

1. Data for 146 LPGA golfers from the year 2009 were collected. The following variables from this data set were considered for a regression analysis:

y = scoring average x_1 = average driving distance x_2 = % fairways hit

x_3 = % greens hit x_4 = average putts per round x_5 = % sand saves

x_6 = no. of tournaments x_7 = putts per greens hit x_8 = no. of tournaments completed.

The accompanying output (available at the back of this exam as a 3-page document) shows information and plots obtained from regression models fit to these data in an “R” session.

Use the information and the attachment above to answer the following 8 questions (i.e. the current one and the next 7 questions).

The value of R^2 for the model containing independent variables x_3 , x_4 and x_6 is closest to:

- (a) 0.655.
- (b) 0.941.
- (c) 0.902.
- (d) 0.945.
- (e) 0.952.

2. Considering AIC, BIC, R^2 and the principle of parsimony, the best choice of the model is:

- (a) Probably the one containing only x_3 .
- (b) Probably the one containing all 8 independent variables.
- (c) Probably the one containing x_1 , x_3 , x_4 , x_5 , x_6 and x_8 .
- (d) Any of the models that contain x_8 .

3. A sports reporter decides to use the simple model containing only x_3 , x_4 and x_8 to predict next year's scoring average for a promising rookie, Jane Doe. Assume that Jane's percentage of greens hit, average putts per round and number of tournaments completed are 70, 29.5 and 18, respectively. The prediction of her scoring average along with a rough measure of the standard error of the prediction would be:

- (a) 70.1 ± 1.07 .
- (b) $70.1 \pm \sqrt{0.2539}$.
- (c) 71.2 ± 0.2539 .
- (d) $71.2 \pm \sqrt{0.2539}$.
- (e) 73.5 ± 0.2539 .

4. It is of interest to test the hypothesis that the variables x_2 , x_6 and x_7 are not needed in the same model with the other five independent variables. If we test this null hypothesis using $\alpha = 0.05$, then which of the following is correct?

- (a) The F -statistic is 4.245 and we would conclude that x_2 , x_6 and x_7 are needed in the model if $4.245 > F_{3,137;0.05}$.
- (b) The F -statistic is 4.245 and we would conclude that x_2 , x_6 and x_7 are *not* needed in the model if $4.245 > F_{3,137;0.05}$.
- (c) The F -statistic is 308.64 and we would conclude that x_2 , x_6 and x_7 are needed in the model if $308.64 > F_{3,137;0.05}$.
- (d) The F -statistic is 308.64 and we would conclude that x_2 , x_6 and x_7 are *not* needed in the model if $308.64 > F_{3,137;0.05}$.
- (e) The F -statistic is 308.64 and we would conclude that x_2 , x_6 and x_7 are needed in the model if $308.64 > F_{5,137;0.05}$.

5. A plot of the residuals for the full model (i.e., the one containing all 8 independent variables) is provided for you in the last page of the R output attachment. Which of the following is the best conclusion based on this plot?

- (a) An assumption of Normally distributed error terms seems reasonable.
- (b) There is a clear decrease in the variance of residuals as the predicted value increases.
- (c) There appears to be no outliers that have a large influence on the fitted model.
- (d) These residuals do not give us any reason to believe that the model assumptions have been violated.
- (e) Both options (b) and (c) are true.

6. Use the output for the model with variables x_3 , x_4 and x_8 to answer this question. Consider a group of golfers who all have the same percentage of greens hit and the same average number of putts per round. If player A in this group completed 5 more tournaments than player B, which of the following would be the best guess as to how their scoring averages compare?

- (a) Player A's scoring average is about 1 higher than that of player B.
- (b) Player A's scoring average is about 1 lower than that of player B.

- (c) Player A's scoring average is about 0.16 higher than that of player B.
 - (d) Player A's scoring average is about 0.16 lower than that of player B.
 - (e) Player A's scoring average is about 0.033 lower than that of player B.
7. Given that $\sum_{i=1}^{146} (y_i - \bar{y})^2 = 189.4666$, the estimate of the error variance using the full model is:
- (a) $189.4666/145$.
 - (b) $189.4666/137$.
 - (c) $189.4666(1 - 0.960028)/137$.
 - (d) $189.4666(0.960028)/137$.
 - (e) 0.960028 .
8. A plot of Cook's D for the full model (i.e., the one containing all 8 independent variables) is provided for you in the last page of the R output attachment. Which of the following is the best conclusion based on this plot?
- (a) An assumption of Normally distributed error terms seems reasonable.
 - (b) There are no extremely large residuals.
 - (c) There appear to be no outliers that have a large influence on the fitted model.
 - (d) An assumption of constant variance for the error terms seems reasonable.
 - (e) Both options (b) and (c) are true.

9. When analyzing a set of data using multiple regression, which of the following is true?
- (a) AIC and BIC will always choose the same subset of independent variables.
 - (b) AIC always chooses a model containing more independent variables than are in the model chosen by BIC.
 - (c) AIC will sometimes choose a model containing more independent variables than are in the model chosen by BIC.
 - (d) BIC will sometimes choose a model containing more independent variables than are in the model chosen by AIC.
 - (e) Both (c) and (d) are true.

10. A group of entomologists were studying data involving spruce moths. The numbers of moths caught in 60 different traps were recorded. The traps varied according to where they were located on a tree: top, middle, lower or ground. Fifteen traps were used for each location, and an analysis of variance was conducted to determine what effect, if any, the locations had on the number of moths caught. The following partial ANOVA table was determined from the data. (The lower case italicized letters in the table denote entries that are not given to you.)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Location	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
Error	56	3261.6	<i>e</i>	
Total	<i>f</i>	<i>g</i>		

Use the information and the table above to answer the following 4 questions (i.e. the current one and the next 3 questions).

The number of moths caught in different traps varies even when all the traps are in the same location. We term this standard deviation σ . The best estimate of σ from the ANOVA table is:

- (a) $(g - 3261.6)/56$.
- (b) g/f .
- (c) $\sqrt{g/f}$.
- (d) $3261.6/56$.
- (e) $\sqrt{3261.6/56}$.

11. Let μ_t , μ_m , μ_ℓ and μ_g denote the average number of moths caught in the top, middle, lower and ground parts of a tree, respectively. The value of the F -statistic for testing $H_0 : \mu_t = \mu_m = \mu_\ell = \mu_g$ is:

- (a) $(a/b)/58.24$.
- (b) $b/3261.6$.
- (c) $(b/3)/58.24$.
- (d) $(b/a)/3261.6$.
- (e) $(a/b)/(g/f)$.

12. It turns out that the value of the F -statistic is 11.34. If we test the null hypothesis $H_0 : \mu_t = \mu_m = \mu_\ell = \mu_g$ using $\alpha = 0.05$, then which of the following is correct? (Remember, if you can't find a certain degree of freedom on the F -table, choose the next smaller one on the table.)

- (a) Since F is smaller than the appropriate table value, we cannot reject H_0 .
- (b) Since F is larger than the appropriate table value, we may conclude that $\mu_t \neq \mu_m$, $\mu_m \neq \mu_\ell$ and $\mu_\ell \neq \mu_g$.
- (c) We cannot reject equality of the four means since $F > F_{3,50;0.05} = 2.79$.
- (d) It is reasonable to conclude that the four means are not all the same since $F > F_{3,50;0.05} = 2.79$.
- (e) Since we do not know the P -value it is impossible to draw a conclusion.

13. The four sample means were as follows:

$$\bar{x}_t = 23.33, \quad \bar{x}_m = 31, \quad \bar{x}_\ell = 33.33, \quad \bar{x}_g = 19.07.$$

Then, using Tukey's HSD procedure, a valid 95% confidence interval for $\mu_t - \mu_g$ is: (Remember, if you can't find a certain degree of freedom on the Q -table, choose the next smaller one on the table.)

- (a) 4.26 ± 7.47 , and hence we should not conclude that μ_t and μ_g are different.
- (b) 4.26 ± 7.47 , and hence it is reasonable to conclude that μ_t and μ_g are different.
- (c) 4.26 ± 3.73 , and hence we should not conclude that μ_t and μ_g are different.
- (d) 4.26 ± 3.73 , and hence it is reasonable to conclude that μ_t and μ_g are different.
- (e) Cannot be determined because we do not know the sum of squares for locations.

14. In class we learned that when performing a one-way analysis of variance, an F -statistic equal to (or less than) 1 is:

- (a) Always strong evidence that there are differences among the treatment means.
- (b) Sometimes strong evidence that there are differences among the treatment means.
- (c) Never strong evidence that there are differences among the treatment means.
- (d) Indication that our estimate of the error variance is equal to 1.

15. A multiple linear regression model relating Y with x_1 and x_2 has the form:

$$Y = 10 + x_1 - 3x_2 + 0.9x_1x_2 + \epsilon,$$

where, for every choice of (x_1, x_2) , ϵ has a Normal distribution with mean 0 and standard deviation 1. The expected value of Y when $x_1 = 1$ and $x_2 = 2$ is:

- (a) 6.8.
- (b) 7.1.
- (c) 8.6.
- (d) 10.
- (e) Cannot be determined from the information given.

16. Variance between populations is measured in a one-way ANOVA by:

- (a) The total sum of squares.
- (b) The error sum of squares.
- (c) The treatment sum of squares.
- (d) The grand mean.
- (e) The National Bureau of Standards.

17. Suppose we want to test 5 null hypotheses: $H_0^{(1)}, \dots, H_0^{(5)}$, and for simplicity, assume that they are tested using 5 independent datasets. Suppose each hypothesis is now tested using a procedure that controls the respective Type I error rate at a level 0.1. Then, the experimentwise error rate for testing all these hypotheses simultaneously is given by: (**Note:** this question carries **4 points**.)

- (a) 0.1.
- (b) $(0.1)^5$.
- (c) 0.5905.
- (d) 0.9999.
- (e) 0.4095.

```

1 X=cbind(x1,x2,x3,x4,x5,x6,x7,x8)
2
3 leaps(X,y,method='r2',nbest=2)$which
4
5      1      2      3      4      5      6      7      8      AIC      BIC
6 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE    269.69    278.64
7 1 FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE    302.86    311.81
8 2 FALSE FALSE TRUE   TRUE FALSE FALSE FALSE FALSE    48.05     59.98
9 2 FALSE FALSE TRUE  FALSE FALSE FALSE TRUE  FALSE    121.99    133.92
10 3 FALSE FALSE TRUE   TRUE FALSE FALSE FALSE TRUE    20.04     34.96
11 3 FALSE FALSE TRUE   TRUE FALSE TRUE  FALSE FALSE    39.80     54.72
12 4 FALSE FALSE TRUE   TRUE TRUE  FALSE FALSE TRUE    15.29     33.19
13 4 TRUE  FALSE TRUE   TRUE FALSE FALSE FALSE TRUE    16.29     34.19
14 5 TRUE  FALSE TRUE   TRUE TRUE  FALSE FALSE TRUE     9.29     30.18
15 5 FALSE FALSE TRUE   TRUE TRUE  FALSE TRUE  TRUE    10.83     31.72
16 6 TRUE  FALSE TRUE   TRUE TRUE  TRUE  FALSE TRUE     5.83     29.70
17 6 TRUE  TRUE  TRUE   TRUE TRUE  TRUE  FALSE TRUE     7.05     30.92
18 7 TRUE  TRUE  TRUE   TRUE TRUE  TRUE  FALSE TRUE     4.06     30.91
19 7 TRUE  FALSE TRUE   TRUE TRUE  TRUE  TRUE  TRUE     5.11     31.98
20 8 TRUE  TRUE  TRUE   TRUE TRUE  TRUE  TRUE  TRUE     2.32     32.16
21
22 $label
23 [1] "(Intercept)" "1"          "2"          "3"          "4"
24 [6] "5"              "6"          "7"          "8"
25
26 $size
27 [1] 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9
28
29 $r2
30 [1] 0.7253822 0.6553326 0.9406433 0.9015029 0.9516711 0.9446654 0.9538540
31 [8] 0.9535363 0.9563137 0.9558510 0.9579181 0.9575638 0.9589913 0.9586936
32 [15] 0.9600280
33
34
35 > fit=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
36 > anova(fit)
37
38 Analysis of Variance Table
39
40 Response: y
41      Df Sum Sq Mean Sq  F value    Pr(>F)
42 x1      1 41.861  41.861 757.2479 < 2.2e-16 ***
43 x2      1 37.150  37.150 672.0372 < 2.2e-16 ***
44 x3      1 46.334  46.334 838.1606 < 2.2e-16 ***
45 x4      1 53.633  53.633 970.2110 < 2.2e-16 ***
46 x5      1  0.912   0.912  16.5013 8.144e-05 ***
47 x6      1  0.444   0.444   8.0349 0.005283 **
48 x7      1  0.333   0.333   6.0215 0.015387 *
49 x8      1  1.226   1.226  22.1824 6.008e-06 ***
50 Residuals 137  7.573   0.055
51 ---
52
53
54
55 > fit1=lm(y~x2+x6+x7)
56 > anova(fit1)
57
58 Analysis of Variance Table
59
60 Response: y
61      Df Sum Sq Mean Sq  F value    Pr(>F)
62 x2      1 10.159  10.159  24.553 2.028e-06 ***
63 x6      1 63.442  63.442 153.329 < 2.2e-16 ***
64 x7      1 57.110  57.110 138.025 < 2.2e-16 ***
65 Residuals 142 58.755   0.414
66 ---

```

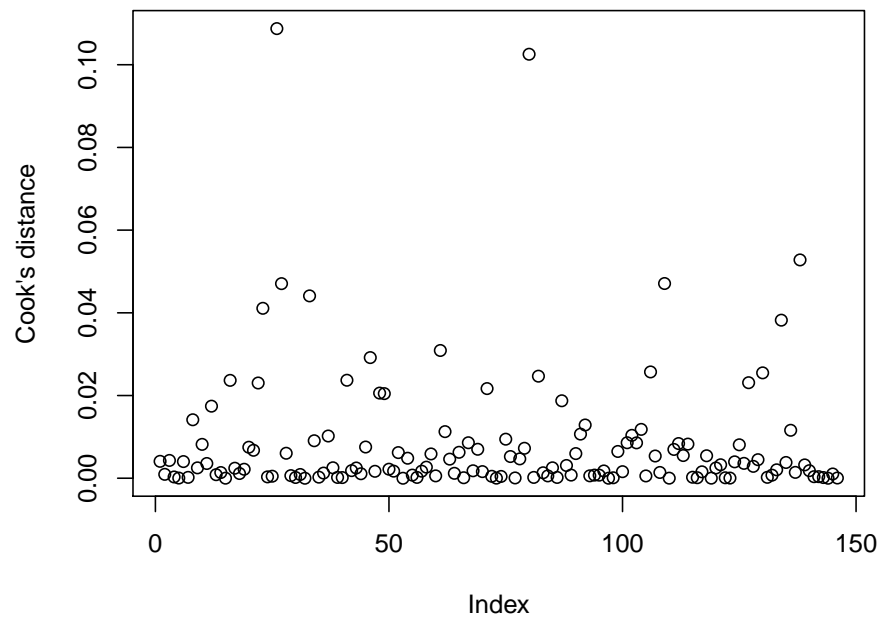


```

67
68
69
70 > fit2=lm(y~x1+x3+x4+x5+x8)
71 > anova(fit2)
72
73 Analysis of Variance Table
74
75 Response: y
76      Df Sum Sq Mean Sq  F value    Pr(>F)
77 x1      1 41.861  41.861   708.037 < 2.2e-16 ***
78 x3      1 82.765  82.765 1399.901 < 2.2e-16 ***
79 x4      1 54.243  54.243  917.479 < 2.2e-16 ***
80 x5      1  0.830   0.830   14.041 0.0002607 ***
81 x8      1  1.490   1.490   25.205 1.542e-06 ***
82 Residuals 140  8.277   0.059
83 ---
84
85
86
87 > fit3=lm(y~x3+x4+x8)
88 > summary(fit3)
89
90 Call:
91 lm(formula = y ~ x3 + x4 + x8)
92
93 Residuals:
94      Min       1Q   Median       3Q      Max
95 -0.66244 -0.17566  0.02268  0.16736  0.84612
96
97 Coefficients:
98             Estimate Std. Error t value Pr(>|t|)
99 (Intercept) 61.240140   1.071359  57.161 < 2e-16 ***
100 x3          -0.195544   0.007981 -24.500 < 2e-16 ***
101 x4           0.820476   0.041074  19.976 < 2e-16 ***
102 x8          -0.032671   0.005740  -5.692 6.93e-08 ***
103 ---
104
105 Residual standard error: 0.2539 on 142 degrees of freedom
106 Multiple R-squared:  0.9517,    Adjusted R-squared:  0.9507
107 F-statistic: 932.1 on 3 and 142 DF,  p-value: < 2.2e-16
108
109 > anova(fit3)
110
111 Analysis of Variance Table
112
113 Response: y
114      Df Sum Sq Mean Sq  F value    Pr(>F)
115 x3      1 124.164 124.164 1925.500 < 2.2e-16 ***
116 x4      1  54.057  54.057  838.301 < 2.2e-16 ***
117 x8      1   2.089   2.089   32.402 6.926e-08 ***
118 Residuals 142   9.157   0.064
119
120

```

Cook's distance for full model



Residuals from full model

