ST590

Fall Semester, 2019

Exam 1

Name:_____

Directions: Answer questions as directed. Please show work. Partial credit may be awarded for correct expressions given in incomplete answers. To save time, it is perfectly ok to give answers as numeric expressions without carrying out every last operation on the calculator. If the question asked for the standard error of the estimated mean of a population at $x = 14$ in a simple linear regression, the following response would receive full credit:

$$\widehat{SE}(\hat{\beta}_0 + 14\hat{\beta}_1) = \sqrt{16.4\left(\frac{1}{38} + \frac{(14 - 10.8)^2}{190}\right)}$$

1. (21 pts) For $n = 9$ randomly sampled school districts, a simple linear regression model of eighth grade math score (y) on per-pupil expenditures (x, in \$K) was fit with partial output below. The model is

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ with } E_i \overset{iid}{\sim} N(0, \sigma^2).$$

The SAS System
The REG Procedure

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		172.80000	172.80000		
Error		126.00000	18.00000		
Corrected Total	8	298.80000			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	230.00000	13.24387	17.37	<.0001
x	1	6.00000	1.93649	3.10	0.0174

(a) Report the coefficient of determination for this regression model.

(b) Report the p-value and F-ratio from a test of $H_0 : \beta_1 = 0$.

(c) Obtain the fitted value, $\widehat{y}_3$ for the third observation below ($x_3 = 6.8 = \overline{x}$). Also, obtain residuals for the third and last observations in the table below. Write them in the blank provided or in the space to the right.

x	y	fitted	resid
5.6	260.6	263.6	-3
6.4	262.4	268.4	-6
6.8	270.8	_____	_____
7.2	279.2	273.2	6
8.0	275.0	278.0	-3
6.0	269.0	266.0	3
6.4	271.4	268.4	3
7.2	276.2	273.2	3
7.6	272.6	275.6	_____

(e) Describe briefly how to obtain a quantile-quantile plot to check for normality. In particular, identify the two residuals at the extreme top and bottom of the plot, and what values against which they would be plotted. (You may refer to quantiles from known distributions without looking them up in a table.)

(f) The mean expenditure in the sample was $\overline{x} = 6.8$ and the estimated mean response at this value is $\widehat{\mu}(x = \overline{x}) = \overline{y}$. It can be shown that $SD(\widehat{\mu}(x = \overline{x})) = \sigma/\sqrt{n}$. Report an estimate, $\widehat{SE}$ of this standard error.

(g) The 97.5^{th} percentile from the appropriate t distribution is $t = 2.36$. Use it, along with the $\widehat{SE}$ term from problem (f) to construct a 95% confidence interval for the mean math score among districts who spend the average amount ($\overline{x}$) per pupil.

(h) Obtain a 95% prediction interval for the grades from one such school district with $x = \overline{x}$ sampled at random.

2. (15 pts) A bivariate random sample $(x_1, y_1), \ldots, (x_{16}, y_{16})$ led to an observed average of $\overline{y} = 10$. The observed standard deviations were $s_x = 3$ and $s_y = 2$. The observed correlation between x and y was $r = 0.6$

   (a) Report the least squares estimate of the slope in a simple linear regression of y on x.

   (b) Estimate the population mean of the response y when x is one standard deviation below its average from the sample, $\overline{x}$, which doesn't need to be given to solve this problem.

   (c) Estimate the difference in the mean estimated in part (b) and the estimated mean when $x = \overline{x}$.

3. (15 pts) An experiment in veterinary medicine involves $N = 20$ cats suffering from the same degenerative hip condition. They are assigned to $t = 4$ treatment groups at random:

- FHO Surgery, plus physical therapy
- FHO Surgery, no physical therapy
- THR Surgery, plus physical therapy
- THR Surgery, no physical therapy

A primary outcome (y) from the surgery is leg extension on the operated side after 12 months. The means, standard deviations and variances from these four treatment groups are given below:

Analysis Variable : extension

trt	N Obs	N	Mean	Std Dev	Variance
FHO,PT	5	5	146.1000000	5.7706152	33.3000000
FHO,no PT	5	5	142.0000000	6.3146655	39.8750000
THR,PT	5	5	149.5000000	3.9051248	15.2500000
THR,no PT	5	5	143.8000000	7.9733933	63.5750000

(a) What is the name of this experimental design ? _____

(b) Complete the ANOVA table below and construct an F-ratio for testing for an effect of the surgery-by-PT treatment combination:

Source	*df*	Sum of squares	Mean square	F-ratio
Treatments	3	157.05		
Error		608		
Total		765.05		

(c) Under the model $Y_{ij} = \mu + \tau_i + E_{ij}$ with $E_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, what is the expectation of the error mean square: $E(MS(E)) =$ _____?

4. (19 pts) Consider further the experiment with cats undergoing hip surgery. Two solution vectors for the parameters $(\mu, \tau_1, \ldots, \tau_4)$ for the model $Y_{ij} = \mu + \tau_i + E_{ij}$ are given as output below.

```
The GLM Procedure (top)
                                              Standard
Parameter                  Estimate              Error    t Value    Pr > |t|

Intercept              143.8000000 B       2.75680975      52.16      <.0001
trt       FHO,PT         2.3000000 B       3.89871774       0.59      0.5635
trt       FHO,no PT     -1.8000000 B       3.89871774      -0.46      0.6505
trt       THR,PT         5.7000000 B       3.89871774       1.46      0.1631
trt       THR,no PT      0.0000000 B        .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used
      to solve the normal equations.  Terms whose estimates are followed by the letter
      'B' are not uniquely estimable.

The GLM Procedure (bottom)
                                              Standard
Parameter                  Estimate              Error    t Value    Pr > |t|

Intercept              146.1000000 B       2.75680975      53.00      <.0001
trt       FHO,PT         0.0000000 B        .                .         .
trt       FHO,no PT     -4.1000000 B       3.89871774      -1.05      0.3086
trt       THR,PT         3.4000000 B       3.89871774       0.87      0.3961
trt       THR,no PT     -2.3000000 B       3.89871774      -0.59      0.5635

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used
      to solve the normal equations.  Terms whose estimates are followed by the letter
      'B' are not uniquely estimable.
```

(a) Use this output to complete the table of estimates below:

	Parameter		
Parameterization	$\tau_3 - \tau_4$	$\mu + \tau_3$	τ_3
top	$5.7 - 0 = 5.7$		
bottom		$146.1 + 3.4 = 149.5$	
Label			

(b) In the table above, write UE or NUE beneath each column to indicate whether the linear combination of parameters in that column is uniquely estimable (UE) or not (NUE).

5. (30 points) Three measurements are made on each of $n = 31$ trees randomly sampled from a population of interest: $y = volume$ (in cubic ft), $x_1 = girth$ (in inches), $x_2 = height$ (in feet). Let X denote the design matrix for a multiple linear regression model of y on x_1 and x_2:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i$$

The partial output at the end of the exam was generated using the SAS code below. Use it to answer the given questions.

```
proc reg;
   model volume=girth height/xpx i ss1 ss2;
run;
```

(a) Enter here the hidden part of the output labelled AAA. _____

(b) Enter here the hidden part of the output labelled BBB. _____

(c) Enter here the hidden part of the output labelled CCC. _____

(d) Specify the dimension of the design matrix X._____

(e) Give the matrix product $(X'X)^{-1}X'Y$

(f) Give the extra sum of squares for girth after controlling for height, $R(\beta_1|\beta_0, \beta_2)$.

(g) Give the regression sum of squares for a simple linear regression of volume on height only.

(h) What is the squared correlation between the observed values (y_i) and the fitted values ($\widehat{y}_i$) from the multiple linear regression? What is this coefficient called?

(i) When the product girth $\times$ height is added to the model, the least squares regression equation becomes

$$\hat{y} = 69.4 - 5.86x_1 - 1.3x_2 + 0.135x_1x_2$$

and the unexplained error is quantified by $SS[E] = 198$ on 27 df. Formulate a test comparing the additive model with the interactive model. Specify a null hypothesis H_0 and report an F-ratio, along with associated degrees of freedom.

(j) Consider trees with $x_1 = 10$. Use the fitted model from (i) to report the least squares regression line for estimated volume as a function of x_2 for these trees:

$$\mu(x_2) = \_\_\_\_\_ + \_\_\_\_\_ x_2$$

The SAS System
The REG Procedure

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	girth	height	volume
Intercept	AAA_____	410.7	2356	935.3
girth	BBB_____	5736.55	31524.7	13887.86
height	2356	31524.7	180274	72962.6
volume	935.3	13887.86	72962.6	36324.99

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	girth	height	volume
Intercept	4.9519429276	0.028680223	-0.069732257	-57.98765892
girth	0.028680223	CCC_____	-0.001185265	4.708160503
height	-0.069732257	-0.001185265	0.0011241461	0.3392512342
volume	-57.98765892	4.708160503	0.3392512342	421.92135922

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7684.16251	3842.08126	254.97	<.0001
Error	28	421.92136	15.06862		
Corrected Total	30	8106.08387			

Root MSE	3.88183	R-Square	0.9480
Dependent Mean	30.17097	Adj R-Sq	0.9442

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-57.98766	8.63823	-6.71	<.0001	28219	679.04025
girth	1	4.70816	0.26426	17.82	<.0001	7581.78133	4782.97364
height	1	0.33925	0.13015	2.61	0.0145	102.38118	102.38118