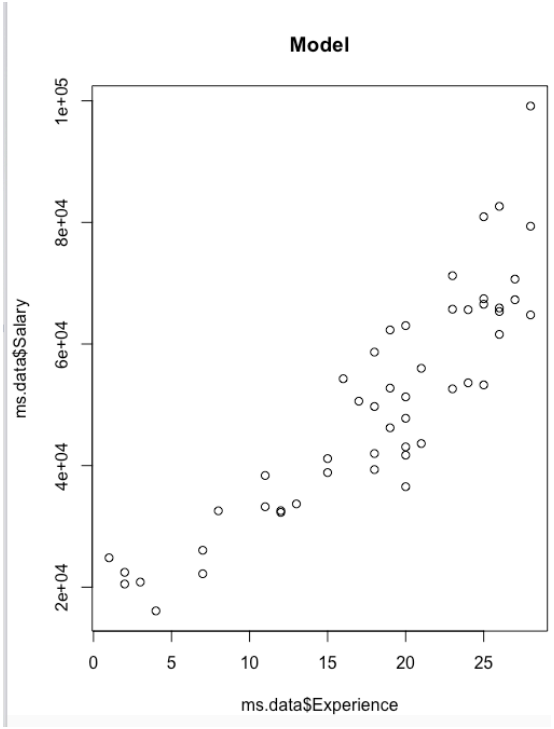
Overview:

In this example there are 2 variables: Salary = the salary of the managers in a company and Experience = the years of experience for the manager. The objective is to fit a model for the Salary in terms of the Experience. The data is in a file named "ManSalary.csv."

a) Plot a scatter plot of the Salary on the vertical axis and Experience on the horizontal axis. Comment on this plot. Is there a relationship between these two variables? How would you describe this relationship? **There appears to be a linear relationship between experience and salary, as the experience increases the salary also increases. There may be a problem with the variation getting larger as the salary level increases.**



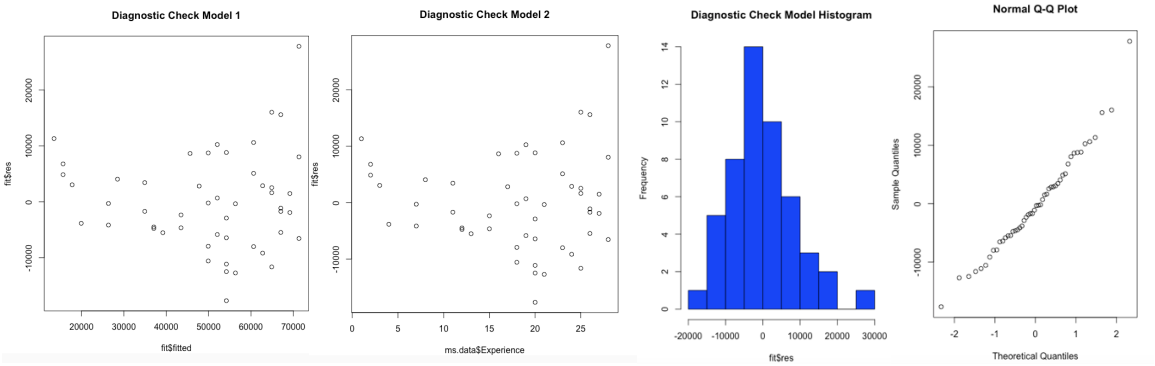
b) Estimate a simple linear model with the Y variable equal to Salary and the X variable equal to Experience. Print out a summary of this model.

```
Call:
lm(formula = Salary ~ Experience, data = ms.data)

Residuals:
     Min       1Q   Median       3Q      Max
-17665.4  -5497.5   -725.5   4667.4  27813.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11369.2     3160.3   3.598 0.000757 ***
Experience    2141.3      160.8  13.314  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8642 on 48 degrees of freedom
Multiple R-squared:  0.7869,    Adjusted R-squared:  0.7825
F-statistic: 177.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

c) Perform the diagnostic checks of the residuals for the model you estimated in part b). For each plot is there any indication of a problem with the model?

What do we expect if data has a normal distribution?

- Histogram: The histogram should look like a symmetric bell-shaped curve.
- Normal probability plot: The normal probability plot should follow a straight line.




```
        Shapiro-Wilk normality test

data:  fit$res
W = 0.97359, p-value = 0.3219
```

```
> confint(fit)
                2.5 %    97.5 %
(Intercept) 5015.095 17723.352
Experience  1817.927  2464.693
```

- **The first plot is not consistent with what is expected if the model assumptions are correct. This plot looks like the variance may be increasing as the values of the fitted values increase. There is a problem with this model that needs to be corrected.**
- **The second plot is not consistent with what is expected if the model assumptions are correct. This plot also suggests that there may be an issue with the variance increasing.**
- **The Histogram is consistent with what we should expect. If looks like a normal distribution.**
- **The normal probability plot and the p-value for the test of the normality are both consistent with the assumption that the errors follow a normal distribution.**

d) Test the hypothesis that the slope of the regression line equal 0 versus the alternative hypothesis that the slope of the regression line is not equal to 0. What can you conclude from this test?

P-Value for the intercept: 3.598 0.000757 which is < 5%

P-Value for Experience: 13.314 < 2e-16 which is < 5%

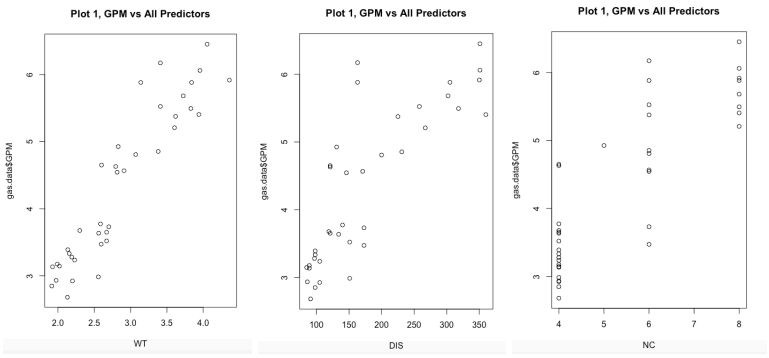
We reject null hypothesis. Experience helps explain salary.

e) What is the value of R-squared for this example? What does this value tell you about this model?

78.69%. This tells us that 78.69% of the data falls on the goodness of the line. Which tell us that our model is relevant.

2) In this example we are interested in predicting the gas consumption of an automobile based on characteristics of the automobile. The data consists of 38 cars with measurements on fuel efficiency, weight of the car, engine displacement, and number of cylinders. The data is in the file gasconsumption.csv. The names of the variables are GPM = gallons used per 100 miles, WT = weight of the car in 1000 pounds, DIS = displacement of the engine in cubic inches, and NC = the number of cylinders in the engine. You are to develop a model for GPM based on the characteristics of the car, WT, DIS, and NC.

a) Read the data into R Studio and plot scatter plots of GPM versus all the X variables. Comment on these plots. Is there a relationship between the GPM and the possible X variables? How would you describe the relationship?



- **Looks like there is a linear relationship between GPM and WT.**
- **There is a linear relationship with GPM and DIS. As DIS increases the GPM increases.**
- **Third plot, as NC increases the GPM increases in a roughly linear manner.**

b) Estimate a multiple regression model with the Y variable equal to GPM and the X variables equal to WT, DIS, and NC. Print out a summary of this model.

```
Call:
lm(formula = GPM ~ WT + DIS + NC, data = gas.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63737 -0.30455  0.00313  0.23899  0.64454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.639152   0.486450  -3.370 0.001886
WT           2.332838   0.288895   8.075 2.05e-09
DIS         -0.010637   0.002697  -3.943 0.000381
NC           0.218151   0.115982   1.881 0.068572

(Intercept) **
WT          ***
DIS         ***
NC          .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.376 on 34 degrees of freedom
Multiple R-squared:  0.9028,    Adjusted R-squared:  0.8942
F-statistic: 105.3 on 3 and 34 DF,  p-value: < 2.2e-16
```

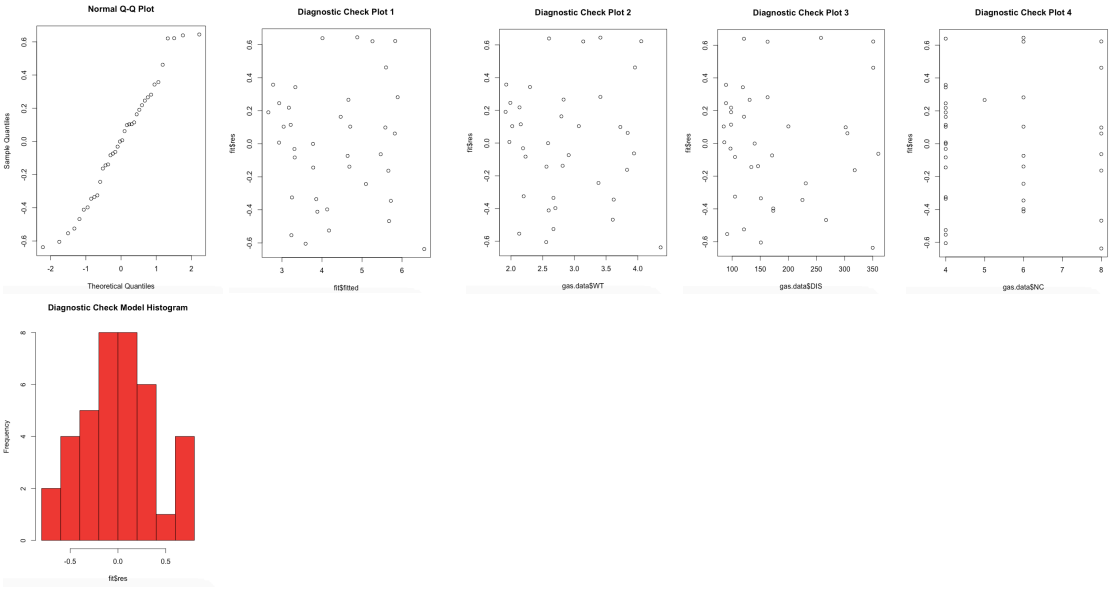
c) Use the output for this model to test the null hypothesis: $\beta_1 = \beta_2 = \beta_3 = 0$ versus the alternative hypothesis: not all β_i's are equal to 0. What does your result mean?

- **We would reject the null hypothesis that B1 = B2 = B3 = 0 because the P-Value "2.2e-16 is < 0.05.**
- **Than we look at P-Value for WT which is 2.05e-09 which is < 0.05. So we reject this hypothesis. This means that the WT variable is helping explain the variable GPM.**
- **Null Hypothesis B2 = 0 vs Alternative Hypothesis: B2 $\neq$ 0. P-Value is 0.000381 < 0.05 so this hypothesis is rejected. This means that the variable DIS is helping explain the variable GPM.**
- **Null Hypothesis B3 = 0 vs Alternative Hypothesis: B3 $\neq$ 0. P-Value is 0.068572 > 0.05, so this hypothesis is not rejected since it's bigger than 0.05. This means that the variable DIS is not helping explain the variable GPM.**

d) Based upon the output, are all of the 3 variables WT, DIS, and NC needed in the model? Justify your answer. **WT and DIS are needed but NC is not since it's P-Value is greater than 0.05. This will only make our model more complicated.**

- **Null Hypothesis B3 = 0 vs Alternative Hypothesis: B3 $\neq$ 0. P-Value is 0.068572 > 0.05, so this hypothesis is not rejected since it's bigger than 0.05. This means that the variable DIS is not helping explain the variable GPM.**

e) Perform the usual diagnostic checks for this model. Based upon your analysis is there any problems with this model?



- Normal Q-Q plot looks fine. This plot is consistent with what is expected if the model assumptions are correct. This does not suggest any problems with the model.
- The plot for residuals vs fitted values is consistent with what is expected if the model assumptions are correct. Nothing suggests that there are any problems with this model.
- The residuals vs WT plot is also fine and is consistent with what is expected if the model assumptions are correct.
- The residuals vs DIS plot is fine, this plot is consistent with what is expected if the model assumptions are correct.
- The plot of the residual's vs NIC is fine, this plot is also consistent with what is expected if the model assumptions are correct.
- Our histogram is also fine, this follows a normal distribution.