Stat 135, Fall 2006

HOMEWORK 10 (due Friday 12/1)

In what follows, you will have to do lots of regressions and pick one or two that are better than the others. **DO NOT** turn in all the regression output. The only computer output I want to see are the relevant plots, along with a clear discussion of what they imply. The rest of each answer should consist of a carefully written paragraph, including numerical values from your computer work as appropriate. **You** will have to work out what's appropriate.

1. The dataset `chicks` was obtained from *BLSS: The Berkeley Interactive Statitstical System* by Abrahams and Rizzardi. Each observation corresponds to an egg (and the resulting chick) of a bird called the Snowy Plover. The data were taken at Point Reyes Bird Observatory. Column 1 contains the egg length in millimeters, Column 2 the egg breadth in millimeters, Column 3 the egg weight in grams, and Column 4 the chick weight in grams. The object is to estimate the size of the chick based on dimensions of the egg.

a) First you are going to regress chick weight on egg length. Write out the usual regression model in terms of these two variables. Plot the data and comment on whether the assumptions of the model are reasonable. Find the means and SDs of both variables, as well as the correlation between them. Use the formulas you derived in HW 9 to find the slope and the intercept of the regression line. Provide units for the slope and the intercept, and write the equation of the regression line. Draw the regression line on your plot.

b) Now use R to regress chick weight on egg length. Check that R produces the same slope and intercept that you got in **a**. For each of the t and F statitics in the R output, state the null and alternative hypotheses that are being tested, and state the conclusion of the test.

c) Which of the three variables egg length, egg breadth, and egg weight is most highly correlated with chick weight? Call this one the "best" predictor for now. Draw the scatter diagram of chick weight versus this "best" predictor and put the regression line through it. Draw the residual plot. Is there any noticeable non-linearity?

d) If possible, construct a 95%-confidence interval for the mean weight of Snowy Plover chicks that hatch from eggs weighing 8.5 grams.

e) I have a Snowy Plover egg that weighs 8.5 grams. If possible, construct a 95%-prediction interval for the weight of the chick that will hatch from this egg. It's a prediction interval, rather than a confidence interval, because it's trying to predict the value of a random variable instead of estimating a fixed parameter.

f) Repeat **d** and **e** when the egg weight is 12 grams instead of 8.5 grams.

2 (Continuing problem 1). The object is still to find a good way to predict the weight of a chick given measurements on the egg, using linear regression as the only tool. The difference between this problem and Problem 1 is that now you are going to use a combination of variables to estimate the weights of the chicks.

a) Regress the weights of the chicks on the lengths and breadths of the eggs. Assess the

regression. Compare it with the "best" simple regression you performed Problem 1. Is one noticeably better than the other?

b) Regress egg weight on egg length and egg breadth. Assess this regression. Use this regression to explain the similarity (or difference) between the two regressions in that were compared in **a** .

c) Now regress the weights of the chicks on all three predictor variables: egg length, egg breadth, and egg weight. How do you reconcile the result of the F test in this regression with the results of the t-tests? Explain why this regression is not as impressive as either of the two you compared in **a**, even though it has a higher R^2.

d) Perform all possible regressions of chick weight using combinations of the three predictor variables used in **c**. **Do not turn in all the results.** Are there any that are clearly better than the others? Which ones, and why?

3. This problem concerns the dataset `tox`. The data are observations on a simple random sample of Hodgkin's disease patients at Stanford Hospital, taken as part of a study of the toxicity of the treatment to the patients' lungs.

Column 1 is the patient's height in centimeters.

Column 2 is a measure of the amount of radiation to the patient's lungs.

Column 3 is a measure of the amount of chemotherapy the patient received.

Column 4 is a score that measures how well the patient's lungs were doing before the start of treatment (also called a "baseline" score). Large scores are good.

Column 5 is a score that measures how well the patient's lungs were doing 15 months after treatment. Large scores are good.

a) Is the treatment toxic in the short term? That is, are patients' lungs worse 15 months after treatment than they were before the start of the treatment? Answer this question by carrying out statistical tests, both parametric and non-parametric. State your null and alternative hypotheses in detail, and justify your choice of procedure.

[Note: the long–term effects are not very severe. The study shows that three years after treatment the patient's lungs are almost back to normal.]

b) Use linear regression as your tool to decide which combination of variables in Columns 1 through 4 should be used to predict a patient's score 15 months after treatment. Justify your answer.

4. The dataset `baby` contains observations on mothers and their newborns at Kaiser Hospital (data courtesy of D. Nolan).

Column 1: baby's weight at birth, to the nearest ounce

Column 2: gestation days (that is, total number of days of pregnancy)

Column 3: mother's age in completed years

Column 4: mother's height to the nearest inch

Column 5: mother's pregnancy weight to the nearest pound

Column 6: indicator of whether the mother smoked (1) or not (0) during her pregnancy

a) Assess the normality of the babies' birthweights by looking at the histogram and the normal q-q plot.

b) Draw the histogram of the mothers' pregnancy weights, and draw the normal q-q plot. What feature of the q-q plot corresponds to the skewness of the distribution? What would the q-q plot look like if the skewness were in the other direction?

c) Which subset of the variables in Columns 2 through 6 should be used as predictors for birthweight? Justify your answer.

d) Interpret the coefficient of the indicator variable. How can it be used? What conclusion does it allow you to make in the present case?

5. The dataset `women` contains the average weight in pounds (Column 2) for American women whose heights, correct to the nearest inch, are given in Column 1.

a) Perform the linear regression of weight on height. What is the value of R? Assess the regression.

b) Suppose the data consisted of the heights and weights of American women. That is, suppose there was a point for each woman, with one co-ordinate representing her weight and the other her height. Would the correlation be the same as that in **a**, or more, or less? Justify your choice.

c) Fit a polynomial model to the data in `women`. Justify your choice of degree, and assess the fit of your model.

6. 14.52.

7. 14.6. The matrix formalism may be overkill in such a simple set-up. You don't have to use it if you don't want to.

8. 14.16.

9. 14.17.

10. Do `help.search("regression")` in R. That should show you that what you've learned about regression in Stat 135 is the tip of a very large iceberg. No, there's nothing to turn in for this problem. You've done enough.