

ST512 - Summer Session, 2009, Midterm exam - solutions

1.

$$\begin{aligned}\bar{x} &= 0.752 & s_x &= 0.064 \\ \bar{y} &= 49.2 & s_y &= 9.4\end{aligned}$$

In a simple linear regression of y on x , the least squares estimate of the slope is $\hat{\beta}_1 = 103.8$.

- (a) Report the least squares estimate of the intercept. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -28.9$
- (b) Report the corrected total sum of squares, $SS[Total]$ for the longevities.
 $\sum(y_i - \bar{y})^2 = (n - 1)s_y^2 = (10 - 1)(9.4)^2 = 795$
- (c) Report the regression sum of squares, $SS[R]$.
 $SS[R] = \hat{\beta}_1^2 S_{xx} = (103.8)^2(10 - 1)(0.064)^2 = 397$
- (d) Report the coefficient of determination. $r^2 = SS[R]/SS[Total] \approx 0.5$
- (e) The error mean square is approximately $MS(E) \approx 50$. Show how this could have been calculated from the information given above.
 $MS(E) = SS(E)/(n - 2) = SS[Total] - SS[R]/8 \approx 50$
- (f) Explain the assumption of homogeneity of variance, in the context of this experiment. Is it possible to assess the validity of this assumption from the summary statistics given? If so, how? **Homogeneity of variance means that variability in longevity is constant across sizes of the insect (even though mean longevity increases with size). No way to assess this assumption without the residuals, or any way to compute them.**
- (g) Use the model to estimate the mean and standard deviation among all such insects living in a similar environment who have a thorax size of $x_0 = 0.85mm$.
 $\hat{\mu}(x = .85) = \hat{\beta}_0 + \hat{\beta}_1(0.85) = -28.9 + 103.8(.85) = 59 \text{ days}$
 $\widehat{SD}(Y|x = .85) = \hat{\sigma} = \sqrt{MS(E)} = \sqrt{50} \approx 7 \text{ days}$
- (h) Assuming longevities are normally distributed, use the estimated parameters to approximate the proportion of such insects with thorax $x_0 = 0.85$ who will live longer than 73 days.

$$\begin{aligned}\Pr(Y > 73; x = .85) &= \Pr\left(\frac{Y - \mu}{\sigma} > \frac{73 - \mu}{\sigma}; x = 0.85\right) \\ &\approx \Pr\left(Z > \frac{73 - \hat{\mu}}{\hat{\sigma}}; x = 0.85\right) \\ &\approx \Pr\left(Z > \frac{73 - 59}{7}\right) \\ &\approx \Pr(Z > 2) \\ &\approx .025\end{aligned}$$

- (i) Use the model to estimate the mean longevity among all such insects living in a similar environment who have a thorax size exactly equal to the average from the sample, \bar{x} . Report a 95% confidence interval for this mean longevity.

$$\hat{\mu}(x = \bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} = 49.2 \text{ days}$$

$$\widehat{SE}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \sqrt{\frac{MS(E)}{10}} = \sqrt{5} = 2.2$$

$$95\% \text{c.i. } 49.2 \pm t(.025, 8)(2.2) \text{ or } 49.2 \pm 2.31(2.2) \text{ or } 49.2 \pm 5.2$$

- (j) If we sample another insect from a population of insects with thorax size very close to \bar{x} , can we be approximately 95% confident that this individual will have a longevity in this interval? **no, we cannot.**
2. An experiment is conducted to study the potential costs of reproductive activity on lifespan in a particular insect. Twenty-five individual male specimens are randomized to the five treatment groups described below ($n = 5$ per group):

Treatment Group	description	sample mean	sample std. deviation
1	control, no companions	43.0	4.4
2	one pregnant companion per day	51.4	9.7
3	one receptive companion per day	39.0	12.1
4	eight pregnant companions per day	46.0	11.2
5	eight receptive companions per day	23.0	(hidden)

There is some heterogeneity of variance, but you may ignore it for this exam.

- (a) Specify a probability model for this experiment that allows for potential treatment effects. State all the usual assumptions required for the analysis of variance.

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

E_{ij} assumed a random sample from mean 0 normal distribution w/ constant variance, σ^2 .

- (b) The error mean square is given by $MS(E) \approx 90$. Deduce the sample variance of treatment group 5.

$$MS(E) = \frac{1}{5}(s_1^2 + s_2^2 + \dots + s_5^2)$$

implies $s_5^2 = 5MS(E) - \sum_1^4 s_i^2 = 64.7$ (with a sample std. deviation of $\sqrt{64.7} = 8$).

- (c) Using the summary statistics above, complete an appropriate ANOVA table:

Source	df	Sum of squares	Mean square
Treatment	4	2181.76	545
Error	20	1800	90
Total	24	3982	

- (d) Formulate and test the hypothesis that the mean longevity is constant across the five treatment groups. Use level $\alpha = .05$ and draw a brief conclusion.

$$H_0 : \tau_1 = \cdots = \tau_5 = 0 \text{ vs } H_1 : \text{not all } 0$$

$$F = \frac{MS(trt)}{MS(E)} = \frac{545}{90} = 6.1$$

critical value is $F(.05, 4, 20) = 2.87$, so H_0 may be rejected and we may conclude that there is evidence of a treatment effect on longevity.

- (e) Construct and estimate a contrast (θ_1) that compares mean longevity among males exposed to eight receptive females a day with mean longevity among males exposed to eight pregnant females a day. Report a standard error. Do the two sample mean longevities differ significantly? $\hat{\theta}_1 = \bar{y}_{4+} - \bar{y}_{5+} = 46 - 23 = 23$ with $\widehat{SE} = \sqrt{MS(E) \frac{2}{5}} = 6$. Since $\hat{\theta}$ is almost 4 SE's bigger than 0, it is significant.
- (f) Construct and estimate a contrast (θ_2) that compares mean longevity between males exposed to one receptive female a day with mean longevity among males exposed to one pregnant females a day. Report a standard error. Do the two sample mean longevities differ significantly? $\hat{\theta}_2 = \bar{y}_{2+} - \bar{y}_{3+} = 51.4 - 39 = 12.4$ with $\widehat{SE} = \sqrt{MS(E) \frac{2}{5}} = 6$. $t = \hat{\theta}/SE = 2.07$ but $t(.025, 20) = 2.09$, so the two sample mean longevities are not quite significantly different.
- (g) Construct and estimate a contrast (θ_3) that compares mean longevity between males exposed to eight receptive females a day with the average mean longevity among males living in the four treatment conditions (which would be appropriate if you believed longevity was constant among these four conditions). Report a standard error.

$$\hat{\theta}_3 = \bar{y}_{5+} - \frac{1}{4} \sum_1^4 \bar{y}_{i+} = -21.9$$

$$\widehat{SE} = \sqrt{MS(E) \left(\frac{1}{5} + \frac{(1/4)^2}{5} + \frac{(1/4)^2}{5} + \frac{(1/4)^2}{5} + \frac{(1/4)^2}{5} \right)} = 4.7$$

- (h) Are the contrasts θ_1, θ_2 and θ_3 estimated parts (e), (f) and (g) mutually orthogonal? No. θ_1 and θ_3 are not orthogonal.

3. Class Level Information

Class	Levels	Values				
diet	3	Pesticide 1 Pesticide 2 control				
Source		DF	Type I SS	Mean Square	F Value	Pr > F
diet		2	10.85243333	5.42621667	559.03	<.0001
diam		1	15.07962072	15.07962072	1553.56	<.0001
Source		DF	Type III SS	Mean Square	F Value	Pr > F
diet		2	1.44042757	0.72021378	74.20	<.0001
diam		1	15.07962072	15.07962072	1553.56	<.0001
Parameter		Estimate		Standard Error	t Value	Pr > t
Intercept		0.244240536 B		0.28560733	0.86	0.4026
diet	Pesticide 1	-0.560178309 B		0.05269297	-10.63	<.0001
diet	Pesticide 2	-0.599402767 B		0.05459908	-10.98	<.0001
diet	control	0.000000000 B		.	.	.
diam		4.124876590		0.10465170	39.42	<.0001

Least Squares Means

diet	y LSMEAN	Standard Error	Pr > t
Pesticide 1	10.3021820	0.0351379	<.0001
Pesticide 2	(hidden)	0.0360949	<.0001
control	(hidden)	0.0375723	<.0001

- (a) Consider a test of the hypothesis of no diet effects, after controlling for diameter.
- Report the F -ratio for this test. $F = 74.2, df = 2, 20$
 - Give the nested models implicitly being compared when one conducts this F -test.

$$M_1 : \mu(x_{i1}, x_{i2}, d_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_d d_i$$

$$M_2 : \mu(x_{i1}, x_{i2}, d_i) = \beta_0 + \beta_d d_i$$

- (b) Report the least squares estimate of mean calcium content as a function of diameter d , separately for each diet.

$$\hat{\mu}(d) = \begin{cases} .24 + 4.12d & \text{if control} \\ -.32 + 4.12d & \text{if Pesticide 1} \\ -.36 + 4.12d & \text{if Pesticide 2} \end{cases}$$

- (c) Use the model to obtain adjusted mean calcium contents for a femur of average diameter ($d = \bar{d}$). LSMEANS are given below:

diet	y LSMEAN	Standard Error	Pr > t
Pesticide 1	10.3021820	0.0351379	<.0001
Pesticide 2	10.3 - .04	0.0360949	<.0001
control	10.3 + .56	0.0375723	<.0001

- (d) Report standard errors for each of these adjusted means. [see above](#)
- (e) Was there a significant decline in calcium content from the Pesticide 1 group to the control group exhibited by the data? Obtain a 95% confidence interval for this decline. The estimated decline from control to pesticide is $\hat{\beta}_{P1} = -.56 (SE = .05)$. The t -statistic for a test of $H_0 : \beta_{P1} = 0$ is $t = -.56/.05 = -10.6 (df = 20)$, which is highly significant.
- (f) What population quantity is being estimated by the difference of the first two regression coefficients in the output, $-0.560 - (-0.599) = .039$? (Please answer using greek symbols.) $\beta_{P1} - \beta_{P2}$, which amounts to the difference in estimated mean calcium, adjusted to any common diameter d .
- (g) Indicate [how](#) the estimated standard error of the quantity in part (g) depends on the matrix X , the error mean square from the ANOVA, and the observed average diameter, \bar{d} .

$$\begin{aligned}
 (\hat{\beta}_{P1} - \hat{\beta}_{P2}) &= (0, 1, -1, 0, 0)\hat{\beta} \\
 \widehat{SE}(\hat{\beta}_{P1} - \hat{\beta}_{P2}) &= (0, 1, -1, 0, 0)\hat{\Sigma}(0, 1, -1, 0, 0)' \\
 &= (0, 1, -1, 0, 0)XX'X)^{-1}(0, 1, -1, 0, 0)' \times MS(E)
 \end{aligned}$$

- (h) Reconstruct the ANOVA table omitted from the output. It should contain three rows: Model, Error and Total.

Source	df	SS	MS	F
Model	3	26	8.6	≈ 890
Error	20	.194	.0097	
Total	23	26.2		

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	304.16617	152.08309	31.47	0.0003
Error	7	33.83383	4.83340		
Corrected Total	9	338.00000			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS	Type II SS
Intercept	1	49.54308	2.07245	23.91	<.0001	39690	2762.18029
fert	1	0.08913	(hidden)	(hidden)	0.4598	271.68391	2.95621
rain	1	0.51397	(hidden)	(hidden)	0.0358	32.48226	32.48226

Covariance of Estimates

Variable	Intercept	fert	rain
Intercept	4.2950391029	0.0905517881	-0.291004511
fert	0.0905517881	0.0129887966	-0.020523223
rain	-0.291004511	-0.020523223	0.0393084434

- (a) Use the multiple linear regression model to estimate the mean yield when fert = 30lbs/acre is used and under rain = 25 inches.

$$\hat{\mu}(f = 30, r = 25) = 64.9$$

- (b) Estimate the mean increase in yld had an additional 10lbs/acre been used under the same rainfall as part (a). Report a standard error of this estimated benefit.

$$10\hat{\beta}_f = 0.89 \quad \widehat{SE}(10\hat{\beta}_f) = 10\widehat{SE}(\hat{\beta}_f) = 10\sqrt{.01299} = 1.14$$

- (c) Report “r-square”, the multiple coefficient of determination for this model.

$$SS[R]/SS[Total] = 304.2/338 = 0.9$$

- (d) Let $\hat{\beta}$ denote the vector of regression coefficients, $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_f, \hat{\beta}_r)$. Report the estimated variance-covariance matrix of this vector, $\widehat{Var}(\hat{\beta})$.

This is just $\hat{\Sigma}$ as given in the output:

Covariance of Estimates

Variable	Intercept	fert	rain
Intercept	4.2950391029	0.0905517881	-0.291004511
fert	0.0905517881	0.0129887966	-0.020523223
rain	-0.291004511	-0.020523223	0.0393084434

- (e) Use $\widehat{Var}(\hat{\beta})$ to obtain the standard error of the partial slope for rainfall.

$$\widehat{SE}(\hat{\beta}_r) = \sqrt{.039} \approx .2$$

- (f) Report the t -statistic for a test that the partial slope for additional fertilizer is 0, $H_0 : \beta_f = 0$. At level $\alpha = .05$, draw a brief conclusion.

$$t = \frac{\hat{\beta}_f}{\widehat{SE}} = \frac{.089}{\sqrt{.01299}} = 0.78 (p = .4598, df = 7)$$

After accounting for rainfall, the effect of fertilizer no longer significant. (Note that this question could have been answered by inspection of the answer to part (b), where the gain per 10 pounds of fertilizer was not significant.)

- (g) Consider a simple linear regression of yield on fertilizer, without consideration of rainfall:

$$Y_i = \beta_0 + \beta_f f_i + E_i$$

Construct an ANOVA table for this model, and report a test statistic for a test of $H_0 : \beta_f = 0$. Is there evidence of linear association between fertilizer and yield?

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Model	1	271.7	271.7	≈ 33
Error	8	66.3	8.3	
Total	9	338		

The F -ratio is huge, indicating significant linear association between fertilizer and yield.

- (h) Consider a simple linear regression of yield on rainfall, without consideration of fertilizer:

$$Y_i = \beta_0 + \beta_r r_i + E_i$$

Construct an ANOVA table for this model, and report a test statistic for a test of $H_0 : \beta_r = 0$. Is there evidence of linear association between rainfall and yield?

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Model	1	301.2	301.2	≈ 65
Error	8	36.8	4.6	
Total	9	338		

The F -ratio is huge, indicating significant linear association between fertilizer and yield.

- (i) Draw a brief conclusion regarding the potential benefit of the fertilizer additive. Looks like the apparent benefit of fertilizer additive may really be ascribed to rainfall. A better predictor of yield is one based on rainfall alone. The significance of the fertilizer term in the simple linear regression is really due to the correlation between fertilizer and rain. High applications of fertilizer tended to be used, by chance, where rainfall was high. Look at the correlation between $\hat{\beta}_f$ and $\hat{\beta}_r$ and you'll see evidence of strong multicollinearity.