Project 3

AUTHOR
Raymond Fleming

```
library( tidyverse )
library( gsheet )
library( ggplot2 )
library( DescTools )
library( lindia )
library( car )
library( ggpubr )
```

Consider the insurance cost data availabile here.

1. Import the data. Extra credit will be given to students that figure out how to directly import from GitHub.

```
datafile<- 'https://raw.github.com/stedy/Machine-Learning-with-R-datasets/mast

data <- read.csv( datafile )
```

2. Fill in the following table:

		Mean (Standard Deviation) or Number (Percent)
Charges		13270.42 (12110.01)
Age		39.21 (14.05)
BMI		30.66 (6.10)
Sex	Male	676 (50.52%)
	Female	662 (49.48%)
Smoking Status	Yes	274 (20.48%)

		Mean (Standard Deviation) or Number (Percent)
	No	1064 (79.52%)
Number of Children	0	574 (42.90%)
	1	324 (24.22%)
	2	240 (17.94%)
	3	157 (11.73%)
	4	25 (1.87%)
	5	18 (1.35%)

```
data %>%

  summarize(
    NumObs          = nrow( data ),
    MinNumChildren = min ( children ),
    MaxNumChildren = max ( children ),
    ChildRange     = max ( children ) - min( children ),
    MeanCharges    = mean( charges ),
    SDCharges      = sd  ( charges ),

    AgeRange = max ( age ) - min( age ),
    MeanAge  = mean( age ),
    SDAge    = sd  ( age ),

    MaxBMI        = max    ( bmi ),
    MinBMI        = min    ( bmi ),
    BMIRange      = max    ( bmi ) - min( bmi ),
    MeanBMI       = mean   ( bmi ),
    SDBMI         = sd     ( bmi ),
    NumFemale     = length ( which( data$sex == "female" )),
    PercentFemale = NumFemale / NumObs * 100,

    NumMale     = length ( which ( data$sex == "male" )),
```

```
    PercentMale = NumMale / NumObs * 100,

    NumSmokers     = length ( which ( data$smoker == "yes" )),
    PercentSmokers = NumSmokers / NumObs * 100,

    NumNonSmokers     = length ( which ( data$smoker == "no" )),
    PercentNonSmokers = NumNonSmokers / NumObs * 100,

    NoChildren        = length ( which ( data$children == 0 )),
    PercentNoChildren = NoChildren / NumObs * 100,

    OneChild        = length ( which ( data$children == 1 )),
    PercentOneChild = OneChild / NumObs * 100,

    TwoChildren        = length ( which ( data$children == 2 )),
    PercentTwoChildren = TwoChildren / NumObs * 100,

    ThreeChildren        = length ( which ( data$children == 3 )),
    PercentThreeChildren = ThreeChildren / NumObs * 100,

    FourChildren        = length ( which ( data$children == 4 )),
    PercentFourChildren = FourChildren / NumObs * 100,

    FiveChildren        = length ( which ( data$children == 5 )),
    PercentFiveChildren = FiveChildren / NumObs * 100)
```

```
  NumObs MinNumChildren MaxNumChildren ChildRange MeanCharges SDCharges
1   1338              0              5          5    13270.42  12110.01
  AgeRange  MeanAge    SDAge MaxBMI MinBMI BMIRange MeanBMI    SDBMI NumFemale
1       46 39.20703 14.04996  53.13  15.96    37.17 30.6634 6.098187       662
  PercentFemale NumMale PercentMale NumSmokers PercentSmokers NumNonSmokers
1      49.47683     676    50.52317        274       20.47833          1064
  PercentNonSmokers NoChildren PercentNoChildren OneChild PercentOneChild
1          79.52167        574          42.89985      324        24.21525
  TwoChildren PercentTwoChildren ThreeChildren PercentThreeChildren
1         240           17.93722           157             11.73393
  FourChildren PercentFourChildren FiveChildren PercentFiveChildren
1           25             1.86846           18            1.345291
```

There are a total of 1338 data points in this dataset. The maximum number of children is 5 and minimum is 0, indicating a range of 5. The mean charge for the dataset is

13170.42 with a standard deviation of 12110.01. The minimum age is 18, and maximum is 64 indicating a range of 46. The mean age is 39.21 with a standard deviation of 14.50. The max BMI is 53.13 and minimum is 15.96 indicating a range of 37.17. The mean BMI is 30.66 with a standard deviation of 6.10. There are a total of 662 (49.48%) females, and 676 (50.52%) male. There are 274 (20.48%) smokers and 1064 (79.52%) nonsmokers. 574 (42.90%) have no children, 324 (24.22%) have one child, 240 (17.94%) have two children, 157 (11.73%) have three children, 25 (1.87%) have four children, and 18 (1.35%) have five children.

3. Model insurance charges as a function of age, BMI, and number of children. Remember to state the resulting model.

```r
m <- lm( charges ~ age + bmi + children,
        data = data )

summary( m )
```

```
Call:
lm(formula = charges ~ age + bmi + children, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-13884  -6994  -5092   7125  48627

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6916.24    1757.48  -3.935 8.74e-05 ***
age           239.99      22.29  10.767  < 2e-16 ***
bmi           332.08      51.31   6.472 1.35e-10 ***
children      542.86     258.24   2.102   0.0357 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11370 on 1334 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.1181
F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

The model was found to be $\hat{y} = -6916.24 + 239.99$ * age $+332.08$ * bmi $+542.86$ * children

4. Provide brief and appropriate interpretations for all regression coefficients.

Age, BMI, and Number of Children were all found to have a positive β indicating that as age, number of children, and BMI increase, so do insurance charges. Of these, number of children increases the fastest at 542.86, followed by BMI at 332.08, and age at 239.99. Its important to note that the possible ranges for these predictors is not even While the number of children has a high coefficient at 542.86, the range is only 5. Age has a much smaller coefficient at 239.99, but the range is 46.

5. Fill in the following table:

Predictor	$\hat{\beta}_i$ (95% CI)	*p*-Value
Age	239.99 (196.27, 283.72)	< 0.001
BMI	332.08 (231.43, 432.74)	< 0.001
Number of Children	542.86 (36.26, 1049.47)	0.0357

```
confint( m )
```

```
                  2.5 %     97.5 %
(Intercept) -10363.96835 -3468.5183
age            196.26940   283.7195
bmi            231.42538   432.7414
children        36.26142  1049.4679
```

(yes, I am asking you to place both the estimated β as well as the corresponding 95% CI in the cell.)

6. Which, if any, are significant predictors of insurance charges? Test at the $\alpha = 0.05$ level. You do not need to state all hypothesis test pieces, but you must provide appropriate justification for your conclusions.

Age, BMI, and Number of Children are all significant predictors of insurance charges at $\alpha = 0.05$ level. All three have $p < 0.05$ with Number of Children at $p = 0.036$, and both Age and BMI at $p < 0.001$

7. Use the appropriate hypothesis test to determine if this is a significant regression line. Test at the $\alpha = 0.05$ level.

Hypotheses

$H_0 : \beta_{Age} = \beta_{BMI} = \beta_{NumberOfChildren} = 0$

H_1 : At least one β is different

Test Statistic

$F_0 = 60.69$

p-Value

$p < 0.001$

Conclusion

Reject H_0 at $\alpha = .05$. There is sufficient evidence to suggest that at least one variable is a significant predictor of Insurance Charges.

8. Construct the correlation matrix for the variables in the regression model. Are any suspiciously high?

```
dataFiltered <- data %>%

  select( charges,
          age,
          bmi,
          children )

cor( dataFiltered,
     use = "complete.obs" )
```

```
            charges       age       bmi   children
charges  1.00000000 0.2990082 0.1983410 0.06799823
```

```
age      0.29900819 1.0000000 0.1092719 0.04246900
bmi      0.19834097 0.1092719 1.0000000 0.01275890
children 0.06799823 0.0424690 0.0127589 1.00000000
```

No correlations are suspiciously high for this model, the highest correlation is Age and Charges with 0.299. Considering health with age, this makes sense. While there are no suspiciously high correlations, Charges and Children is an interestingly low correlation.

9. Check for outliers. How many are there?

```
dataFiltered <- dataFiltered %>%
  mutate( outlier = abs( rstandard( m )) > 2.5 )

dataFiltered %>% count( outlier )
```

```
  outlier    n
1   FALSE 1327
2    TRUE   11
```

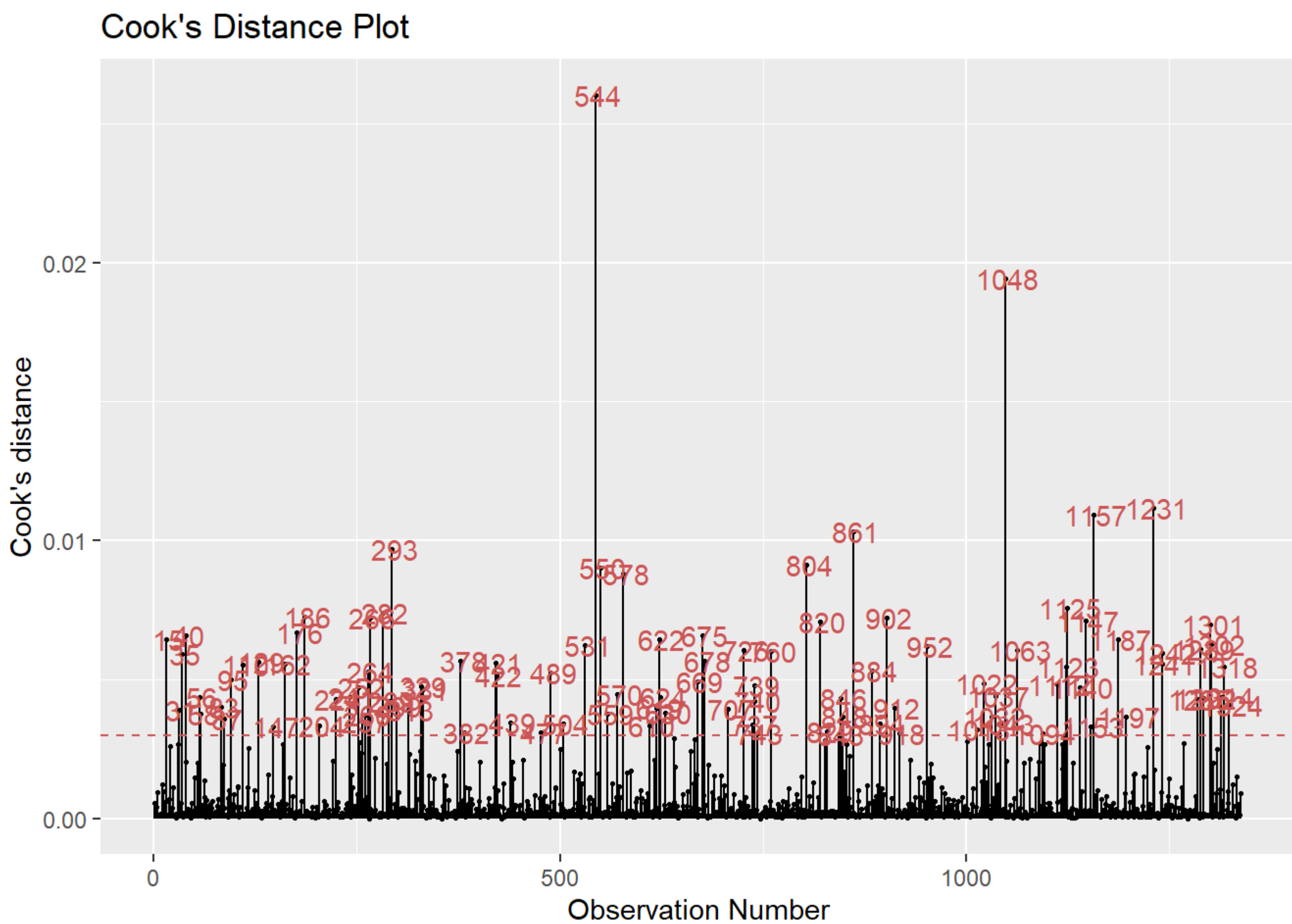
```
Outliers <- dataFiltered %>% filter( outlier == TRUE )

head( Outliers )
```

```
   charges age    bmi children outlier
1 51194.56  28 36.400        1    TRUE
2 48885.14  44 38.060        0    TRUE
3 63770.43  54 47.410        0    TRUE
4 58571.07  31 38.095        1    TRUE
5 43943.88  34 30.210        1    TRUE
6 44585.46  29 35.500        2    TRUE
```

There are a total of eleven outliers when using the data relevant to the model. In this case, the filtered data is used so that outliers which are not relevant to the model are not included.

10. Check for influential/leverage points. How many are there?

```
gg_cooksd( m )
```



The Cook’s Distance Plot shows that there are two influence and leverage points in this model, located at observation 544 and 1048. There are a number of other points which could be considered as well, however since the data is quite variable, these points are not large enough spikes to be considered as influence/leverage points.

11. Check for multicollinearity. Do the results surprise you?

```
vif( m )
```

```
     age      bmi  children
1.013816 1.012152 1.001874
```

There is no multicolinearity. This result is not surprising given the correlation matrix in question 8, none of the three predictors were highly correlated with one another. Looking at it from what I know about healthcare however, this is a surprising result. I would expect age and number of children to show multicolinearity since a person cannot start with 5 children. Its likely that because people have children in a smaller window of their lifetime, but age constantly the correlation doesn't meet multicolinearity.

12. Construct a graph to aid with explanation of the regression model. Create lines for 0, 2, and 4 children. You pick what goes on the *x*-axis and what is plugged in for the remaining variable. Extra credit if you make the outlier dots a different color than the non-outlier dots. $\hat{y} = -6916.24 + 239.99$ * age $+332.08$ * bmi $+542.86$ * children

```
c1 <- coefficients( m )

dataFiltered <- dataFiltered %>%

  mutate( Children0 = c1[1] + c1[2] * age + c1[3] * 31 + c1[4] * 0,
          Children2 = c1[1] + c1[2] * age + c1[3] * 31 + c1[4] * 2,
          Children4 = c1[1] + c1[2] * age + c1[3] * 31 + c1[4] * 4 )
1
```

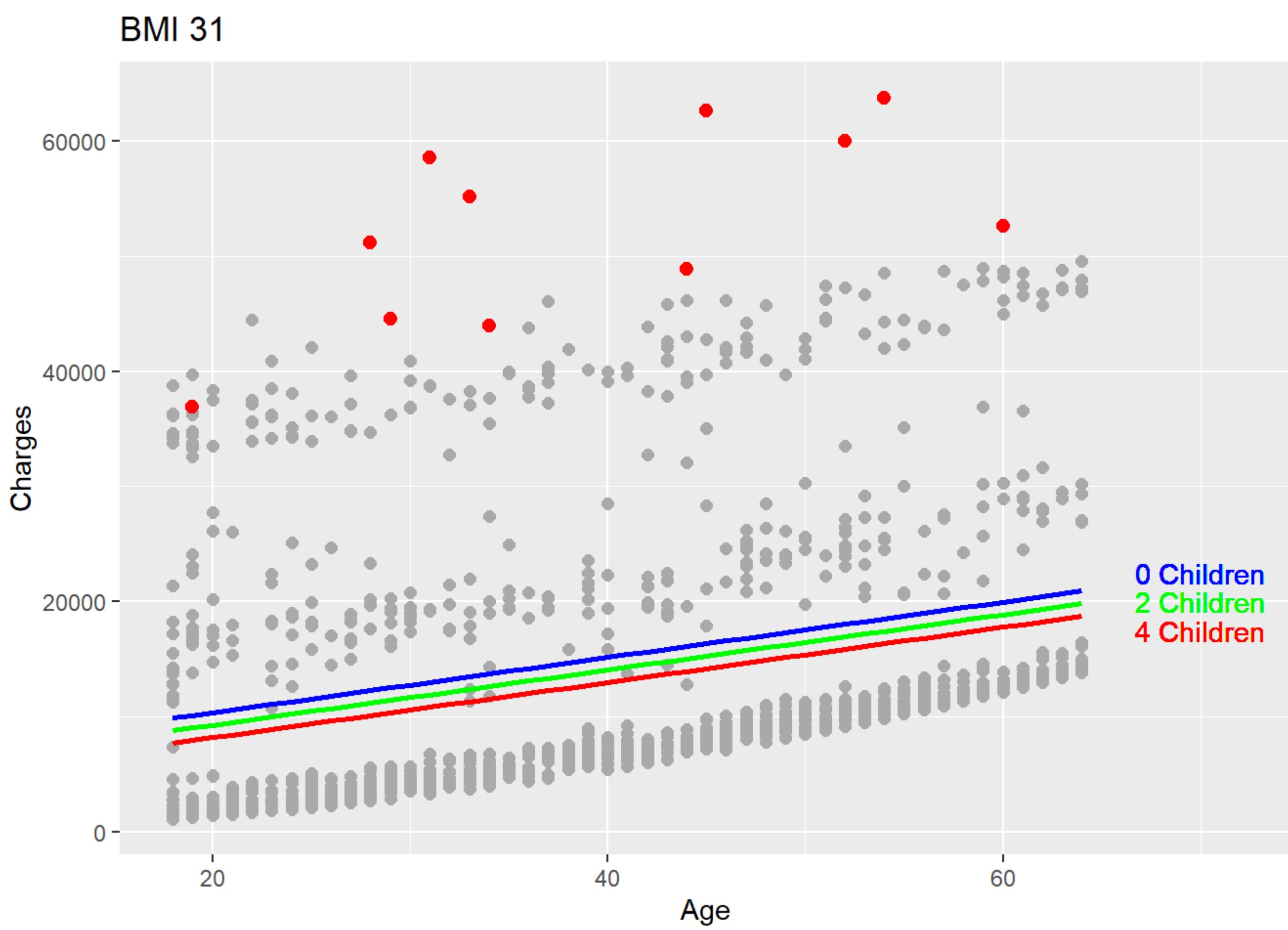
```
[1] 1
```

```
quantile( data$children,
          na.rm = TRUE )
```

```
  0%  25%  50%  75% 100%
   0    0    1    2    5
```

```
dataFiltered %>% ggplot(aes( x = age,
                    y = charges )) +

  geom_point( size = 2,
              color = "darkgrey" ) +

  geom_point( size=2.1,
              color = "red",
              data = Outliers )+
  geom_line( aes(y=Children0),size = 1, color = "red" )+
  geom_line( aes(y=Children2),size = 1, color = "green" )+
  geom_line( aes(y=Children4),size = 1, color = "blue" )+
```

```r
  geom_text( aes(x = 70, y = 22500, label = "0 Children" ), color = "blue" ) +
  geom_text( aes(x = 70, y = 20000, label = "2 Children" ), color = "green" )
  geom_text( aes(x = 70, y = 17500, label = "4 Children" ), color = "red" ) +
  geom_text( aes(x = 72, y = 17500, label = " " ), color = "white" ) + #this i

  labs(x = "Age",
       y = "Charges",
       title = "BMI 31" )
```



13. Write a short paragraph to accompany your results, appropriate for your supervisor who is not a statistician or data scientist. Outline your modeling technique as well as the summary of the data (i.e., the first table) and results.

A model demonstrating how Number of Children, BMI, and Age predict insurance charges was created which showed that all three of these variables are significant predictors of charges. To create the results, the data was first summarized as shown in the table shown above in Section 2. Following this, the model was generated then verified after. The confidence intervals were checked as well, which provide the range of how much each item affects charges to a 95% level, meaning that there is a 5% likelihood that the real value is outside of the interval. Following this, it was found that there are a total of eleven outliers in the dataset. The data was filtered before checking for these, so that outliers in other categories would not appear since only the outliers in the relevant portion of the entire dataset are useful in this case. In the graph above, these data points are marked as red, all are near towards the top section indicating that the outliers are unusually high charges, rather than low. If the red points were removed, the lines would appear to fit the data better than they already do. Two data points were identified which by themselves change the model, if these two points were to be removed the model would be noticably different. However, as there was not statistical reason to remove these and there appears to be nothing indicating that these are erroneous data points, they were left in the model. The three predictors weren't found to be highly correlated to the point where one could be used to predict another. Since this is not the case, the model is stronger than it otherwise would be. While it would seem on first glance that age and number of children would be highly correlated, since its not possible to start with 5 children (except in very rare cases). Its important to note that the reason is likely that children are often born in a small window of an adult's age range, so a window in higher ages would be relatively constant. In the graph, BMI was chosen as the X axis rather than age. This was chosen because age is a discrete number, which makes the data form columns rather than being scattered. However, it could also be graphed with BMI on the x axis, and this was done during the model to determine which graph would be better to report.