

Homework 6 - handed out Weds, 3 Oct 2007

DUE DATE: on campus: Weds, 10 Oct, by 4 pm, to Norma Elwick in 115 Snedecor.

off campus: Monday, 15 Oct, by 4 pm to Ying Shi

1. Analysis of paired data —

The data in platelet.txt are from a study of the physiological effects of smoking. The subjects are 11 non-smokers. The response is the extent of platelet aggregation. Aggregation normally happens at the site of a cut. That’s what forms the scab. Platelet aggregation inside blood vessels is bad. That increases the probability of heart attacks.

In this study, blood was drawn and platelet aggregation measured. Then the subject smoked a cigarette. Thirty minutes later, blood was again drawn and platelet aggregation measured. There are two responses for each subject (before and after).

- (a) Estimate the mean difference between before and after smoking. Use t quantiles to estimate a 95% confidence interval for the difference.
- (b) Is it appropriate to use a paired-t test to test the hypothesis of no difference? If not, what test might be appropriate? Explain.  
Hint: I’m not asking about paired or 2-sample test is more appropriate.
- (c) Test the hypothesis of no difference, using the test you consider appropriate.
- (d) Find the variance among the before observations and the variance among after observations. Use these to calculate the variance of the mean difference if you ignored the pairing. (you may use the unequal variance formula or compute  $s_p^2$  then use the equal variance formula). Compare this to the correct variance of the mean difference. Is a paired design a good idea for this sort of experiment? Explain why or why not.

2. Analysis of data from a block design – Fat in Diets

Problem 21.7 and 21.8 from the text. Do parts:

21.7: a, b, c (omit d)

21.8: a, c, d.(omit b, e, f)

The data are in dietfat.txt; a “copy-able” copy of the problem is available in Snedecor115.

3. 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.

Note: I won't ask you to do the calculations because this is a non-computer problem, but here is what happens if you imagine redoing the analysis 10 years from now with two additional observations: (-20, 0.135) and (25, 0.763). The prediction intervals are slightly larger because  $\sqrt{\text{MSE}}$  is slightly larger and the  $t$  quantile hasn't shrunk enough to compensate.  $R^2$  is now 84.6%! Lots of people are impressed with that size of  $R^2$ . I don't see why, at least when the purpose of the regression is to make predictions or to estimate slopes.