Final Exam

Sasha Uspenskaya

Stat 615

Table of Contents

```
ir<- read.table("~/Downloads/SENIC_2023Fall.txt", header=T)
ir <- ir[ , c(2, 3, 4, 5,6)]
ir

##     InfectionRisk LengthOfStay RoutineCulturingRatio Region
AverageDailyCensus
## 1             4.1         7.13                   9.0      4
207
## 2             1.6         8.82                   3.8      2
51
## 3             2.7         8.34                   8.1      3
82
## 4             5.6         8.95                  18.9      4
53
## 5             5.7        11.20                  34.5      1
134
## 6             5.1         9.76                  21.9      2
147
## 7             4.6         9.68                  16.7      3
151
## 8             5.4        11.18                  60.5      2
399
## 9             4.3         8.67                  24.4      3
```

```
130
## 10            6.3         8.84                  29.6      1
59
## 11            4.9        11.07                  28.5      1
591
## 12            4.3         8.30                   6.8      3
105
## 13            7.7        12.78                  46.0      1
252
## 14            3.7         7.58                  20.8      2
59
## 15            4.2         9.00                  14.6      3
61
## 16            5.5        11.08                  18.6      3
326
## 17            4.5         8.28                  26.0      4
84
## 18            6.4        11.62                  25.5      1
113
## 19            4.2         9.06                   6.9      2
103
## 20            4.1         9.35                  15.9      3
547
## 21            4.2         7.53                  23.1      4
47
## 22            4.8        10.24                  36.3      2
163
## 23            5.0         9.78                  17.6      1
240
## 24            4.8         9.84                  12.0      3
468
## 25            4.0         9.20                  17.5      4
244
## 26            3.9         8.28                  12.0      2
413
## 27            4.5         9.31                  30.2      1
124
## 28            3.2         8.19                  10.8      1
156
## 29            4.4        11.65                  18.6      1
217
## 30            4.9         9.89                  17.7      2
113
## 31            5.0        11.03                  19.7      1
270
## 32            5.2         9.84                  17.7      2
200
## 33            5.3        11.77                  17.3      1
164
## 34            6.1        13.59                  24.2      1
```

```
258
## 35            6.3         9.74                  11.4      2
170
## 36            5.0        10.33                  21.2      1
181
## 37            2.8         9.97                  16.5      2
69
## 38            4.6         7.84                   7.1      3
50
## 39            4.1        10.47                   5.7      2
168
## 40            1.3         8.16                   1.9      3
49
## 41            3.7         8.48                  12.1      3
145
## 42            4.7        10.72                  23.2      3
90
## 43            3.0        11.20                   7.0      3
95
## 44            5.6        10.12                  14.9      3
313
## 45            5.5         8.37                  15.1      2
96
## 46            4.6        10.16                   8.4      4
581
## 47            6.5        19.56                  17.2      1
273
## 48            5.5        10.90                  10.6      2
446
## 49            1.8         7.67                   2.5      3
93
## 50            4.2         8.88                  10.1      3
238
## 51            5.6        11.48                  20.3      1
207
## 52            4.3         9.23                  11.6      2
413
## 53            7.6        11.41                  16.6      3
330
## 54            7.8        12.07                  52.4      2
115
## 55            3.1         8.63                   8.4      1
39
## 56            3.9        11.15                   7.7      1
217
## 57            3.7         7.14                   2.6      4
37
## 58            4.3         7.65                  16.4      4
265
## 59            3.9        10.73                  19.3      2
```

```
374
## 60            4.5        11.46                  15.6      3
153
## 61            3.4        10.42                   8.0      1
67
## 62            5.7        11.18                  18.8      2
546
## 63            5.4         7.93                   7.5      4
42
## 64            4.4         9.66                   9.9      2
66
## 65            5.0         7.78                  20.9      3
391
## 66            4.3         9.42                  24.8      1
421
## 67            4.4        10.02                   8.3      2
191
## 68            3.7         8.58                   7.4      3
248
## 69            4.5         9.61                   6.9      3
404
## 70            3.5         8.03                  24.3      1
65
## 71            4.2         7.39                  14.6      2
38
## 72            2.0         7.08                  12.3      3
52
## 73            5.2         9.53                  15.0      3
241
## 74            4.5        10.05                  36.7      1
144
## 75            3.4         8.45                  12.9      2
143
## 76            4.5         6.70                  13.0      4
51
## 77            2.9         8.90                  12.7      1
37
## 78            4.9        10.23                   9.9      2
595
## 79            4.4         8.88                  14.1      2
165
## 80            5.1        10.30                  27.8      2
113
## 81            2.9        10.79                   2.6      2
320
## 82            3.5         7.94                   6.2      2
139
## 83            5.5         7.63                  11.6      4
109
## 84            4.7         8.77                   5.2      4
```

```
85
## 85            1.7         8.09                  7.6      3
61
## 86            4.1         9.05                 20.5      3
127
## 87            2.9         7.91                 11.9      3
349
## 88            4.3        10.39                 14.0      2
223
## 89            4.8         9.36                 18.3      1
127
## 90            5.8        11.41                 23.8      3
359
## 91            2.9         8.86                  9.5      3
65
## 92            2.0         8.93                  6.2      3
59
## 93            1.3         8.92                  2.2      2
40
## 94            5.3         8.15                 12.3      4
55
## 95            5.3         9.77                 15.7      2
123
## 96            2.5         8.54                 27.0      1
57
## 97            3.8         8.66                  6.8      3
178
## 98            4.8        12.01                 10.8      1
237
## 99            2.3         7.95                  4.6      3
128
## 100           6.2        10.15                 16.4      3
452
## 101           2.6         9.76                  6.9      4
47
## 102           4.3         9.89                 11.8      1
141
## 103           2.7         7.14                 13.1      4
40
## 104           6.6        13.95                 15.6      1
308
## 105           4.5         9.44                 10.9      3
230
## 106           2.9        10.80                  1.6      3
69
## 107           1.4         7.14                  4.1      3
90
## 108           2.1         8.02                  3.8      2
44
## 109           5.7        11.80                  9.1      2
```

```
441
## 110           5.8         9.50                  42.0      3
68
## 111           4.4         7.70                  12.2      4
85
## 112           5.9        17.94                  26.4      1
791
## 113           3.1         9.41                  20.6      3
20
```

```
colnames(ir) <- c("infection_risk", "length", "ratio", "region", "census")
summary(ir)
```

```
##  infection_risk      length           ratio           region     
##  Min.   :1.300   Min.   : 6.700   Min.   : 1.60   Min.   :1.000  
##  1st Qu.:3.700   1st Qu.: 8.340   1st Qu.: 8.40   1st Qu.:2.000  
##  Median :4.400   Median : 9.420   Median :14.10   Median :2.000  
##  Mean   :4.355   Mean   : 9.648   Mean   :15.79   Mean   :2.363  
##  3rd Qu.:5.200   3rd Qu.:10.470   3rd Qu.:20.30   3rd Qu.:3.000  
##  Max.   :7.800   Max.   :19.560   Max.   :60.50   Max.   :4.000  
##      census     
##  Min.   : 20.0  
##  1st Qu.: 68.0  
##  Median :143.0  
##  Mean   :191.4  
##  3rd Qu.:252.0  
##  Max.   :791.0  
```

```
colnames(ir)
```

```
## [1] "infection_risk" "length"         "ratio"          "region"        
## [5] "census"
```

```
dim(ir)
```

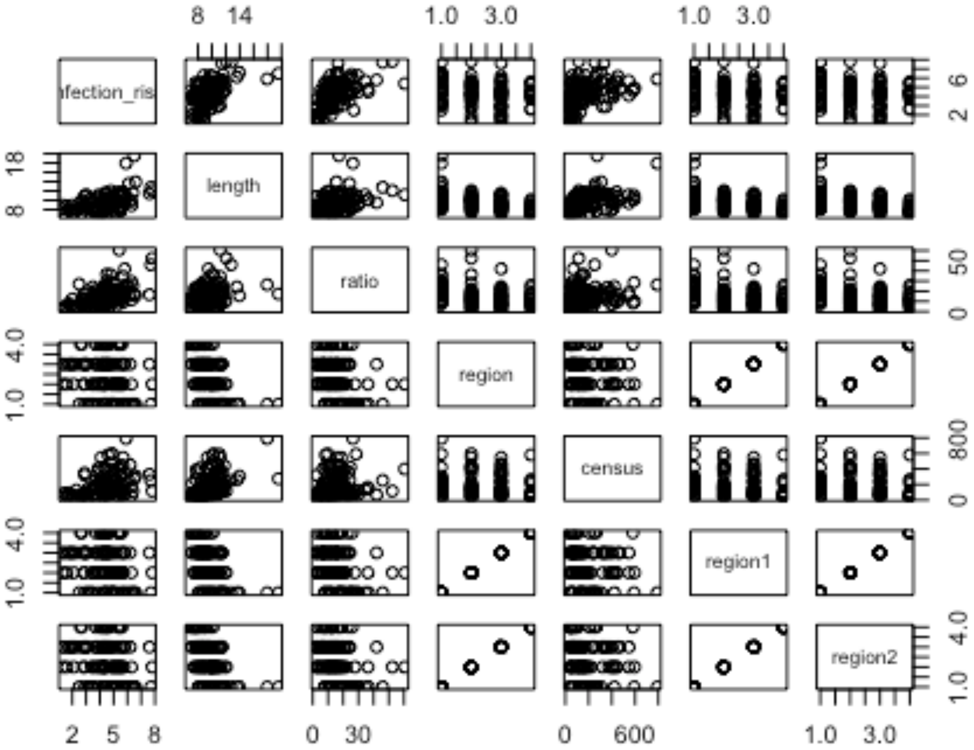
```
## [1] 113   5
```

```
ir$region1 <- as.factor(ir$region)
ir$region2 <- factor(ir$region, labels = c("NE", "NC", "S", "W"))
head(ir)
```

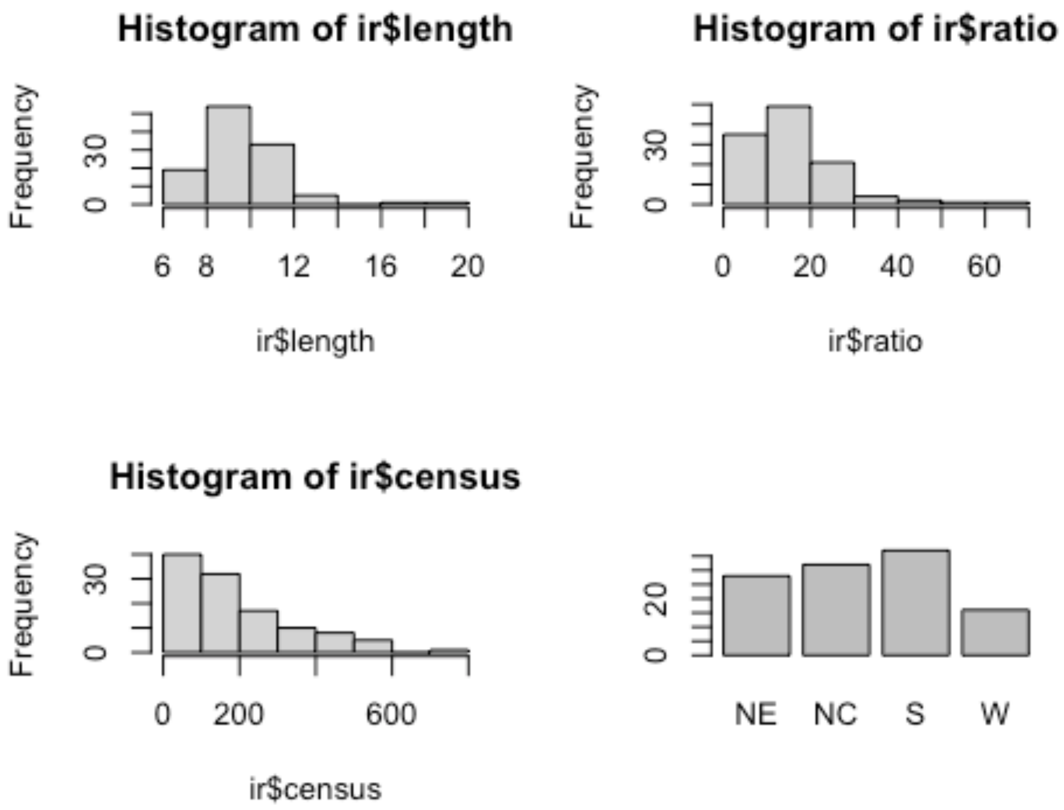
```
##   infection_risk length ratio region census region1 region2
## 1            4.1   7.13   9.0      4    207       4       W
## 2            1.6   8.82   3.8      2     51       2      NC
## 3            2.7   8.34   8.1      3     82       3       S
## 4            5.6   8.95  18.9      4     53       4       W
## 5            5.7  11.20  34.5      1    134       1      NE
## 6            5.1   9.76  21.9      2    147       2      NC
```

1.

```
pairs(ir)
```



```
par(mfrow = c(2,2))
hist(ir$length)
hist(ir$ratio)
hist(ir$census)

barplot(table(ir$region2))
```



We can look at the graphs an conclude that The histograms of the predictors called length, ratio and frequency are all skewed right. But it is difficult to claim outliersfrom the above plots.

2.

```
ir$region2 <- relevel(ir$region2, ref="W")
irregW <- lm(infection_risk ~ length + ratio + region2+census, data=ir)
irregW

##
## Call:
## lm(formula = infection_risk ~ length + ratio + region2 + census,
##     data = ir)
##
## Coefficients:
## (Intercept)       length         ratio     region2NE     region2NC
region2S
##    1.196423     0.278602     0.058033     -0.962847     -0.730743    -
0.855428
##      census
##    0.001461

summary(irregW)
```

```
##
## Call:
## lm(formula = infection_risk ~ length + ratio + region2 + census,
##     data = ir)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2742 -0.6121  0.1412  0.6080  2.6348
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.196422   0.527344   2.269  0.02531 *
## length       0.278602   0.063040   4.419 2.40e-05 ***
## ratio        0.058033   0.009569   6.064 2.07e-08 ***
## region2NE   -0.962847   0.345675  -2.785  0.00633 **
## region2NC   -0.730743   0.305595  -2.391  0.01856 *
## region2S    -0.855428   0.292627  -2.923  0.00424 **
## census       0.001461   0.000678   2.154  0.03348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9565 on 106 degrees of freedom
## Multiple R-squared:  0.5184, Adjusted R-squared:  0.4912
## F-statistic: 19.02 on 6 and 106 DF,  p-value: 6.273e-15
```

The estimated regression function is $\widehat{\text{infection\_risk}} = 1.196423 + 0.278602(length) + 0.058033(ratio) - 0.962847(NE) - 0.730743(NC) - 0.855428(S) + 0.001461(census)$

3.

We will revise the reference level and rerun the analysis by setting “NC” as the reference level The slopes for the dummy variable of the remaining regions is the difference of physicians between NC and the rest of the regions. Then I will use the confint() function to get the confidence intervals.

```
irregNC <- lm(infection_risk ~ length + ratio + relevel(region2,
ref="NC")+census, data=ir)
irregNC

##
## Call:
## lm(formula = infection_risk ~ length + ratio + relevel(region2,
##     ref = "NC") + census, data = ir)
##
## Coefficients:
##                     (Intercept)                          length
##                        0.465679                        0.278602
##                           ratio   relevel(region2, ref = "NC")W
##                        0.058033                        0.730743
## relevel(region2, ref = "NC")NE   relevel(region2, ref = "NC")S
##                       -0.232104                        -0.124685
##                          census
##                        0.001461
```

```
confint(irregNC, level = 0.90)
```

```
##                                           5 %        95 %
## (Intercept)                     -0.4892650624 1.420623654
## length                           0.1739951892 0.383208373
## ratio                            0.0421536524 0.073911481
## relevel(region2, ref = "NC")W    0.2236520381 1.237834437
## relevel(region2, ref = "NC")NE -0.6711321789 0.206924502
## relevel(region2, ref = "NC")S  -0.5123983394 0.263028747
## census                           0.0003355673 0.002585616
```

At 90% confidence, the mean infection risk in NC region differs from W region by (0.2236520381, 1.237834437). At 90% confidence, the mean infection risk in NC region differs from NE region by (-0.6711321789.0.206924502). At 90% confidence, the mean infection risk in NC region differs from S region by (-0.5123983394, 0.263028747).

4.

Zero is included in all of the intervals that compare the infection risk difference between NC the rest of the regions means we can conclude that there is not a significant differenct between NC and the regions. 0 being in the interval suggests that the difference in the means of the NC regions could be zero, implying no significant difference and that we can drop region from the model.

5.

```
irreg.fit <- lm(infection_risk ~ length + ratio +
region2+census+census:length+census:ratio, data=ir)
irreg.fit
```

```
##
## Call:
## lm(formula = infection_risk ~ length + ratio + region2 + census +
##     census:length + census:ratio, data = ir)
##
## Coefficients:
##   (Intercept)        length         ratio      region2NE      region2NC
##    -0.0959197     0.3689018     0.0900904     -1.0298746     -0.8338318
##      region2S        census  length:census   ratio:census
##    -0.9377797     0.0070193     -0.0002599     -0.0001727
```

$$\widehat{\text{infection\_risk}} = 1.196423 + 0.278602(length) + 0.058033(ratio) - 0.962847(NE) - 0.730743(NC) - 0.855428(S) + 0.001461(census) - 0.0002599(census * length) - 0.0001727(census * ratio)$$

6.

First we will perform a partial f-test to test whether or not we can drop the interaction terms from our fitted model in question 5. $H_0: \beta_7 = \beta_8 = 0$ \$H_a:ALOI \$ (At least one inequality)

```
anova(irregW,irreg.fit)

## Analysis of Variance Table
##
## Model 1: infection_risk ~ length + ratio + region2 + census
## Model 2: infection_risk ~ length + ratio + region2 + census +
census:length +
##     census:ratio
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    106 96.981
## 2    104 85.177  2    11.803 7.2058 0.001173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting partial F-statistic is 7.2058, with df1 = 2, df2 = 104. The resulting p-value = 0.001173 Since p-value= 0.001173 < α = 0.05, we reject H0. At 0.05 level, we do not drop the interactions terms from the model.

```
selcri<-function(lmout)
{
n <- length(lmout$fit)
rsq <- summary(lmout)$r.sq
adj.rsq <- summary(lmout)$adj.r.sq
aic <- extractAIC(lmout)[2]
bic <- extractAIC(lmout, k = log(n))[2]
press <- sum((lmout$residuals/(1 - hatvalues(lmout)))^2)
cbind(rsq, adj.rsq, aic, bic, press)
}

selcri(irregW)

##            rsq   adj.rsq       aic     bic    press
## [1,] 0.5184188 0.4911594 -3.274917 15.8168 114.2829

selcri(irreg.fit)

##           rsq  adj.rsq       aic      bic    press
## [1,] 0.577031 0.544495 -13.93962 10.60687 104.4905
```

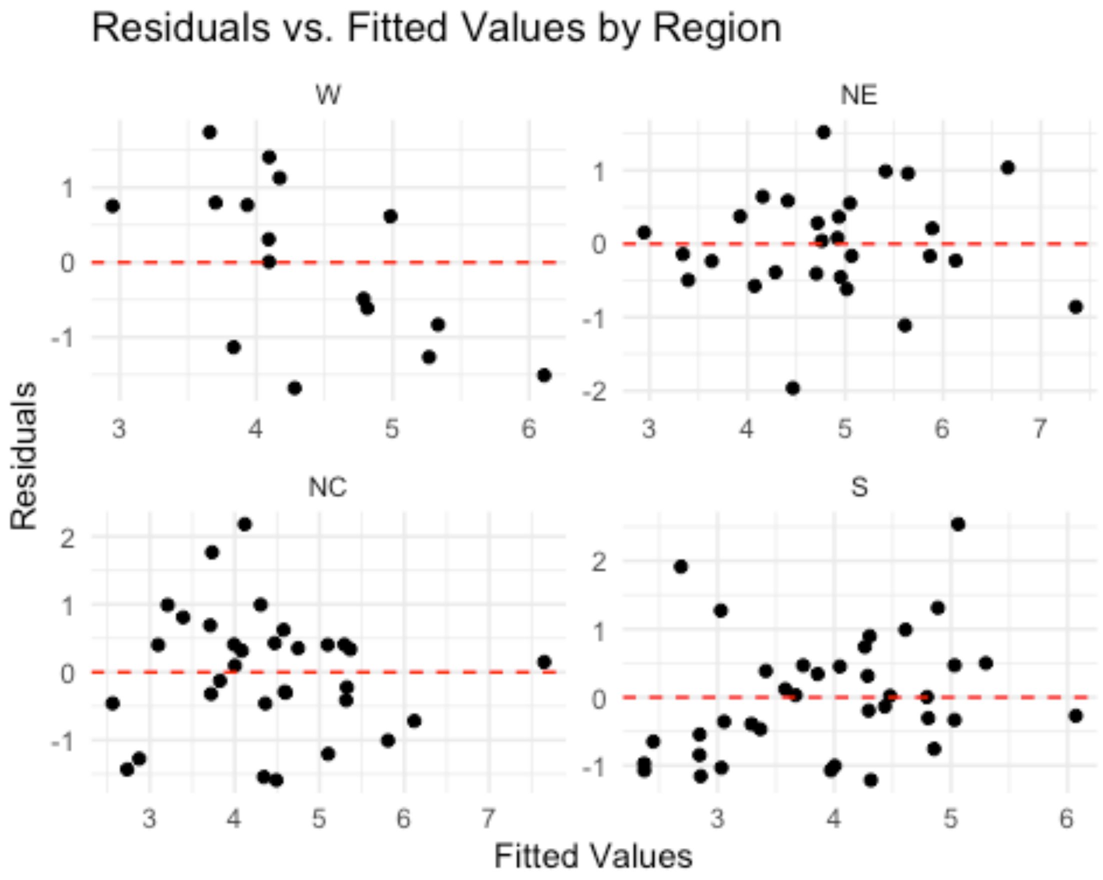
We typically look for models with smaller AIC or BIC. in this case the fitted model with the interactive values have lower AIC and BIC values with AIC=-13.93962<-3.274917 and BIC=10.60687<15.8168.

The PRESS (prediction sum of squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i. A model with a small press is prefered so in this case the model with the interactive terms has the smaller press value.

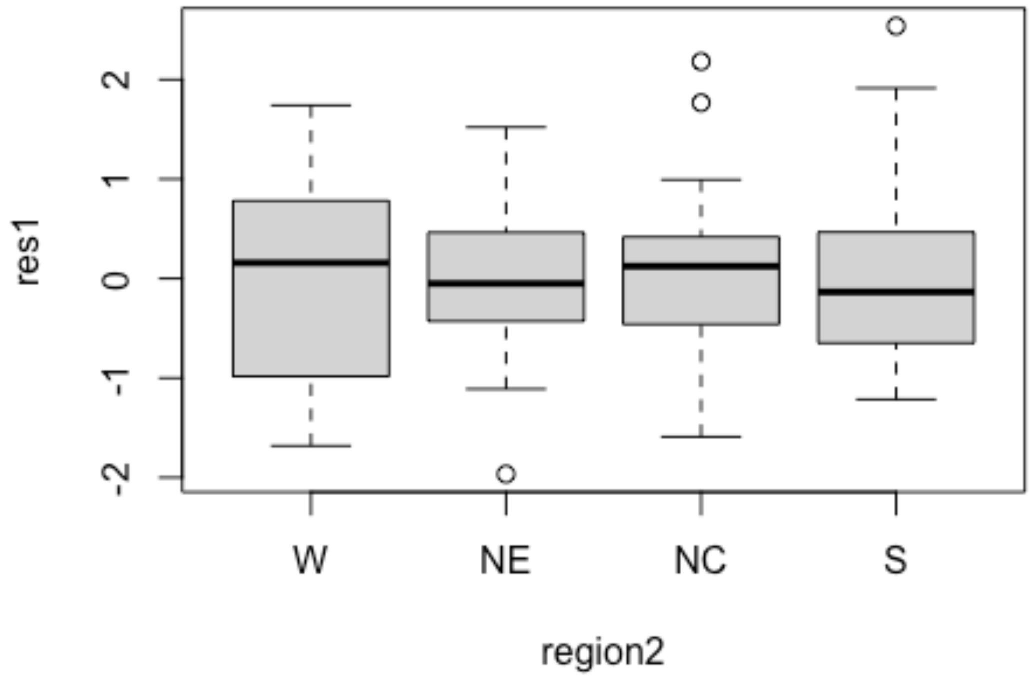
Based on these values I would recommend to use the irreg.fit model which is the one with the interactive terms added to it.

7.

```
ir$resid1 <- irreg.fit$residuals
ir$fitted <- fitted(irreg.fit)

library(ggplot2)
ggplot(ir, aes(x = fitted, y = resid1)) +
  geom_point() + geom_hline(yintercept = 0, linetype = "dashed", color =
"red") + labs(title = "Residuals vs. Fitted Values by Region",
       x = "Fitted Values", y = "Residuals") +
  facet_wrap(~region2, scales = "free") +
  theme_minimal()
```



```
ir$res1<-irreg.fit$resi
plot(res1~region2, data=ir)
```



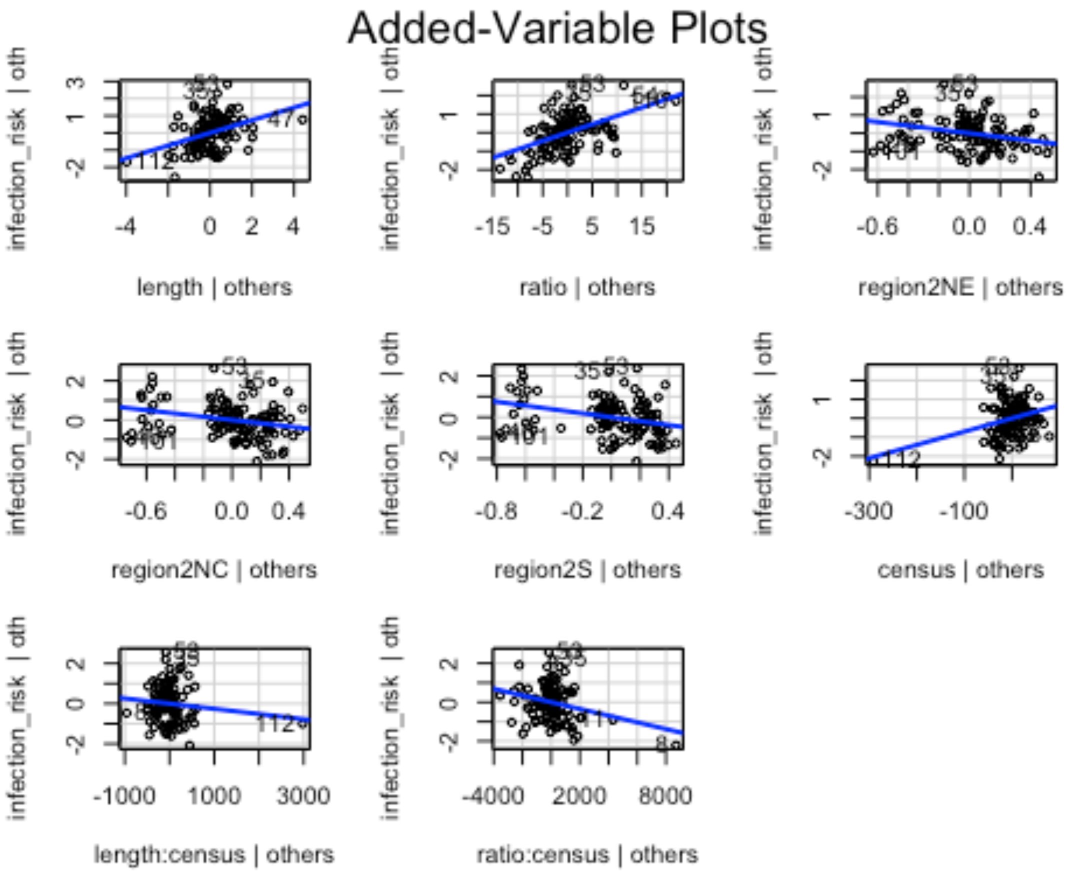
Based off of looking at the graphs, the variance of residuals S and W stay relatively constant while the residuals are slightly non-constant for NE and NC. More specifically they tend to vary less as the fitted values increase for NC.

8.

In order to revise and improve the model if the residuals of observations in some region(s) spread wider than those from other region(s) we would have to perform a box cox transformation. The transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variances, and nonlinearity of the regression function. The Box-Cox procedure automatically identifies a transformation from the family of power transformations on Y, which is of the form $Y' = Y^{\lambda}$ where λ is the parameter to be determined.

9.

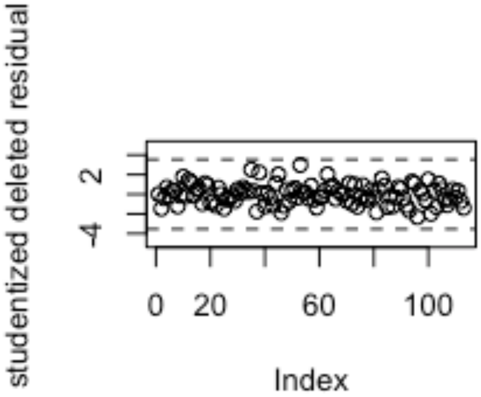
```
library(car)

## Loading required package: carData

avPlots(irreg.fit)
```



The first step I took was plotting the AV plots to see if the regression relationships in the fitted regression function in Question 5 are inappropriate for any of the predictor variables. We can see that there doesnt seem to be any curvature in any of the plots.

```
af <- rstudent(irreg.fit)
head(af)

##            1            2            3            4            5 
6
##  0.008178396 -1.463344586 -0.399018594  0.705764996 -0.189014027 
0.396307320

n <- 199
p <- 9
fam <- 0.05
tcrit1 <- qt(1-fam/(2*n), n-p-1)
tcrit1

## [1] 3.731929

af[abs(af) > tcrit1]

## named numeric(0)
```

```
par(mfrow=c(2,2))
plot(rstudent(irreg.fit), ylab="studentized deleted residual", ylim = c(-
5,5))
abline(qt(1-0.1/(2*n), df=n - p - 1), 0, lty=2)
abline(qt(0.1/(2*n), df=n - p - 1), 0, lty=2)
```



Next we check the studentized residuals to see if there are any outliers. Decision rule: if $|t_i| > t(1-\alpha/(2n), n-p-1)$, conclude observation i is an outlier with respect to the response variable. Otherwise the i-th observation is not an outlier on the response variable. Conclusion: According to the studentized deleted residual, we can find the numbers are smaller than t = 3.73. We can conclude that there is no outlying Y observation in the model.

```
step(irreg.fit, direction="both")

## Start:  AIC=-13.94
## infection_risk ~ length + ratio + region2 + census + census:length +
##     census:ratio
##
##                  Df Sum of Sq    RSS      AIC
## - length:census   1    1.0958 86.273 -14.4951
## <none>                        85.177 -13.9396
## - ratio:census    1    6.2673 91.445  -7.9168
## - region2         3   10.2698 95.447  -7.0761
##
```

```
## Step:  AIC=-14.5
## infection_risk ~ length + ratio + region2 + census + ratio:census
## 
##                  Df Sum of Sq     RSS      AIC
## <none>                         86.273 -14.4951
## + length:census   1    1.0958  85.177 -13.9396
## - region2         3    9.3442  95.617  -8.8747
## - ratio:census    1   10.7075  96.981  -3.2749
## - length          1   19.8016 106.075   6.8535

## 
## Call:
## lm(formula = infection_risk ~ length + ratio + region2 + census + 
##     ratio:census, data = ir)
## 
## Coefficients:
##  (Intercept)        length         ratio     region2NE     region2NC  
##    0.4628978     0.2940274     0.0970332    -0.9744121    -0.7659808  
##     region2S        census  ratio:census  
##   -0.8703405     0.0048456    -0.0002038
```

The stepwise selection algorithm uses AIC as the selection criterion, so based on AIC, the predictors length, ratio, NE,NC,S,census, Ratio*census are selected to fit a regression model for job proficiency. The estimated regression function is:

$$\widehat{\text{infection\_risk}} = 0.4628978 + 0.2940274(length) + 0.0970332(ratio) - 0.9744121(NE) - 0.7659808(NC) - 0.8703405(S) + 0.0048456(census) - 0.0002038(census * ratio)$$

```
irregg<-lm(infection_risk ~ length + ratio + region2+census+census:ratio, 
data=ir)
selcri(irreg.fit)

##           rsq  adj.rsq       aic      bic    press
## [1,] 0.577031 0.544495 -13.93962 10.60687 104.4905

selcri(irregg)

##            rsq   adj.rsq       aic      bic    press
## [1,] 0.5715894 0.5430287 -14.49512 7.323981 100.8548
```

Based on the AIC, we should choose to use the stepwise selection function that does not inculde the census*length interaction term.

10.

The first 5 rows of the design matrix for the model in question 5 is,

$$\begin{aligned}&\widehat{\text{infection\_risk}}\\&= 1.196423 + 0.278602(length) + 0.058033(ratio) - 0.962847(NE) - 0.730743(NC)\\&\quad - 0.855428(S) + 0.001461(census) - 0.0002599(census * length)\\&\quad - 0.0001727(census * ratio)\end{aligned}$$

$$X = \begin{pmatrix} 1 & 7.13 & 9.0 & 0 & 0 & 0 & 207 & 207 * 7.13 & 207 * 9.0 \\ 1 & 8.82 & 3.8 & 0 & 1 & 0 & 51 & 51 * 8.82 & 61 * 3.8 \\ 1 & 8.34 & 8.1 & 0 & 0 & 1 & 82 & 82 * 8.34 & 82 * 8.1 \\ 1 & 8.95 & 18.9 & 0 & 0 & 0 & 53 & 53 * 8.95 & 53 * 18.9 \\ 1 & 11.20 & 34.5 & 1 & 0 & 0 & 134 & 134 * 11.20 & 134 * 34.5 \end{pmatrix}$$