Chimdi_Homework6

Chimdi Chikezie

5/30/2023

```
library(leaps)
library(Sleuth3)
library(ggplot2)
```

###Problem 1 Suppose that X1, X2 and X3 are three explanatory variables in a multiple linear regression model with n = 28 observations. The following table shows the residual sum of squares and degrees of freedom for all the possible models: ModelVariables | SS(Residual) | df (Residual) None | 8100 | 27 X1 | 6240 | 26 X2 | 5980 | 26 X3 | 6760 | 26 X1 , X2 | 5500 | 25 X1 , X3 | 5250 | 25 X2 , X3 | 5750 | 25 X1 , X2 , X3 | 5160 | 24

(a)

```
SS <- c(8100, 6240, 5980, 6760, 5500, 5250, 5750, 5160)
dfr <- c(27, 26, 26, 26, 25, 25, 25, 24)
p <- c(0, 1, 1, 1, 2, 2, 2, 3)
p <- (p+1) #SinceIamusingformulasgiveninlab
p_str <- c("None", "X1", "X2", "X3", "X1, X2", "X1, X3", "X2, X3", "X1, X2, X3")
MSE <- SS/dfr
MSE
```

```
## [1] 300 240 230 260 220 210 230 215
```

(b)

```
n = 28
Cp <- (n - p)*MSE/MSE[length(MSE)] + 2*p - n
p_str
```

```
## [1] "None"       "X1"         "X2"         "X3"         "X1, X2"
## [6] "X1, X3"     "X2, X3"     "X1, X2, X3"
```

```
Cp
```

```
## [1] 11.674419  5.023256  3.813953  7.441860  3.581395  2.418605  4.744186
## [8]  4.000000
```

(c)

```
BIC <- n*log(MSE) + (p+1)*log(n)
p_str
```

```
## [1] "None"       "X1"         "X2"         "X3"         "X1, X2"
## [6] "X1, X3"     "X2, X3"     "X1, X2, X3"
```

```
BIC
```

```
## [1] 166.3703 163.4545 162.2628 165.6957 164.3504 163.0478 165.5950 167.0389
```

(d)

```
AIC <- n*log(MSE) + 2*(p+1)
p_str
```

```
## [1] "None"       "X1"         "X2"         "X3"         "X1, X2"
## [6] "X1, X3"     "X2, X3"     "X1, X2, X3"
```

```
AIC
```

```
## [1] 163.7059 159.4579 158.2662 161.6991 159.0216 157.7190 160.2662 160.3779
```

(e)

```
sprintf("Smallest Cp is X1,X3 cp = %0.3f", min(Cp))
```

```
## [1] "Smallest Cp is X1,X3 cp = 2.419"
```

```
sprintf("Smallest BIC is X2, BIC = %0.3f", min(BIC))
```

```
## [1] "Smallest BIC is X2, BIC = 162.263"
```

```
sprintf("Smallest AIC is X1,X3, AIC = %0.3f", min(AIC))
```

```
## [1] "Smallest AIC is X1,X3, AIC = 157.719"
```

###Problem 2

```
head(ex1220)
```

```
##         Island Total Native  Area Elev DistNear DistSc AreaNear
## 1       Baltra    58     23 25.09  332      0.6    0.6     1.84
## 2    Bartolome    31     21  1.24  109      0.6   26.3   572.33
## 3     Caldwell     3      3  0.21  114      2.8   58.7     0.78
## 4     Champion    25      9  0.10   46      1.9   47.4     0.18
## 5      Coamano     2      1  1.05  130      1.9    1.9   903.82
## 6 Daphne Major    18     11  0.34  119      8.0    8.0     1.84
```

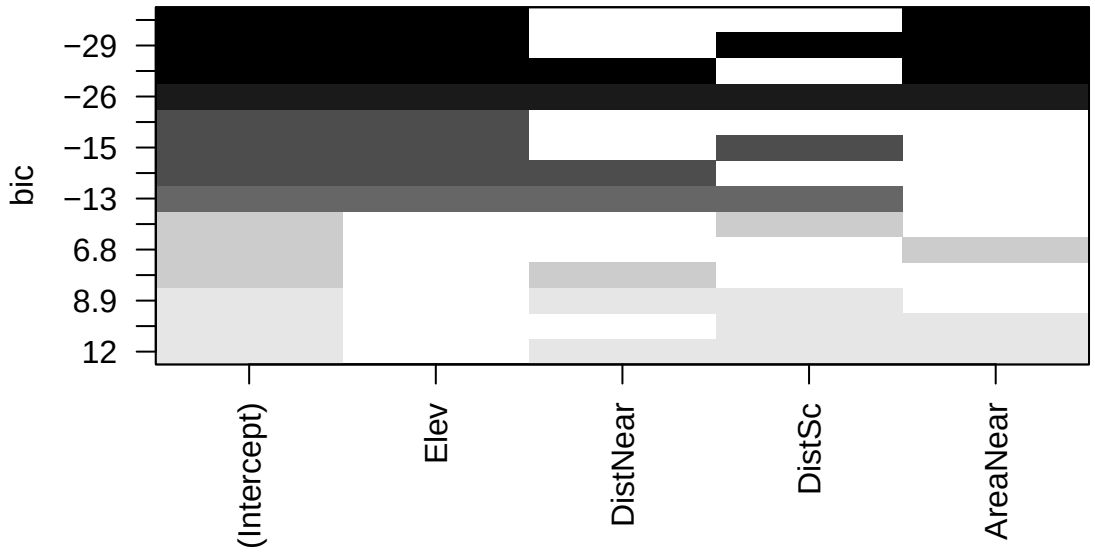
```
data1 <- ex1220
```

a) With total number of species (Total) as the response, based on Cp and BIC, select the five best fitting regression models involving all the explanatory variables except the island area (Area)

```
all_mod <- regsubsets(Total ~ Elev + DistNear + DistSc + AreaNear, data = data1,
                  nbest = 5, method = "exhaustive")
summary(all_mod)
```

```
## Subset selection object
## Call: regsubsets.formula(Total ~ Elev + DistNear + DistSc + AreaNear,
##     data = data1, nbest = 5, method = "exhaustive")
## 4 Variables  (and intercept)
##          Forced in Forced out
## Elev         FALSE      FALSE
## DistNear     FALSE      FALSE
## DistSc       FALSE      FALSE
## AreaNear     FALSE      FALSE
## 5 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          Elev DistNear DistSc AreaNear
## 1  ( 1 ) "*"  " "      " "    " "
## 1  ( 2 ) " "  " "      "*"    " "
## 1  ( 3 ) " "  " "      " "    "*"
## 1  ( 4 ) " "  "*"      " "    " "
## 2  ( 1 ) "*"  " "      " "    "*"
## 2  ( 2 ) "*"  " "      "*"    " "
## 2  ( 3 ) "*"  "*"      " "    " "
## 2  ( 4 ) " "  "*"      "*"    " "
## 2  ( 5 ) " "  " "      "*"    "*"
## 3  ( 1 ) "*"  " "      "*"    "*"
## 3  ( 2 ) "*"  "*"      " "    "*"
## 3  ( 3 ) "*"  "*"      "*"    " "
## 3  ( 4 ) " "  "*"      "*"    "*"
## 4  ( 1 ) "*"  "*"      "*"    "*"
```

```
plot(all_mod)
```

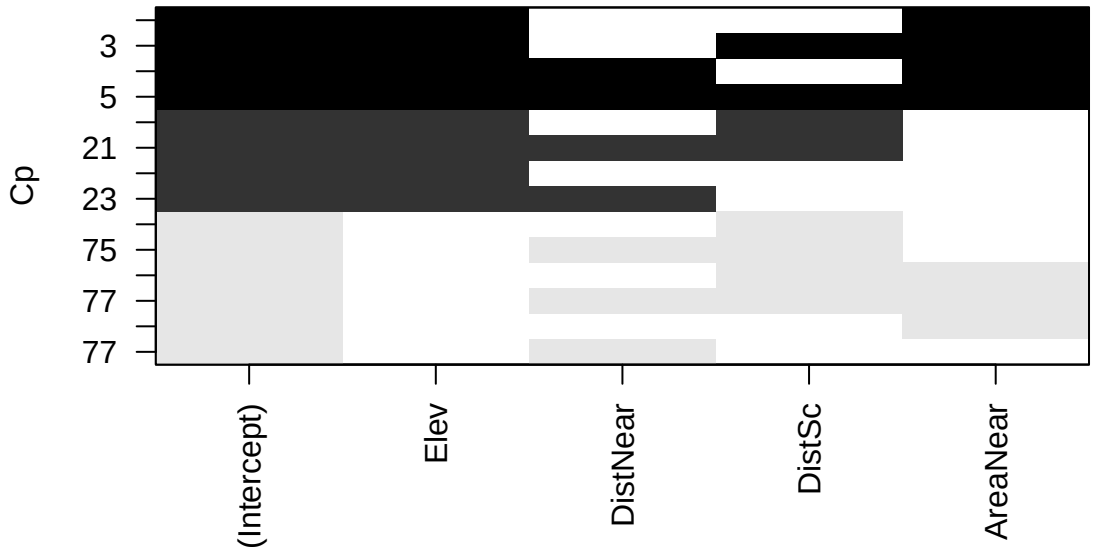



```
summary(all_mod)$cp
```

```
##  [1] 21.106287 74.408188 77.367034 77.417308  2.735411 20.480386 23.102820
##  [8] 75.022509 76.281082  3.045589  4.305329 21.063546 76.676287  5.000000
```

Best 5 models from BIC:
1. $\beta_0 + \beta_1$Elev $+ \beta_2$AreaNear
2. $\beta_0 + \beta_1$Elev $+ \beta_2$DistSc $+ \beta_3$AreaNear
3: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$AreaNear
4: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$DistSc $+ \beta_4$AreaNear
5: $\beta_0 + \beta_1$Elev

```
plot(all_mod, scale = 'Cp')
```



Best 5 models from Cp:
1. $\beta_0 + \beta_1$Elev $+ \beta_2$AreaNear
2. $\beta_0 + \beta_1$Elev $+ \beta_2$DistSc $+ \beta_3$AreaNear
3: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$AreaNear
4: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$DistSc $+ \beta_4$AreaNear
5: $\beta_0 + \beta_1$Elev $+ \beta_3$DistSc

(b) To the model with the lowest Cp, add the island area (Area) variable and obtain the p-value from the extra-sum-of-squares F -test due to its addition

```
modwithout <- lm(Total ~ Elev + AreaNear, data = data1)
modwith <- lm(Total ~ Elev + Area + AreaNear, data = data1)
anova(modwithout, modwith)
```

```
## Analysis of Variance Table
##
## Model 1: Total ~ Elev + AreaNear
## Model 2: Total ~ Elev + Area + AreaNear
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     27 98497
## 2     26 94791  1    3706.8 1.0167 0.3226
```

H_0: Model 1 is significant H_A: Model 2 is significant

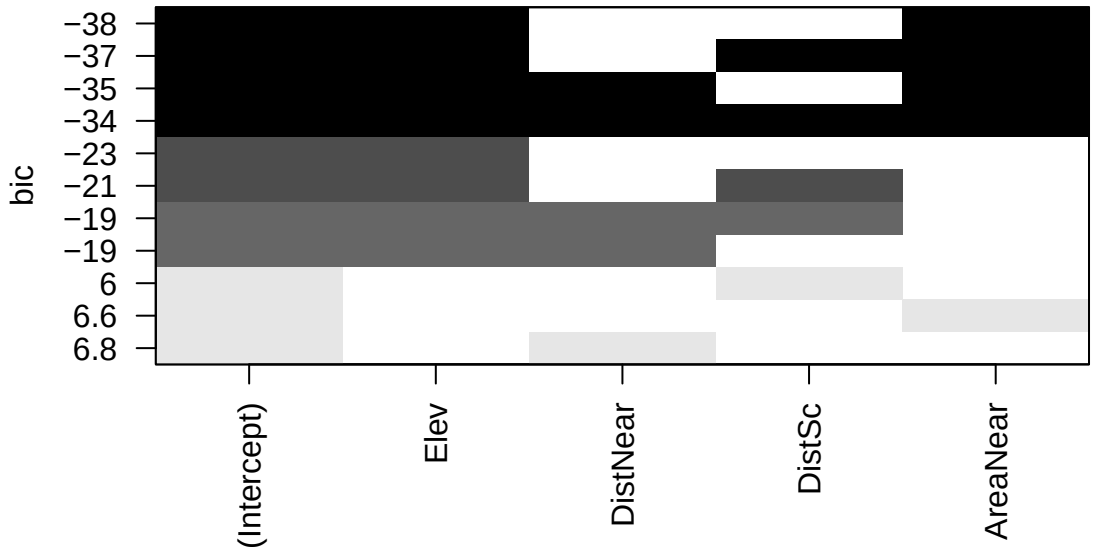
The p-value (0.3226) from the extra-sum-of-squares F-test is less than $\alpha = 0.05$ which means it is not statistically significant. We therefore fail to reject the null hypothesis that Model 1 is significant and conclude that the simpler model fits the data better.

(c) With total native number of species (Native) as the response, find the best fitting regression model based on sequential variable selection technique - forward selection and backward elimination involving all the explanatory variables except the island area (Area).

```
foward_mod <- regsubsets(Native ~ Elev + DistNear + DistSc + AreaNear, data = data1,
                  nbest = 5, method = "forward")
summary(foward_mod)
```

```
## Subset selection object
## Call: regsubsets.formula(Native ~ Elev + DistNear + DistSc + AreaNear,
##     data = data1, nbest = 5, method = "forward")
## 4 Variables  (and intercept)
##          Forced in Forced out
## Elev         FALSE      FALSE
## DistNear     FALSE      FALSE
## DistSc       FALSE      FALSE
## AreaNear     FALSE      FALSE
## 5 subsets of each size up to 4
## Selection Algorithm: forward
##          Elev DistNear DistSc AreaNear
## 1  ( 1 ) "*"  " "      " "    " "
## 1  ( 2 ) " "  " "      "*"    " "
## 1  ( 3 ) " "  " "      " "    "*"
## 1  ( 4 ) " "  "*"      " "    " "
## 2  ( 1 ) "*"  " "      " "    "*"
## 2  ( 2 ) "*"  " "      "*"    " "
## 2  ( 3 ) "*"  "*"      " "    " "
## 3  ( 1 ) "*"  " "      "*"    "*"
## 3  ( 2 ) "*"  "*"      " "    "*"
## 3  ( 3 ) "*"  "*"      "*"    " "
## 4  ( 1 ) "*"  "*"      "*"    "*"
```

```
plot(foward_mod)
```



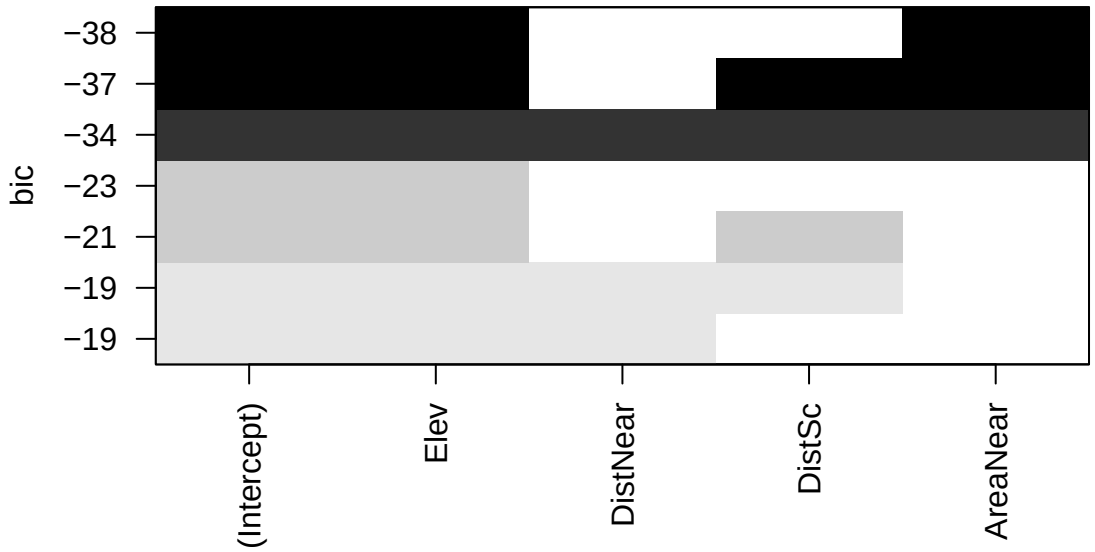
Best 5 forward models from BIC:

1. $\beta_0 + \beta_1 \text{Elev} + \beta_2 \text{AreaNear}$
2. $\beta_0 + \beta_1 \text{Elev} + \beta_2 \text{DistSc} + \beta_3 \text{AreaNear}$
3: $\beta_0 + \beta_1 \text{Elev} + \beta_2 \text{DistNear} + \beta_3 \text{AreaNear}$
4: $\beta_0 + \beta_1 \text{Elev} + \beta_2 \text{DistNear} + \beta_3 \text{DistSc} + \beta_4 \text{AreaNear}$
5: $\beta_0 + \beta_1 \text{Elev}$

```
backward_mod <- regsubsets(Native ~ Elev + DistNear + DistSc + AreaNear, data = data1,
                  nbest = 5, method = "backward")
summary(backward_mod)
```

```
## Subset selection object
## Call: regsubsets.formula(Native ~ Elev + DistNear + DistSc + AreaNear,
##     data = data1, nbest = 5, method = "backward")
## 4 Variables  (and intercept)
##          Forced in Forced out
## Elev         FALSE      FALSE
## DistNear     FALSE      FALSE
## DistSc       FALSE      FALSE
## AreaNear     FALSE      FALSE
## 5 subsets of each size up to 4
## Selection Algorithm: backward
##          Elev DistNear DistSc AreaNear
## 1  ( 1 ) "*"  " "      " "    " "
## 2  ( 1 ) "*"  " "      " "    "*"
## 2  ( 2 ) "*"  " "      "*"    " "
## 2  ( 3 ) "*"  "*"      " "    " "
## 3  ( 1 ) "*"  " "      "*"    "*"
## 3  ( 2 ) "*"  "*"      "*"    " "
## 4  ( 1 ) "*"  "*"      "*"    "*"
```

```
plot(backward_mod)
```



Best 5 backward models from BIC:

1. $\beta_0 + \beta_1$Elev $+ \beta_2$AreaNear
2. $\beta_0 + \beta_1$Elev $+ \beta_2$DistSc $+ \beta_3$AreaNear
3: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$AreaNear
4: $\beta_0 + \beta_1$Elev $+ \beta_2$DistNear $+ \beta_3$DistSc $+ \beta_4$AreaNear
5: $\beta_0 + \beta_1$Elev $+ \beta_2$DistSc

(d) To the best fitting model from forward regression, add the island area (Area) variable and obtain the p-value from the extra-sum-of-squares F -test due to its addition

```
modwithout2 <- lm(Native ~ Elev + AreaNear, data = data1)
modwith2 <- lm(Native ~ Elev+ Area + AreaNear, data = data1)
anova(modwithout2, modwith2)
```

```
## Analysis of Variance Table
##
## Model 1: Native ~ Elev + AreaNear
## Model 2: Native ~ Elev + Area + AreaNear
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 4288.5
## 2     26 3815.1  1    473.41 3.2263 0.0841 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: Model 1 is significant H_A: Model 2 is significant

The p-value (0.0841) from the extra-sum-of-squares F-test is less than $\alpha = 0.05$ which means it is not statistically significant. We therefore fail to reject the null hypothesis that Model 1 is significant and conclude that the simpler model fits the data better.

###Problem 3 Pollution and Mortality. Look at the description of the data set in Problem 15.18 (page 473-474) of Sleuth. Each part carries three marks. (a) Fit a regression model of the number of Cases on Year, Vaccine and their interaction. Is there any effect of Vaccine and the interaction on Cases?

```
#?ex1518
head(ex1518)
```

```
##   Year  Cases Vaccine
## 1 1950 319124      no
## 2 1951 530118      no
## 3 1952 683077      no
## 4 1953 449146      no
## 5 1954 682720      no
## 6 1955 555156      no
```

```
data3 <- ex1518
```

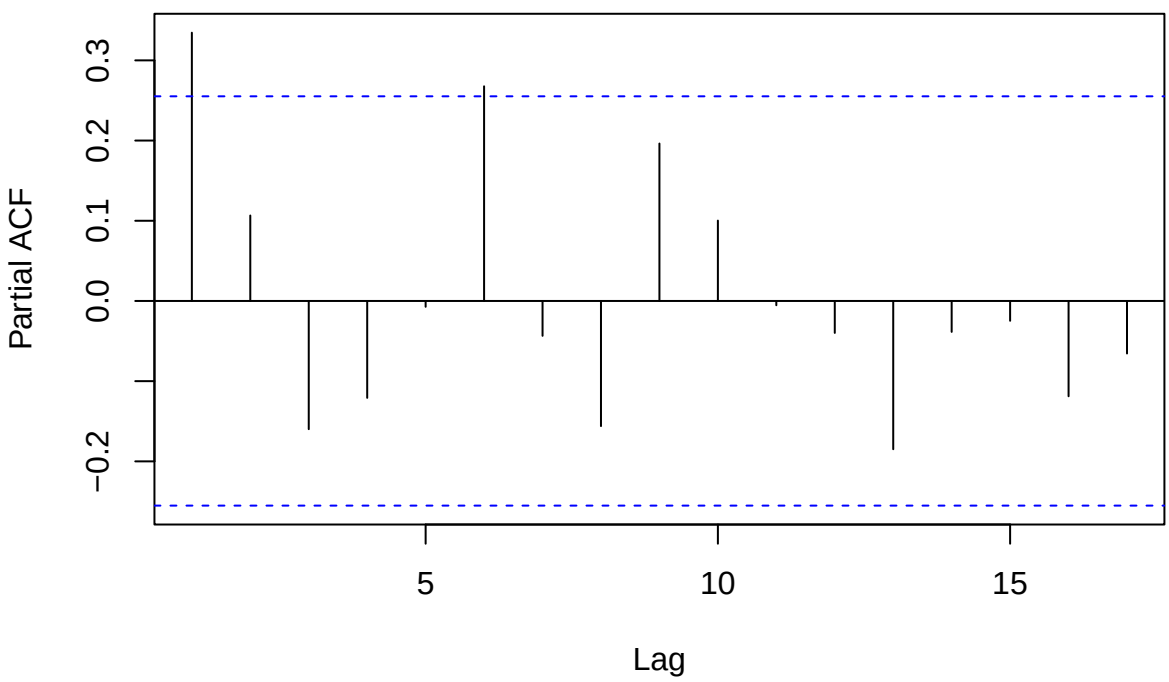
```
linmod3 <- lm(Cases ~ Year * Vaccine, data = data3)
summary(linmod3)
```

```
##
## Call:
## lm(formula = Cases ~ Year * Vaccine, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -225021  -54267  -11590   27198  327124
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6529317   13382183   0.488    0.628
## Year               -3069       6842  -0.449    0.655
## Vaccineyes       1815479   13536073   0.134    0.894
## Year:Vaccineyes    -1113       6918  -0.161    0.873
##
## Residual standard error: 92300 on 55 degrees of freedom
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8349
## F-statistic:  98.8 on 3 and 55 DF,  p-value: < 2.2e-16
```

The parameter that measures the change in mean of vaccines is β_2 which has a p-value (0.894) greater than $\alpha = 0.05$, which means it is not statistically significant and we conclude that there is no difference in mean between the yes or no. Therefore, we conclude that whether the measles vaccine had been licensed or not, it does not affect the number of measles cases. The parameter that measures the interaction of vaccines and year is β_3 which has a p-value (0.873) greater than $\alpha = 0.05$, which means it is not statistically significant and we conclude that the interaction term does not have an effect on the number of measles cases.

(b) Adjust the standard errors of the estimates using autocorrelation of the residuals. Do the p-values of the tests in part (a) change after standardization of the standard errors?

```
linmod3 <- lm(Cases ~ Year * Vaccine, data = data3)
pacf(residuals(linmod3))
```




```
pacf(residuals(linmod3), plot = F)$acf[1]
```

```
## [1] 0.3345066
```

```
r1 <- acf(residuals(linmod3), plot = F)$acf[2]
SE_adj <- sqrt((1+r1)/(1-r1))*summary(linmod3)$coef[,2]
SE_adj
```

```
##     (Intercept)            Year      Vaccineyes Year:Vaccineyes
##     18950272.13         9688.26     19168193.50         9796.40
```

Then, we can use the adjusted standard errors, to both do t-tests and construct confidence intervals.

```
n <- nrow(data3)
t_stat <- abs(summary(linmod3)$coef[,1])/SE_adj
p_value <-2 * pt(-abs(t_stat), n-4, lower.tail = TRUE)
summary(linmod3)$coef
```

```
##                    Estimate   Std. Error    t value  Pr(>|t|)
## (Intercept)     6529316.879 13382182.898  0.4879112 0.6275522
## Year              -3069.319     6841.594 -0.4486262 0.6554639
## Vaccineyes      1815479.026 13536073.231  0.1341215 0.8937961
## Year:Vaccineyes   -1112.895     6917.959 -0.1608704 0.8727850
```

```
round(cbind(SE_adj, t_stat, p_value), 4)
```

```
##                       SE_adj t_stat p_value
## (Intercept)     18950272.126 0.3446  0.7317
## Year                9688.261 0.3168  0.7526
## Vaccineyes      19168193.501 0.0947  0.9249
## Year:Vaccineyes     9796.400 0.1136  0.9100
```

Yes, the p-values changed after standardization of the standard errors but the parameters are still not statistically significant.