Name: Key

STAT 704 –Test 2 – Fall 2026

1. A sample of trees was obtained and various measurements were taken on each tree. A multiple regression was performed in order to predict the weight of the tree based on the diameter at breast height (dbh); the height; the age; and the specific gravity (grav) of the wood. SAS output from this regression is given in the Appendix.

(a) For this part (a), please assume that the usual model assumptions are satisfied. Note that it is much easier for scientists to measure the diameter at breast height (dbh) and the height of a tree than is it to age and specific gravity of the tree. Partial output from a model with only the two predictors "height" and "dbh" is also given in the Appendix. Conduct a formal test to determine whether the model with only these two predictors is sufficient, compared to the model with all four predictors. Use $\alpha = 0.05$. State the null and alternative hypotheses, the test statistic value (showing work), the rejection rule, and your conclusion. $H_0: \beta_3 = \beta_4 = 0$ vs. $H_a: \beta_3, \beta_4$ not both zero.

$F^* = \frac{(662861 - 556076)/2}{13240} = 4.03$. Reject H_0 if $F^* > F_{.95,2,42} \approx 3.22$

Reject H_0. Conclude the reduced model is not sufficient.

(b) For the 4-predictor model, examine the given diagnostics and decide whether any model assumptions are violated. If so, which ones? And if so, suggest possible remedial action for the violation, being as specific as possible with your suggestion.

Model misspecification – the linearity assumption is doubtful based on the residual plot. Recommend that "dbh" and "height" should enter the model in a nonlinear fashion.

(c) For the 4-predictor model, are any of the observations are flagged as high-leverage observations? If so, which ones? Justify your answer numerically.

$\frac{2p}{n} = \frac{2(5)}{47} = 0.213$. Observations 6, 7, 14, 16, 17 are high-leverage obs. with $h_{ii} > 0.213$.

(d) For the 4-predictor model, are any of the observations are flagged as influential observations based on the DFFITS criterion? If so, which ones? Justify your answer numerically.

$2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{5}{47}} = 0.652$. Observations 5, 6, 7, 17, 19 have $|DFFITS| > 0.652$ and are thus influential.

(e) Based on the output, how well do you expect that the 4-predictor model would perform when predicting the weights of a new data set consisting of 50 different trees of this type? Justify your answer numerically.

$\frac{PRESS}{SSE} = \frac{774895}{556076} = 1.39$. Not ideal, but should be reasonable predictive performance on new data.

2. A multiple regression analysis was performed in which a sample of 34 observations was used to fit the model. There were 8 candidate predictor variables under consideration. Consider selecting the best model based on the adjusted R^2 criterion. Consider the following four candidate models (all of these models include an intercept term):

Full Model (with 8 predictor variables) has $R^2 = 0.7824$
Medium Model (with 6 predictor variables) has $R^2 = 0.7717$
Small Model (with 5 predictor variables) has $R^2 = 0.7191$

(a) Find the adjusted R^2 for each of these models, and report the best model according to the adjusted R^2 criterion. $n=34$ and $\frac{SSE}{SSTO} = 1-R^2$.

Full $(p=9)$
$R^2_a = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SSTO}\right)$
$= 1 - \left(\frac{33}{25}\right)(.2176)$
$= .7128$

Medium $(p=7)$
$R^2_a = 1 - \left(\frac{33}{27}\right)(.2283)$
$= .7210$

Small $(p=6)$
$R^2_a = 1 - \left(\frac{33}{28}\right)(.2809)$
$= .6689$

$\Rightarrow$ The Medium model is best, having the highest adjusted R^2.

(b) The Mallows' Cp for the Small Model is 10.28, and the Cp for the Medium Model is 6.23. Explain carefully what this indicates about the respective quality of these two models.

Small: $C_p = 10.28 >> p = 6$. Medium: $C_p = 6.23 \approx p = 7$.
$\Rightarrow$ The medium model is preferred, having less bias relative to the full model.

(c) What is the numerical value of the Cp for the Full Model? $C_p = 9$

(d) The AIC of the Medium Model is 247.9836 and the AIC of the Full Model is 250.3508. *First, without calculating the BIC for each of these*, can you be 100% sure whether the Medium Model or the Full Model will have a lower **BIC**? Briefly explain.

Yes. The Medium Model will have a lower BIC. It already has a lower AIC, and BIC penalizes larger models more than AIC does, when $n \geq 8$.

(e) The AIC of the Medium Model is 247.9836 and the AIC of the Full Model is 250.3508. Find the numerical values of the BICs of the Medium Model and the Full Model, and state which model is preferred according to the BIC criterion.

$BIC - AIC = [\ln(n)]p - 2p \Rightarrow BIC = AIC + \ln(n)p - 2p$
$\Rightarrow$ Medium: $BIC = 247.9836 + [\ln(34)]7 - 2(7) = 258.668$
Full: $BIC = 250.3508 + [\ln(34)]9 - 2(9) = 264.088$
$\Rightarrow$ Medium model is preferred, having a lower BIC.

3. (a) State one advantage for each of: (i) using graphical methods, and (ii) using formal tests to detect non-normal errors and non-constant error variance.

(i) Easier to identify the nature of the violation and decide on the right remedy.
(ii) Less subjective, more clear-cut

(b) Plotting residuals against fitted values may reveal whether the error variance is not constant.

(c) Plotting sorted residuals against their expected quantiles under normality may reveal whether the errors are not normally distributed.

(d) The Breush-Pagan (Brown-Forsythe or) test is a formal test for non-constant error variance while the Shapiro-Wilk test is a formal test for non-normal errors.

4. What is the difference between multicollinearity and interaction in a regression model?

Multicollinearity is correlation among the X variables, while interaction is when the effect of one X variable on Y depends on the value of other X variable(s).

5. Why use PRESS or MSPR as measures of a model's predictive ability? Why not simply look at the SSE or MSE or R^2 to determine whether a model is predicting the response variable well?

PRESS and MSPR show how well the model predicts for observations outside the sample the model was fit with.

6. Suppose for a particular model that we refit it using a transformation of Y, say $\sqrt{Y}$. Would the VIFs change in value? Briefly explain.

No - the VIFs are only related to the X variables, not the response.

7. (a) TRUE or FALSE: Multicollinearity is a concern in both simple linear regression and multiple linear regression.

(b) TRUE or FALSE: Among all possible models, the model with maximum adjusted R^2 will always be the model with minimum MSE.

(c) TRUE or FALSE: Among all possible models, the model with maximum R^2 will always be the model with maximum SSR (i.e., regression sum of squares).

(d) TRUE or FALSE: An advantage of centering predictor variables in polynomial regression is that the resulting prediction intervals will be narrower.

8. Note that the $DFFITS_i$ criterion can be rewritten mathematically as:

$$DFFITS_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

where t_i is the externally studentized residual for observation i and h_{ii} is the leverage value for the i-th observation. Explain (specifically based on parts appearing in this representation) why the $DFFITS_i$ criterion is a reasonable measure of an observation's influence on the fitted regression equation.

If the observation is an outlier in Y (high t_i) and has high leverage (large h_{ii}), it is likely to be influential on the regression fit.

9. What does a partial regression (added-variable) plot (say for Y and X_2) tell you that you would not learn from the corresponding scatter plot (of Y versus X_2)? That is, what is the additional information carried within the partial regression plot?

It assesses the relationship between Y and X_2 given the presence of other predictors in the model.

10. Consider the multiple linear regression model with normal errors,

where the (independent) $\varepsilon_i \sim N(0,\sigma^2), i = 1,\ldots,n.$

(a) Let the random variable W represent the sample mean of the n fitted values.

Show that $W = \mathbf{A}\hat{\mathbf{Y}}$ where $\mathbf{A}$ is the $1 \times n$ matrix $\left[\frac{1}{n}\ \frac{1}{n}\ \cdots\ \frac{1}{n}\right]$.

$$\underset{\sim}{A}\,\underset{\sim}{\hat{Y}} = \left[\tfrac{1}{n}\ \cdots\ \tfrac{1}{n}\right]\begin{bmatrix}\hat{Y}_1\\ \vdots\\ \hat{Y}_n\end{bmatrix} = \tfrac{1}{n}\hat{Y}_1 + \cdots + \tfrac{1}{n}\hat{Y}_n = \tfrac{1}{n}\sum \hat{Y}_i = W$$

(b) Using the fact from part (a) and results about random vectors, find both E(W) and Var(W). You may leave your answers in terms of **A**, but otherwise simplify as much as possible (your final answers may be in matrix/vector notation but should contain only constants and model parameters).

$$E(W) = E\left(A\,\underset{\sim}{\hat{Y}}\right) = A\,E\left(\underset{\sim}{\hat{Y}}\right) = A\,E\left(X\underset{\sim}{b}\right) = AX\,E\left(\underset{\sim}{b}\right)$$

$$= AX\underset{\sim}{\beta}$$

$$\text{var}(W) = \text{var}\left(A\,\underset{\sim}{\hat{Y}}\right) = A\,\text{var}\left(\underset{\sim}{\hat{Y}}\right)A' = A\,\text{var}\left(X\underset{\sim}{b}\right)A'$$

$$= AX\,\text{var}\left(\underset{\sim}{b}\right)X'A' = AX\sigma^2(X'X)^{-1}X'A'$$

$$= \sigma^2 AHA'$$

(could also use the relation $\underset{\sim}{\hat{Y}} = H\underset{\sim}{Y}$ to show these)

(c) W is, in general, an unbiased estimator of what parameter?

(A) the sample mean of the values $E(Y_1), \ldots, E(Y_n)$

(B) the sample mean of the values $Y_1, \ldots, Y_n$

(C) The expected value of the distribution of the response variable Y