STAT 350

Assignment 6 : FINAL version. Question 2 fixed.

NOTE: Due by 3 PM Friday 1 August in my mailbox in the Statistics and Actuarial Science Department or by email to lhzhao@cs.sfu.ca by 8PM Friday.

1. For the Nitrogen output in Wallabies data set from Assignment 3 do forward, backward, stepwise and all subsets regression.

2. Suppose X_1, X_2, X_3 are independent $N(\mu, \sigma^2)$ random variables, so that $X_i = \mu + \sigma Z_i$ with Z_1, Z_2, Z_3 independent standard normals.

   (a) If $X^T = (X_1, X_2, X_3)$ and $Z^T = (Z_1, Z_2, Z_3)$ express X in the form $AZ + b$ for a suitable matrix A and vector b.

   (b) Show that X is $MVN_3(\mu_X, \Sigma_X)$ and identify μ_X and Σ_X.

   (c) Let $Y_i = X_i - \bar{X}$ for $i = 1, 2, 3$ and $Y_4 = \bar{X}$. Show that $Y \sim MVN_4(\mu_Y, \Sigma_Y)$ and find μ_Y and Σ_Y.

   (d) In class I may have stated that the if the covariance between two components of a multivariate normal vector is 0 then the components are independent, but I indicated a proof only when the multivariate normal distribution in question has a density. In this case the variance matrix is singular so there is no density. However, in terms of the original Z it is possible to find two independent functions of Z such that Y_1, Y_2, Y_3 are a function of the first function while Y_4 is a function of the second.

      i. Let $U_1 = (Z_1 - Z_2)/\sqrt{2}$, $U_2 = (Z_1 + Z_2 - 2Z_3)/\sqrt{6}$ and $U_3 = (Z_1 + Z_2 + Z_3)/3$. Show that $U = (U_1, U_2, U_3)^T$ has a multivariate normal distribution and identify the mean and variance of U.

      ii. Use the result in class, for multivariate normals which have a density to show that (U_1, U_2) is independent of U_3.

      iii. Express Y_3 as a function of U.

      iv. Use the fact that if X_1 and X_2 are independent then so are $G(X_1)$ and $H(X_2)$ for any functions G and H to show that Y_1, Y_2, Y_3 is independent of Y_4.

v. Express the sample variance of the $X_i, i = 1, 2, 3$ in terms of U and use this to show that $(n-1)s_X^2/\sigma^2$ has a χ^2 distribution on 2 degrees of freedom (with $n = 3$). Note: in fact the sample variance of X_1, X_2 is a function of U_1. Generalizations of this idea can be used to develop an identity of the form

$$(n-1)s_n^2 = (n-2)s_{n-1}^2 + U_n^2$$

for a suitable U_n where s_n^2 is the sample variance for $X_1, \ldots, X_n$.

3. In class I discussed the general formula for a multivariate normal density. Suppose that Z_1 and Z_2 are independent standard normal variables. Assume that $X_1 = aZ_1 + bZ_2 + c$ and $X_2 = dZ_1 + eZ_2 + f$. Find the joint density of X_1 and X_2 by evaluating the formulas I gave in class. Express $P(X_1 \le t)$ as a double integral. I want to see the integrand and the limits of integration but you need not try to do the integral.

4. Power and sample size calculations must be done before the data are gathered. However: pilot studies are often used to determine the size of the unknown parameters which are needed for these calculations. Use the Sand and Fibre Hardness data discussed in class as follows.

   (a) Consider the model

   $$Y_i = \beta_0 + \beta_1 S_i + \beta_2 F_i + \beta_3 F_i^2 + \epsilon_i$$

   Fit this model to get estimates of all the βs and of σ. Use these fitted values as if they were the true parameter values in the following.

   i. Compute the power of a two sided t test (at the 5% level) of the hypothesis that $\beta_3 = 0$ for

   $$\beta_3 \in \{-0.006, -0.003, 0, 0.003, 0.006\}.$$

   ii. Find a number m of copies of the basic design (each combination of Sand and Fibre tried twice) to guarantee that the power of the t-test of the hypothesis $\beta_3 = 0$ is 0.9 when the true parameter values are as in the fit above.

(b) Now consider the model

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 F_i + \beta_3 F_i^2 + \beta_4 S_i^2 + \beta_5 S_i F_i + \epsilon_i$$

Fit this model to get estimates of all the βs and of σ. Use these fitted values as if they were the true parameter values in the following.

Find a number m of copies of the basic design (each combination of Sand and Fibre tried twice) to guarantee that the power of the F-test of the hypothesis $\beta_3 = \beta_4 = \beta_5 = 0$ is 0.9 when the true parameter values are as in this fit.