Stat/For/Hort 572 — Midterm I, Spring 2004 — Solutions

1. (a) From the R `predict` output, $\hat{y}_{\text{pred}} = 59.94$ and s.e.$(\hat{y}_{\text{est}}) = 1.331315$. Thus the confidence interval is: $\hat{y}_{\text{pred}} \pm t_{6,0.025} \times \text{s.e.}(\hat{y}_{\text{est}}) = 59.94 \pm 2.447 \times 1.331315 = (56.68, 63.20)$.

   (b) From the R output `anova(tree.lm)` and `anova(tree.aov)`, obtain SSErr = 84.4 on df = 6 and SSPE = 84.0 on df = 4. Thus $F = \frac{\text{SS}_{\text{LOF}}/\text{df}_{\text{LOF}}}{\text{SSPE}/\text{df}_{\text{PE}}} = \frac{(84.4-84.0)/2}{84.0/4} = 0.00952$ on df = (2,4). The p-value = $P(F_{2,4} \geq 0.00952) > 0.25$. There is no evidence of a lack of fit.

2. (a) The slope b_3 represents the difference between the expected seedling height for species D and that for species A. To test $H_0 : [b3 = 0|b_0, b_1, b_2]$, use either t-test or f-test. From the R output, $t = -1.420$ on df = 16 or $f = 2.0177$ on df = (1,16). The p-value is 0.17466 and thus do not reject H_0 at 5% level. There is no evidence of a difference between species A and D.

   (b) The test of interest is $H_0 : [b_2 = 0|b_0, b_1]$. Apply the additional sum of squares principle. The full model is $y = b_0 + b_1 w_1 + b_2 w_2 + e$ and the reduced model is $y = b_0 + b_1 w_1 + e$. From the ANOVA table, the additional SS is 546.99 on df = 1. $\text{SSE}_{(F)} = 131.77 + 1044.88 = 1176.65$ on df = 17. Thus the observed $f = \frac{546.99/1}{1176.65/17} = 7.91$ on df = (1, 17). The p-value is between 0.01 and 0.05 and thus reject H_0 at 5% level. There is evidence of a difference in height between species A and C, given that there is no difference between species A and D.

3. (a) The test of interest is $H_0 : [b_1 = b_2 = b_3 = 0|b_0]$. From the R output, the observed $f = 297$ on df = (3, 21). The p-value is less than 0.001 and thus reject H_0 at 5% level. There is very strong evidence that there is an overall effect of all the independent variables on y.

   (b) Start with the full model $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$. Since the t-values for b_1, b_2, b_3 are all greater than 2, stop and the final model is the full model.

   (c) Apply the additional sum of squares principle. The full model is $y = b_0 + b_1 x_1 + b_3 x_3 + e$ and the reduced model is $y = b_0 + b_3 x_3 + e$. From the ANOVA table, $\text{SSE}_{(\text{R})} = 665.93 + 2398.6 + 134.55 = 3199.08$ on df $= 1 + 1 + 21 = 23$. To find $\text{SSE}_{(\text{F})}$, consider the fitting order of x_3, x_1, x_2 with

Source	df	SS
x_3	1	$\text{SS}(b_3\|b_0)$
x_1	1	$\text{SS}(b_1\|b_0, b_3)$
x_2	1	$\text{SS}(b_2\|b_0, b_1, b_3)$
Error	21	SSE = 134.55

   Since $F = \frac{SS(b2|b0,b1,b3)/1}{134.55/21} = t^2 = (11.517)^2$, $SS(b2|b0, b1, b3) = 850.231$. Thus $\text{SSE}_{(\text{F})} = 850.231 + 134.55 = 984.781$ on df $= 21 + 1 = 22$. To test $H_0 : [b_1 = 0|b_0, b_3]$, the observed $f = \frac{(3199.08-984.781)/1}{984.781/22} = 49.467$ on df $= (1, 22)$. The p-value is less than 0.001 and we reject H_0 at 5%. There is very strong evidence against H_0.

   (d) In the first step, consider selecting x_3, as the correlation between y and x_3 is the highest. The f-value associated with x_3 is $\frac{2643.38/1}{3199.08/23} = 19.00$, which is greater than 4. Thus select x_3. In the second step, consider selecting x_1, as the R^2 for $y = b_0 + b_3 x_3 + b_1 x_1 + e$ is $\frac{\text{SS}(b_3, b_1|b_0)}{\text{SSTotal}} = \frac{4855.08}{5842.46} = 0.831$ and and the R^2 for $y = b_0 + b_3 x_3 + b_2 x_2 + e$ is $\frac{\text{SS}(b_3, b_2|b_0)}{\text{SSTotal}} = \frac{3309.31}{5842.46} = 0.566$. To see this, note that SSTotal $= 2643.38 + 665.93 + 2398.60 + 134.55 = 5842.46$, $\text{SS}(b_3, b_1|b_0) = \text{SS}(b_1, b_2, b_3|b_0) - \text{SS}(b_2|b_0, b_1, b_2) = 2643.38 + 665.93 + 2398.60 - 850.231 = 4855.08$, and $\text{SS}(b_3, b_2|b_0) = 2643.38 + 665.93 = 3309.31$. By (c), the f-value associated with x_1 is 49.67, which is greater than 4. Thus select x_1. In the third step, consider selecting x_2. Since the t-value is 11.517, which is greater than 2, select x_2. The final model is $y = b_0 + b_3 x_3 + b_1 x_1 + b_2 x_2 + e$.

Grade Distribution

```
90-99:9
80-89:22
70-79:28          mean = 75, median = 77
60-69:13
50-59:2
  <50:3
```