total points = 247+3 = 250

STA 6167 – Exam 1 – Fall 2019 – **PRINT** Name ANSWER KEY

For all significance tests, use $\alpha = 0.05$ significance level.

Q.1. Two multiple linear regression models were fit relating price of art works (Y = log(sale price)) to the following predictors: surface area (SA) of the object, the medium of the object (collage, drawing, painting*, photograph, print, sculptures). There were 5 dummy variables for medium ($M_1,\ldots,M_5$), with painting being the reference category. The first model had a linear trend for year (t), while the second model had 12 dummy variables ($Yr_1,\ldots,Yr_{12}$) for the 13 individual years (thus not forcing the trend to be linear). The models and results are given below, based on a sample of n = 518 artworks sold during the 13 year period 1997-2009.

Model 1: $E\{Y\} = \beta_0 + \beta_{SA}SA + \sum_{i=1}^{5}\beta_{M_i}M_i + \beta_t t \quad R_1^2 = .502$

Model 2: $E\{Y\} = \beta_0 + \beta_{SA}SA + \sum_{i=1}^{5}\beta_{M_i}M_i + \sum_{i=1}^{12}\beta_{Yr_i}Yr_i \quad R_2^2 = .555$ 2 2

p.1.a. Give the number of parameters for the models. Model 1: 8 Model 2: 19

p.1.b. For Model 1, test H_0: $\beta_{SA} = \beta_{M1} = \beta_{M2} = \beta_{M3} = \beta_{M4} = \beta_{M5} = \beta_t = 0$

8

$$F = \frac{R^2/p}{(1-R^2)/(n-p')} = \frac{.502/7}{.498/510} = \frac{510(.502)}{7(.498)} = 73.44$$

$F_{.05,7,510} \approx 2.010$

73.44

Test Statistic $F = 73.44$ Rejection Region $F \geq 2.040$ P(<) or > 0.05

3 2

p.1.c. Model 1 is a special case of Model 2, with the yearly trend being a straight line, while Model 2 allows any structure for the year effects. Based on comparing Complete and Reduced models, test between the following hypotheses.

H_0: Model 1 is appropriate (linear trend) versus H_A: Model 2 is appropriate (trend is not linear)

9

$$F = \frac{\frac{.555-.502}{19-8}}{\frac{1-.555}{499}} = \frac{.053(499)}{.445(11)} = 5.40$$

$F_{.05,11,499} \approx 1.789$

Test Statistic $F = 5.40$ Rejection Region $F \geq 1.789$ P(<) or > 0.05

30 3 2

Q.2. A regression model was fit, relating the heat capacity of solid hydrogen bromide (Y, in cal/(mol*K)) to Temperature (X, in degrees Kelvin) based on n=18 experimental runs. The temperatures were centered (for computational reasons), but this has no effect on predicted values or Sums of Squares. The following 3 models are fit where the mean temperature was 145.16.

Model 1: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X})$ $\hat{Y}^1 = 11.2756 + 0.0216(X - 145.16)$ $SSE_1 = 0.1945$ $SSR_1 = 3.3889$

Model 2: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2$ $\hat{Y}^2 = 11.1596 + 0.0192(X - 145.16) + 0.00029(X - 145.16)^2$ $SSE_2 = 0.0370$ $SSR_2 = 3.5465$

Model 3: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2 + \beta_3(X - \overline{X})^3$

$\hat{Y}^3 = 11.1718 + 0.0155(X - 145.16) + 0.00021(X - 145.16)^2 + 0.0000059(X - 145.16)^3$ $SSE_3 = 0.0172$ $SSR_3 = 3.5662$

p.2.a. For Model 3, Test $H_0: \beta_1 = \beta_2 = \beta_3 = 0$.

$$F = \frac{3.5662/3}{.0172/(18-4)} = \frac{14(3.5662)}{.0172(3)} = 961.98$$ 8

$F_{.05,3,14} = 3.344$

3 2

Test Statistic: $F = 961.98$ Rejection Region: $F \geq 3.344$ P-value > or < .05

p.2.b. What proportion of the total variation in Y is "explained" by the predictors in Model 2.

5 $$\frac{3.5465}{3.5465 + .0370} = \frac{3.5465}{3.5835} = \cancel{.9897}\ .9897$$

p.2.c. Give the predicted heat capacities for temperatures X=125.16, 145.16, and 165.16 for each Model.

1: $11.2756 \pm 20(.0216) = 11.2756 \pm .\cancel{2}$ .4320

2. $11.1596 \pm 20(.0192) + 400(.00029) = 11.1596 \pm .3840 + .1160$

3. $11.1718 \pm 20(.0155) + 400(.00021) \pm 8000(.0000059)$

$= 11.1718 \pm .3100 + .0840 \pm .0472$

3 M1: 125.16: ~~1~~ 10.8436 145.16: 11.2756 165.16: ~~11~~ 11.7076 3

3 M2: 125.16: 10.8916 145.16: 11.1596 165.16: 11.6596 3

3 M3: 125.16: 10.8986 145.16: 11.1718 165.16: 11.6130 3

42

2 each

Q.3. An experiment was conducted relating viscosity of flour used in baking ice cream cones (Y, in degrees MacMichael) to the conents of moisture (X_M, in %), protein (X_P, in %), and ash (X_A, in percent) for $n = 39$ flours obtained from different flour mills. The following models were fit, with the results for Model 3 given below. All models assume errors are independent and normally distributed.

Model 1: $Y = \beta_0 + \beta_A X_A + \varepsilon$ Model 2: $Y = \beta_0 + \beta_P X_P + \beta_A X_A + \varepsilon$

Model 3: $Y = \beta_0 + \beta_M X_M + \beta_P X_P + \beta_A X_A + \varepsilon$

ANOVA	3 each		2 each	2	~~3~~ 4			Coefficient	tandard Err	t Stat (2 each)	t(.025) (3)	
	df	SS	MS	F	F(.05)		Intercept	-115.36	63.29	-1.82	≈ 2.03	
Regression	3	24094.91	8631.64	27.65	~~0.0000~~		moisture	4.15	4.38	.95		NS
Residual	35	10164.83	290.42		≈ 2.88		protein	19.99	2.76	7.24		✓
Total	38	34259.74					ash	-128.86	15.37	-8.38		✓

p.3.a Compute the coefficient of determination, R^2 for the above model (Model 3).

$$\frac{24094.91}{34259.74} = .7033$$

(4)

p.3.b. Complete the ANOVA and regression coefficients tables and test 1) whether the Viscosity is related to any of the content variables $H_0{:}\beta_M = \beta_P = \beta_A = 0$ and 2) whether Viscosity is related to the individual content variables, controlling for the others H_0: $\beta_i = 0$.

(4) Reject $H_0: \beta_M = \beta_P = \beta_A = 0$ Reject $H_0: \beta_P = 0$, $H_0: \beta_A = 0$ Not $H_0: \beta_M = 0$

p.3.c. The Regression Sums of Squares for Models 1 and 2 are $SSR_1 = 8869.33$ and $SSR_2 = 23834.15$, respectively. Give the following sequential sums of squares.

23834.15 − 8869.33 = 14964.82

24094.91 − 23834.15 = 260.76

(3) (4) (4)

$SSR(X_A) =$ 8869.33 $SSR(X_P \mid X_A) =$ 14964.82 $SSR(X_M \mid X_A, X_P) =$ 260.76

p.3.d. Compute $R^2_{YX_P \bullet X_A}$ (the coefficient of partial determination between Y and X_P, given X_A).

(6)

$$\frac{14964.82}{34259.74 - 8869.33} = \frac{14964.82}{25390.41} = .5894$$

50

Q.4. A model was fit, relating US annual energy consumption to the following set of predictors: X_1 = **GDP**, X_2 = price of electricity (**pElec**), X_3 = **pop**ulation, X_4 = price of natural gas (**pNatGas**), and X_5 = price of heating oil (**pHeatOil**). A second model is fit, with only **GDP** (X_1) and **pElec** (X_2). The models were fit for the years 1984-2010.

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \quad SSE_1 = 2.860 \quad SSR_1 = 100.752 \quad \sum_{t=2}^{27}(e_t - e_{t-1})^2 = 5.608$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad SSE_2 = 3.038 \quad SSR_2 = 100.574$

p.4.a. The critical values for the Durbin-Watson statistic for $n = 27$ and $p = 5$ are $d_L = 1.01$ and $d_U = 1.86$. Compute the Durbin-Watson statistic for testing H_0: the errors are not autocorrelated and circle the best conclusion.

$$DW = \frac{5.608}{2.860} = 1.96$$

(4) (3)

D-W Statistic: 1.96 Conclude: Reject H_0 Accept H_0 Inconclusive

p.4.b. Compute SSR($X_3,X_4,X_5 \mid X_1,X_2$)

$$100.752 - 100.574 = 0.178$$

(4)

p.4.c. Test H_0: $\beta_3 = \beta_4 = \beta_5 = 0$

(8)

$$F = \frac{0.178/3}{2.860/(27-6)} = \frac{21(.178)}{3(2.860)} = 0.4357$$

$$F_{.05,3,21} = 3.072$$

(3) (2)

Test Statistic: $F = .4357$ Rejection Region: $F \geq 3.072$ P-value > or < .05

24

Q.5. An experiment was conducted relating energy consumption (Y, in MJ) to fiber space velocity (X, in m/h) in a carbon fiber production process. There were $c = 4$ distinct fiber space velocity "groups", with varying n_j runs per group. The lack-of-fit test for a linear relation is:

$$H_0 : E\{Y_{ij}\} = \mu_j = \beta_0 + \beta_1 X_j \quad i = 1,\dots,n_j;\ j = 1,\dots,4 \quad H_A : E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

ANOVA											
	df	SS	MS	F	gnificance F		fsv	n_grp	yhat_grp	ybar_grp	s_grp
Regressio	1	47.1060	47.1060	809.1265	0.0000		20	8	7.5625	7.7913	0.1283
Residual	28	1.6301	0.0582				25	9	6.4686	6.2143	0.0784
Total	29	48.7361					30	5	5.3747	5.1922	0.0802
							35	8	4.2801	4.4523	0.0735
	Coefficients	tandard Err	t Stat	P-value							
Intercept	11.9381	0.2135	55.9060	0.0000							
fsv	-0.2188	0.0077	-28.4451	0.0000							

p.5.a. Give the fitted value for the linear regression for the 4th group ($X_4 = 35$).

$\hat{Y}_{35} = 11.9381 - .2188(35) = 11.9381 - 7.658 = 4.2801$

3

p.5.b. Compute the Pure Error Sum of Squares, degrees of freedom and Mean Square.

$SSPE = (8-1)(.1283)^2 + (9-1)(.0784)^2 + (5-1)(.0802)^2 + (8-1)(.0735)^2$

$= .1152 + .0492 + .0257 + .0378 = .2279$

8

$df = (8+9+5+8) - 4 = 26$

SS_{PE} = .2279 df_{PE} = 26 3 MS_{PE} = .0088 2

p.5.c. Compute the Lack-of-Fit Sum of Squares, degrees of freedom and Mean Square.

$8(7.5625 - 7.7913)^2 + 9(6.4686 - 6.2143)^2 + 5(5.3747 - 5.1922)^2$

$+ 8(4.2801 - 4.4523)^2 = .4188 + .5820 + .1665 + .2372 = 1.4045$

$df = 4 - 2 = 2$

SS_{LF} = 1.4045 8 df_{LF} = 2 3 MS_{LF} = .7023 2

p.5.d. Give the Test Statistic, Rejection Region, and P-value relative to .05 for the Lack-of-Fit test.

$F = \frac{MSLF}{MSPE} = \frac{.7023}{.0088} = 79.81 \qquad F_{.05,2,26} = 3.369$

Test Statistic: $F = 79.81$ Rejection Region: $F \geq 3.369$ P-value > or < .05

4 3 2

38

Q.6. A study related height (Y, in cm) to foot length (X, in cm) among n = 5195 adult South Koreans of ages 20 to 59. A dummy variable (M = 1 if male, 0 if female) is created to reflect subject's gender. Three models are fit (each assuming independent, normally distributed errors with constant variance).

Model 1: $E\{Y\} = \beta_0 + \beta_1 X$ $\quad \hat{Y}^1 = 45.609 + 4.947X \quad SSE_1 = 116921 \quad SSR_1 = 313347$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X + \gamma_1 M$ $\quad \hat{Y}^2 = 65.574 + 4.031X + 3.857M \quad SSE_2 = 108416.5 \quad SSR_2 = 321851.5$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X + \gamma_1 M + \delta_1 XM$ $\quad \hat{Y}^3 = 66.910 + 3.972X + 1.577M + 0.096XM \quad SSE_3 = 108404 \quad SSR_3 = 321864$

p.6.a. Give the predicted heights for females and males with foot lengths of 23 and 25 cm based on model 3.

F/23 : $66.910 + 3.972(23) = 66.910 + 91.356 = 158.266$

F/25 : $158.266 + 2(3.972) = 158.266 + 7.944 = 166.210$

M/23 : $(66.910 + 1.577) + (3.972 + 0.096)X^{(23)} = 68.487 + 4.068(23)$

$= 68.487 + 93.564 = 162.051$ $\quad$ M/25: $162.051 + 2(4.068) = 170.187$

4 each

F/23: 158.266 $\quad$ F/25: 166.210 $\quad$ M/23: 162.051 $\quad$ M/25: 170.187

p.6.b. Based on models 1 and 2 test whether males and females differ in mean height, controlling for foot length.

H

H_0: $\gamma_1 = 0$ $\quad H_A$: $\gamma_1 \neq 0$

8

$$F = \frac{\frac{SSE_1 - SSE_2}{1}}{\frac{SSE_2}{5195-3}} = \frac{\frac{116921 - 108416.5}{1}}{\frac{108416.5}{5192}} = \frac{8504.5}{20.88} = 407.28$$

$F_{.05,1,5192} \approx 3.841$

3 $\quad$ 2

Test Statistic: $F = 407.28$ $\quad$ Rejection Region: $F \geq 3.841$ $\quad$ P-value > or < .05

p.6.c. Based on models 2 and 3 test whether the slopes with respect to foot length differ for males and females.

H_0: $\delta_1 = 0$ $\quad H_A$: $\delta_1 \neq 0$

$$F = \frac{\frac{108416.5 - 108404}{1}}{\frac{108404}{5195-4}} = \frac{12.5}{20.88} = 0.60$$

8

3 $\quad$ 2

Test Statistic: $F = 0.60$ $\quad$ Rejection Region: $F \geq 3.841$ $\quad$ P-value > or < .05

42

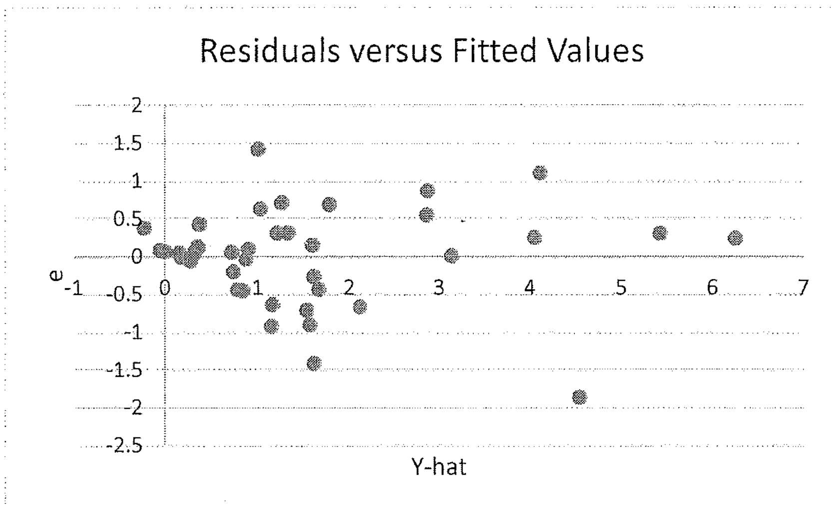
Q.7. A regression model is fit, relating mobility (Y) to six predictor variables: GDP (X_1), vehicles/km of road (X_2), population density (X_3), percent urban population (X_4), land area (X_5), and population (X_6) for $n = 38$ island nations. The Analysis of Variance for the multiple linear regression model is given below.

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	6	87.57	14.60	28.83	0.0000
Residual	31	15.69	0.51		
Total	37	103.26			

p.7.a. A plot of the residuals versus predicted values is given below. It demonstrates which possible violations of assumptions (circle all that apply).

Non-normal Errors Unequal Variance Serial Correlation of Errors Non-linear Relation between Y and X

2 each



p.7.b. A second regression model is fit, relating the squared residuals (Y) to the 6 predictors ($X_1,\ldots,X_6$). Conduct the Breusch-Pagan test to test whether the equal variance assumption is reasonable. The sums of squares are given below.

ANOVA		
	df	*SS*
Regressio	6	269.47
Residual	31	166.87
Total	37	436.34

$$X^2_{BP} = \frac{SS_{Reg^*}/2}{(SSE/n)^2} = \frac{269.47/2}{(15.69/38)^2} = \frac{134.735}{.1705}$$

$$= 790.32 \qquad \chi^2_{.05,6} = 12.592$$

(8)

(3)

(2)

Test Statistic: $X^2_{BP} = 790.32$ Rejection Region: $X^2_{BP} \geq 12.592$ P-value > or < .05

21