STATISTICS 3A03

Fall 2017 Dr. Angelo Canty

TERM TEST 1 October 24, 2017.

DAY CLASS
DURATION OF EXAMINATION: 50 Minutes

THIS EXAMINATION PAPER INCLUDES 5 PAGES AND 3 QUESTIONS. YOU ARE RESPONSIBLE FOR ENSURING THAT YOUR COPY OF THE PAPER IS COMPLETE. BRING ANY DISCREPANCY TO THE ATTENTION OF YOUR INVIGILATOR.

Instuctions:

1. Use of the Casio FX-991 calculator only is allowed.
2. Each question is worth 20 marks. Marks for parts of questions are given in the question.
3. Show **all of your work** for full marks!
4. Some useful results and formulae are given on Page 3.
5. A table of critical values of the Student's t distribution is given on Page 4.
6. A table of the 5% critical values of the F distribution is given on Page 5.

Q. 1 Sometimes we wish to fit the following model which has intercept 0 and so is forced to go through the origin

$$Y = \beta_1 X + \varepsilon$$

where ε has mean 0 and variance σ^2 for any value of X.

a) Derive the least squares estimator, $\hat{\beta}_1$ of the slope β_1. **[8]**

b) Show that $\mathrm{E}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right) = \beta_1$. **[6]**

c) Find an expression for $\mathrm{Var}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right)$ in terms of $x_1, \ldots, x_n$ and σ^2. **[6]**

Q. 2 A study was conducted to examine the relationship between height (X) and weight (Y). A sample of $n = 10$ 18 year-old girls was taken and the following summary statistics found:

$$\begin{aligned} \overline{x} &= 165.52 & \overline{y} &= 59.47 \\ S_{xx} &= 472.076 & S_{yy} &= 731.961 \\ S_{xy} &= 274.786 \end{aligned}$$

a) Calculate the correlation coefficient, r, between height and weight. **[4]**

b) Consider the model to predict weight from height. Find point estimates of the intercept and slope. **[6]**

Continued on Page 2

c) Show that the error sum of squares can be written as

$$SSE = S_{yy}(1 - r^2)$$

and hence give an unbiased estimator of the error variance. **[8]**

d) What proportion of the variability in weight can be explained by the model? **[2]**

Q. 3 A study published in 1983 gathered data on 25 patients with Cystic Fibrosis. A variable, called `pemax` measures the maximum lung function in the patients. A possible model for this uses age, height, weight and total lung capacity (`tlc`) as covariates. The following is partial output from SAS for this model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	a	d	2853.05	g
Error	b	e	f	
Corrected Total	c	26833		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	23.99755	102.42243	0.23	0.8171
age	1	1.89829	3.22053	0.59	0.5622
height	1	-0.02630	0.81636	-0.03	0.9746
weight	1	0.83297	0.87192	0.96	0.3508
tlc	1	0.26021	0.37987	0.69	0.5012

a) Fill in the 7 missing numbers a–g in the ANOVA table. **[7]**

b) Give the null and alternative hypotheses that the F Value is testing. Do we reject this null hypothesis at the 5% significance level? **[4]**

c) Does the result in part b) contradict the results of the t-tests in the Parameter Estimates table? Explain. **[4]**

d) Below is another ANOVA table for the model that uses only age and tlc as covariates. Test the null hypothesis that height and weight are not needed in the model and state your conclusion using the 5% significance level. **[5]**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10488	5243.79552	7.06	0.0043
Error	22	16345	742.95677		
Corrected Total	24	26833			

Continued on Page 3

Some Useful Results and Formulae

- $S_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$. $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{x}\,\overline{y}$.
- The population and sample correlation coefficients are

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \qquad r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- The least squares estimates for the simple linear model $Y = \beta_0 + \beta_1 x + \varepsilon$ are

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

- An unbiased estimator of the error variance σ^2 in a simple linear model is

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-2} = \frac{\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

- The standard error of the estimators in a simple linear model are

$$\text{s.e.}\left(\hat{\beta}_0\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} \qquad \text{s.e.}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

- For multiple regression with design matrix $\boldsymbol{X}$ the least squares estimators are $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y}$.
- The variance-covariance matrix of the estimators in a multiple regression is $\mathrm{Var}(\boldsymbol{\beta}) = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\sigma^2$
- An unbiased estimator of σ^2 is $\hat{\sigma}^2 = \mathrm{SSE}/(n-p-1)$.
- For any coefficient in a linear model with $p \geqslant 1$ covariates

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim t_{n-p-1} \qquad j = 0, 1, \ldots p$$

- The coefficient of determination is $R^2 = 1 - \mathrm{SSE}/S_{yy}$.
- To test if a reduced model is sufficient compared to a full model we use the test statistic

$$F = \frac{(\mathrm{SSE}_{\text{red}} - \mathrm{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\mathrm{SSE}_{\text{full}}/df_{\text{full}}}$$

  where the degrees of freedom are the error degrees of freedom in the models. This statistic has an F distribution with $df_{\text{red}} - df_{\text{full}}$ and df_{full} degrees of freedom if the reduced model is sufficient.

Continued on Page 4