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:

$$\mathbf{y} = \text{scoring average} \quad \mathbf{x_1} = \text{average driving distance} \quad \mathbf{x_2} = \%\ \text{fairways hit}$$

$$\mathbf{x_3} = \%\ \text{greens hit} \quad \mathbf{x_4} = \text{average putts per round} \quad \mathbf{x_5} = \%\ \text{sand saves}$$

$$\mathbf{x_6} = \text{no. of tournaments} \quad \mathbf{x_7} = \text{putts per greens hit} \quad \mathbf{x_8} = \text{no. of tournaments completed.}$$

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.

Using the leaps output, the model containing x_3, x_4 and x_6 is the sixth one given, and the sixth R^2 value is 0.945.

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.

The six-variable model containing x_1, x_3, x_4, x_5, x_6 and x_8 has the smallest BIC value and to two decimal places it has the same R^2 as the full model, which has the smallest AIC value.

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.

The prediction using the three-variable model with x_3, x_4 and x_8 is:

$$\hat{y} = 61.240140 - 0.195544(70) + 0.820476(29.5) - 0.032671(18) = 71.2.$$

A 'rough' measure of the standard error of the prediction is just the estimated standard error of the residuals from the fitted model, i.e. $\hat{\sigma}$, which from the output is given by $\hat{\sigma} = 0.2539$. The reason why this is a 'roughly' acceptable estimate of the error of the prediction lies in the formula of prediction intervals in pg. 95 of Chapter 1C (for multiple linear regression), and in the formulae on pg. 37 of Chapter 1A (for the special case of simple linear regression).

In the formula of the prediction interval, the standard error of the prediction is determined by a sum of two terms out of which the second term is the standard error of $\hat{\mu}(x_0)$ itself, while the first term is the one that is added to account for the extra uncertainty due to the noise ϵ and is an essential modification when dealing with prediction intervals. As discussed (many times) in class, especially in the context of simple linear regression, the standard error of $\hat{\mu}(x_0)$ itself *always* decreases with n and converges to 0 as n increases. The same argument also holds for multiple linear regression. On the other hand, the first term will *never* go to 0 as n increases but will go to the constant $\sigma > 0$. Hence, for moderately large n, it is only the first term (which is simply $\hat{\sigma}$) that dominates the other term and serves as a 'rough' but reasonably accurate estimate of the uncertainty of the prediction. Remember this whole argument is applicable for prediction intervals *only* where the extra $\hat{\sigma}$ needs to be added, but not in case of confidence intervals for the mean of Y given x. For the latter the width of the confidence interval indeed goes to 0 as n increases, but never for prediction intervals!

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 $308.64 > F_{5,137;0.05}$.

You should use the reduction method here. The full model has an SSE of 7.573 and the reduced model, the one with x_1, x_3, x_4, x_5 and x_8, has an SSE of 8.277. Therefore the F-statistic is:

$$F = \frac{(8.277 - 7.573)/3}{7.573/137} = 4.245.$$

We would conclude that the variables x_2, x_6 and x_7 are needed in the model if 4.245 is larger than $F_{3,137,;0.05}$.

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.

The residuals appear to be randomly scattered, as they should be when the model assumptions are met.

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.

Using the model with x_3, x_4 and x_8, a player's scoring average is predicted to be:

$$61.240140 - 0.195544x_3 + 0.820476x_4 - 0.032671x_8.$$

If players A and B have the same values for x_3 and x_4, then the difference in their predicted scoring averages is:

$$-0.032671(x_{8A} - x_{8B}),$$

and since the difference $x_{8A} - x_{8B}$ is 5, it is predicted that the scoring average of player A is $5(0.032671) = 0.16$ 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.

We know that:

$$\begin{aligned}\hat{\sigma}^2 &= SSE/(n-k-1) \\ &= (SST - SSR)/(n-k-1) \\ &= (SST - SST \cdot R^2)/(n-k-1) \\ &= SST(1-R^2)/(n-k-1) \\ &= 189.4666(1-0.960028)/137.\end{aligned}$$

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.

The Cook's D values are not the same thing as residuals. They indicate whether leaving out a data value changes the estimated regression coefficients very much. Since no Cook's D value is larger than 1.5 (in fact, all are less than 0.15), there are no influential observations.

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.

I mentioned several times in class that AIC sometimes chooses a larger model than BIC, but it will never choose a smaller model. At the same time, it is also not true that AIC necessary always chooses a larger model either. So the only correct option is (c).

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	a	b	c	d
Error	56	3261.6	e	
Total	f	g		

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) $\sqrt{3261.6/56}$.

We estimate σ^2 using the MSE, in this case 3261.6/56. So, the estimate of σ is $\sqrt{3261.6/56}$.

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) $(a/b)/(g/f)$.

The value of the F-statistic is: $MSTr/MSE = (b/a)/(3261.6/56) = (b/3)/58.24$.

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 the know the P-value it is impossible to draw a conclusion.

The correct degrees of freedom are 3 and 56, but 56 is not in the table. So we use the next smaller degrees of freedom, 50. We would reject the hypothesis of equal means if $F \geq F_{3,50;0.05} = 2.79$, and since $11.34 > 2.79$, it is reasonable to conclude that the four means are not all the same.

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.

A 95% confidence interval for $\mu_t - \mu_g$ using Tukey's HSD method is given by:

$$(\bar{x}_t - \bar{x}_g) \pm Q_{4,56;0.05}\sqrt{\frac{MSE}{15}} \approx$$
$$(\bar{x}_t - \bar{x}_g) \pm Q_{4,40;0.05}\sqrt{\frac{MSE}{15}} =$$
$$(23.33 - 19.07) \pm 3.79\sqrt{\frac{58.24}{15}} =$$
$$4.26 \pm 7.47.$$

Since the interval includes 0, we should not conclude that the two means are different.

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.

I did say in class more than once that an F-statistic of 1 or less is never a strong evidence of a difference in means. This was also formally verified in details in class while going through the calculations in pg. 126-130 of Chapter 2A notes.

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.

The expected value of Y given $x_1 = 1$ and $x_2 = 2$ is:

$$E(Y) = 10 + x_1 - 3x_2 + 0.9x_1x_2 = 10 + 1 - 6 + 0.9(2) = 6.8.$$

16. Variance <u>between</u> 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.

See pg. 120-121 of the Chapter 2A notes.

17. Suppose we want to test 5 null hypotheses: $H_0^{(1)}, \ldots, H_0^{(5)}$, 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.

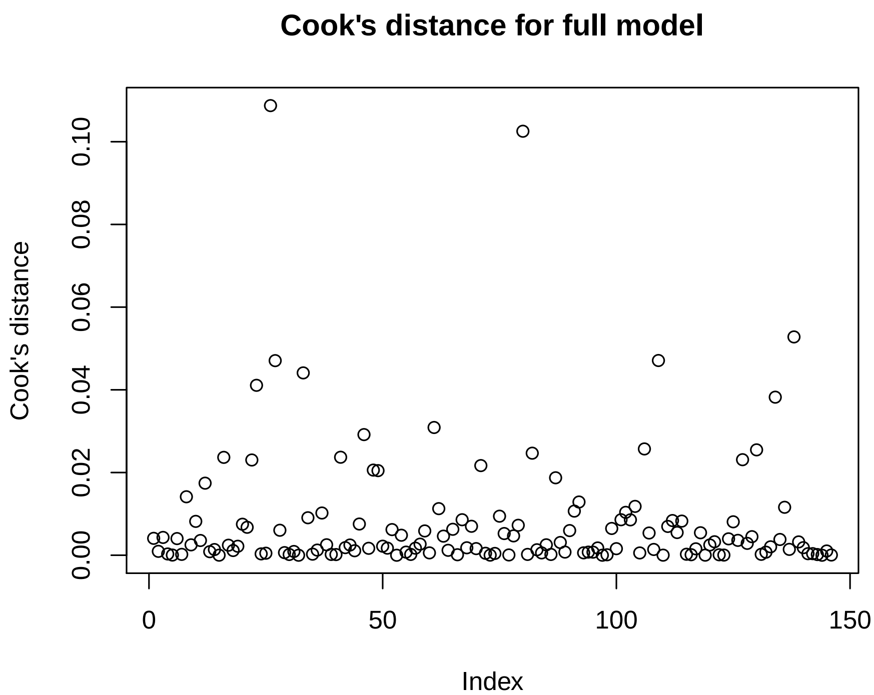
(d) 0.9999.

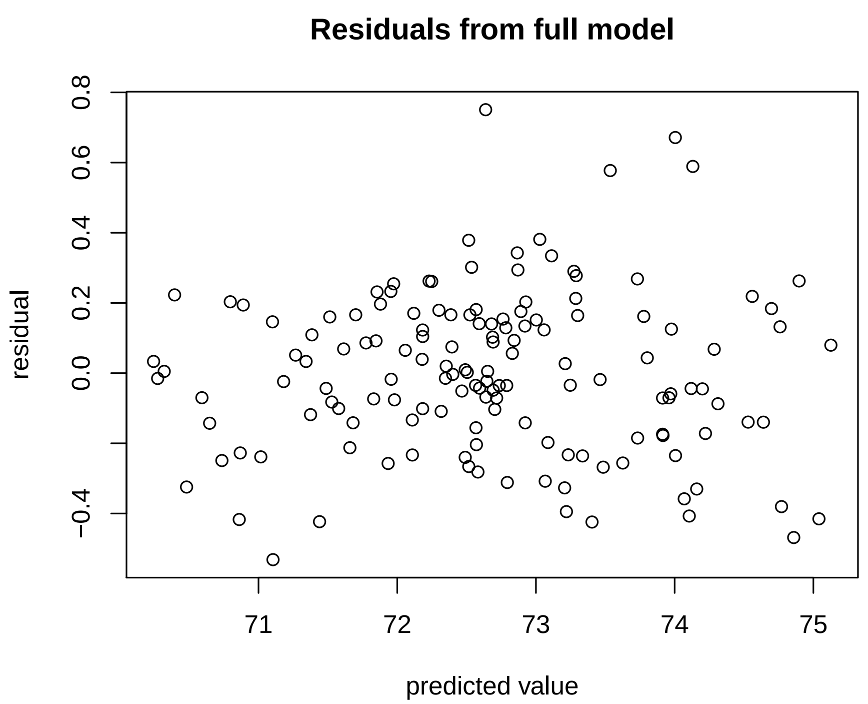
(e) 0.4095.

Using the formula and calculations on pg. 135-136 of Chapter 2A notes, the experimentwise error rate (EWER) is given by: $\alpha_E = 1 - (1 - \alpha)^m$. Here, there are $m = 5$ hypotheses (tested independently using 5 different datasets) and for each test, the Type I error rate is controlled at a level $\alpha = 0.1$. Hence, the EWER $= 1 - (1 - 0.1)^5 = 1 - 0.9^5 = 1 - 0.5905 = 0.4905$.

```
X=cbind(x1,x2,x3,x4,x5,x6,x7,x8)

leaps(X,y,method='r2',nbest=2)$which

      1     2     3     4     5     6     7     8            AIC     BIC
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE         269.69  278.64
1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE         302.86  311.81
2 FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE          48.05   59.98
2 FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE         121.99  133.92
3 FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE          20.04   34.96
3 FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE          39.80   54.72
4 FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE          15.29   33.19
4  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE          16.29   34.19
5  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE           9.29   30.18
5 FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE          10.83   31.72
6  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE           5.83   29.70
6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE           7.05   30.92
7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE           4.06   30.91
7  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE           5.11   31.98
8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE           2.32   32.16

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"           "6"           "7"           "8"

$size
 [1] 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9

$r2
 [1] 0.7253822 0.6553326 0.9406433 0.9015029 0.9516711 0.9446654 0.9538540
 [8] 0.9535363 0.9563137 0.9558510 0.9579181 0.9575638 0.9589913 0.9586936
[15] 0.9600280


> fit=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
> anova(fit)

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq  F value    Pr(>F)
x1          1 41.861  41.861 757.2479 < 2.2e-16 ***
x2          1 37.150  37.150 672.0372 < 2.2e-16 ***
x3          1 46.334  46.334 838.1606 < 2.2e-16 ***
x4          1 53.633  53.633 970.2110 < 2.2e-16 ***
x5          1  0.912   0.912  16.5013 8.144e-05 ***
x6          1  0.444   0.444   8.0349  0.005283 **
x7          1  0.333   0.333   6.0215  0.015387 *
x8          1  1.226   1.226  22.1824 6.008e-06 ***
Residuals 137  7.573   0.055
---



> fit1=lm(y~x2+x6+x7)
> anova(fit1)

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x2          1 10.159  10.159  24.553 2.028e-06 ***
x6          1 63.442  63.442 153.329 < 2.2e-16 ***
x7          1 57.110  57.110 138.025 < 2.2e-16 ***
Residuals 142 58.755   0.414
---
```

```
> fit2=lm(y~x1+x3+x4+x5+x8)
> anova(fit2)

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq  F value    Pr(>F)
x1          1 41.861  41.861  708.037 < 2.2e-16 ***
x3          1 82.765  82.765 1399.901 < 2.2e-16 ***
x4          1 54.243  54.243  917.479 < 2.2e-16 ***
x5          1  0.830   0.830   14.041 0.0002607 ***
x8          1  1.490   1.490   25.205 1.542e-06 ***
Residuals 140  8.277   0.059
---



> fit3=lm(y~x3+x4+x8)
> summary(fit3)

Call:
lm(formula = y ~ x3 + x4 + x8)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66244 -0.17566  0.02268  0.16736  0.84612

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.240140   1.071359  57.161  < 2e-16 ***
x3          -0.195544   0.007981 -24.500  < 2e-16 ***
x4           0.820476   0.041074  19.976  < 2e-16 ***
x8          -0.032671   0.005740  -5.692 6.93e-08 ***
---

Residual standard error: 0.2539 on 142 degrees of freedom
Multiple R-squared:  0.9517,    Adjusted R-squared:  0.9507
F-statistic: 932.1 on 3 and 142 DF,  p-value: < 2.2e-16

> anova(fit3)

Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq  F value    Pr(>F)
x3          1 124.164 124.164 1925.500 < 2.2e-16 ***
x4          1  54.057  54.057  838.301 < 2.2e-16 ***
x8          1   2.089   2.089   32.402 6.926e-08 ***
Residuals 142   9.157   0.064
```


Cook's distance for full model
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
Index


Residuals from full model
residual
−0.4
0.0
0.2
0.4
0.6
0.8
71
72
73
74
75
predicted value