STA 302/1001 Summer 2001 Assignment 4

Quiz on June 13th. Do this assignment in preparation for the quiz. Bring a calculator. Bring printouts of the log and list files to the quiz. Note that there will be no formula sheet on Quiz 4. Some of the questions below are intended to provide information about what facts and formulas you need to store in your brain (there are not many). Questions of this sort say "Just write down the answer."

Please begin by reading Chapter 6. You may skip or skim Section 6.8 on diagnostic and remedial measures for now. We'll come back to this material later. Section 6.9 is especially useful because it's based upon the data used in our first SAS lesson on multiple regression. Therefore, you can tell exactly what SAS is doing by comparing the output with Section 6.9. Incidentally, doing a textbook example is *always* a good idea when you're getting acquainted with a new piece of statistical software.

1. Do questions 6.1 and 6.4, and exercises 6.22 and 6.25.
2. Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{c}$ be a vector of constants. What is the distribution of $\mathbf{X} + \mathbf{c}$? Just write down the answer.
3. Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{A}$ be a matrix of constants. What is the distribution of $\mathbf{AX}$? Just write down the answer.
4. Let X_1 be Normal(μ_1, σ_1^2), and X_2 be Normal(μ_2, σ_2^2), independent of X_1. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? What is required for Y_1 and Y_2 to be independent?
5. State the general linear regression model in matrix terms. Just write down the answer. You are being asked for expression (6.19) in the text, except please say $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.
6. Give the formula for the least squares (and maximum likelihood) estimator $\mathbf{b}$. You are being asked for expression (6.25) in the text. Just write down the answer.
7. Give the formulas for $\hat{\mathbf{Y}}$ and $\mathbf{e}$. Just write down the answers.
8. For the general linear regression model in matrix terms, derive the distributions of $\mathbf{Y}$, $\mathbf{b}$, $\hat{\mathbf{Y}}$ and $\mathbf{e}$. These derivations are very short.
9. Assuming that the errror variance σ^2 in the general linear regression model is *known*,
   (a) Derive a $(1-\alpha) * 100\%$ confidence interval for $\mathbf{c}'\boldsymbol{\beta}$, where $\mathbf{c}$ is a $p \times 1$ vector of constants.
   (b) State the statistic that you would use for testing $H_0 : \mathbf{c}'\boldsymbol{\beta} = h$. Remember, you know the value of σ^2, so you're *not* using the t-distribution.
10. Prove
    (a) $\mathbf{e}'\mathbf{X} = \mathbf{0}$
    (b) $\mathbf{e}'\hat{\mathbf{Y}} = \mathbf{0}$

(c) If the model has an intercept, the least squares plane goes through $(\overline{x}_1, \ldots, \overline{x}_{p-1}, \overline{Y})$. You may use $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$ without proof. Hint: what is $\frac{1}{n}\mathbf{X}'\mathbf{1}$?

11. Prove that if the columns of $\mathbf{X}$ are linearly dependent, $(\mathbf{X}'\ \mathbf{X})^{-1}$ does not exist.

12. The following question guides you through the derivation of $\mathbf{b}$ for the general linear regression model. Much of it was done in lecture.

    (a) Here is a substitute for the "umbrella argument." Show that for any $\sigma^2 > 0$, an estimator $\hat{\boldsymbol{\beta}}$ minimizes $Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ if and only if it minimizes the likelihood function over $\hat{\boldsymbol{\beta}}$. Begin as follows. $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \leq (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \Leftrightarrow \ldots$

    (b) Show $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) + [\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})]'[\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})]$. In showing the middle term zero, use $\mathbf{e}'\mathbf{X} = \mathbf{0}$, which you proved above.

    (c) How do you know the expression you just derived is minimized for $\boldsymbol{\beta} = \mathbf{b}$?

    (d) Why does the linear independence of the columns of $\mathbf{X}$ ensure that the minimum is unique?

13. In the last question, you showed that for every fixed $\sigma^2 > 0$, $\mathbf{b}$ maximizes the likelihood function. Complete the derivation of the MLE by setting $\boldsymbol{\beta} = \mathbf{b}$ in the likelihood function, and maximizing over σ^2 to obtain $\hat{\sigma}^2$. Don't forget the second derivative test.

14. The SMSA data set (see p. 704 in your text for a description) is available on the data disk that came with your text. For your convenience, it is also on the course Web site. On the data disk, it is called APC2.DAT; on the Web, it's smsa.dat. Please use SAS to fit a multiple regression model in which the dependent variable is Total Serious Crimes, and the independent variables are Variables 2 through 10. Even though Geographic Region is interesting, I have not yet showed you how to deal with independent variables whose values represent unordered categories. So we'll come back to it later. Use the `simple` option to get simple descriptive statistics for all the variables in your model.

    Your SAS command file must read all the data (including Geographic Region) and provide labels. I mean the `label` statement. You don't need to label the values of Geographic Region using `proc format` (yet). Note that if your output is turned in, you can lose marks by not having labels. Based on your output, be ready to answer questions like the following.

    (a) Which variables have a statistically significant relationship (at $\alpha = 0.05$) to number of crimes, once you control for all the other independent variables in the model? For each one, is the relationship positive or negative?

    (b) Give the value of the test statistic for simultaneously testing whether *any* of the independent variables is useful. The aswer is a number.

    (c) At $\alpha = 0.05$, is that last test significant?

    (d) Give an estimate of the expected number of serious crimes for a SMSA that is average on all the independent variables.

    (e) What proportion of the variation in number of serious crimes is explained by the independent variables in the model?

    (f) Give a 95% confidence interval for the increase in expected number of serious crimes when the number of active physicians in a SMSA goes up by one. You will need to use `proc iml` to get the critical value.