Department of Statistics

Paper 475.330 **Term Test Solutions** **Date: Sept. 23**

1. (a) Height and weight are supplying essentially the same information about catheter length. The results of the t-tests for model (i) indicate that if either one is in the model the other is not necessary. The results for the t-tests for models (ii) and (iii) indicate that on their own both are useful for predicting catheter length.

   (b) i. This indicates that 82.5% of the variability in catheter length (around its mean) can be explained by model (i).

   ii. The F-test is for the null hypothesis that the coefficients for both height and weight are zero. The results give very strong evidence against this hypothesis indicating that at least one of these coefficients is non-zero.

   (c) i. This tells us that observation 6 has the highest leverage.

   ii. The "hats" value is a measure of the potential that an observation has to affect the fitted model. The "cooks.d" value is a measure of the amount the observation actually does affect the fitted model.

   iii. This indicates that the fitted coefficient for weight will change (decrease) by 2.056 standard errors if observation 6 is deleted.

   iv. This indicates that the fitted value for the regression surface at observation 6 (height$=0.98$, weight$=7.7$) would change (increase) by 2.214 standard errors if observation 6 is deleted.

2. (a) Multicollinearity means that there is a (near) linear relationship between some of the explanatory variables. Some of the problems that can occur are (1) the fitted regression surface is unstable–a small change in the data can cause a big change in the fitted surface, (2) the standard errors for the fitted coefficients are inflated making it difficult to evaluate the effects of individual regressors on the response, (3) there may be a number of different subset models that all work approximately as well as each other so selecting the best one is difficult, (4) we need to be very careful not to extrapolate when using the fitted model for prediction. (note: you only needed to give 1 problem).

   (b) The diagnostic quantities used are the variance inflation factors (VIF's) which indicate how much the estimated variances of the fitted coefficients are increased due to multicollinearity. Values near 1 indicate no multicollinearity while larger values indicate that that regressor is involved in a multicollinearity. Any value of approximately 10 or more indicates a serious problem.

3. (a) The value of R^2 always increases when an additional regressor is added to a model which makes it unsuitable for comparing subsets of different sizes - the model containing all the regressors always has the largest R^2. The value of Mallow's C_p does always improve (decrease) when an additional variable is added so it is useful for comparing models with different numbers of regressors.

   (b) For Mallow's C_p any model that contains all the important regressors should have a value near $p+1$. Good models will be those that have a small values of C_p that are close to $p + 1$. A plot C_p versus p is a useful method of picking out the good subset models.

4. (a) To get a model that has the same slope but different intercepts include both lathe speed (numerical regressor) and tool type (factor) in the model.

   (b) To get a model that has different slopes and different intercepts include the lathe speed by tool type interaction as well as lathe speed and tool type.

   (c) Testing whether different slopes are required is the same as testing whether the interaction term is needed in the model. We can use a standard F-test where the full model is the model in (b) and the reduced model is the one in (a). This test is done automatically in S-plus if you apply the anova function to a regression object containing the model from (b).