STAT8011 Assignment 2 (2022)

John Fitzgerald R00156081

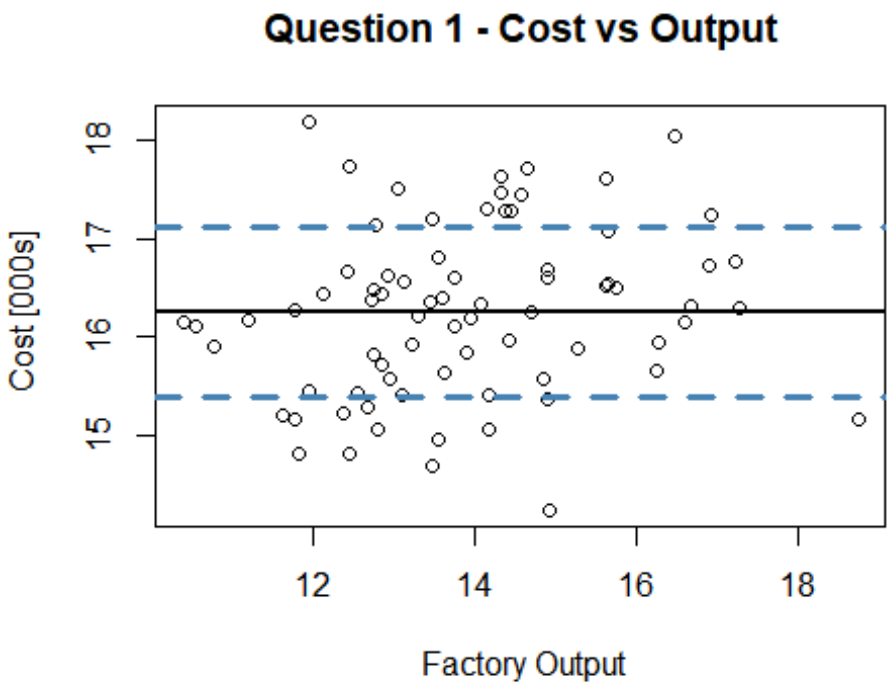
2022-12-28

Question 1

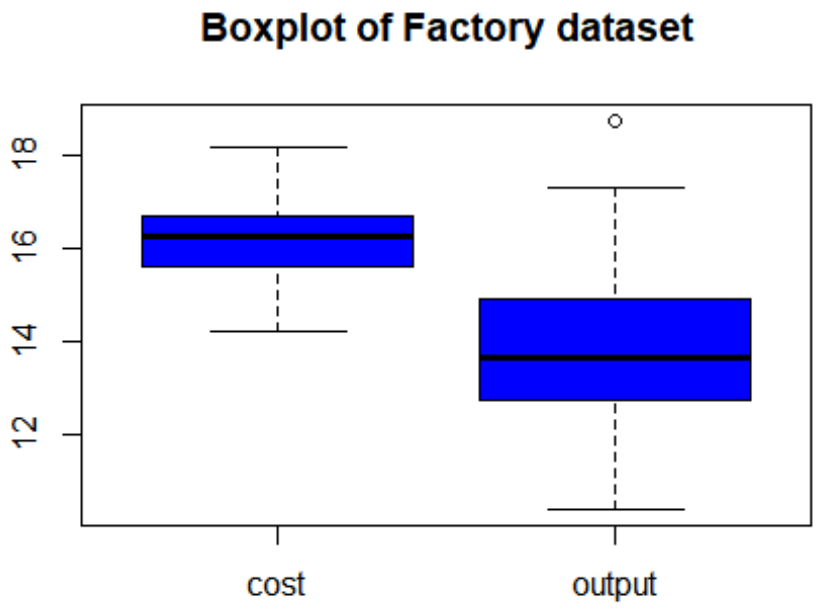
It is thought that the production costs of a factory rise with the monthly production output. The data set consists of 2 variables, cost and output. The variable cost is measured in ($000s) and the variable output is measured in Stock keeping units (SKU). I will find out whether the variable output, is of use in predicting the value of cost.

(a) Make a numerical and graphical summary of the data, commenting on the results.

Linear regression requires the relationship between the dependent (cost) and the independent (output) variables to be linear. The linearity assumption can be easily tested with a scatter plot.The black line represents the mean value of 'output' and the blue lines show the standard deviation above and below the mean value. We can see from the plot that most of the observations for the variables are within the +/- SD bounds. The following example shows there is little or no linearity present, also, there is no obvious pattern in the plot.



The boxplot below shows that the 'output' varaible contains an outlier. I will have to ensure that this does not influence my results. There does not seem to be much skew and both variables appear to be fairly symmetrical.



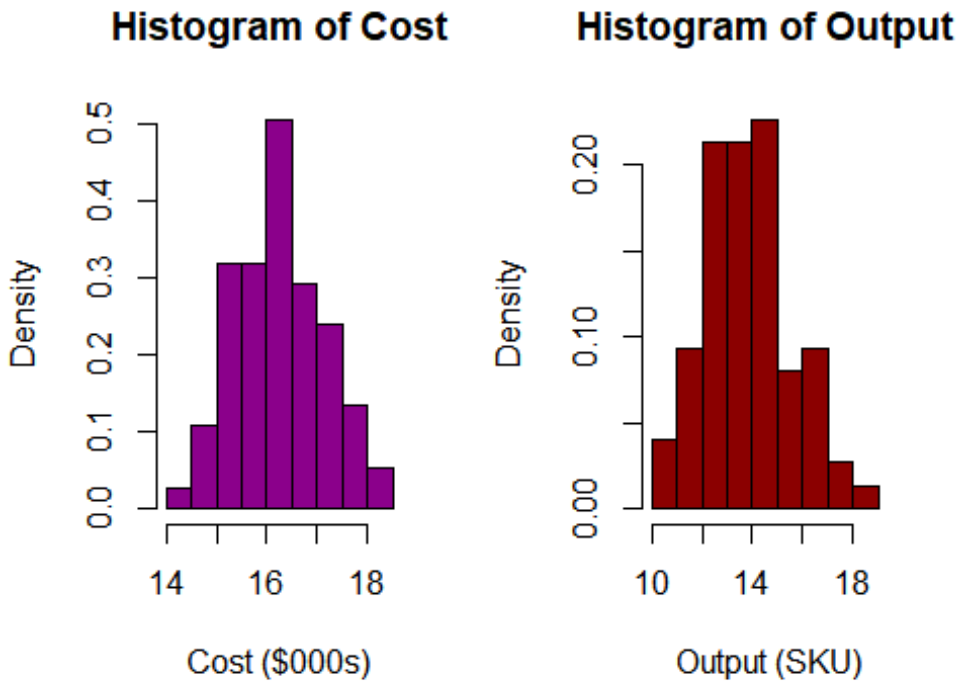
The following table shows the minimum, maximum and interquartile range for both variables. There is not much difference between the mean and median for both variables which indicates a normal distribution for both variables.

```
##       cost           output     
##  Min.   :14.23   Min.   :10.40  
##  1st Qu.:15.60   1st Qu.:12.76  
##  Median :16.27   Median :13.64  
##  Mean   :16.25   Mean   :13.90  
##  3rd Qu.:16.70   3rd Qu.:14.91  
##  Max.   :18.19   Max.   :18.75
```

The second assumption requires all variables to be multivariate normal. I checked this assumption using Shapiro-wilk test and view histograms for both variables.

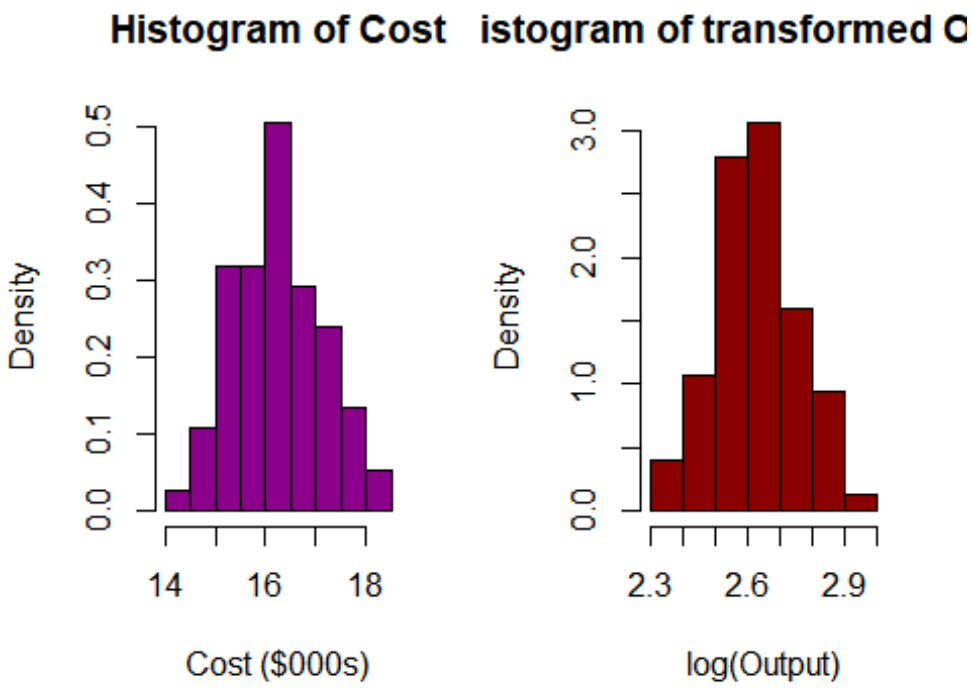
```
##
##  Shapiro-Wilk normality test
##
## data:  factory$output
## W = 0.9823, p-value = 0.3785

##
##  Shapiro-Wilk normality test
##
## data:  factory$cost
## W = 0.98814, p-value = 0.714
```

The p-values for both variables are greater than the significance level 0.05 implying that the distribution of the data are not significantly different form normal distribution. We can



assume normality.

Since 'output' is not very symmetric, a non-linear transformation might make the distribution more symmetric. I am going to apply log values to this variable and assess the impact:



I am going to used the transformed values of the 'output' variable for this test.

I want to know if there is a significant correlation between the ranks of the two variables. I also want some measure of the strength of this correlation. I wont use Pearson's correlation as the relationship does not follow a straight line even though we have continuous data for the variables. I will use the Spearman's rank correlation test as its more suited if the data doesn't follow a straight line and contains continuous data.

```
##
##  Spearman's rank correlation rho
##
## data:  factory$cost and log(factory$output)
## S = 52960, p-value = 0.03315
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2466572
```

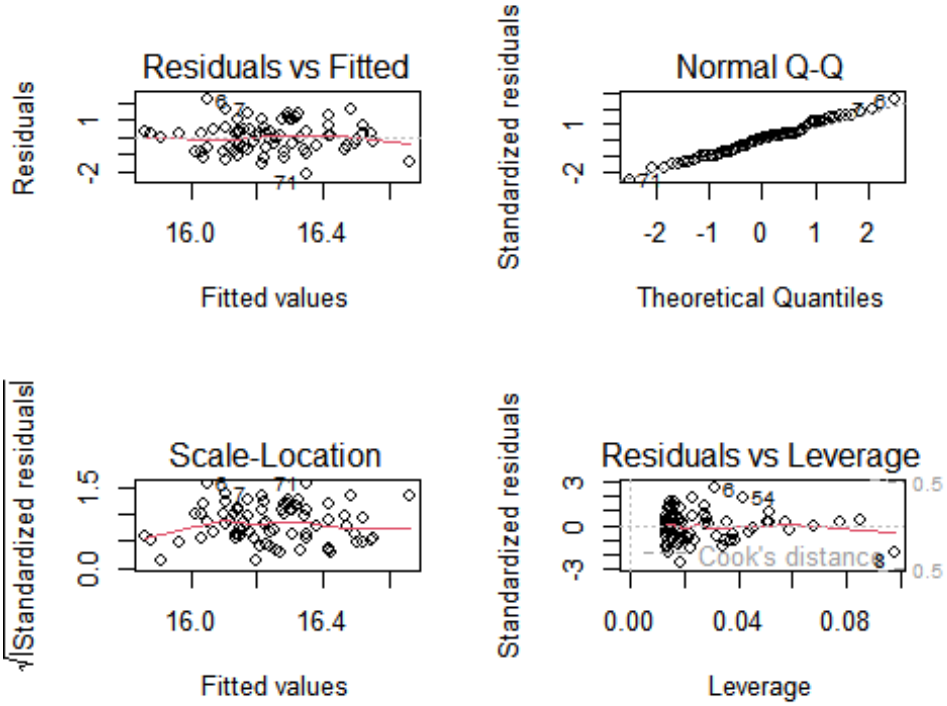
The p-value of the test is 0.03315, which is less than the significance level alpha = 0.05. We can conclude that log(output) and cost are correlated with a positive correlation coefficient (rho) of .24665 and p-value of 0.03315. It is not a hugely significant correlation - I would have expected the rho to be higher.

There are no units attached to ranks, so the results are less useful in some ways. All I can say is that an increase in 'output' leads to some increase in 'cost'.

Correlation Coefficient between the variables:

```
## [1] 0.1934734
```

The correlation coefficient of cost and output is 0.193. As this value is much closer to 0 than 1, I can conclude that the variables are not linearly related and it again shows a not significantly strong relationship between the variables.



I don't see a distinctive pattern in the 'residuals vs fitted' plot, this plot does not meet the regression assumptions very well. The 'Normal-Q-Q' plot shows a relatively straight line which indicates normal distribution, although observations 6 and 71 are off line and might be a potential problem. In the 'Scale-Location' plot, the line is horizontal and the observations seem randomly spread satisfying the assumption of homoscedasticity. The 'Residuals vs Leverage' plot is the typical look when there is no influencial case or cases. You can barely see Cook's distance lines because all cases reside well within the lines. From the plots I can conclude there is a linear relationship between the variables.

There seems to be some positive linear relationship between *cost* and *output*

(b) Fit a model of the form $\hat{\beta}_0+\hat{\beta}_1$output+e and interpret the value of $\hat{\beta}_1$.

Cost = $\hat{\beta}_0+\hat{\beta}_1$output

Cost = 12.6742058+B1$*$1.3607752

To interpret the coefficient β_1 we can say: a one unit increase in cost (in this case thousand dollars) increases log(output) by 1.3607752 units. i.e. For every thousand dollars that Cost increases, the Output increases by log 1.36 SKU.

By hand:

$$SS_{xx} = \sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n}$$

$$SS_{xx} = 517.6315842 - \frac{(196.8213834)^2}{75}$$

$$SS_{xx} = 1.11616$$

$$SS_{xy} = \sum_1^n x\,y - \frac{(\sum_1^n x)(\sum_1^n y)}{n}$$

$$SS_{xy} = 3198.93493 - \frac{(1218.395088)(196.8213834)}{75}$$

$$SS_{xy} = 1.518840221$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{1.518840221}{1.11616} = 1.360775$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = 16.24526784 - 1.360775 \times 2.624285112 = 12.674206$$

Using r:

```
##
## Call:
```

```
## lm(formula = cost ~ log(output), data = factory)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11878 -0.60180  0.06474  0.41766  2.13444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6742     2.1218   5.973 7.78e-08 ***
## log(output)   1.3608     0.8076   1.685   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8533 on 73 degrees of freedom
## Multiple R-squared:  0.03743,    Adjusted R-squared:  0.02425
## F-statistic: 2.839 on 1 and 73 DF,  p-value: 0.09628
```

(c) Calculate a 95% confidence interval for the $\hat{\beta}_1$ coefficient.

Calculations by hand:

For the slope:

$\hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\text{s.e.}(\hat{\beta}_1)$

$1.3607752 \pm t_{\frac{0.05}{2},75-2}\frac{\hat{\sigma}}{\sqrt{SS_{xx}}}$

$1.3607752 \pm t_{\frac{0.05}{2},73}\frac{0.8532618}{\sqrt{1.11616}}$

$1.3607752 \pm 1.9929971 * 0.807642 = 1.602962$

95% confidence interval for the $\hat{\beta}_1$ coefficient is: [-0.2488448, 2.970405]

For the intercept:

$\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\text{s.e.}(\hat{\beta}_0)$

$12.6742058 \pm t_{\frac{0.05}{2},75-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_{xx}}}$

$12.6742058 \pm t_{\frac{0.05}{2},73}0.8532618\sqrt{\frac{1}{75}+\frac{2.624285112^2}{1.11616}}$

$12.6742058 \pm 1.9929971 * 2.120171 = 4.225495$

[8.448711, 16.899701]

Using R:

```
##                   2.5 %   97.5 %
## log(output) -0.2488551 2.970405

##                 2.5 %  97.5 %
## (Intercept) 8.445516 16.9029
```

Please note that there are small differences due to rounding.

(d) Test the following hypothesis(and what do results imply for the regression model):

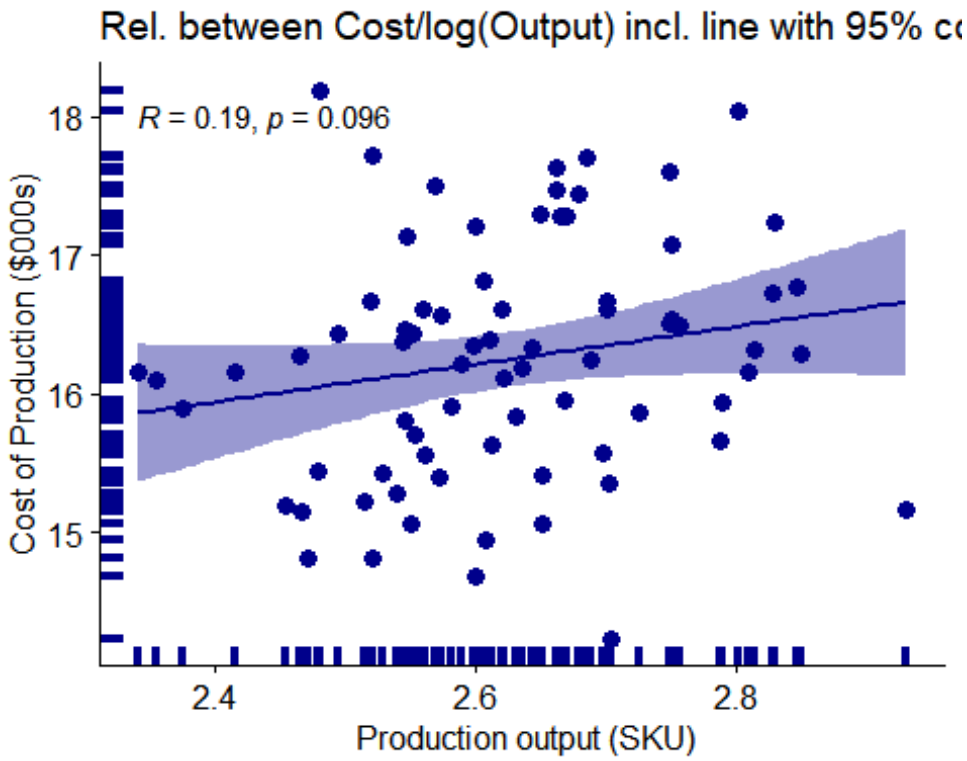
$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 12.674206   2.121774 5.973399 7.782209e-08
## log(output)  1.360775   0.807643 1.684872 9.628418e-02
```

I will fail to reject the H_0 as p-value= 0.096284 > 0.05 .I can conclude that the coefficient (1.36) is not significantly different from zero. In other words, I accept the hypothesis that the Output size has no influence on the Cost at the 5% level. Though I should acknowledge that there is some positive correlation as the coefficient indicates (see regression line of scattperplot).

(e) Plot the regression line onto a scatterplot of the data with a 95% confidence band

```
## `geom_smooth()` using formula = 'y ~ x'
```

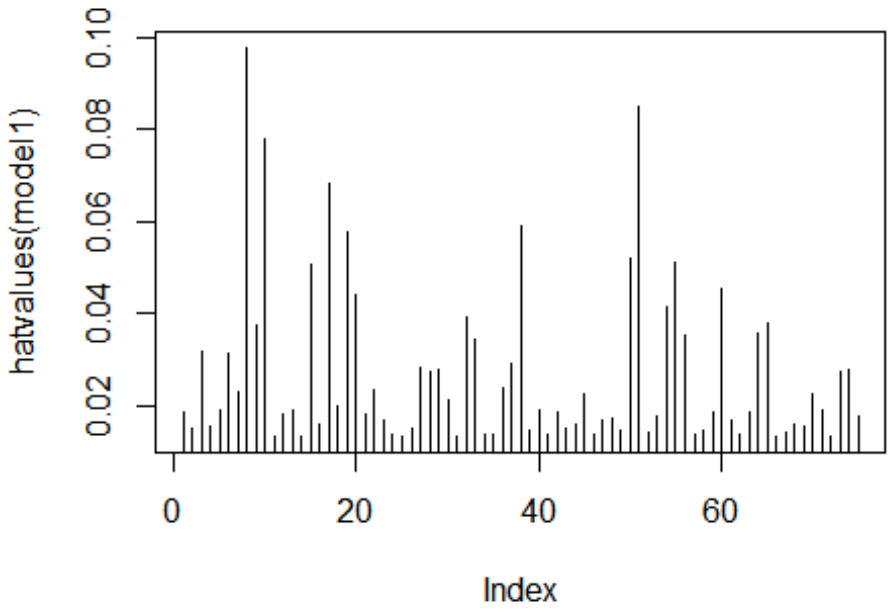


In the above plot we see the slightly positive regression line discussed earlier, proving that increasing output does have an impact on cost of production.

(f) Plot the studenized residuals against the fitted values and identify any outliers

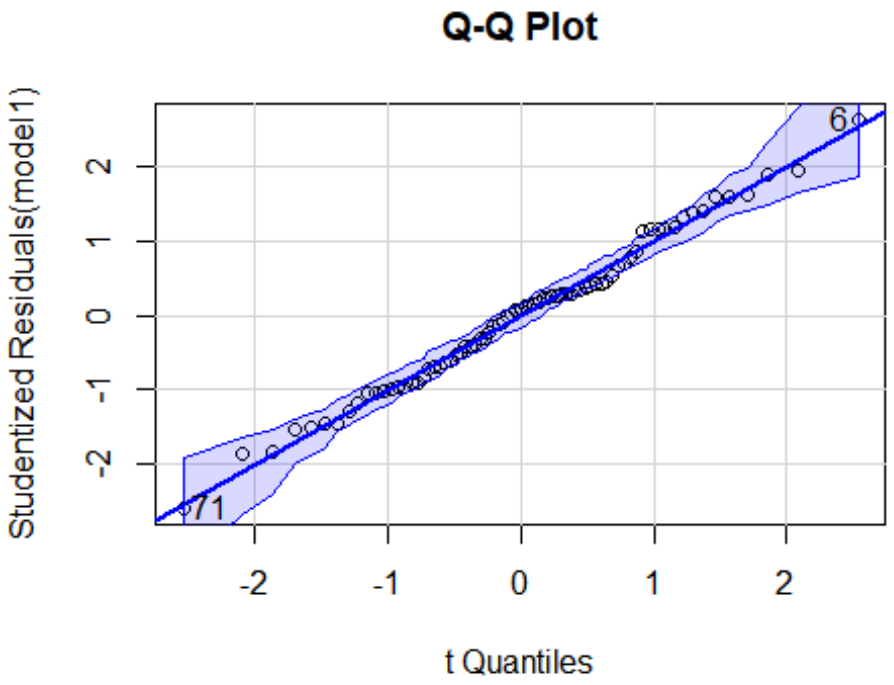
From the plot we can see that 2 observations (6 and 71) have a studentized residual with an absolute value near to 3. A result of 3 or higher would be classed as an outlier. As all 75 residuals fall between 2 and -2 there are no outliers in the model.

(g) Plot the leverage of each case and identify any observations that have a high leverage



The largest leverage value of the 75 observations is 0.09779745. Since it isn't greater that 2, we know that none of the observations in the dataset have high leverage.

(h) Identify the observation that has the largest influence on the estimate of the $\hat{\beta}_1$ coefficient.Explain why this observation has a large influence.




```
## [1]  6 71
```

Again 6 and 71 are identified as potential outliers (as seen in Q-Q plot above). Applying an outlier test to help confirm if they are outliers.

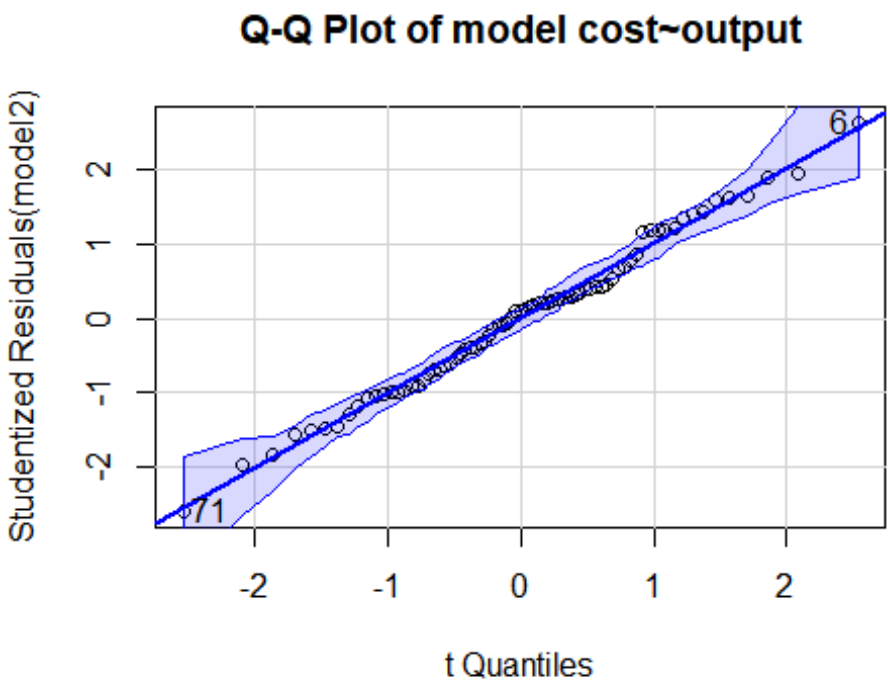
I remove the log and create a new fit on the raw data:

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 6 2.644181           0.010045      0.75338
```

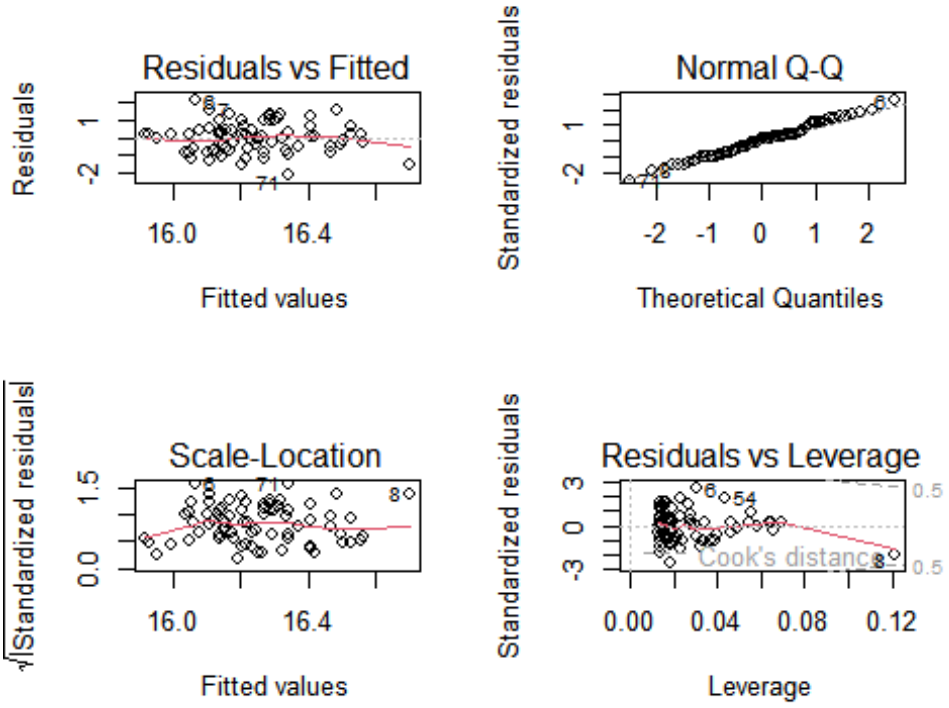
The 6th observation is returned as an outlier. The 6th observation is returned as the only outlier which could influence $\hat{\beta}_1$. I am going to remove the log(output) and create a model on cost~output to check again for any outlier or impact on model:

```
## Rows: 75 Columns: 2
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (2): cost, output
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
```

```
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```




```
## [1]  6 71
```

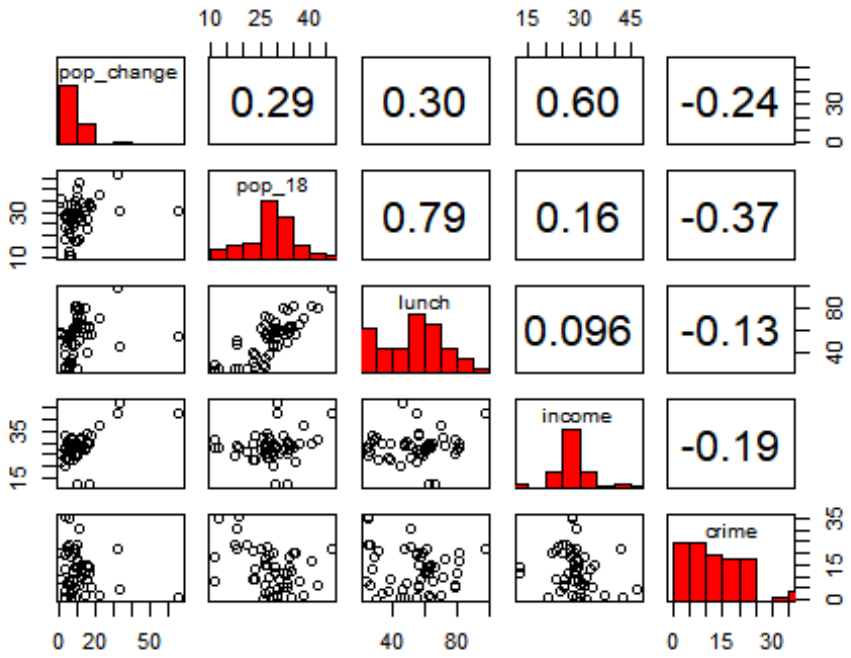


There is not any significant change to plots fitted with original data. There is no impact of outliers influencing the results.

Question 2

Using the Crime dataset which consists of n observations and 5 variables. Each observation represents the crime rate and demographics from cities in the US.

(a) Make a numerical and graphical display of the data, commenting on the results.



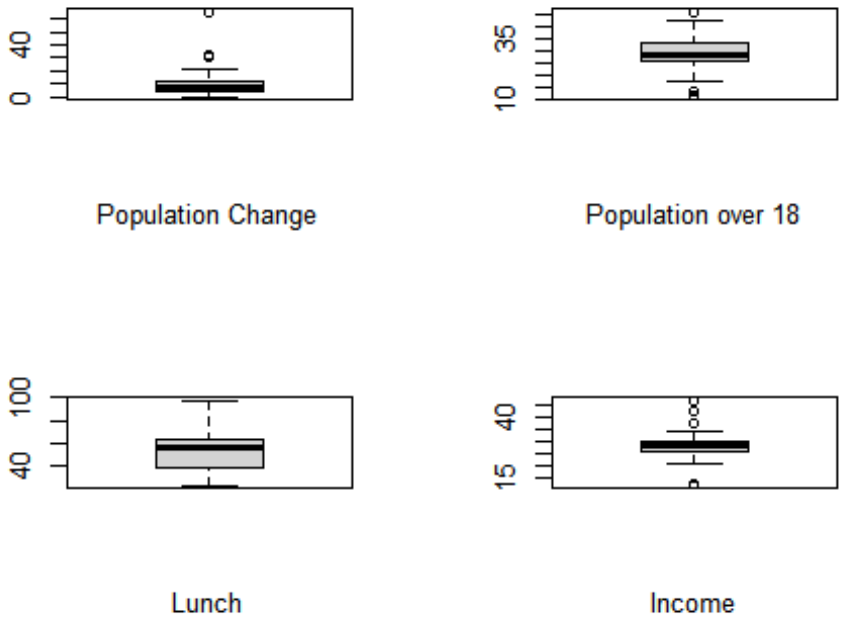
The plot(s) above show a strong correlation between 'pop_18' and 'lunch' (0.79), there is also a strong linear relationship between both variables. There is an another positive relationship between 'pop_change' and 'income' (0.60) with a positive linear relationship between both these variables also.

All variables are showing a negative relationship with the 'crime' - indicating that all 4 demographic variables decreasing as crime increases (which is as I would have expected).

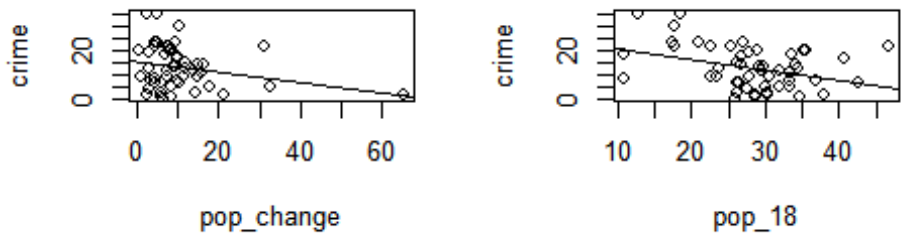
The summary statistics below concur out the histograms shown above - 'pop_change' shows a positive skew and the mean is greater than the median shown. This indicates outliers for these data. Crime also shows a positive skewness with a mean greater than median, again, indicating outliers are present. 'pop_18' and 'income' are relatively symmetrical with similar mean and median values. 'lunch' indicates a negative skewness with a mean values less than the median.

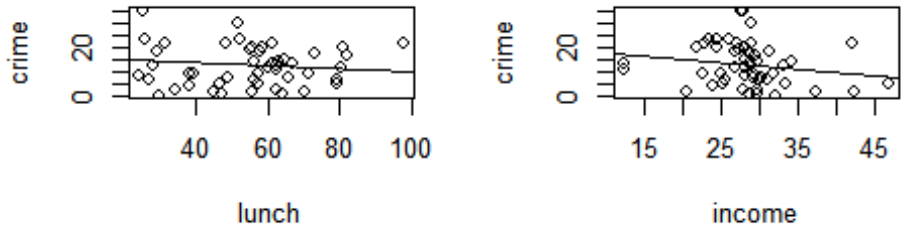
```
##    pop_change         pop_18          lunch           income
##  Min.   : 0.440   Min.   :10.88   Min.   :23.99   Min.   :12.37
##  1st Qu.: 4.836   1st Qu.:25.58   1st Qu.:38.51   1st Qu.:25.66
##  Median : 8.049   Median :28.73   Median :56.14   Median :28.47
##  Mean   :10.355   Mean   :28.24   Mean   :53.45   Mean   :28.45
##  3rd Qu.:11.671   3rd Qu.:33.39   3rd Qu.:63.64   3rd Qu.:29.88
##  Max.   :65.459   Max.   :46.69   Max.   :97.65   Max.   :46.82
##      crime
##  Min.   : 0.123
##  1st Qu.: 6.247
##  Median :11.792
##  Mean   :12.818
##  3rd Qu.:19.651
##  Max.   :35.529
```

The boxplots below shows the outliers present in the independent variables: 'pop_change' [3 outliers], 'pop_18' [3] and 'income' [4]. The presence of outliers on both sides of the 'income' and 'pop_18' distribution call into question my earlier assertion of centrality - they may be having an influence on the symmetric shape.



Plotting the regression lines for the 4 variables against crime, we can see the negative linear relationships. In all cases, as crime increases the value of the explanatory variable drops, though the outlier in pop_change may be having an unwanted influence on the regression line. I may have to correct for this.





(b) Fit the model:

$y=\hat{\beta}_0+\hat{\beta}_1 popchange+\hat{\beta}_2 pop18+\hat{\beta}_3 lunch+\hat{\beta}_4 income$

Crime=25.8518416+b1-0.154335+b2-0.8368053+b30.2360671+b4-0.0148923

```
##
## Call:
## lm(formula = crime ~ pop_change + pop_18 + lunch + income, data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.335  -4.685  -1.169   5.672  19.966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.85184    7.66954   3.371  0.00155 **
## pop_change  -0.15433    0.14822  -1.041  0.30333
## pop_18      -0.83681    0.25792  -3.244  0.00222 **
## lunch        0.23607    0.10726   2.201  0.03292 *
## income      -0.01489    0.24735  -0.060  0.95226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.103 on 45 degrees of freedom
## Multiple R-squared:  0.2414, Adjusted R-squared:  0.174
## F-statistic:  3.58 on 4 and 45 DF,  p-value: 0.01284
```

(i) Intepret the coefficient for pop_18

A 1% increase in crime causes a reduction in the % population of children by 84% adjusting for variables 'pop_change', 'lunch' and 'income'.

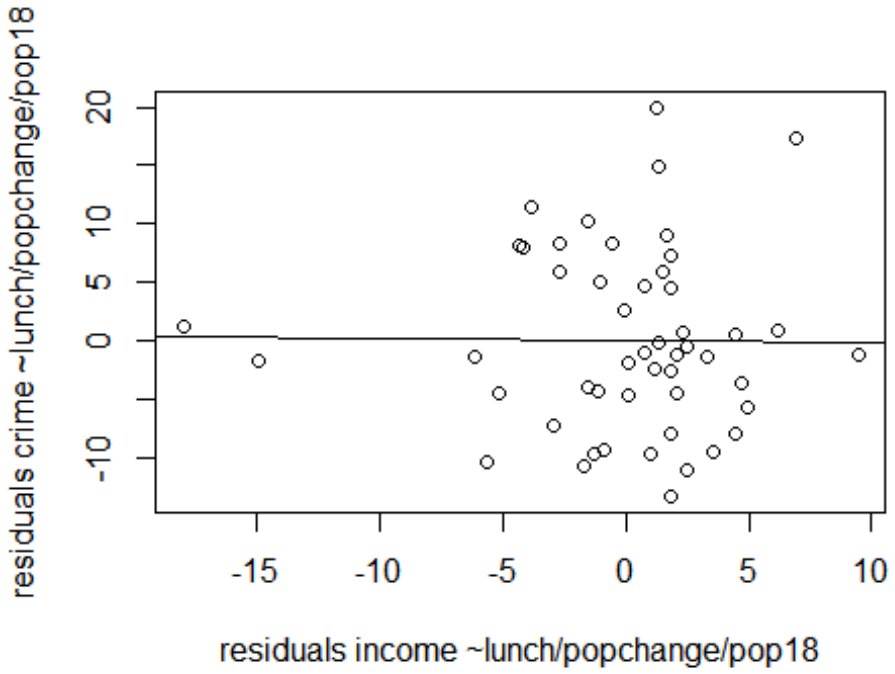
(ii) Calculate the variance inflation factors for this model and discuss their implications for collinearity in the model

```
## pop_change     pop_18      lunch     income
##   1.737596   2.743910   2.810836   1.616292
```

The VIFs lie between 1 and 3 for all independent variables indicating that collinearity is not having a large impact on the coefficient estimates for this model as it is less than 5. I will be analysing the full model in the diagnostics section below.

(iii) Create a partial regression plot to examine the relationship between 'income' and 'crime' adjusted for 'pop_change', 'lunch' and 'pop_18'

I fitted a regression model to the residuals from both models :




```
##
## Call:
## lm(formula = modelc$res ~ modeld$res)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.335  -4.685  -1.169   5.672  19.966
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.257e-15  1.110e+00   0.000    1.000
## modeld$res  -1.489e-02  2.395e-01  -0.062    0.951
##
## Residual standard error: 7.846 on 48 degrees of freedom
## Multiple R-squared:  8.055e-05,  Adjusted R-squared:  -0.02075
## F-statistic: 0.003867 on 1 and 48 DF,  p-value: 0.9507
```

(iv) Test the following hypothesis and What do the results of this test imply for the regression model? :

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

H_A: at least one of the $\beta_i \neq 0$

```
##
## Call:
## lm(formula = crime ~ pop_change + pop_18 + lunch + income, data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.335  -4.685  -1.169   5.672  19.966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.85184    7.66954   3.371  0.00155 **
## pop_change  -0.15433    0.14822  -1.041  0.30333
## pop_18      -0.83681    0.25792  -3.244  0.00222 **
## lunch        0.23607    0.10726   2.201  0.03292 *
## income      -0.01489    0.24735  -0.060  0.95226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.103 on 45 degrees of freedom
## Multiple R-squared:  0.2414, Adjusted R-squared:  0.174
## F-statistic:  3.58 on 4 and 45 DF,  p-value: 0.01284

##
## Call:
## lm(formula = crime ~ 1, data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.695  -6.571  -1.026   6.833  22.711
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)   12.818      1.261   10.17 0.000000000000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

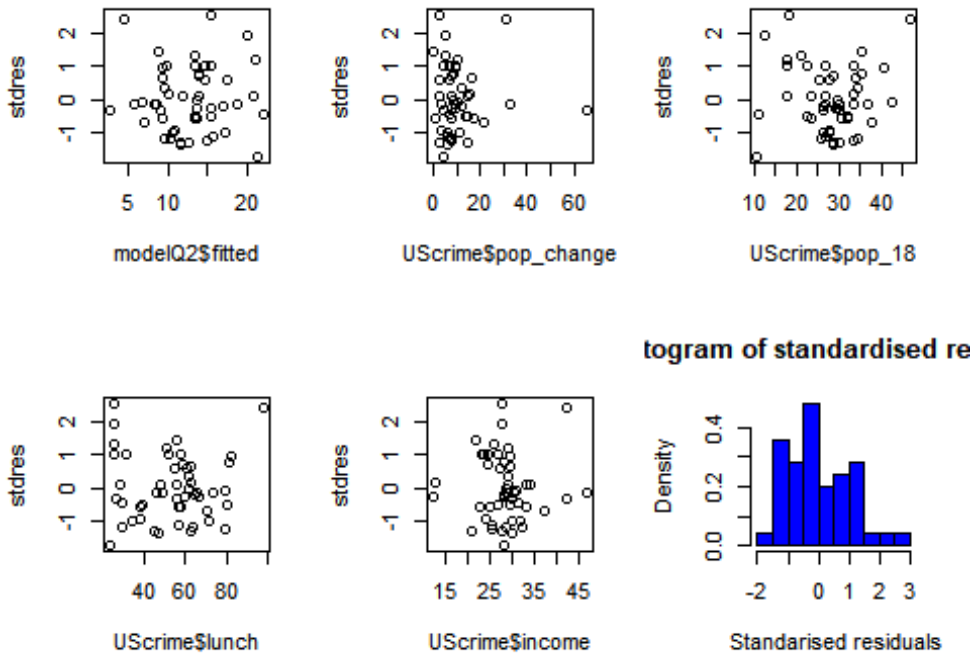
```
##
## Residual standard error: 8.916 on 49 degrees of freedom
```

The first model above contains all the explanatory variables and the second is an intercept only model which contains no explanatory variables. Given the significant difference between both models and the p-value of < 0.001 in the intercept only model - this indicates that at least one of the explanatory variables is of value in explaining the variation in crime.

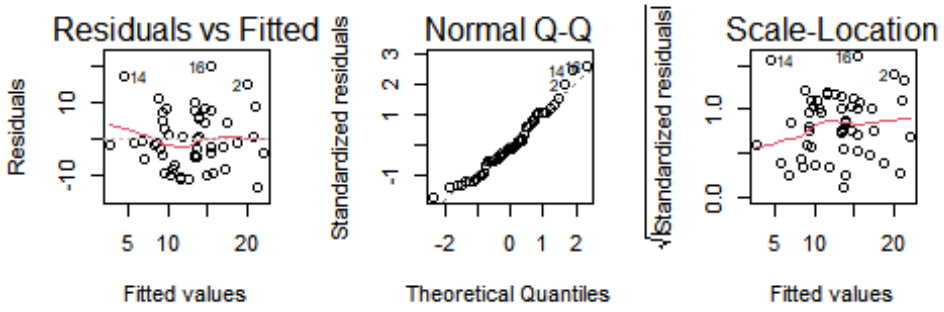
Examining the output from the full model, I can see that the global F-statistic is F(4, 45) = 3.58 and p =0.01284. In this instance I may reject the null hypothesis at the 5% confidence level and conclude that at least one of the β coefficients is not 0 and is associated with crime.

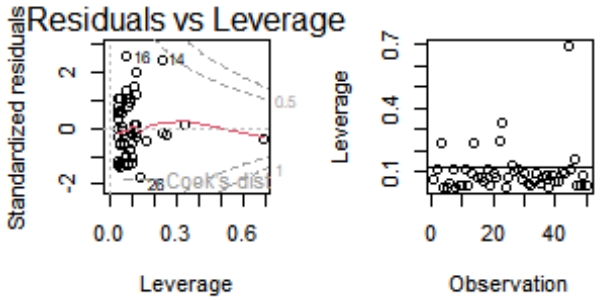
(v) Assess the fit of the model using diagnostic plots, commenting on the assumptions of the regression model and influencial points

Below, I have calculated the standardized residuals and plotted them against the fitted values and also to each of the explanatory variables. I included a histogram of the standardized residuals that show a normal distribution.




```
## integer(0)
```





(c) Use the predict function to calculate the expected crime rate when pop_change =20, pop_18=17, lunch=20 and income=30

Plugging the values into the prediction function, I get the predicted expected crime rate of:

```
##        1
## 12.81403
```

which means there will be a reported violent crime rate of 12.9 per 100,000 residents based on the above figures applied to the independent variables.

(d) Compare the full model (model y) to the model where income and lunch are excluded (model x) using 50 repeats of 10-fold cross validation. Which model is best to predict the crime rate?

```
## Linear Regression
##
## 50 samples
##  2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 50 times)
## Summary of sample sizes: 45, 45, 45, 44, 45, 46, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   8.445295  0.3651745  7.134267
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

## Linear Regression
##
## 50 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 50 times)
## Summary of sample sizes: 46, 44, 46, 44, 45, 46, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   8.194757  0.3601204  6.801687
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

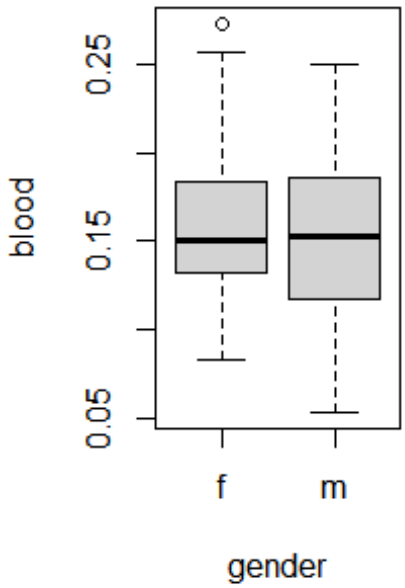
There is no significant difference between the prediction accuracy of the 2 model - with both performing reasonably well. The root mean squared error for both are 8.194 and 8.445, the lower RMSE is for the model with 4 predictors. R-Squared for both are 0.365 and 0.360 - slightly higher for the 2 predictor model. The mean absolute error 8.445 and 8.195 - again, the 4 predictor model has the lower value.

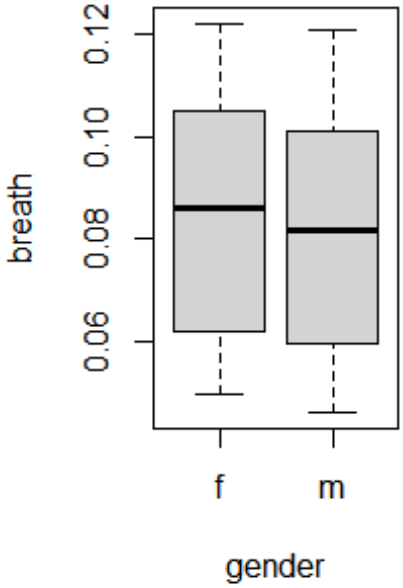
There is not a significant difference in the results. I would choose the model with 4 predictors (pop_change, pop_18, lunch and income) as both od their error results are slightly lower than the 2 predictor model (pop-change, pop_18).

Question 3

The alcohol dataset consists of 80 observations and 3 variables. Find out whether breath and gender are of use in predicting the value of Blood.

(a) Make a numerical and graphical summary of the data, commenting on the results.





```
##
##  Welch Two Sample t-test
##
## data:  blood by gender
## t = 0.99209, df = 77.902, p-value = 0.3242
## alternative hypothesis: true difference in means between group f and group
m is not equal to 0
## 95 percent confidence interval:
##  -0.01051283  0.03139732
## sample estimates:
## mean in group f mean in group m
##       0.1599802       0.1495379

## [1] 0.7548458

##
## Call:
## lm(formula = blood ~ breath * gender, data = alkie)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.084723 -0.022075  0.000134  0.017411  0.096371
##
## Coefficients:
##                Estimate Std. Error t value        Pr(>|t|)
## (Intercept)     0.01939    0.01861   1.042           0.301
## breath          1.65538    0.21133   7.833 0.0000000000229 ***
## genderm         0.01946    0.02601   0.748           0.457
## breath:genderm -0.28731    0.30230  -0.950           0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03121 on 76 degrees of freedom
## Multiple R-squared:  0.577,  Adjusted R-squared:  0.5603
## F-statistic: 34.55 on 3 and 76 DF,  p-value: 0.00000000000003415
```

(b) Fit the model

y=Breath+Gender+Breath:Gender

or

$y=\hat{\beta}_0+\hat{\beta}_1 Breath+\hat{\beta}_2 Z_1+\hat{\beta}_3 Years Z_1$+e

```
##
## Call:
## lm(formula = blood ~ breath * gender, data = alkie)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.084723 -0.022075  0.000134  0.017411  0.096371
##
## Coefficients:
##                Estimate Std. Error t value        Pr(>|t|)
## (Intercept)     0.01939    0.01861   1.042           0.301
## breath          1.65538    0.21133   7.833 0.0000000000229 ***
## genderm         0.01946    0.02601   0.748           0.457
## breath:genderm -0.28731    0.30230  -0.950           0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03121 on 76 degrees of freedom
## Multiple R-squared:  0.577,  Adjusted R-squared:  0.5603
## F-statistic: 34.55 on 3 and 76 DF,  p-value: 0.00000000000003415
```

(c) using the output from(b) write down the equation of the fitted regression line of blood on breath when separate regression lines are fitted for each level of gender.

$\hat{Y}$ = 0.0193921+ 1.6553796+0.0194591Z_1-0.2873127Z_1

(d) Use the predict function to calculate the expected blood alcohol measurement for a female who has a breath alcohol measurement of 0.078 mg/l

$\hat{Y}$ = 0.0193921+ 1.6553796(0.078)+0.0194591(0.78)(1) -0.2873127(0.078)(0)

$\hat{Y}$ = 0.0193 + .12911 + 0.001518 = 0.15

$\hat{Y}$ = 0.15

(e) Test the hypothesis that hte true regression lines for each gender are parallel

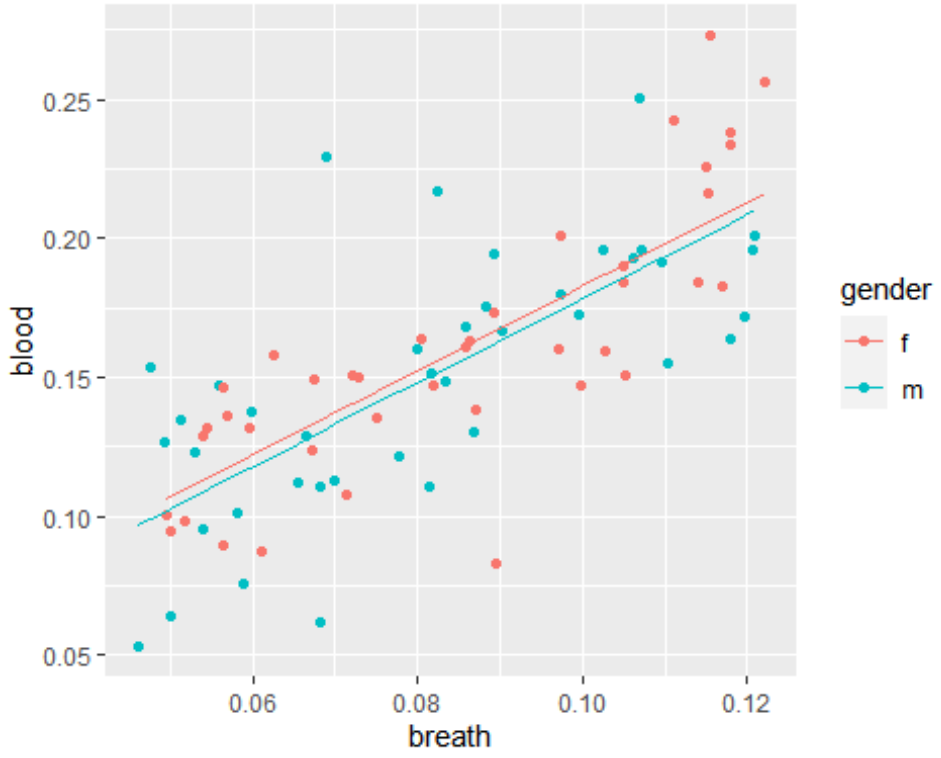
$$H_0: \beta_2 = \beta_3 = 0$$

H_A: β_2, β_3 not both 0

```
## Analysis of Variance Table
##
## Model 1: blood ~ breath + gender
## Model 2: blood ~ breath + gender + breath:gender
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     77 0.074914
## 2     76 0.074034  1 0.00087995 0.9033 0.3449
```

Based on the output shown above where I have created a model (m2) containing an interaction with breath and gender and a model that does not contain the interaction (m1), the results show that the m2 model explains significantly more variance in blood level than the the m1 model that does not include the interaction p=0.345 - therefore I am rejecting the null hypothesis and am concluding that the interaction ocefficients are different to 0 implying that the regression lines are not parallel.

(f) Plot the data using a different colour for each level of gender and plot the fitted regression lines.



(g) What do the results of this analysis tell you about the impact of gender and breath alcohol measurements on expected blood alcohol mesasurements?

From the results - females seem to have a slightly higher levels.

#ENDS