Fall 2008 Eric Slud

Stat 430 Fall 2008 Sample Test Problems

The sample problems on the Fall 2006 sample test are all of the type I said in class I might ask this year.

Of the sample problems on the sample test from 2001, the parts of question II [parts c, e,f] which are based on the multiple regression output are not in scope for this year's mid-term because we have not covered any multiple regression topics yet.

Here are a few more sample problems appropriate for this year's test.

(A). Suppose that you have a `SAS` dataset ABCdat.sas7bdat in your home directory with three numeric columns named A, B, and C. Explain clearly, either by writing SAS code or by verbally summarizing the `SAS` data-steps and `PROC`'s you would use to find

(i) how many records have at least one missing column entry, and

(ii) the mean of C and the correlation of variables A and B over all records which have no missing values.

(B). Four vectors $\mathbf{y}, \mathbf{x}, \mathbf{w}$, and $\mathbf{e}$ of dimension 20 have inner-product values $\mathbf{w}'\mathbf{e} = \mathbf{x}'\mathbf{e} = 0$ along with

$$\mathbf{x}'\mathbf{1} = \mathbf{w}'\mathbf{1} = \mathbf{e}'\mathbf{1} = 0 \;, \quad \mathbf{x}'\mathbf{x} = 5 \;, \quad \mathbf{w}'\mathbf{w} = 10 \;, \quad \mathbf{x}'\mathbf{w} = 5 \;, \quad \mathbf{e}'\mathbf{e} = 6$$

and the vector $\mathbf{y}$ is related to the other vectors by:

$$\mathbf{y} = 2\mathbf{x} + \mathbf{w} + \mathbf{e}$$

(a) Find s_x^2, s_y^2, and the sample correlation $\hat{\rho}_{yx}$ of $\mathbf{y}$ on $\mathbf{x}$..

(b) Find an expression for the residuals of $\mathbf{y}$ and of $\mathbf{x}$ from their simple linear regressions on $\mathbf{w}$.

(c) Find the (sample) partial correlation of $\mathbf{y}$ on $\mathbf{x}$ after removing the effect of $\mathbf{w}$.

(C). Define the following statistical concepts for a univariate column $x_1, \ldots, x_n$ of data values:

(i) *empirical distribution function*.

(ii) *upper quartile* or *75th percentile*.

(iii) *scaled relative frequency histogram* on the range (a,b) (assumed to contain all of the data points).

(D). Suppose you have a numeric column named Y in a SAS dataset XYdat.sas7bdat, together with a group-label column GP which takes on values 1 and 2. If you wanted to explore how the distribution of Y-values differs between the GP=1 and GP=2 groups, explain how to get information on this question using each of the following `SAS PROC`'s, saying as clearly as possible what options you would use and what you would look for in the output and what it could tell you:

(i) `PROC UNIVARIATE` or `PROC BOXPLOT` side-by-side boxplots for Y-values for the two GP-defined groups.

(ii) `PROC FREQ` after breaking Y down into categories YCATG defined by intervals of values in a DATA step.

(iii) `PROC TTEST` on the two GP-defined samples of Y values.