Statistics 201b Midterm Exam

Rick Paik Schoenberg, 2/23/09, 3:00pm-4:15pm, MS 5128.

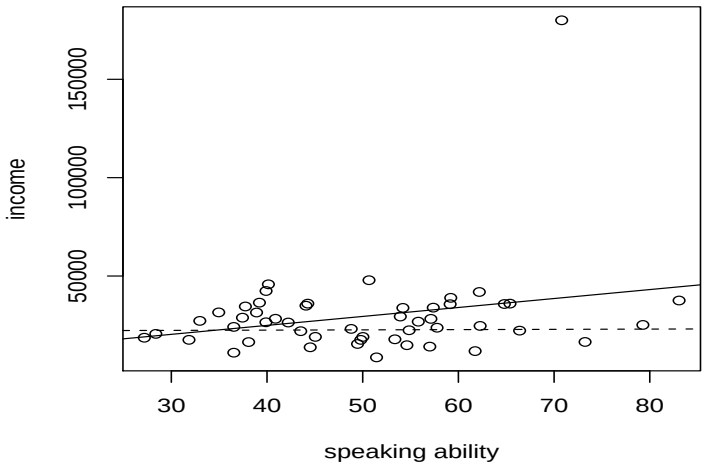
PRINT YOUR NAME:

Do not turn the page and begin the exam until you are instructed to do so.

Absolutely NO looking at other students' exams during the test!

1. Suppose that $Y_i = (X\beta)_i + \epsilon_i$, for $i = 1, 2, ..., n$, where X is fixed (non-random), $E(\epsilon_i) = 0$ and $cov(\epsilon) = V$, where $V \neq \sigma^2 I$. That is, the errors are correlated, and have a non-constant variance. Show that the OLS estimate $\hat{\beta}$ is nevertheless unbiased.

2. Researchers interviewed a simple random sample of 50 Americans and rated their speaking ability (X) and asked them their annual income (Y). A scatterplot of is given below. The solid line is the OLS regression line, and the dashed line is the least trimmed squares regression line.



Why do the two lines differ? What can you conclude? Does improving one's speaking ability seem like a good way to ensure a higher salary?

3. Suppose that Poisson regression with log link function is used to model the relationship between X and Y.

a) Write the equation for predicting Y from X.

b) Suppose that X is univariate, and that an intercept term is used, so that $\eta_i = \beta_0 + \beta_1 X_i$. Suppose also that $\beta_0 = 1$ and $\beta_1 = 2$, and that for observation $i = 17$, $X_{17} = 3$. Under the Poisson regression model with log link function, what is $var(Y_{17})$?

4. Suppose you fit a logistic regression model to predict Y = the probability that a student has taken a statistics course in college, based on X = the student's high school grade point average (GPA), using a sample of 3rd-year students at a certain college. Suppose that the estimated parameters are $\hat{\beta}_0 = 0.20$ for the intercept term and $\hat{\beta}_1 = 0.030$ for the slope.

a) Write out the full model for predicting Y as a function of X.

b) In 1-2 sentences, interpret the estimated slope $\hat{\beta}_1$.

5. Suppose you are trying to predict Y = the speed of a dog, as a function of several physical variables X_i, such as the size of its paws, the size of its head, the length of its tail, the length of its legs, etc. A sample of 50 dogs is used, and for each dog, 100 such physical variables are recorded, as well as its speed.

a) Describe the main 2-3 problems you might expect to face if you are trying to model the relationship between Y and $X_1, X_2, ..., X_{100}$ using linear regression.

b) Suppose that the (sorted) eigenvalues of $X^T X$ are $\lambda_1 = 50.0, \lambda_2 = 27.2, \lambda_3 = 15.9, ..., \lambda_{101} = 0.0010$. What does this indicate? Why?

c) Suppose a group of researchers restrict their attention to only the few most significant explanatory variables, and find that the OLS estimate of the coefficient corresponding to *length of tail* is approximately -0.5. Suppose also that the GLS estimate, the WLS estimate, ridge regression estimates, and m-estimates of this parameter are all approximately -0.5. The partial residual and added variable plots also resulted in a slope of approximately -0.5 as well. Would this represent strong evidence in favor of the claim that "cutting two inches off of a dog's tail would cause the dog to run an additional mph faster"?

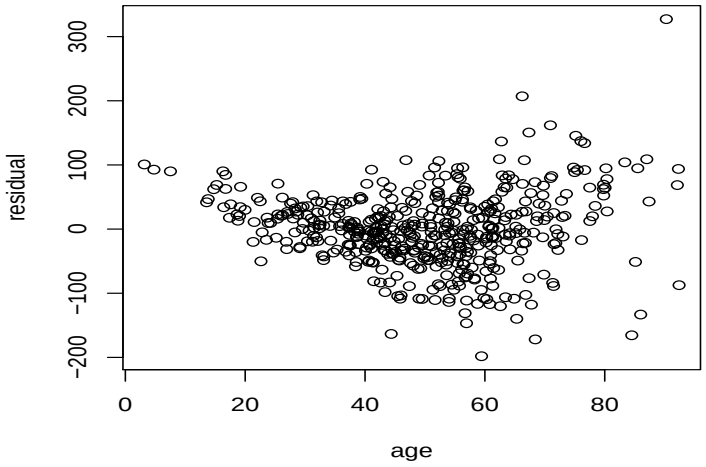
6. Suppose that the OLS model is used to predict Y using X, where X is an $(n \times p)$-matrix, and $n = 10, p = 4$. Suppose that the diagonal elements of H are:

$\{0.4, 0.3, 0.4, 0.9, 0.1, 0.2, 0.4, 0.6, 0.4, 0.3\}$.

Which of the 10 data points have the highest leverage? Would any of them be considered to have high leverage?

7. Suppose researchers study the relationship between the age (X) and efficiency (Y) for a sampling of various washing machines.

For the OLS model $Y = \beta_0 + \beta_1 X + \epsilon$, $E(\epsilon) = 0$, $var(\epsilon) = \sigma^2 I$,
the parameters are estimated by least squares regression, and the residuals are plotted against X below.



What does the residual plot suggest? Which aspects of the model seem to be invalid?

8. Assume the usual assumptions of the linear model, i.e. $Y_i = (X\beta)_i + \epsilon_i$, for $i = 1, 2, ..., n$, where X is fixed (non-random), $E(\epsilon_i) = 0$ and $cov(\epsilon) = \sigma^2 I$. Suppose also that X and Y have mean 0. Show that $cov\{\hat{\epsilon}, \hat{Y}\} = 0$, where $\hat{\epsilon}$ is the vector of residuals in OLS regression, and $\hat{Y} = X\hat{\beta}$ are the fitted Y-values in OLS regression.