

ST512
Fall Quarter, 2005
Exam 1 Solutions

1. (42 points) A random sample of $n = 30$ NBA basketball players is drawn and heights and weights are recorded. The sample means, variances and corrected sums of squares for x and y are given in the table below:

Variable	Mean	St. Dev.	Corr. sum of squares
x : height	79 <i>in</i>	4.2 <i>in</i>	507 <i>in</i> ²
y : weight	224.6 <i>lbs.</i>	32 <i>lbs</i>	29623 <i>lbs</i> ²

The sample covariance is $s_{xy} = 110.3$ (lb-inches). In answering the questions below, make the usual assumption that errors about the mean are iid normal.

- (a) Report the sample correlation coefficient for height and weight.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{110.3}{4.2 \times 32} = 0.82$$

- (b) Give the least squares regression line for a regression of weight on height.

$$\begin{aligned}\hat{\beta}_1 &= r_{xy} \frac{s_y}{s_x} \\ &= 0.82 \frac{32}{4.2} = 6.25 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= -269 \\ \hat{y} &= -269 + 6.25x\end{aligned}$$

- (c) Report the MSE from the regression analysis.

$$MSE = \frac{SS[E]}{n-2} = \frac{SS[Tot] - SS[R]}{n-2} = \frac{29623 - 6.25^2(507)}{28} = 350.6$$

- (d) Use the model to estimate the mean weight among players who are $x = 80$ inches tall.

$$\hat{y} = -269 + 6.25(x = 80) = 231$$

- (e) Using the fact that $\sqrt{(1, 80)(X'X)^{-1}(1, 80)'MSE} = 3.46$, report a 95% confidence interval for this mean.

$$231 \pm 2.05(3.46) \text{ or } 231 \pm 7.1$$

- (f) The width of a 95% confidence interval for the mean among players of height $x = 84$ is (bigger/smaller/equal to) the width of the interval in part (e). (Circle one.)
- (g) A player has a weight clause in his contract that says that if he is over the 97.5th percentile of the league weight among players with similar height, he is subject to a fine. Danny Fortson (Seattle Supersonics) weighs $y = 260\text{lbs}$ and is $x = 80$ inches tall. The team wants to impose the fine on him arguing that he is outside the 95% confidence interval computed in part (e). Do you agree? Support your argument briefly, with appropriate calculations if necessary.

I disagree. They are misinterpreting the confidence interval for the mean. The estimated 97.5th percentile of weights among these players is $231 + 1.96\sqrt{MSE} = 267.7$ and while Mr. Fortson might be a little overweight, he is well below the estimated percentile of interest.

2. (18 points) A simple linear regression model for a random sample of n measurements assumes that given x_1, \dots, x_n , the mean of the response y is a linear function of x :

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, n.$$

- (a) Briefly list all of the additional assumptions about E_i that we make that are necessary for small-sample inference about the regression coefficients.

$$E_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

- (b) Let \hat{y}_i denote the fitted value on the least squares regression line. Label the sums of squares below as corrected total (TOT), or being due to Regression(R) or Error(E). (Fill in the blanks.)

- $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ $SS[E]$
- $\sum_{i=1}^n (y_i - \bar{y})^2$ $SS[Tot]$
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ $SS[R]$

Also, express one sum of squares as the sum of the other two.

$$SS[Tot] = SS[R] + SS[E]$$

3. (40 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) Fill in the blank space in the output labelled AAA. $n = 31$
- (b) Fill in the blank space in the output labelled BBB. 0.02868 by symmetry
- (c) Fill in the blank space in the output labelled CCC. $SE(\hat{\beta}_1)^2/MSE = 0.0046$
- (d) Specify the dimension of the design matrix X .
 $dim(X) = (31 \times 3)$
- (e) Give the matrix product $(X'X)^{-1}X'Y$

$$\hat{\beta} = \begin{pmatrix} -57.99 \\ 4.71 \\ 0.34 \end{pmatrix}$$

- (f) Give the extra sum of squares for girth after controlling for height, $R(\beta_1|\beta_0, \beta_2)$.
From output, type II SS for girth is 4783
- (g) Give the regression sum of squares for a simple linear regression of volume on height only.

$$R(\beta_2|\beta_0) = R(\beta_1, \beta_2|\beta_0) - R(\beta_1|\beta_2, \beta_0) = 7684 - 4783 = 2901$$

- (h) 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. Using $\alpha = 0.05$, is there evidence to select a model on the basis of this test?

$$H_0 : \beta_3 = 0, F = \frac{SSE_{red} - SSE_{full}}{SSE/(n-4)} = \frac{422 - 198}{198/27} = \frac{224}{7.3} = 30.5$$

(bigger than $F(0.05, 1, 27) = 4.21$, so choose interaction model)

- (i) Use the complex model to estimate the mean volume among trees that are $x_1 = 15$ inches around and $x_2 = 80$ feet tall. Estimate the standard deviation of volumes among such a population of trees.

$$\begin{aligned}\hat{\mu} &= 69.4 - 5.86(15) - 1.3(80) + 0.135(15)(80) \\ &= 39.5 \text{ ft}^3 \\ \widehat{SD}(y|x_1 = 15, x_2 = 80) &= \sqrt{MSE} \\ &= 2.7 \text{ ft}^3\end{aligned}$$

- (j) Going back to the additive model, it may be shown from the output that the correlation between the slopes is $\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -0.52$. True or false: if another response, such as crown width, z , were measured on each of these trees and z were regressed on x_1 and x_2 , then the correlation between the slopes $\hat{\beta}_1$ and $\hat{\beta}_2$ would still be -0.52 . Explain briefly.

True. The covariance of the slopes depends on MSE and hence on the response, z , but in the computation of the correlation, the MSE cancels out, and the correlation structure of the regression coefficients depends only on the design matrix, X

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	girth	height	volume
Intercept	AAA_____	410.7	2356	935.3
girth	410.7	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	BBB_____	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

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

Variable	DF	Type II SS
Intercept	1	679.04025
girth	1	4782.97364
height	1	102.38118