

ST512  
 Fall Semester, 2008  
 Quiz 1- solution

1. In NFL games from a sample of  $n = 682$  games, two measurements are made: the published point spread  $x$  and margin of victory  $y$  for the favored team. A simple linear regression of  $y$  on  $x$  was fit with SAS. Code and *selected* output are given below. For all significance tests, use  $\alpha = .05$ .

```
proc reg data=spreads ;
  model outcome=spread;
run;
```

The SAS System 1  
 The REG Procedure

Root MSE	13.26051	R-Square	0.0755
Dependent Mean	6.09673	Adj R-Sq	0.0742
Coeff Var	217.50217		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	0.02755	0.96668	0.03	0.9773
spread	1	1.14291	0.15446	7.40	<.0001

- (a) Obtain a 95% confidence interval for the slope term ( $\beta_1$ ) from the regression model. At level  $\alpha = .05$ , is it plausible that  $\beta_1 = 0$ ? How about  $\beta_1 = 1$ ?

$\hat{\beta}_1 \pm 1.96\widehat{SE}(\hat{\beta}_1)$  or  $1.14 \pm 0.3$  or  $(.84, 1.44)$  Since  $\beta_1 = 0$  is not in this region, it is not a plausible value of the slope in the population of all games. Conversely,  $\beta_1 = 1$  is in the region, and it is plausible that for every point that the published spread increases, the mean outcome also increases by a point.

- (b) Is  $\mu(x = 0) = 0$  a plausible value for the mean outcome when  $x = 0$ ? Conduct an appropriate test.

$\mu(x = 0) = \beta_0$  and inference for  $\beta_0$  can be based directly on the output. A  $t$ -statistic for  $H_0 : \beta_0 = 0$  is  $t = .03$  with a  $p$ -value of  $.9773$ , indicating that a zero-intercept is entirely plausible. It would be interesting to see if confidence bands for  $\mu(x)$  contain the hypothetical function  $\mu(x) = x$ .

- (c) Let the average point spread and outcome be denoted by  $\bar{x}$  and  $\bar{y}$ , respectively. (Note that the latter is given in the output as  $\bar{y} = 6.10$ .) Consider games in which one team is favored by  $x = \bar{x}$  points.
- Report an estimate of the mean margin of victory for the favored team in these ( $x = \bar{x}$ ) games, along with a standard error.

$$\hat{\mu}(x = \bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} = 6.1$$

$$\widehat{SE}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \sqrt{MSE\left(\frac{1}{n} + 0^2\right)} = 13.26/\sqrt{682} = .5$$

- Report an estimate of the standard deviation of the margin of victory for the favored teams in these ( $x = \bar{x}$ ) games.

$$\sqrt{MS(E)} = 13.26$$

- Obtain a 95% prediction interval for the amount by which a team favored by this amount ( $x = \bar{x}$ ) will win in an upcoming game.

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} \pm 1.96\sqrt{MS(E)(1 + 1/n)}$$

or

$$6.10 \pm 1.96(13.26)\sqrt{1 + 1/682} \quad \text{or} \quad 6.1 \pm 26$$

- (d) By writing (up,down, same) indicate how the answers to questions i) and ii) in part (c) would change if games where  $x = \bar{x} + 1$  were considered:
- mean ? **up**                      standard error ? **up**
  - same**
- (e) **PROBLEM NOT ASKED:** report the regression sum of squares for this analysis.

2. An experiment with  $n = 20$  observations is run to minimize peanut kernel damage  $y$ , during the shelling process. Consider the two multiple linear regression models below. The predictor variables are spacing between sheller bars ( $x$ ) and stroke frequency, ( $x_h = 0$  for low frequency and  $x_h = 1$  for high frequency):

$$\begin{aligned}\mu_1(x, \text{high}) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x_h + \beta_4 x_h x + \beta_5 x_h x^2 \\ \mu_2(x, \text{high}) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x_h\end{aligned}$$

The models are fit using the SAS code on the next page, where predictors are given in the model statements in the same order as above. Assume kernel damage varies normally about its mean, with constant variance.

- (a) Report an  $F$ -ratio (and degrees of freedom) for a test comparing the two models.

$$F = \frac{R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) / 2}{MS(E)_{\mu_2}} = \frac{(62.5 - 54.3) / 2}{.39} = \frac{8.2 / 2}{.39} = 10.5, df = 2, 14$$

- (b) Estimate the mean kernel damage when a spacing of  $x = 1$  in. is used at high frequency. Give two answers, one for each model.

$$\hat{\mu}_1(x = 1, x_h = 1) = 24.2 - 38.0 + 17.8 - .75 = 3.25$$

$$\hat{\mu}_2(x = 1, x_h = 1) = 46.2 - 87.2 + 43.7 - 25.5 + 53.6 - 27.3 = 3.5$$

- (c) **PROBLEM NOT ASKED:** For a spacing of  $x = 1$  in., estimate the difference in mean damage across the two frequencies. Give two answers, one for each model.
- (d) Let the estimated covariance matrices of  $\hat{\beta}$  under models  $\mu_1$  and  $\mu_2$  be denoted by  $\hat{\Sigma}_1$  and  $\hat{\Sigma}_2$ , respectively. Give matrix expressions for the standard errors of each of the two estimated means in part (b).

$$\begin{aligned}\widehat{SE}_1 &= \sqrt{(1, 1, 1, 1) \hat{\Sigma}_1 (1, 1, 1, 1)'} \\ \widehat{SE}_2 &= \sqrt{(1, 1, 1, 1, 1, 1) \hat{\Sigma}_1 (1, 1, 1, 1, 1, 1)'}\end{aligned}$$

- (e) **PROBLEM NOT ASKED:** Report the standard error of the difference in part (c) under the simple complex model.

```

proc reg data=goobers;
  model damaged=x x2 high highx highx2 ;
  model damaged=x x2 high ;
run;

```

The REG Procedure  
Model: MODEL1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	62.49729	12.49946	32.18	<.0001
Error	14	5.43868	0.38848		
Corrected Total	19	67.93598			

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	46.21242	9.26303	4.99	0.0002
x		1	-87.17297	22.69791	-3.84	0.0018
x2		1	43.66730	13.17293	3.31	0.0051
high		1	-25.53382	9.58744	-2.66	0.0185
highx		1	53.62141	23.42934	2.29	0.0382
highx2		1	-27.29604	13.59290	-2.01	0.0643

Model: MODEL2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	54.31576	18.10525	21.27	<.0001
Error	16	13.62021	0.85126		
Corrected Total	19	67.93598			

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	24.16788	3.48654	6.93	<.0001
x		1	-38.03888	8.31808	-4.57	0.0003
x2		1	17.81967	4.80912	3.71	0.0019
high		1	-0.75035	0.47862	-1.57	0.1365

3. A dentistry experiment randomizes subjects to two groups ( $n = 20$  each). One group receives a control toothpaste, the other an experimental fluoride toothpaste. Average daily plaque scores,  $y$ , are measured as the response, along with a covariate of brushing frequency called *compliance*, which is centered about its observed mean (1.984) leading to a mean-zero covariate,  $z = \text{compliance} - 1.984$ . SAS code and output pertinent to this problem are given on the next page. In testing for an effect of the fluoride treatment, consider the model in which mean plaque score depends linearly on *compliance* (and hence on  $z$ ), and in which this dependence is constant across treatments.

(a) Report a statistic,  $p$ -value, and associated degrees of freedom for a test for a fluoride treatment effect on plaque score after controlling for *compliance*.

Directly from output,  $t = -2.35, p = .0240, df = 1, 37$

(b) Report the mean plaque score for each treatment, adjusted to the overall mean compliance of 1.984 (or  $z = 0$ ).

Group	Adjusted mean
control	$\hat{\beta}_0 = 7.78$
fluoride	$\hat{\beta}_0 + \hat{\beta}_F = 7.78 - .74 = 7.04$

(c) Estimate the difference between mean plaque score for the two treatments at a given compliance. Report a standard error.

Directly from output,  $\hat{\beta}_F = -.74 (SE = .315)$

(d) Give the proportion of observed variation in plaque score explained by a simple linear regression on  $z$ , where the treatment (toothpaste type) is ignored.

$$r^2 = \frac{R(\beta_z|\beta_0)}{SS(Tot)} = 28.7/70.6 = .40$$

(e) Tough problem: given that the unadjusted mean plaque score was higher for the control group than for the fluoride group, recover the unadjusted means for each group using the output.

The ANCOVA model may be written

$$E(Y_i|F_i, z + i) = \beta_0 + \beta_F F_i + \beta_z z_i$$

where  $i$  indexes all 40 observations.

The reduced one-factor model may be written

$$E(Y_{ij}) = \mu + \tau_i$$

where  $i$  indexes the fluoride treatment and  $j = 1, \dots, 20$  indexes the subjects receiving a given treatment.

The treatment sum of squares is the sum of squared differences of the unadjusted means from the grand mean,

$$SS(Trt) = \sum_{i=1}^2 \sum_{j=1}^{20} (\bar{y}_{i+} - \bar{y}_{++})^2$$

and can therefore be used to recover  $\bar{y}_{1+}$  and  $\bar{y}_{2+}$ . Using the output,

$$\begin{aligned} SS(Trt) + R(\beta_z|\beta_1, trt) &= SS(full\ model) \\ SS(Trt) &= SS(full\ model) - R(\beta_z|\beta_1, trt) \\ &= 34.1 - 26.6 \\ &= 7.5 \\ &= \sum_i \sum_j (\bar{y}_{i+} - \bar{y}_{++})^2 \\ &= 20 \sum_i (\bar{y}_{i+} - \bar{y}_{++})^2 \\ &= 20[(\bar{y}_{1+} - \bar{y}_{++})^2 + (\bar{y}_{2+} - \bar{y}_{++})^2] \\ &= 40(\bar{y}_{1+} - \bar{y}_{++})^2 \quad (\bar{y}_{i+} \text{ equidistant from } \bar{y}_{++}) \\ \frac{SS(trt)}{40} &= (\bar{y}_{1+} - \bar{y}_{++})^2 \\ \sqrt{\frac{SS(trt)}{40}} &= (\bar{y}_{1+} - \bar{y}_{++}) \\ 0.433 &= (\bar{y}_{1+} - \bar{y}_{++}) \end{aligned}$$

so that the two treatment means are  $\bar{y}_{++} \pm 0.433$  or

$$\bar{y}_c = 6.976, \bar{y}_F = 7.84$$

```

proc reg data=teeth2;
  model plaque=z fluoride /ss1 ss2; /* fluoride is an indicator for fluoride group */
run;

```

The SAS System  
 The REG Procedure  
 Model: MODEL1  
 Dependent Variable: plaque

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	34.13180	17.06590	17.30	<.0001
Error	37	36.50552	0.98664		
Corrected Total	39	70.63731			

Root MSE	0.99330	R-Square	0.4832
Dependent Mean	7.40873	Adj R-Sq	0.4553
Coeff Var	13.40709		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS	Type II SS
Intercept	1	7.77946	0.22243	34.97	<.0001	2195.57355	1206.88547
z	1	-0.90661	0.17450	-5.20	<.0001	28.66618	26.63263
fluoride	1	-0.74145	0.31502	-2.35	0.0240	5.46562	5.46562