Funmi - HW6

2023-05-31

#Question 1

```
n <-28
SSR <- c(8100,  6240, 5980, 6760, 5500, 5250, 5750, 5160)
df <- c(27, 26, 26, 26, 25, 25, 25, 24)
p <- c(0, 1, 1, 1, 2, 2, 2, 3)
sigma2.full <- 215

MSE <- SSR/df
Cp <- (SSR/sigma2.full) -n + 2* (p+1)
BIC <- n * log(SSR/n) + log(n) * (p + 2)
AIC <- n * log(SSR/n) + 2 * (p + 2)
MSE

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

Cp

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

BIC

## [1] 165.3520 161.3795 160.1878 163.6207 161.1772 159.8746 162.4218
162.7227

AIC

## [1] 162.6876 157.3829 156.1912 159.6241 155.8484 154.5458 157.0930
156.0616

which.min(Cp)

## [1] 6

which.min(BIC)

## [1] 6

which.min(AIC)

## [1] 6
```

The output is presented in the table below;

Model	***SS*(*Residual*)**	***df*(*Residual*)**	**σ^2**	**Cp**	**BIC**	**AIC**
None	8100	27	300	11.67	165.35	162.69
X_1	6240	26	240	5.02	161.38	157.38
X_2	5980	26	230	3.81	160.19	156.19
X_3	6760	26	260	7.44	163.62	159.62
X_1,X_2	5500	25	220	3.58	161.18	155.85
X_1,X_3	5250	25	210	2.42	159.87	154.55
X_2,X_3	5750	25	230	4.74	162.42	157.09
X_1,X_2,X_3	5160	24	215	4.00	162.72	156.06

The model with the smallest Cp from the R-output above is the 6th model, the smallest BIC is the 6th model, and the smallest AIC is also the 6th model, which is the model with explanatory variables X_1 and X_3.

#Question 2

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

data <- ex1220
```

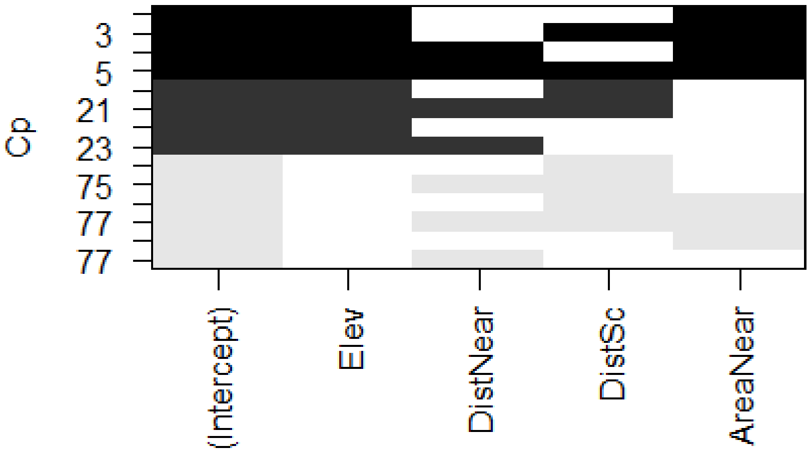
#Question 2a

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 <- regsubsets(Total ~  Elev + DistNear + DistSc + AreaNear,
                  data = subset(ex1220),
                  nbest = 5, method = "exhaustive")
summary(all)
```

```
## Subset selection object
## Call: regsubsets.formula(Total ~ Elev + DistNear + DistSc + AreaNear,
##     data = subset(ex1220), 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 ) "*"  "*"      "*"    "*"
```

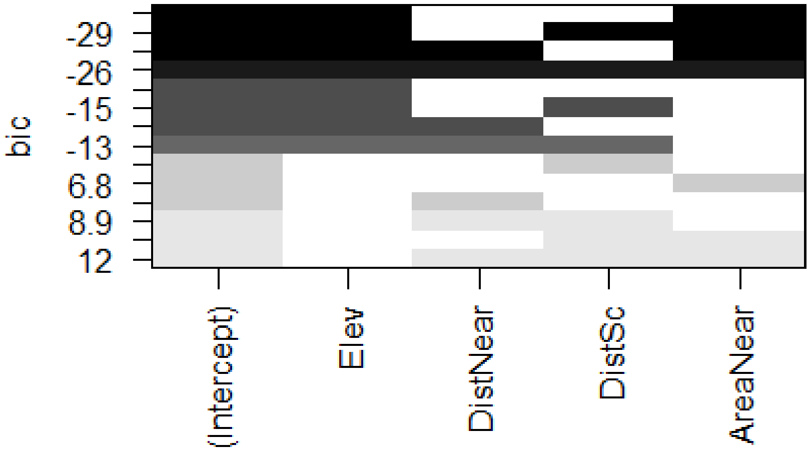
```
plot(all, scale="Cp")
```



Based on Cp, i.e. the models with lowest cp values, the five best fitting regression models are;

1. Total (Intercept) + Elev + AreaNear
2. Total (Intercept) + Elev + DistSc + AreaNear
3. Total (Intercept) + Elev + DistNear + AreaNear
4. Total (Intercept) + Elev + DistNear + DistSc + AreaNear
5. Total (Intercept) + Elev + DistSc

```
plot(all)
```



Based on BIC, i.e., the models with lowest BIC values, the five best fitting regression models are;

1. Total (Intercept) + Elev + AreaNear
2. Total (Intercept) + Elev + DistSc + AreaNear
3. Total (Intercept) + Elev + DistNear + AreaNear
4. Total (Intercept) + Elev + DistNear + DistSc + AreaNear
5. Total (Intercept) + Elev

#Question 2b

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.

```
cp <- summary(all)$cp
results <- list(summary(all)$which[which.min(cp),],
                cp = summary(all)$cp[which.min(cp)], bic =
summary(all)$bic[which.min(cp)])
results

## [[1]]
## (Intercept)        Elev    DistNear      DistSc    AreaNear
##         TRUE        TRUE       FALSE       FALSE        TRUE
##
## $cp
## [1] 2.735411
##
## $bic
## [1] -30.38587
```

```
lin_mod <- lm(Total ~ Elev + AreaNear, data = data)
summary(lin_mod)
```

```
##
## Call:
## lm(formula = Total ~ Elev + AreaNear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.257  -33.946   -8.186   23.082  202.485
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.88675   14.93454   0.059 0.953090
## Elev         0.27855    0.03166   8.797 2.06e-09 ***
## AreaNear    -0.06997    0.01543  -4.536 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.4 on 27 degrees of freedom
## Multiple R-squared:  0.7415, Adjusted R-squared:  0.7224
## F-statistic: 38.73 on 2 and 27 DF,  p-value: 1.168e-08
```

```
lin_mod1 <- lm(Total ~ Elev + AreaNear + Area, data = data)
summary(lin_mod1)
```

```
##
## Call:
## lm(formula = Total ~ Elev + AreaNear + Area, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.631  -34.770   -5.518   28.981  194.043
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.89694   16.80744  -0.410 0.684911
## Elev         0.32021    0.05205   6.152 1.67e-06 ***
## AreaNear    -0.07702    0.01693  -4.549 0.000111 ***
## Area        -0.02187    0.02169  -1.008 0.322585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 60.38 on 26 degrees of freedom
## Multiple R-squared:  0.7513, Adjusted R-squared:  0.7226
## F-statistic: 26.18 on 3 and 26 DF,  p-value: 5.12e-08
```

```
anova(lin_mod, lin_mod1)
```

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

The p-value from the extra-sum-of-squares F -test due to the addition of variable Area is 0.3226.

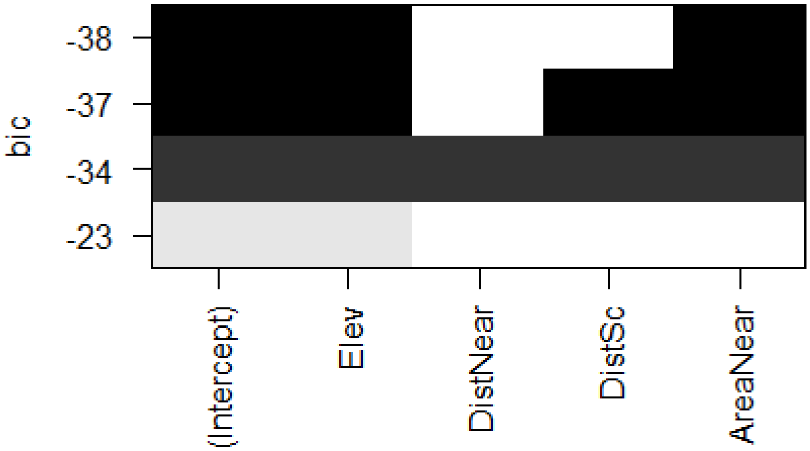
#Question 2c

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).

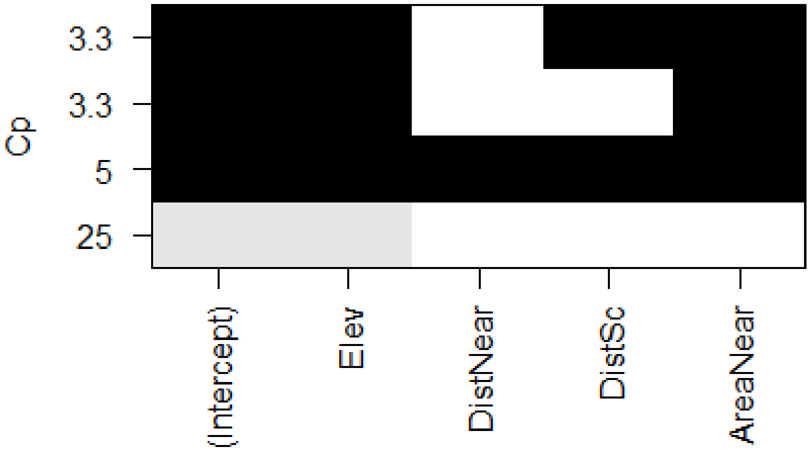
```
forward <- regsubsets(Native ~ Elev + DistNear + DistSc+ AreaNear,
                      data = subset(ex1220), nbest = 1, method = "forward")
summary(forward)
```

```
## Subset selection object
## Call: regsubsets.formula(Native ~ Elev + DistNear + DistSc + AreaNear,
##     data = subset(ex1220), nbest = 1, method = "forward")
## 4 Variables  (and intercept)
##          Forced in Forced out
## Elev         FALSE      FALSE
## DistNear     FALSE      FALSE
## DistSc       FALSE      FALSE
## AreaNear     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: forward
##          Elev DistNear DistSc AreaNear
## 1  ( 1 ) "*"  " "      " "    " "
## 2  ( 1 ) "*"  " "      " "    "*"
## 3  ( 1 ) "*"  " "      "*"    "*"
## 4  ( 1 ) "*"  "*"      "*"    "*"
```

```
plot(forward)
```



```
plot(forward, scale= "Cp")
```



The best fitting regression model from the forward selection is different for Cp and BIC. The best fitting regression model using the Cp has more variables than the model using the BIC. Following the principle of parsimony, I would prefer a more simpler model, so I'm going with the model with BIC.

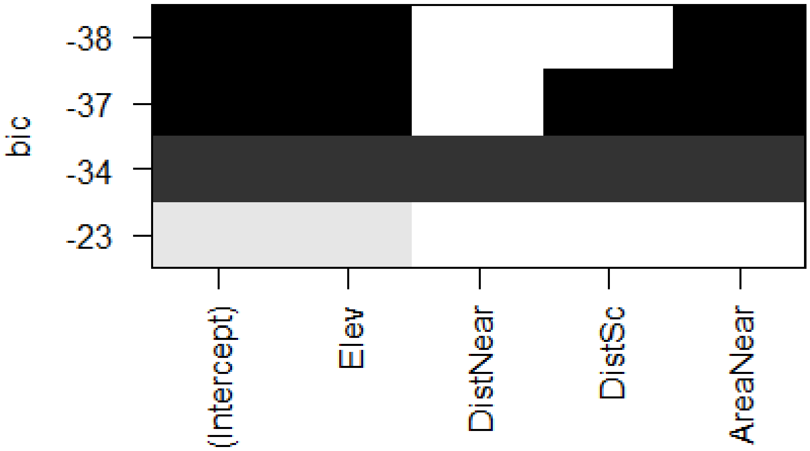
Therefore, the best fitting regression model for forward selection is ;

Native (Intercept) + Elev + AreaNear

```
backward <- regsubsets(Native ~ Elev + DistNear + DistSc + AreaNear,
                       data = subset(ex1220), nbest = 1, method = "backward")
summary(backward)

## Subset selection object
## Call: regsubsets.formula(Native ~ Elev + DistNear + DistSc + AreaNear,
##     data = subset(ex1220), nbest = 1, method = "backward")
## 4 Variables  (and intercept)
##          Forced in Forced out
## Elev         FALSE      FALSE
## DistNear     FALSE      FALSE
## DistSc       FALSE      FALSE
## AreaNear     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          Elev DistNear DistSc AreaNear
## 1  ( 1 ) "*"  " "      " "    " "     
## 2  ( 1 ) "*"  " "      " "    "*"     
## 3  ( 1 ) "*"  " "      "*"    "*"     
## 4  ( 1 ) "*"  "*"      "*"    "*"     

plot(backward)
```



The same selection I did in forward selection also applies here. The best fitting model for backward elimination is; Native (Intercept) + Elev + AreaNear

#Question 2d

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.

The best fitting model from the forward selection above is Native (Intercept) + Elev + AreaNear

```
lin_mod2 <- lm(Native ~ Elev + AreaNear, data = data)
summary(lin_mod2)

##
## Call:
## lm(formula = Native ~ Elev + AreaNear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5749  -8.7105  -0.8665   8.0576  30.8510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.528316   3.116261   1.774   0.0873 .
## Elev         0.067857   0.006607  10.271 7.98e-11 ***
## AreaNear    -0.015423   0.003219  -4.791 5.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.6 on 27 degrees of freedom
## Multiple R-squared:  0.7975, Adjusted R-squared:  0.7825
## F-statistic: 53.16 on 2 and 27 DF,  p-value: 4.339e-10
```

```
lin_mod3 <- lm(Native ~ Elev + AreaNear + Area, data = data)
summary(lin_mod3)
```

```
##
## Call:
## lm(formula = Native ~ Elev + AreaNear + Area, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2521  -7.4062  -0.9017   7.2945  27.8342
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.746671   3.371890   0.815   0.4227
## Elev         0.082747   0.010442   7.924 2.11e-08 ***
## AreaNear    -0.017942   0.003397  -5.282 1.60e-05 ***
## Area        -0.007817   0.004352  -1.796   0.0841 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.11 on 26 degrees of freedom
## Multiple R-squared:  0.8198, Adjusted R-squared:  0.799
## F-statistic: 39.44 on 3 and 26 DF,  p-value: 8.047e-10
```

```
anova (lin_mod2, lin_mod3)
```

```
## Analysis of Variance Table
##
## Model 1: Native ~ Elev + AreaNear
## Model 2: Native ~ Elev + AreaNear + Area
##   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
```

The p-value from the extra-sum-of-squares F -test due to the addition of variable Area to this model is 0.08.

#Question 3

(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?

```
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
```

```
data2 <- ex1518
```

```
mod_fit <- lm(Cases ~ Year*Vaccine, data=data2)
summary(mod_fit)
```

```
## 
## Call:
## lm(formula = Cases ~ Year * Vaccine, data = data2)
## 
## 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
```

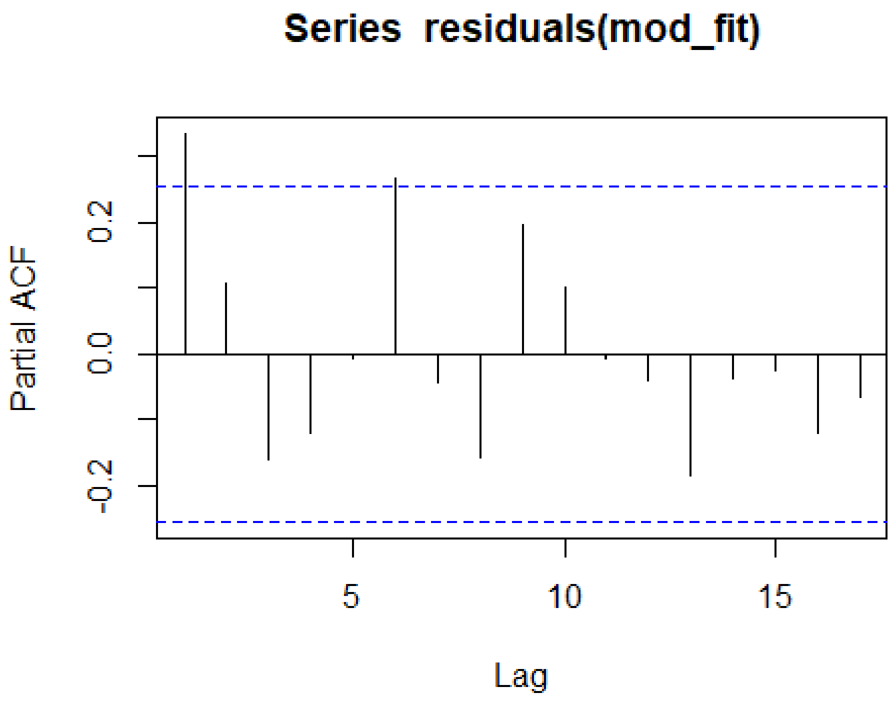
From the output above, there seems to not be an effect of Vaccine on cases with the p-value of 0.894 which is greater than 0.05 and 0.10 at 5% and 10% level respectively which shows the variable is not significant.

Also, the interaction between year and vaccine appears not to be significant in the model with its high p-value, showing that it does not have effect on cases.

#Question 3b

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?

```
pacf(residuals(mod_fit))
```



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

## [1] 0.3345066

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

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

n <- nrow(data2)
t_stat <- abs(summary(mod_fit)$coef[,1])/SE_adj
p_value <- 2*pt(-abs(t_stat), df=n-4, lower.tail=TRUE)

summary(mod_fit)$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 for the test increased after adjusting the standard errors.