.4470	-5.69
1928	0.4118	-0.63	1956	0.4224	-0.69	1984	0.4083	3.04
1932	0.5916	-13.98	1960	0.5009	-1.92	1988	0.4610	2.14
1936	0.6246	13.41	1964	0.6134	2.38			
1940	0.5500	6.97	1968	0.4960	4.00			

Two years, 1932 and 1972, are explainable anomalies (depression, Watergate). They are omitted; the following summaries are computed from the remaining data:

$$\sum_{i=1}^{17} x_i = 32.18 \quad \sum_{i=1}^{17} x_i^2 = 446.171 \quad \sum_{i=1}^{17} (x_i - \bar{x})^2 = 385.256 \quad \sum_{i=1}^{17} x_i y_i = 19.838$$

$$\sum_{i=1}^{17} y_i = 8.288 \quad \sum_{i=1}^{17} y_i^2 = 4.121 \quad \sum_{i=1}^{17} (y_i - \bar{y})^2 = 0.0812 \quad \sum_{i=1}^{17} (x_i - \bar{x})(y_i - \bar{y}) = 4.149$$

- (a) You want to use growth rate to predict the proportion of Democratic vote. Estimate b_0 and b_1 .
- (b) The sum of squared residuals around the regression line is .03653. Find the t-statistic for testing the hypothesis that the slope is equal to 0. What conclusion do you draw from this?
- (c) Give a 99% CI for the expected Democratic vote in a year with 5% growth.
- (d) Suppose the economic growth in 1992 was 5%. Predict the Democratic share of the vote for that year and give a 99% prediction interval. How does this interval compare with the previous confidence interval? Explain the difference.
- (e) The model SS is 0.04468. You already have the error SS. Calculate the remaining values in the regression ANOVA table, including the total d.f. and SS. Calculate the F statistic for testing $H_0: \beta_1 = 0$. This F statistic should be the square of the t statistic from part b. Is it?
- (f) Calculate R^2 . Many social scientists are impressed by models with R^2 larger than 50%. Is this such a model?
- (g) The narrowest possible prediction interval is found when you predict at the mean X. Here, that is 1.9% growth. Calculate a 95% prediction interval at $x=1.9\%$.

(h) Is the prediction you just made sufficiently precise to be useful? Answers to this question depend on criteria for “sufficiently precise” that are subject-specific. I can think of two ways to quantify sufficiently precise. You may adopt either of these, or devise your own criterion.

a) Compare the width of your prediction interval to the range of observed values. Here, the maximum proportion Democratic vote is 0.62; the minimum is 0.36.

b) In 9 of the 17 years, the proportion Democratic vote is between 0.45 and 0.55. A reliably correct prediction for any of these 9 years requires a prediction interval width that is less than 0.10.

2. **Regression in SAS** – The file ‘snow1.txt’ on the class web site contains data from a snow gauge calibration study. A snow gauge is an instrument that measures the wetness of snow, which is crucial in western states for predicting water availability. Wet snow is denser than dry snow. Density is time consuming to measure directly; the snow gauge instrument measures a quantity called gain that depends on the density. The data at hand were collected to calibrate the instrument, that is describe how gain is related to density. When the snow gauge is used, the gain is measured and used to predict the snow density.

Polyethylene blocks were used as substitute for snow. These can be manufactured in different densities. The density is set by the process used to manufacture the blocks. The data set (in snow1.txt) includes 9 densities. Ten blocks of each density were measured.

(a) The investigators plan to use regression to describe the relationship between gain and density. Which variable (gain or density) should be used as the X variable? Which is the Y variable? If it doesn’t matter, say so. Briefly explain.

(b) Rightly or wrongly, the investigators decide to use density as the X variable and gain as the Y variable. For all this and subsequent parts of this question, please assume that the usual simple linear regression model is appropriate. Estimate the slope and intercept of the regression of gain on density.

(c) If you assume a linear relationship, is density related to gain? In other words, test whether the slope = 0. Report your p-value and a short conclusion.

(d) Predict the average gain when the density = 0.2 and calculate a 95% confidence interval for the average gain at density = 0.2

(e) Calculate a 95% prediction interval for gain measurements when the density = 0.2. The 95% prediction interval includes 95% of all observations at density = 0.2

3. **Regression Diagnostics** – The data in anscombe.txt on the class web site are classic data sets constructed by Anscombe to illustrate the need for graphical diagnostics. There are four data sets in this single data file. There are three columns: set, x, and y. The observations in data set 1 have set = 1, and so on.

The issue is to decide, for each data set, whether a linear regression is a good description of the relationship between y and x.

(a) Fit a simple linear regression, predicting y using x, separately to each data set. Using only the numerical results (estimates, standard errors, tests, r^2 , whatever else you might think of), is the linear regression a good description for data set 1? for data set 2? for data set 3? for data set 4?

(b) plot Y vs X and residual vs predicted value for each data set. Make sure you understand how the pattern in the residual plot relates to the pattern in the observations (no answer required). After looking at the graphs, is the linear regression a good description for

data set 1? for data set 2? for data set 3? for data set 4? Explain why or why not for each data set.

4. **Diagnostics / Lack of fit / Inverse prediction** – Snow data, part 2.

- (a) Plot the residuals vs the predicted values. Do you have any concerns?
- (b) These data include multiple observations at each density. Hence it is possible to compute the mean and s.d. for each density, then use the Box-Cox method to choose a transformation that equalizes the variances. Calculate the mean and s.d. for each of the 9 densities, then regress $\log \text{sd} (Y)$ on the $\log \text{mean} (Y)$. What transformation (if any) does this suggest?
- (c) Rightly or wrongly, the investigators decide to use a log transformation. Fit the regression of $Y = \log \text{gain}$ on $X = \text{density}$. (You don't have to report anything about this regression). Plot the residuals vs. predicted values. Are there any concerns?
- (d) Lack of fit is an important concern for these investigators. Since there are repeated measurements for each density, it is possible to use the ANOVA lack of fit test. Is a straight line sufficient to describe the relationship between $\log \text{gain}$ and density. Report the p-value and a one-sentence conclusion.
- (e) The investigators measure a snow sample. The measured gain is 152. Predict the density using the regression of $\log(\text{gain})$ on density and give an approximate s.e.