Midterm Exam — Stat 410

Dr. Scott
scottdw@rice.edu

November 10, 2005

Instructions:

1. Open book, notes, homework, and R/Splus programming environment.
2. Problems 1–3 involve data analysis; 4–5 are theoretical.
3. Organize (and staple) your work in the order of the problems. Use a (yellow) highlighter to indicate numerical results. Supporting computer code should be nearby.
4. In problems with multiple parts, you should be able to continue even if you cannot work a particular portion. In particular, assume what you need and continue.
5. Time limit: 4 hours in 4 sittings, not including generic code programming.
6. Work all problems. Liberal partial credit, so attempt all parts of each problem. Maximum score is 100.
7. Record the dates/time of your work. When you are done, write and sign the pledge on the cover page of your exam. Print your name as well.
8. Turn in your exam to my office by noon, 12:00 p.m., Tuesday, November 15.
9. There is no class on Tuesday, November 15. Exams returned Thursday, 11/17.

1. (30 points) Abalone is a gourmet shellfish cherished for its delicate flavor, with a multifaceted colorful shell. The file *fish.txt* contains data on the price of an Abalone dinner in San Francisco dating back to 1924. The prices are all given in 2004 dollars. For numerical stability, transform the years (1924,2005) to the interval $(0, 1)$. *Hint: Do not use any of the time series ideas of Chapter 12 for this problem, i.e. assume the residuals are independent.*

   (a) Plot the data and add a straight-line least-squares fit. Give 95% confidence intervals for the regression coefficients.

   (b) Since multiple prices for several years are available, perform an ANOVA lack-of-fit test for this model. What conclusion do you draw?

   (c) Plot these data again and add quadratic and cubic least-squares fits. For each, is the largest coefficient significantly different than zero?

   (d) Finally, consider the straight-line fit again, but on a transformed y axis. Find the "best" transformation using the Box-Cox technique, and plot as a function of λ. Is the $\log(y)$ a good transformation? Add the best fit to a graph on the transformed scale.

   (e) Plot a histogram of the residuals and a q-q normal plot of the residuals. Do they seem normal?

   (f) Perform the ANOVA lack-of-fit test for these transformed data.

2. (20 points) As a Merrill Lynch credit card holder, every dollar charged results in the accumulation of "signature reward points," which may be used to select merchandise rewards. The "Holiday 2005" rewards catalog arrived in late October. I selected a range of merchandise from the catalog and used Google to find current prices of the identical items. The file *mlol.rewards.txt* contains the item name, points required to redeem, actual dollar cost, and a rough estimate of the shipping expense (1=small/light, ..., 4=big/heavy).

   (a) Using the variables *points* and *price*, plot and display the linear regression lines using *points* as the y variable, and then *price* as y. Test the significance of the regression coefficients.

   (b) If I am interested in the "value" of one reward point, which plot seems more appropriate to answer this question? Since zero reward points corresponds to zero dollars, find the best fitting line with no intercept. Superimpose upon the appropriate plot in part (a). Compute the residuals and make a q-q normal plot. Do the residuals seem normal?

(c) Make 3 indicator columns to represent the shipping component. (Let category 1 be $(0,0,0)$, 2 be $(1,0,0)$, 3 be $(0,1,0)$, and 4 be $(0,0,1)$. Does adding these variables to the model produce b_k's which are significant? Test whether the three variables can all be dropped from the full model.

(d) Stick with your choice for y in part (b). Transform the points and price using a $\log_{10}$ transformation, and plot. Add the 4 regression lines corresponding to the 4 shipping categories. Are they significantly different?

3. (20 points) Home sales prices during the year 2002 in an unnamed midwestern city are described in Appendix C.7; see file *APPENC07.txt*. The second column is the sales price, y. However, we will use the transformation,

$$yt = \log_{10}(y - 76310)\,,$$

as it is more symmetric (the skewness is almost exactly zero). Many of the other variables are categorical, and I have coded these as below. The file *houses.txt* contains the data for the 522×18 matrix X. The columns are:

- 1 - square feet
- 2 - bedrooms = 3 (baseline is bedrooms ≤ 2)
- 3 - bedrooms = 4
- 4 - bedrooms ≥ 5
- 5 - bathrooms = 2 (baseline is bathrooms ≤ 1)
- 6 - bathrooms = 3
- 7 - bathrooms ≥ 4
- 8 - air conditioning (0/1)
- 9 - cars in garage = 2 (baseline is ≤ 1)
- 10 - cars in garage ≥ 3
- 11 - swimming pool (0/1)
- 12 - year built
- 13 - house quality = 2 (baseline is 3, the worst)
- 14 - house quality = 1 (1 is the best)
- 15 - house style = 1 (baseline is the rest)
- 16 - house style = 7 (the most expensive)

- 17 - lot size
- 18 - adjacent to highway (0/1)

(a) Look at the signs and significance levels of the estimated β_k's and interpret. Are there any surprises (in your opinion)? Is the multiple R^2 impressive? *Hint: use scale(X) and scale(yt) in the LS fit.*

(b) Compute the leverages, h_{ii}, and plot. There is an interesting cluster of points with leverage above 0.10. Look at X values for those cases. Can you see what most of them have in common?

4. (15 points) We want to show that the point $(\bar{x}, \bar{y})$ is on the regression surface (hyperplane). Suppose X is an $n \times p$ data matrix (with the first column a vector of 1's for the intercept), and that Y is the $n \times 1$ vector of responses. The least-squares estimate of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$.

(a) Let the vector of ones be denoted by $1_n = (1, 1, \ldots, 1)^T$. Show that the $p \times 1$ vector $\bar{x}$ and the scalar $\bar{y}$ can be computed as

$$\bar{x} = \frac{1}{n} X^T 1_n \qquad \text{and} \qquad \bar{y} = \frac{1}{n} 1_n^T Y \,.$$

(b) Recall the hat matrix $H = X(X^T X)^{-1} X^T$ is idempotent, and that the vector of predictions at the original data points is given by $\hat{Y} = HY$. Show that the vector 1_n is unchanged by H, that is,

$$H\, 1_n = 1_n \,.$$

Hint: Compute the matrix product HX.

(c) The linear prediction at $x = \bar{x}$ is given by

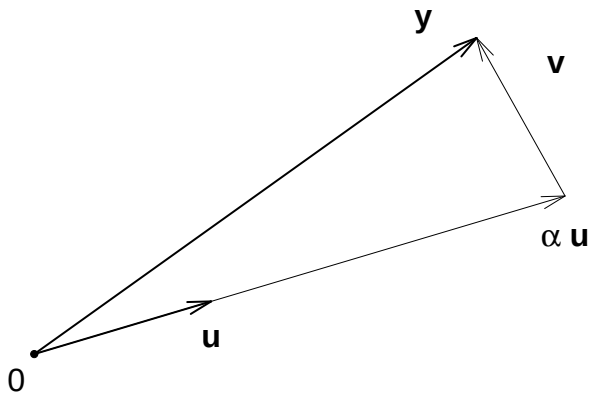
$$\hat{y} = \bar{x}^T \hat{\beta} \,.$$

Show that this $\hat{y}$ is exactly $\bar{y}$.

5. (15 points) A well-known identity used to compute the sample variance of a set of data $y = (y_1, y_2, \ldots, y_n)^T$ is

$$\frac{1}{n} \sum_{i=1}^{n} (y - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \bar{y}^2 \,.$$

This result is easily shown by ordinary algebra, but we want to demonstrate this identity by using the Pythagorean Theorem in n-dimensions, $\Re^n$.



(a) In the figure above, the vector $\mathbf{u}$ is a vector of length one in the direction $1_n = (1, 1, \ldots, 1)^T$. Find $\mathbf{u}$ in terms of the vector 1_n so that $\mathbf{u}^T\mathbf{u} = 1$.

(b) The vector $\mathbf{y}$ can be written as the sum of the two perpendicular vectors

$$\mathbf{y} = \alpha\,\mathbf{u} + \mathbf{v}\,,$$

where $\alpha\,\mathbf{u}$ is a vector in the same direction as $\mathbf{u}$. Note that the length of the vector $\alpha\,\mathbf{u}$ is $|\alpha|$, where α is a scalar. Recall vectors $\mathbf{u}$ and $\mathbf{v}$ are perpendicular if and only if $\mathbf{u}^T\mathbf{v} = 0$. Find the unique value of the scalar, α, that makes $\alpha\,\mathbf{u}$ and $\mathbf{v}$ perpendicular by multiplying both sides of $\mathbf{y} = \alpha\,\mathbf{u} + \mathbf{v}$ by the vector $\mathbf{u}^T$; solve for α.

(c) Compute $\mathbf{v} = \mathbf{y} - \alpha\mathbf{u}$. Show that the Pythagorean Theorem with these 3 vectors proves the variance identity. Note the Pythagorean Theorem states that

$$||\mathbf{y}||^2 = ||\alpha\mathbf{u}||^2 + ||\mathbf{v}||^2$$

or

$$\mathbf{y}^T\mathbf{y} = (\alpha\mathbf{u})^T(\alpha\mathbf{u}) + \mathbf{v}^T\mathbf{v}\,.$$