Exam 1 Solutions

1. For $n = 9$ randomly sampled school districts, a simple linear regression model of eighth grade math NAEP (Natl Assessment of Educ. Progress) score (y) on per-pupil expenditures (x, in \$K) was fit ($\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$) with partial output below:

```
The SAS System
The REG Procedure
                              Analysis of Variance

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

                          Parameter        Standard
Variable      DF           Estimate           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.

   $\boxed{r^2 = 172.8/298.8 = 0.58}$

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

   $\boxed{F = 172.8/(126/7) = 9.5 (= t^2) \text{ with } p = .0174}$

   (c) Obtain the fitted value, $\widehat{y}$ for math scores among schools spending \$6800 per pupil. Also, obtain residuals for the third and last observations in the table below:

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	-----

   The fitted value is $\widehat{\beta}_0 + \widehat{\beta}_1(6.8) = 230 + 6(6.8) = 270.8$. Residuals are $e_3 = 0$ and $e_9 = -3$.

(e) Describe briefly how to obtain a quantile-quantile plot to check for normality. ...

The sorted residuals are plotted against the corresponding quantiles from $N(0, 1)$. The most negative residual, -6, is plotted against the 10^{th} percentile, $z_{.9} = -1.28$ and the largest residual, 6 is plotted against the 10^{th} percentile, $z_{.1} = 1.28$.

(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.

$\sigma/\sqrt{n}$ can be estimated by $SE = \sqrt{MS(E)/9} = \sqrt{2}$

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

$$\overline{y} \pm 2.36SE \text{ or } 270.8 \pm 2.36\sqrt{2} \text{ or } 270.8 \pm 3.34$$

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

$$\overline{y} \pm 2.36\sqrt{SE^2 + MS(E)} \quad \text{or} \quad 270.8 \pm 2.36\sqrt{2 + 18} \text{ or } 270.8 \pm 10.6$$

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. $\widehat{\beta}_1 = r\frac{s_y}{s_x} = 0.6\frac{2}{3} = 0.4$

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

$$\begin{aligned}
\widehat{\mu}(x = \overline{x} - s_x) &= \widehat{\beta}_0 + \widehat{\beta}_1(\overline{x} - s_x) \\
&= \overline{y} - \widehat{\beta}_1\overline{x} + \widehat{\beta}_1(\overline{x} - s_x) \\
&= \overline{y} - \widehat{\beta}_1 s_x \\
&= \overline{y} - (0.4)(3) \\
&= 10 - 1.2 = 8.8
\end{aligned}$$

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

$$\begin{aligned}
\widehat{\mu}(x = \overline{x} - s_x) &= \overline{y} - r\frac{s_y}{s_x}s_x \\
\widehat{\mu}(x = \overline{x}) &= \widehat{\beta}_0 + \widehat{\beta}_1(\overline{x}) \\
&= \overline{y} - \widehat{\beta}_1\overline{x} + \widehat{\beta}_1(\overline{x}) \\
&= \overline{y} \\
\widehat{\mu}(\overline{x}) - \widehat{\mu}(\overline{x} - s_x) &= -rs_y = -0.6(2) = -1.2
\end{aligned}$$

3. 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

                          N
trt                     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? Completely Randomized Design

(b) Complete the ANOVA table below and (b) 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	52.35	1.38
Error	16	608	38	
Total	19	765		

(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: $E(MS(E)) = \underline{\sigma^2}$.

4. 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:

Parameterization	Parameter: $\tau_3 - \tau_4$	$\mu + \tau_3$	τ_3
top	$5.7 - 0 = 5.7$	$143.8 + 5.7 = 149.5$	5.7
bottom	$3.4 - (-2.3) = 5.7$	$146.1 + 3.4 = 149.5$	3.4
Label	UE	UE	NUE

(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. $\boxed{n = 31}$

(b) Enter here the hidden part of the output labelled BBB. $\boxed{\sum x_{i1} = 410.7}$

(c) Enter here the hidden part of the output labelled CCC. Since $SE(\widehat{\beta}_1) = .26$ and $V(\widehat{\beta}) = (X'X)^{-1}MS(E)$, we can solve for the entry in $(X'X)^{-1}$:

$$\text{CCC} \quad MS(E) = 0.26^2 \Longrightarrow CCC = .26^2/MS(E) = .0045$$

(d) Specify the dimension of the design matrix X. $\boxed{31 \times 3}$

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

$$\widehat{\beta} = \begin{pmatrix} -58 \\ 4.7 \\ 0.34 \end{pmatrix}$$

(f) Give the extra sum of squares for girth after controlling for height, $R(\beta_1|\beta_0, \beta_2)$. Type II SS for girth: 4783.

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

$$R(\beta_2|\beta_0) + R(\beta_1|\beta_0, \beta_2) = \text{Model SS} = 7684 \Leftrightarrow R(\beta_2|\beta_0) = 7684 - 4782 = 2902$$

(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? $r^2 = .948$, the multiple coefficient of determination.

(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.

$$F = \frac{SS(E)_r - SS(E)_f}{MS(E)_f} = (422 - 198)/(198/27) = 30.5, df = 1, 27$$

(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:

$$\mu(x_2) = \underline{\hspace{4cm}} + \underline{\hspace{4cm}}x_2$$

Substituting $x_1 = 10$ into the regression equation gives

$$\mu(x_1 = 10, x_2) = 69.4 - 5.86(10) - 1.3x_2 + 0.135(10)x_2 = 10.8 + .05x_2$$

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