Multiple Linear Regression Solution

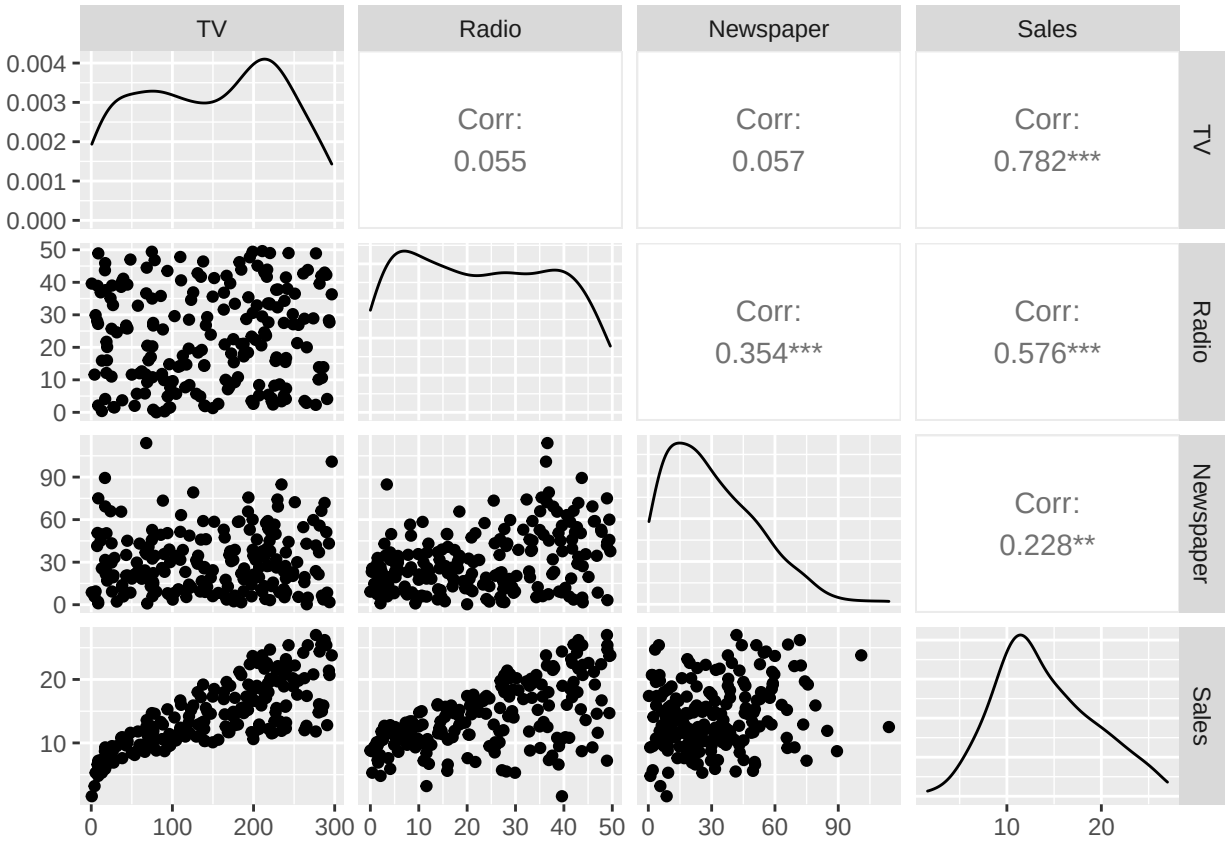
MATH224 - Intro to Stat

Exercise 1 (2 Points)

0.5 Points Sales and TV has the highest correlation with a value of 0.782.

1 Point We can see a positive relationship sales and TV where the scatter points are close to each other but they start to spread out as TV increases. Same case for Sales and Radio. But we can't see an apparent pattern with Sales and Newspaper which is reflected by a low correlation of 0.228.

```
ggpairs(advertising) # 0.5 Points
```



Exercise 2 (2 Points)

Intercept: 7.0326 **0.5 Points** Slope: 0.0475 **0.5 Points**

```r
simple.model = lm(Sales ~ TV, data = advertising) # 0.5 Points
summary(simple.model) # 0.5 Points
```

```
##
## Call:
## lm(formula = Sales ~ TV, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Exercise 3 (3 Points)

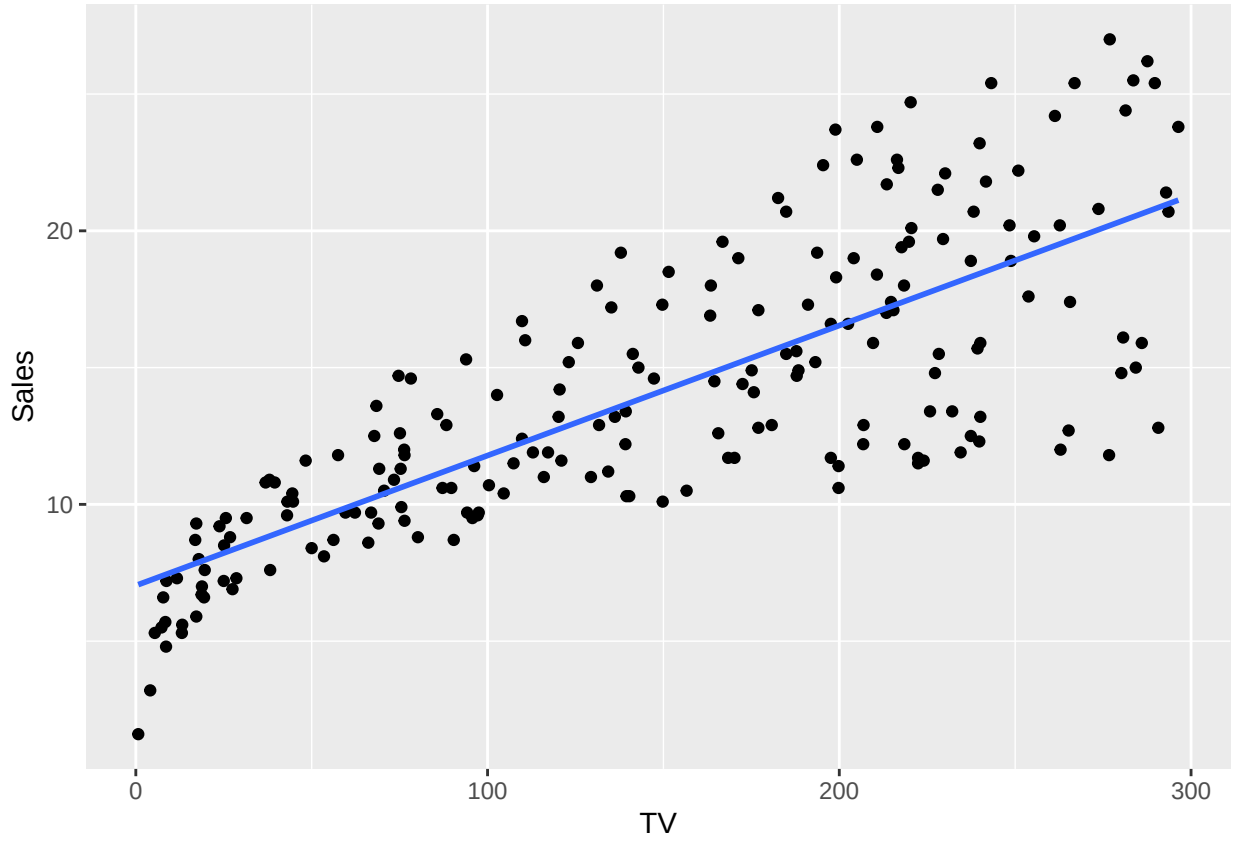
1 Point The variable `TV` is highly significant as the p-value is far less than 0.05.

2 Point The R^2 value is 0.6119. The total variation in the outcome variable is explained by the explanatory variables with 61.19%

Exercise 4 (2 Points)

2 Points The line of best fit matches the linear model results. The intercept seems to start around 7 on the y-axis. For from the x-axis, we can see that for a change of 100, we can observe a change of around 5 on the y-axis. So the slope is $\frac{5}{100} = 0.05$ which is closer to the actual result.

```r
advertising%>%
  ggplot(aes(x = TV, y = Sales))+
  geom_point()+
  geom_smooth(method = "lm", se = F)
```



Exercise 5 (2 Points)

1 Point Newspaper wasn't a significant variable while the others were. Adjusted R^2 value is 0.8956

```
full_model = lm(Sales ~ TV + Radio + Newspaper, data = advertising) #0.5 Points
summary(full_model) #0.5 Points
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Exercise 6 (3 Points)

1 Point each

While holding all other variables constant, a unit increase in TV correlates to a increase of 0.046 units in sales on average.

While holding all other variables constant, a unit increase in Radio correlates to a increase of 0.189 units in sales on average.

While holding all other variables constant, a unit increase in Newspaper correlated to a decrease of 0.001 units in sales on average.

Exercise 7 (2 Points)

1 Points Adjusted R^2 is 0.8962. The final model has a higher Adjusted R^2 than the full model. The final model provides a better fit out of the two.

```
final_model = lm(Sales ~ TV + Radio, data = advertising) # 0.5 Points
summary(final_model) # 0.5 Points
```

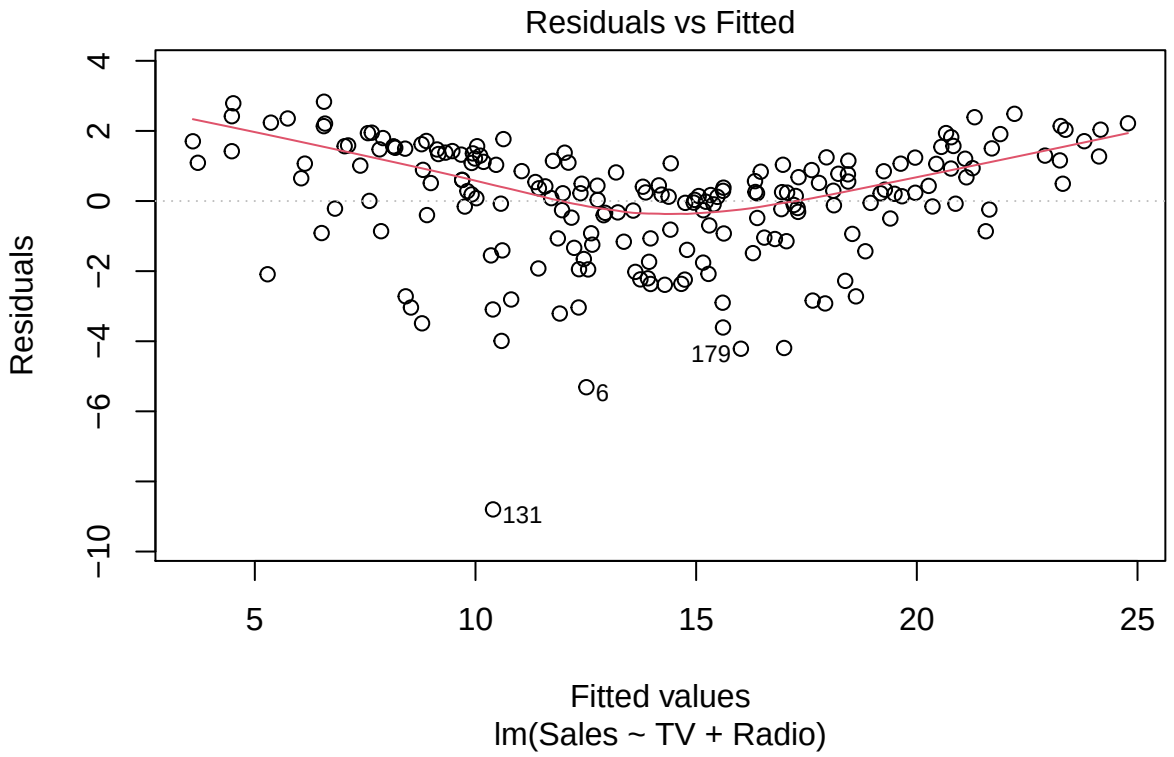
```
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## Radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

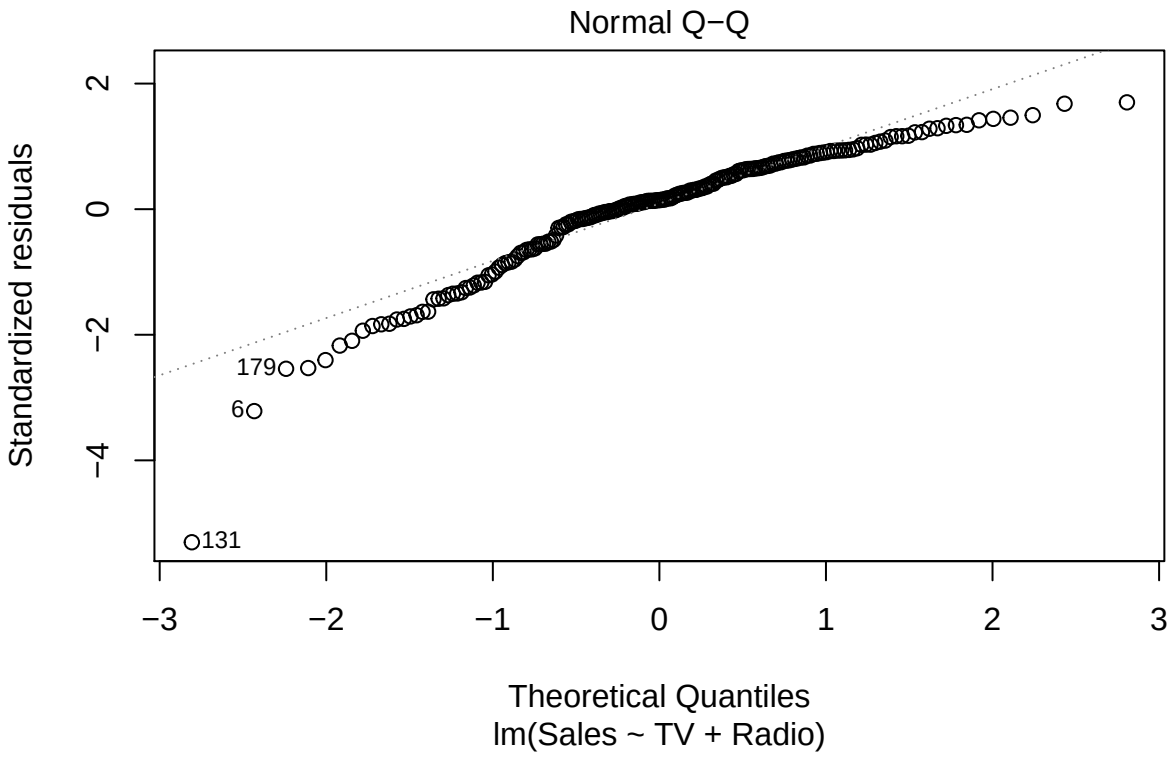
Exercise 8 (4 Points)

1.5 Points The first residual plot seems to slightly violate the assumption of constant variance in residuals regardless of the x-axis.

1.5 Points The second residual plot is a qq plot to check for normality of the residuals from the model. We can clearly see that the residual isn't normal. It seems to be left skewed. Thus, the residual normality assumption has been violated.

```r
plot(final_model, which = 1:2) # 1 Point
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
179
6
131
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Sales ~ TV + Radio)