ST 512-Practice Exam I - Osborne

Directions: Answer questions as directed. For true/false questions, circle either true or false.

1.

$$\begin{aligned} SS[E]_3 &= 2.365 \\ R(\beta_1|\beta_0) &= 0.305 \text{ } (SS[R] \text{ for simple linear regression of } y \text{ } on \text{ } x_1) \\ R(\beta_2|\beta_0,\beta_1) &= 0.488 \\ R(\beta_1|\beta_0,\beta_2) &= 0.595 \end{aligned}$$

Some critical values from F distributions which might be useful here are

$$F(0.95,1,60) = 4.001, \ F(0.95,2,60) = 3.150, \ F(0.95,1,62) = 3.996$$

The completed ANOVA table (for the full model)

Source	SS	df	MS	F
Reg	0.793	2	0.3965	10.1
Error	2.365	60	0.039	
Total	3.158	62		

(a) $MS[E] = SS[E]/(n-3) = 2.365/60 = 0.039$, variance of Y, number of moths caught for fixed x_1, x_2.

(b) $SS[R]_3 = SS[R]_1 + R(\beta_2|\beta_0\beta_1) = 0.305 + 0.488 = 0.793$ (type I SS sum to SS[R])

(c) i. observed F-ratio: $F = \frac{SSR/(p-1)}{MSE} = \frac{0.793/2}{0.039} = 10.1$

ii. the $\alpha = 0.05$ critical value above is $F(0.95,2,60) = 3.15$, which is exceeded by the observed F-ratio, so we reject H_0 and conclude that at least one of β_1 or β_2 is nonzero.

(d) The comparison of models 1 and 3 is appropriate here (model 1 nested in model 3): $F = \frac{R(\beta_2|\beta_0,\beta_1)/1}{MS[E]_f} = \frac{0.488}{0.039} = 12.4$ which exceeds $F(0.95,1,60) = 4.001$ so we reject $H_0 : \beta_{2.1} = 0$ at level $\alpha = 0.05$ and choose the model 3 over model 1. Yes, we should incorporate β_2 into the model.

(e) $r^2 = r^2_{y \cdot x_1, x_2} = \frac{SS[R]_3}{SS[Tot]} = \frac{0.793}{3.158} = 0.25 = 25\%$

(f)

$$r^2_{yx_2 \cdot x_1} = r^2_{2 \cdot 1} = \frac{R(\beta_2|\beta_0,\beta_1)}{SS[E]_1} = \frac{R(\beta_2|\beta_0,\beta_1)}{SS[Tot] - SS[R]_1} = \frac{0.488}{2.853} = .17 = 17\%.$$

This is the proportion of variation in y explained adding x_2 to the model that already has x_1 in it.

(g) $r_{yx_2 \cdot x_1} = \sqrt{r_{yx_2 \cdot x_1}} = \sqrt{0.17} = 0.41$

2. (a) i. $SS[E] = SS[Tot] - SS[R] = 2625168 - 204526 = 2420642$.
ii. $MS[E] = SS[E]/(n-2) = 2420642/14 = 172903$ and $E(MS[E]) = \sigma^2$.

Source	DF	Sum of Squares	Mean Square	$E(MS)$	F Value	p-value
Model	1	204526	204526	$\sigma^2 + \beta_1 \sum(x_i - \bar{x})^2$	1.18	0.2951
Error	14	2420642	172903	σ^2		
Corrected Total	15	2625168				

(b) $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{MS[E]} = 416$

(c) i. Using $t(0.975, 14) = 2.14$, a 95% confidence interval for β_1 is $12.3 \pm 2.14(11.3)$ or 12.3 ± 24.2 or $(-11.9, 36.5)$.
ii. Use $\hat{y} = 2903 + 12.3x$ and $e = y - \hat{y}$. See below:

Obs	x	y	yhat	residual
1	16	2508	3099	-591
2	18	2518	3124	-606
3	20	3304	3148*	156*
4	22	3423	3173	250
5	24	3057	3197	-140
6	26	3190	3222	-32
7	28	3500	3246	254
8	30	3883	3271	612
9	32	3823	3295	528
10	34	3646	3320	326
11	36	3708	3344	364
12	38	3333	3369*	-36
13	40	3517	3393	124
14	42	3241	3418	-177*
15	44	3103	3443	-340
16	46	2776	3467	-691

(d) The p-value for $H_0 : \beta_1 = 0$ is 0.2951. There is no evidence of a linear association.

(e) False

(f) $r^2 = 204526/2625168 = 0.08$ This is the coefficient of determination. It means that 8% of the variation in yield is explained by the linear association with days after harvest x.

(g) False. (Just fails to show evidence of linear assn.)

(h) i. Points at $(3148, 156), (3369, -36), (3418, -177)$ should be circled.

$$E(Y|x_i) = \beta_0 + \beta_1 x + \beta_2 x^2$$

(i)

(j) $r^2 = 2084779/2625168 = 0.79$ this is much higher than for the linear model and provides some evidence of an improved fit.

(k) $\hat{\sigma} = \sqrt{MS[E]} = \sqrt{41568} = 204$

(l) At $x = 30, \hat{y} = -1070.4 + 293.483(30) - 4.536(30)^2 = 3652$ after harvesting. (Completed table below).

Obs	x	y	yhat	residual
1	16	2508	2464	44
2	18	2518	2743	-225
3	20	3304	2985	319
4	22	3423	3191	232
5	24	3057	3361	-304
6	26	3190	3494	-304
7	28	3500	3591	-91
8	30	3883	3652*	231*
9	32	3823	3676	147
10	34	3646	3665	-19
11	36	3708	3617	91
12	38	3333	3532	-199
13	40	3517	3412	105
14	42	3241	3255	-14
15	44	3103	3062	41
16	46	2776	2832	-56

(m)

$$\hat{\mu}(x = 30, x^2 = 900) \pm t(0.975, 13)\sqrt{\text{Var}(\hat{\mu}(x = 30, x^2 = 900))}$$

$$\text{or } 3652 \pm 2.16\sqrt{5840} \text{ or } 3652 \pm 165 \text{ or } (3487, 3817)$$

3.
- x_1: number of fibers/mm^2 of springwood
- x_2: percent springwood
- x_3: percent light absorption of summerwood

PROC GLM in SAS was used to fit the following MLR model

$$F(0.95, 1, 16) = 4.49,\ F(0.95, 2, 16) = 3.63,\ F(0.95, 1, 17) = 3.59.$$

(a)

$$\begin{aligned} \text{model } df &= 3 \\ \text{error } df &= 16 \\ \text{total } df &= 19 \end{aligned}$$

(b) $MS[E] = 0.005662/16 = 0.00035$

(c) $R(\beta_1, \beta_0) = 0.01284$

(d) $R(\beta_1|\beta_0, \beta_2, \beta_3) = 0.009564$

(e)

$$F = \frac{R(\beta_1|\beta_0, \beta_2, \beta_3)}{MS[E]} = \frac{0.009564}{0.00035} = 27.3$$

This exceeds the critical value $F(0.95, 1, 16) = 4.49$, so the reduced model is rejected at $\alpha = 0.05$.

(f)

$$F = \frac{R(\beta_2, \beta_3|\beta_0, \beta_1)}{MS[E]} = \frac{(R(\beta_3|\beta_0, \beta_1, \beta_2) + R(\beta_2|\beta_0, \beta_1))/2}{MS[E]} = \frac{0.00105}{0.00035} = 2.99$$

This does not exceed the critical value $F(0.95, 2, 16) = 3.63$, so the reduced model is not rejected at $\alpha = 0.05$.

(g) x_3 does, but x_2 does not appear to explain variability after accounting for x_1. I'd go with

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3.$$

Call this the reduced (r) model. Then

$$r^2 = \frac{SS[R]_r}{SS[Tot]} = \frac{SS[R]_f - R(\beta_2|\beta_0, \beta_1, \beta_3)}{SS[Tot]} = \frac{0.01494 - 0.0000965}{0.02060} = 0.72$$