Regression

Chetan

11/03/2022

```r
library("rstatix")
```

```
##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter
```

Question 16

The variables y1, y2, y3 are the response variables. Take the time to look at the type of variables in the database. Fit a one (1) way ANOVA between y1 and x1. Comment on your results.

```r
#anova1_2 <- read.csv("./anova1_2.csv")
anova1_2 <- readRDS("my_data.rds")
head(anova1_2)
```

```
##         y1       y2         y3 X1          X2 X3 X4       C_X1        C_X2
## 1 2.456812 4.956812  4.4037980  2 -0.13038998  1  1  1.5384302 -0.99579872
## 2 1.902846 3.902846  4.2601818  0  0.04707731  1  0 -0.1097103 -1.03995504
## 3 1.602922 3.102922  3.4458329  1  0.01559217  1  1  0.5114708 -0.01798024
## 4 2.112113 4.112113  4.2038403  0 -0.19237133  1  1  0.2139580 -0.13217513
## 5 1.849287 3.849287  6.7436999  0 -0.01426162  1  0 -0.1861207 -2.54934277
## 6 2.573954 5.073954 -0.1608652  2  0.28891017  1  0 -0.1203938  1.04057346
##         C_X3       C_X4       C_X5       E_X1 E_X2
## 1  1.0517013 -0.8209867 -0.7015052 -0.5116037    0
## 2  0.6229055 -0.3072572  0.8822346  0.2369379    0
## 3  0.4336204 -0.9020980 -0.1333704 -0.5415892    0
## 4  0.3860844  0.6270687 -1.1206785  1.2192276    1
## 5  1.2913233  1.1203550  0.4611925  0.1741359    0
## 6 -1.0022599  2.1272136  1.5241428 -0.6152683    0
```

```r
summary(anova1_2)
```

```
##        y1              y2              y3           X1          X2
##  Min.   :1.276   Min.   :1.306   Min.   :-1.485   0:77   Min.   :-0.493180
##  1st Qu.:1.584   1st Qu.:2.276   1st Qu.: 2.041   1:94   1st Qu.:-0.118066
##  Median :2.003   Median :3.034   Median : 3.605   2:79   Median : 0.002827
```

```
##  Mean   :2.007   Mean   :3.225   Mean   : 3.642          Mean   : 0.008615
##  3rd Qu.:2.452   3rd Qu.:4.068   3rd Qu.: 5.095          3rd Qu.: 0.140618
##  Max.   :2.924   Max.   :5.424   Max.   : 9.630          Max.   : 0.648208
##  X3             X4              C_X1                C_X2
##  0: 94   Min.   :0.000   Min.   :-2.50792   Min.   :-2.54934
##  1:156   1st Qu.:0.000   1st Qu.:-0.63882   1st Qu.:-0.72941
##          Median :0.000   Median : 0.02479   Median :-0.01404
##          Mean   :0.468   Mean   : 0.02963   Mean   : 0.01546
##          3rd Qu.:1.000   3rd Qu.: 0.64922   3rd Qu.: 0.61662
##          Max.   :1.000   Max.   : 2.68486   Max.   : 3.18404
##       C_X3               C_X4               C_X5               E_X1
##  Min.   :-2.69533   Min.   :-3.04786   Min.   :-2.62933   Min.   :-2.5082
##  1st Qu.:-0.59240   1st Qu.:-0.57962   1st Qu.:-0.65582   1st Qu.:-0.6496
##  Median : 0.09513   Median :-0.00018   Median : 0.08929   Median :-0.1217
##  Mean   : 0.03694   Mean   : 0.04709   Mean   : 0.07037   Mean   :-0.1002
##  3rd Qu.: 0.70798   3rd Qu.: 0.85318   3rd Qu.: 0.85822   3rd Qu.: 0.4551
##  Max.   : 3.39037   Max.   : 3.29052   Max.   : 2.81608   Max.   : 2.2820
##  E_X2
##  0:123
##  1:127
##
##
##
##
```

```
str(anova1_2)
```

```
## 'data.frame':    250 obs. of  14 variables:
##  $ y1  : num  2.46 1.9 1.6 2.11 1.85 ...
##  $ y2  : num  4.96 3.9 3.1 4.11 3.85 ...
##  $ y3  : num  4.4 4.26 3.45 4.2 6.74 ...
##  $ X1  : Factor w/ 3 levels "0","1","2": 3 1 2 1 1 3 2 1 2 2 ...
##  $ X2  : num  -0.1304 0.0471 0.0156 -0.1924 -0.0143 ...
##  $ X3  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...
##  $ X4  : num  1 0 1 1 0 0 0 1 0 0 ...
##  $ C_X1: num  1.538 -0.11 0.511 0.214 -0.186 ...
##  $ C_X2: num  -0.996 -1.04 -0.018 -0.132 -2.549 ...
##  $ C_X3: num  1.052 0.623 0.434 0.386 1.291 ...
##  $ C_X4: num  -0.821 -0.307 -0.902 0.627 1.12 ...
##  $ C_X5: num  -0.702 0.882 -0.133 -1.121 0.461 ...
##  $ E_X1: num  -0.512 0.237 -0.542 1.219 0.174 ...
##  $ E_X2: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 2 2 ...
```

```
##  [1] "y1"   "y2"   "y3"   "X1"   "X2"   "X3"   "X4"   "C_X1" "C_X2" "C_X3"
## [11] "C_X4" "C_X5" "E_X1" "E_X2"
```

```
## # A tibble: 3 x 5
##   X1    variable     n  mean    sd
##   <fct> <chr>    <dbl> <dbl> <dbl>
## 1 0     y1          77  2.02 0.109
## 2 1     y1          94  1.54 0.115
## 3 2     y1          79  2.55 0.128
```

```r
identify_outliers(X1.grouping, y1)
```

```
## # A tibble: 6 x 16
##   X1       y1    y2    y3     X2 X3       X4   C_X1    C_X2    C_X3   C_X4
##   <fct> <dbl> <dbl> <dbl>  <dbl> <fct> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>
## 1 1      1.88  1.88  2.68 -0.330 0         0  1.07   0.183   0.893  -1.24
## 2 2      2.86  5.36  5.59  0.648 1         0 -1.89   0.0549 -1.99    0.217
## 3 2      2.92  5.42  3.76  0.426 1         1  0.231  1.75   -0.0517 -1.55
## 4 2      2.83  5.33  8.53  0.440 1         1 -0.592 -0.451   0.850   0.820
## 5 2      2.26  4.76  4.39 -0.158 1         1  2.55   0.825  -0.171  -1.60
## 6 2      2.88  2.88  6.85 -0.493 0         0 -0.517 -1.82    1.18    0.900
## # ... with 5 more variables: C_X5 <dbl>, E_X1 <dbl>, E_X2 <fct>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

We have no extreme outliers.

```r
shapiro_test(X1.grouping, y1)
```

```
## # A tibble: 3 x 4
##   X1    variable statistic      p
##   <fct> <chr>        <dbl>  <dbl>
## 1 0     y1           0.984 0.429
## 2 1     y1           0.990 0.671
## 3 2     y1           0.967 0.0382
```

Our data is normal as p > 0.05 for all values of X1.

```r
levene_test(anova1_2, y1 ~ X1)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     2   247     0.299 0.742
```

```r
anova_test(anova1_2, y1 ~ X1)
```

```
## Coefficient covariances computed by hccm()

## ANOVA Table (type II tests)
##
##   Effect DFn DFd        F         p p<.05   ges
## 1     X1   2 247 1599.357 4.41e-142     * 0.928
```

```r
# here's a map on how to interpret this:
# Effect = grouping variables (in this case treatment)
# DFn = degree of freedom for your groups (k-1)
# DFd = degree of freedom for your sample (n -k)
#F = your actual ANOVA ratio!
#p = your significance statistics
#p<.05 = how significant is your p in stars?
# ges = generalized eta square (effect size!)
```

```
tukey_hsd(anova1_2, y1 ~ X1)
```

```
## # A tibble: 3 x 9
##   term  group1 group2 null.value estimate conf.low conf.high    p.adj p.adj.si~1
## * <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
## 1 X1    0      1               0   -0.487   -0.530    -0.445 8.84e-14 ****
## 2 X1    0      2               0    0.525    0.481     0.570 8.84e-14 ****
## 3 X1    1      2               0    1.01     0.970     1.05  8.84e-14 ****
## # ... with abbreviated variable name 1: p.adj.signif
```

Question 17

Fit a simple linear regression model between y1 and X1. Comment on your results.

```
mod_lm <- lm(data = anova1_2, y1 ~ X1)
summary(mod_lm)
```

```
##
## Call:
## lm(formula = y1 ~ X1, data = anova1_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28488 -0.07691 -0.00192  0.06988  0.37464
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.02431    0.01337  151.37   <2e-16 ***
## X11         -0.48733    0.01804  -27.02   <2e-16 ***
## X12          0.52528    0.01879   27.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1173 on 247 degrees of freedom
## Multiple R-squared:  0.9283, Adjusted R-squared:  0.9277
## F-statistic:  1599 on 2 and 247 DF,  p-value: < 2.2e-16
```

```
#anova(mod_lm)
```

Question 18

Which among the groups of X1 has significantly the smallest mean (at a threshold of 5%)? How does the answer to the previous question help you answer this question?

Answer 18:

X1 has three category variables (0, 1 and 2). From the previous linear model results we can conclude that the Intercept value is 2.02 which is the Group 0. Group 1 is lower than group 0 by average of -0.48. Group 2 is higher on average by 0.52. Therefore, Group 1 of X1 has the smallest mean at the threshold of 5%.

Question 19:

Fit a simple linear regression model between y1 and x2. Can you deduce that there is no association between y1 and X2?

```
mod_lm_1 <- lm(data = anova1_2, y1 ~ X2)
summary(mod_lm_1)
```

```
##
## Call:
## lm(formula = y1 ~ X2, data = anova1_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73355 -0.41909 -0.00708  0.43861  0.93057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.00613    0.02765   72.54   <2e-16 ***
## X2           0.10852    0.13743    0.79    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4369 on 248 degrees of freedom
## Multiple R-squared:  0.002508,   Adjusted R-squared:  -0.001514
## F-statistic: 0.6235 on 1 and 248 DF,  p-value: 0.4305
```

There is no association as our model is insignificant. $(p = 0.4305) > 0.05$.

Question 20:

Fit a 2-way ANOVA between y2 and X1, X3. Is there interaction?

```
X.grouping <- group_by(anova1_2, X1, X3)

get_summary_stats(X.grouping, y2, type = "mean_sd")
```

```
## # A tibble: 6 x 6
##   X1    X3    variable     n  mean    sd
##   <fct> <fct> <chr>    <dbl> <dbl> <dbl>
## 1 0     0     y2          34  2.03 0.12
## 2 0     1     y2          43  4.02 0.1
## 3 1     0     y2          30  1.54 0.128
## 4 1     1     y2          64  3.04 0.109
## 5 2     0     y2          30  2.55 0.124
## 6 2     1     y2          49  5.05 0.132
```

```
identify_outliers(X.grouping, y2)
```

```
## # A tibble: 8 x 16
##   X1    X3       y1     y2     y3      X2     X4   C_X1     C_X2     C_X3    C_X4
```

```
##   <fct> <fct> <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>
## 1 0     0      2.31  2.31  5.66  0.320     0 -0.132 -1.57    0.621   0.854
## 2 1     0      1.80  1.80  4.36  0.328     0  0.101 -0.0660 -1.17    0.570
## 3 1     0      1.88  1.88  2.68 -0.330     0  1.07   0.183   0.893  -1.24
## 4 2     0      2.88  2.88  6.85 -0.493     0 -0.517 -1.82    1.18    0.900
## 5 2     1      2.86  5.36  5.59  0.648     0 -1.89   0.0549 -1.99    0.217
## 6 2     1      2.92  5.42  3.76  0.426     1  0.231  1.75   -0.0517 -1.55
## 7 2     1      2.83  5.33  8.53  0.440     1 -0.592 -0.451   0.850   0.820
## 8 2     1      2.26  4.76  4.39 -0.158     1  2.55   0.825  -0.171  -1.60
## # ... with 5 more variables: C_X5 <dbl>, E_X1 <dbl>, E_X2 <fct>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

```
shapiro_test(X.grouping, y2)
```

```
## # A tibble: 6 x 5
##   X1    X3    variable statistic      p
##   <fct> <fct> <chr>        <dbl>  <dbl>
## 1 0     0     y2           0.984 0.890
## 2 0     1     y2           0.954 0.0812
## 3 1     0     y2           0.947 0.142
## 4 1     1     y2           0.979 0.348
## 5 2     0     y2           0.976 0.725
## 6 2     1     y2           0.950 0.0383
```

```
levene_test(anova1_2, y2 ~ X1 * X3)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     5   244     0.243 0.943
```

Levene says our variance are homogeneous.

Run Two-way ANOVA

```
(anova_2 <- aov(data = anova1_2, y2 ~ X1 * X3))
```

```
## Call:
##    aov(formula = y2 ~ X1 * X3, data = anova1_2)
##
## Terms:
##                        X1        X3     X1:X3 Residuals
## Sum of Squares  102.84367 227.46098   9.77510   3.39636
## Deg. of Freedom         2         1         2       244
##
## Residual standard error: 0.1179809
## Estimated effects may be unbalanced
```

```r
summary(anova_2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## X1            2 102.84   51.42  3694.2 <2e-16 ***
## X3            1 227.46  227.46 16341.2 <2e-16 ***
## X1:X3         2   9.78    4.89   351.1 <2e-16 ***
## Residuals   244   3.40    0.01                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, first row has a simple effect of X1 on y2. The second row has a simple effect of X3 on y2. Third row denotes the complex effect of X1 and X3 on y2. Here, X1, X2 and X1:X3 are statistically significant as their p-value are less than 0.05. Our model shows interaction between X1 and X3 on y2 as p value of interaction is greater than 0.05. Therefore it is significant.

Question 21:

What are we trying to find with the following R code?

Answer 21:

Here, in this code we are creating a linear model of our complex effect which can be used as a ONE-way ANOVA of a simple effect.

Further, using this linear model we will use error of the complex model to peer inside the interaction effect. There is a significant difference in mean of y2 because of interaction of X1 and X3.

```r
mod3 <- lm(y2 ~ X1 * X3, data = anova1_2)
summary(mod3)
```

```
## 
## Call:
## lm(formula = y2 ~ X1 * X3, data = anova1_2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.28576 -0.07290  0.00336  0.06859  0.37377 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.03321    0.02023  100.49   <2e-16 ***
## X11         -0.49650    0.02955  -16.80   <2e-16 ***
## X12          0.51494    0.02955   17.42   <2e-16 ***
## X31          1.98406    0.02708   73.28   <2e-16 ***
## X11:X31     -0.48367    0.03761  -12.86   <2e-16 ***
## X12:X31      0.51825    0.03849   13.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.118 on 244 degrees of freedom
## Multiple R-squared:  0.9901, Adjusted R-squared:  0.9899 
## F-statistic:  4886 on 5 and 244 DF,  p-value: < 2.2e-16
```

Question 22:

```r
library(emmeans)
emm <- emmeans(mod3, specs = c("X1", "X3"))
emm
```

```
##   X1 X3 emmean     SE  df lower.CL upper.CL
##   0  0    2.03 0.0202 244     1.99     2.07
##   1  0    1.54 0.0215 244     1.49     1.58
##   2  0    2.55 0.0215 244     2.51     2.59
##   0  1    4.02 0.0180 244     3.98     4.05
##   1  1    3.04 0.0147 244     3.01     3.07
##   2  1    5.05 0.0169 244     5.02     5.08
##
## Confidence level used: 0.95
```

Here, for the threshold of 5%, the average response of y2 for the observation belonging to the treatment (1,1) is 3.04 which is less than the average of the average response for observations belonging to the treatment (0,1) i.e. 4.02 and (2,0) 2.55 which turns out to be (4.02 + 2.55)/2 = 3.285

Question 23:

For a threshold of 5%, is the average response for an observation belonging to treatment (2.0) smaller than the average of the average responses for observations belonging to treatments (1.1) and (1.0)?

Answer 23:

The average response for an observation belonging to treatment (2,0) is 2.55 which is greater than the average of the average responses for observations belonging to treatments (1,1) with average of 3.04 and (1,0) with average of 1.54 which turns out to be 2.29

Question 24:

The variables C_X1, C_X3, C_X4; are causes for the treatment X4 and for the response y3. E_X1 and E_X2 are causes for response y3. 1. What are the confounding variables for treatment x4 versus response y3? 2. If you decide to fit multiple models to answer this question, show the AIC of each model.

Answer 24: 1. Confounding variables for treatment X4 and response y3 will be C_X1, C_X3, C_X4. As DAG graph would show the C_X1, C_X3, C_X4 for the increase and decrease of X4 and y3.

2.

```r
mod1 <- lm(data = anova1_2, y3 ~ X4)
summary(mod1)
```

```
##
## Call:
## lm(formula = y3 ~ X4, data = anova1_2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4032 -1.4294 -0.1767  1.5142  5.0702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8342     0.1754   16.16  < 2e-16 ***
## X4            1.7251     0.2563    6.73 1.17e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.022 on 248 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.151
## F-statistic: 45.29 on 1 and 248 DF,  p-value: 1.168e-10
```

```
AIC(mod1)
```

```
## [1] 1065.582
```

Model y3 on X4 without any confounding variables gives us a significant model ($p < 0.05$) and AIC of 1065.58. It means with every unit increase of X4 their will be an increase of 1.72 of y3.

Since, E__X1 and E__X2 are causes for response y3. Adding it in our model.

```
mod2 <- lm(data = anova1_2, y3 ~ X4 + E_X1 + E_X2)
summary(mod2)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + E_X1 + E_X2, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8431 -1.3184 -0.0801  1.2061  5.3677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0963     0.1997  10.495  < 2e-16 ***
## X4            1.6503     0.2387   6.913 4.06e-11 ***
## E_X1         -0.1835     0.1333  -1.377     0.17
## E_X21         1.4852     0.2382   6.235 1.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.879 on 246 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.2668
## F-statistic:  31.2 on 3 and 246 DF,  p-value: < 2.2e-16
```

```
AIC(mod2)
```

```
## [1] 1030.915
```

Here, E__X1 is not significant as p = 0.17 > 0.05, so we will remove it from our linear model equation.

```r
mod3 <- lm(data = anova1_2, y3 ~ X4 + E_X2)
summary(mod3)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + E_X2, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0918 -1.2400 -0.0946  1.2722  5.2295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1162     0.1996  10.603  < 2e-16 ***
## X4            1.6397     0.2390   6.859 5.52e-11 ***
## E_X21         1.4921     0.2386   6.254 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.883 on 247 degrees of freedom
## Multiple R-squared:   0.27,  Adjusted R-squared:  0.2641
## F-statistic: 45.68 on 2 and 247 DF,  p-value: < 2.2e-16
```

```r
AIC(mod3)
```

```
## [1] 1030.834
```

Model y3 on X4 and E__X2 without any confounding variables gives us a significant model ($p < 0.05$) and AIC of 1030. It means with every unit increase of X4 their will be an increase of 1.63 of y3 provided other variables are constant and with every unit increase of E__X21 their will be an increase of 1.49 of y3 provided other variables are constant. We will select it as our base model for further comparisons because every variable is statistically significant, $p < 0.05$.

As provided in question, C__X1, C__X3, C__X4; are causes for the treatment X4 and for the response y3. Therefore treating them as confounding variables.

```r
mod4 <- lm(data = anova1_2, y3 ~ X4 + E_X2 + C_X1 + C_X3 + C_X4)
summary(mod4)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + E_X2 + C_X1 + C_X3 + C_X4, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2150 -1.3146 -0.0701  1.2211  4.5829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9578     0.2070   9.458  < 2e-16 ***
```

```
## X4            1.9656     0.2915   6.742 1.12e-10 ***
## E_X21         1.5126     0.2300   6.576 2.92e-10 ***
## C_X1         -0.5070     0.1132  -4.477 1.16e-05 ***
## C_X3         -0.1022     0.1381  -0.740   0.4599
## C_X4          0.3017     0.1274   2.367   0.0187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.807 on 244 degrees of freedom
## Multiple R-squared:  0.3361, Adjusted R-squared:  0.3225
## F-statistic:  24.7 on 5 and 244 DF,  p-value: < 2.2e-16
```

```
AIC(mod4)
```

```
## [1] 1013.117
```

Here, C_X3 is not significant as p = 0.45 > 0.05, therefore removing it from our model. AIC calculated is 1013.11

```
mod5 <- lm(data = anova1_2, y3 ~ X4 + E_X2 + C_X1 + C_X4)
summary(mod5)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + E_X2 + C_X1 + C_X4, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2217 -1.3247 -0.0253  1.2455  4.5685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9993     0.1991  10.041  < 2e-16 ***
## X4            1.8547     0.2499   7.421 1.91e-12 ***
## E_X21         1.5283     0.2288   6.678 1.61e-10 ***
## C_X1         -0.5095     0.1131  -4.506 1.03e-05 ***
## C_X4          0.2761     0.1226   2.253   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.805 on 245 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3237
## F-statistic:  30.8 on 4 and 245 DF,  p-value: < 2.2e-16
```

```
AIC(mod5)
```

```
## [1] 1011.678
```

This model has an AIC of 1011.67 which is the lowest of all. Therefore, it is the best model. Moreover, our variables and overall model is statistically significant as $p < 0.05$.

Calculate the percentage change in the parameter estimate and determine whether confounding is present

```r
Percentage_Change = (mod5$coefficients[2] - mod3$coefficients[2])/mod3$coefficients[2]*100

#Percentage_Change = (2.100 - 1.035)/2.100 * 100

Percentage_Change
```

```
##      X4
## 13.1163
```

Since the percentage change is 11.32%, which is greater than 10%, this indicates that the association between y3 and X4 is confounded by C_X1 + C_X4.

Also, adding those variables to the model the R square increase from 0.27 to 0.33, which means that these new variables are explaining 6% of the variance.

Since confounding is present, we should present the results from the adjusted analysis.

Question 25:

What are the modifying variables of the effect of X4 on the response y3? If you decide to fit multiple models to answer this question, show the AIC of each model.

```r
mod1 <- lm(data = anova1_2, y3 ~ X4)
summary(mod1)
```

```
##
## Call:
## lm(formula = y3 ~ X4, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4032 -1.4294 -0.1767  1.5142  5.0702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8342     0.1754   16.16  < 2e-16 ***
## X4            1.7251     0.2563    6.73 1.17e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.022 on 248 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.151
## F-statistic: 45.29 on 1 and 248 DF,  p-value: 1.168e-10
```

```r
AIC(mod1)
```

```
## [1] 1065.582
```

Model y3 on X4 without any confounding variables gives us a significant model ($p < 0.05$) and AIC of 1065.58. It means with every unit increase of X4 their will be an increase of 1.72 of y3.

Adding confounding variables.

```
mod6 <- lm(data = anova1_2, y3 ~ X4 + C_X1 + C_X3 + C_X4)
summary(mod6)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + C_X1 + C_X3 + C_X4, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9090 -1.2456 -0.0477  1.3593  5.2817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6539     0.1926  13.777  < 2e-16 ***
## X4            2.1255     0.3146   6.757 1.02e-10 ***
## C_X1         -0.4861     0.1226  -3.966 9.61e-05 ***
## C_X3         -0.1856     0.1489  -1.246   0.2139
## C_X4          0.3021     0.1380   2.190   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.956 on 245 degrees of freedom
## Multiple R-squared:  0.2184, Adjusted R-squared:  0.2057
## F-statistic: 17.12 on 4 and 245 DF,  p-value: 2.15e-12
```

```
AIC(mod6)
```

```
## [1] 1051.907
```

Here, C_X3 is not significant as $p = 0.21 > 0.05$, therefore removing it from our model. AIC calculated is 1051.90

```
mod7 <- lm(data = anova1_2, y3 ~ X4 + C_X1 + C_X4)
summary(mod7)
```

```
##
## Call:
## lm(formula = y3 ~ X4 + C_X1 + C_X4, data = anova1_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1512 -1.2770 -0.0332  1.4176  5.2684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7429     0.1791  15.315  < 2e-16 ***
## X4            1.9256     0.2709   7.108 1.27e-11 ***
```

```
## C_X1         -0.4902     0.1227  -3.997 8.49e-05 ***
## C_X4          0.2555     0.1329   1.922   0.0558 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.958 on 246 degrees of freedom
## Multiple R-squared:  0.2135, Adjusted R-squared:  0.2039
## F-statistic: 22.26 on 3 and 246 DF,  p-value: 8.763e-13
```

```
AIC(mod7)
```

```
## [1] 1051.486
```

Here, This model has an AIC of 1051.90. Moreover, our variables and overall model is statistically significant as $p < 0.05$. Our model improved from 0.15 to 0.21, our new confounding variables show increase in variance by 6%.