Computer Practical 9 and 10 Statistical Report

Alex Goodall, Joe Burnage, Sophia Pym

March 5, 2019

1 Preliminaries

Firstly, before we began and analysis of the data set we created 4 new data frames from the data set to group the patients, T1 and T2 costs by accomodation type using the code below.

It might be useful to take note of the identifiers given to the different data frames as they may appear later on in the document.

```
library(durham)
data(learndis)

domdis <- subset(learndis, ACCOM=="DOM", select = c("PATIENT","COSTS.T1","COSTS.T2"))
hosdis <- subset(learndis, ACCOM=="HOS", select = c("PATIENT","COSTS.T1","COSTS.T2"))
rnhdis <- subset(learndis, ACCOM=="RNH", select = c("PATIENT","COSTS.T1","COSTS.T2"))
sghdis <- subset(learndis, ACCOM=="SGH", select = c("PATIENT","COSTS.T1","COSTS.T2"))
```

2 Numerical Summaries

2.1 Summaries

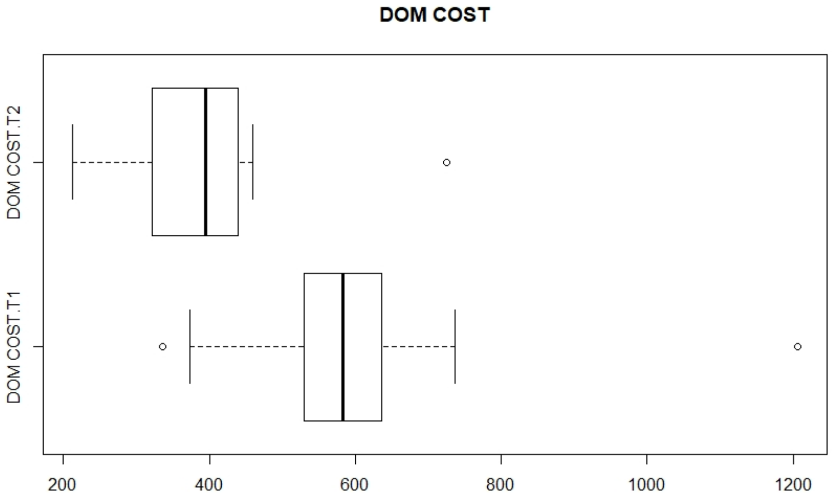
Below is some R output, displaying the numerical summaries for each of the distributions. In the next section we will use boxplots to visually assess the distributions, with T1 and T2 costs side by side for each of the accomodation types.

```
> summary(domdis$COSTS.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  335.2   529.9   583.1   605.6   635.9  1207.0
> summary(domdis$COSTS.T2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  212.2   322.2   394.8   390.4   439.1   725.3
>
> summary(hosdis$COSTS.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  407.1   458.3   799.9   725.2   913.7  1312.7
> summary(hosdis$COSTS.T2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  260.0   452.2   656.0   612.5   712.3  1177.8
>
> summary(rnhdis$COSTS.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  275.2   530.7   798.4   747.5   856.9  1539.1
> summary(rnhdis$COSTS.T2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  221.5   440.6   571.8   590.3   771.2   961.6
>
> summary(sghdis$COSTS.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  383.1   677.8   860.2   853.4   952.7  1655.8
> summary(sghdis$COSTS.T2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  304.4   707.2   846.4   825.3   980.5  1313.4
>
> summary(learndis)
    PATIENT          COSTS.T1          COSTS.T2      ACCOM
 Min.   :  1.00   Min.   : 275.2   Min.   : 212.2   DOM:15
 1st Qu.: 27.00   1st Qu.: 555.5   1st Qu.: 445.2   HOS:21
 Median : 53.00   Median : 783.9   Median : 659.3   RNH:24
 Mean   : 52.59   Mean   : 764.8   Mean   : 660.6   SGH:41
 3rd Qu.: 78.00   3rd Qu.: 922.9   3rd Qu.: 854.4
 Max.   :103.00   Max.   :1655.8   Max.   :1313.4
> summary(learndis$COSTS.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  275.2   555.5   783.9   764.8   922.9  1655.8
> summary(learndis$COSTS.T2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  212.2   445.2   659.3   660.6   854.4  1313.4
```

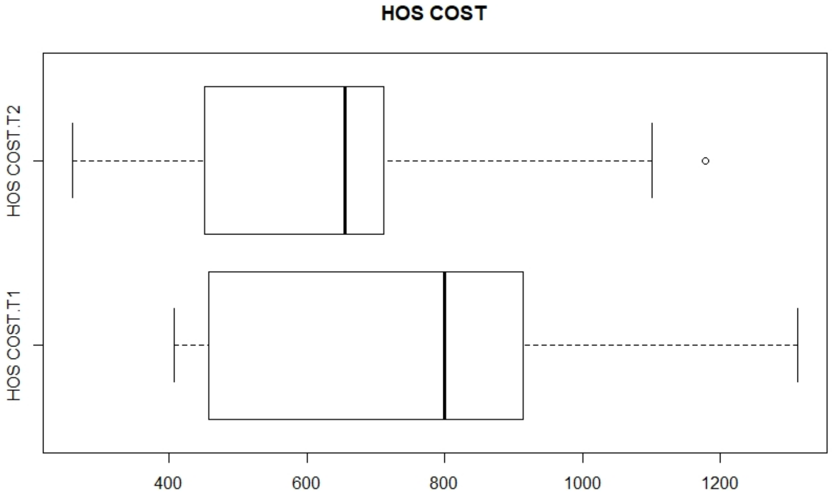
We can see that in general T2 costs are smaller than T1 costs, this statement holds for all accomodation types. We can also see the n values for each accomodation type, with domestic accomodation having the smallest $n = 15$, and professionally staffed homes having the largest $n = 41$. We may wish to consider this later on.

2.2 Boxplots

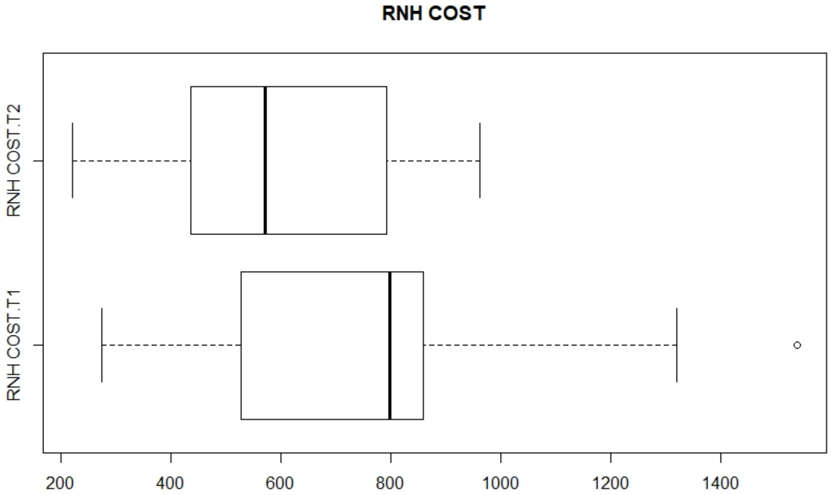
Illustrated by the boxplot below there is an obvious difference between the T1 and T2 costs for domestic accomodation. Both appear to be left skewed, with T2 appearing more left skewed than T1. Both have an outlier but T2 has a very extreme outlier which will need to be considered and may cause problems later on.



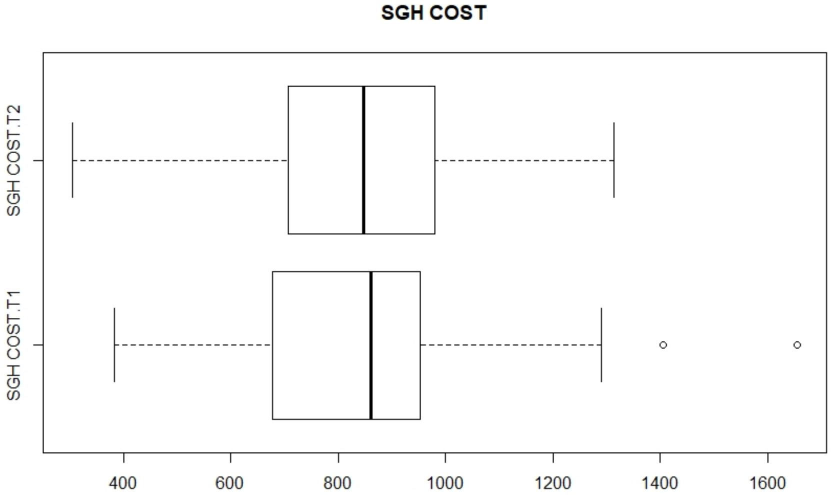
The box plot below shows us that T1 and T2 costs for hostel accomodation are both widely spread. The skew of the either distribution is not obvious from the boxplots, but we can see that in general T1 costs appear to be greater than T2 costs.



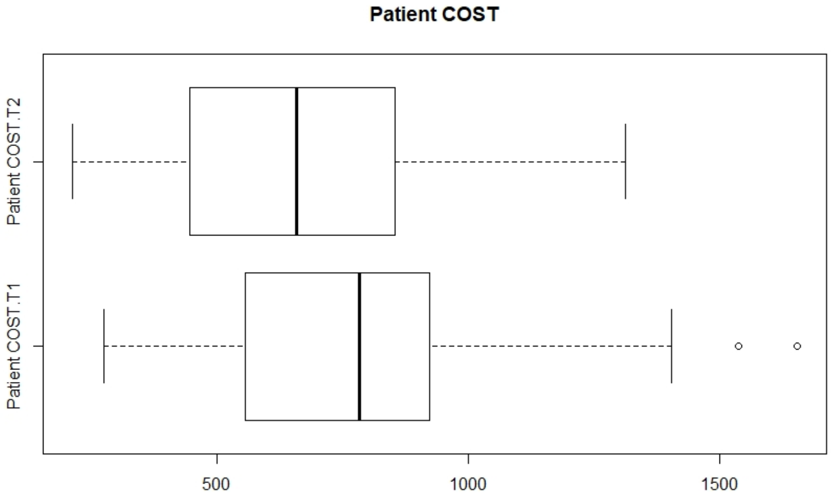
The box plot below shows there is some difference between T1 and T2 costs for residential and nursing homes. T1 costs are alot more widely spread than T2 costs, but T2 costs look very normal at a first glance.



T1 and T2 costs do not seem to differ much for professionally staffed homes. T2 costs are slightly wider spread and T1 costs have 2 outliers, but regardless, they appear to be very similarly distributed.



The box plot below shows the distribution for all T1 and T2 costs (regardless of the accomodation type). We can see that in general T1 costs ares higher than T2 costs, and at a first glance both distributions look reasonably normal.



3 Assesment of Normality

3.1 Stem and Leaf plots

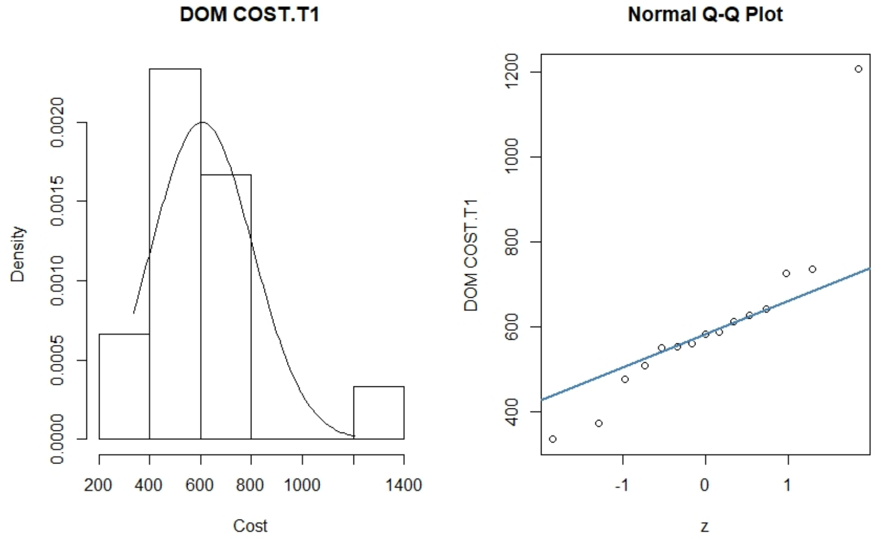
The following R output could be used to asses the normality of the different distributions by looking at the shape of the stem and leaf plots. However, we will fully asses the normaility of each distribution in the next section.

```
> stem(domdis$COSTS.T1)

  The decimal point is 2 digit(s) to the right of the |

   2 | 47
   4 | 8155689
   6 | 13434
   8 |
  10 |
  12 | 1

> stem(domdis$COSTS.T2)

  The decimal point is 2 digit(s) to the right of the |

  2 | 16922479
  4 | 022566
  6 | 3

>
```

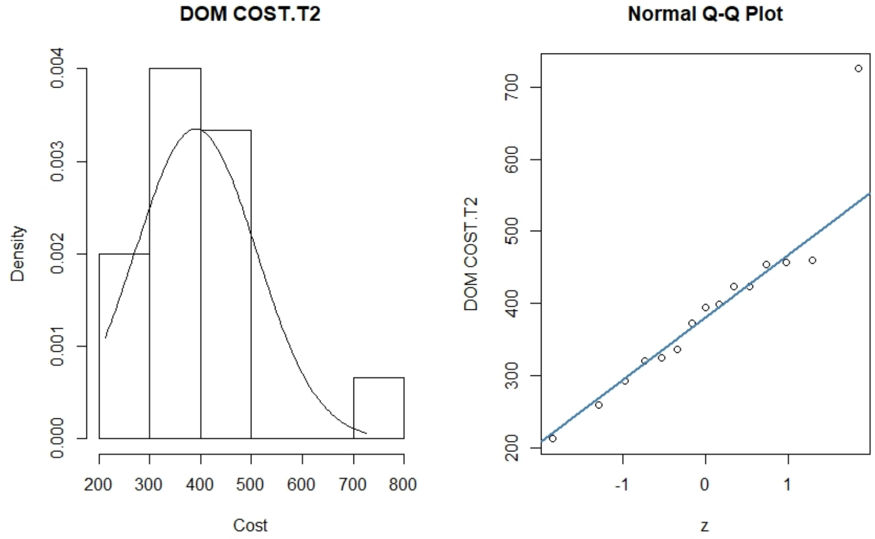
```
> stem(hosdis$COSTS.T1)

  The decimal point is 2 digit(s) to the right of the |

   4 | 11246646
   6 | 03
   8 | 0570113444
  10 |
  12 | 1

> stem(hosdis$COSTS.T2)

  The decimal point is 2 digit(s) to the right of the |

   2 | 646
   4 | 3456678
   6 | 66011145
   8 | 1
  10 | 08

>
> stem(rnhdis$COSTS.T1)

  The decimal point is 3 digit(s) to the right of the |

  0 | 334
  0 | 555567778888889999
  1 | 03
  1 | 5

> stem(rnhdis$COSTS.T2)

  The decimal point is 2 digit(s) to the right of the |

  2 | 28899
  4 | 359167788
  6 | 0335
  8 | 456236

>
> stem(sghdis$COSTS.T1)

  The decimal point is 2 digit(s) to the right of the |

   2 | 8
   4 | 22456
   6 | 1334823788
   8 | 02236890123334588
  10 | 36672
  12 | 9
  14 | 0
  16 | 6

> stem(sghdis$COSTS.T2)

  The decimal point is 2 digit(s) to the right of the |

   2 | 044
   4 | 556
   6 | 337011123579
   8 | 2455993344448
  10 | 055691348
  12 | 1

>
> stem(learndis$COSTS.T1)

  The decimal point is 2 digit(s) to the right of the |

   2 | 88
   3 | 4789
   4 | 11246689
   5 | 01223344555666889
   6 | 011333344678
   7 | 12334788
   8 | 0000222334456667889
   9 | 001112333344444588
  10 | 23667
  11 | 2
  12 | 19
  13 | 12
  14 | 0
  15 | 4
  16 | 6

> stem(learndis$COSTS.T2)

  The decimal point is 2 digit(s) to the right of the |

   2 | 12668899
   3 | 02244446799
   4 | 022334555556666789
   5 | 1667788
   6 | 03333667
   7 | 00111111233455579
   8 | 1244555699
   9 | 2333444468
  10 | 05569
  11 | 013488
  12 |
  13 | 1
```

3.2 Normal Fit

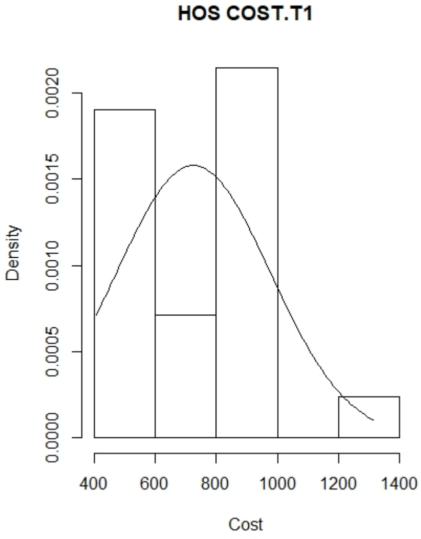
The T1 costs for domestic accomodation appear to loosely follow a normal distribution. We have a small n, and excluding the outlier we can see that the bulk of the data follows a linear trend on the normal quantile plot. The histogram also shows some signs of normality, so we could begin to make probabalistic calculations using the normal distribution.

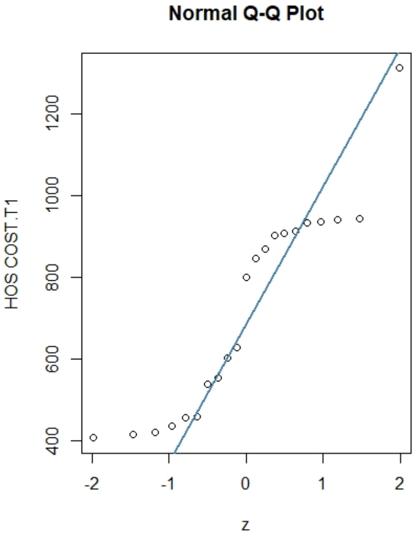


The T2 costs for domestic accomodation give us a similar situation. We have a small n, and an extreme outlier which can be excluded. On the normal quantile plot all the data follows the same linear trend apart from the outlier, so these costs actually appear even more normal and we could safely say they follow a normal distribution.

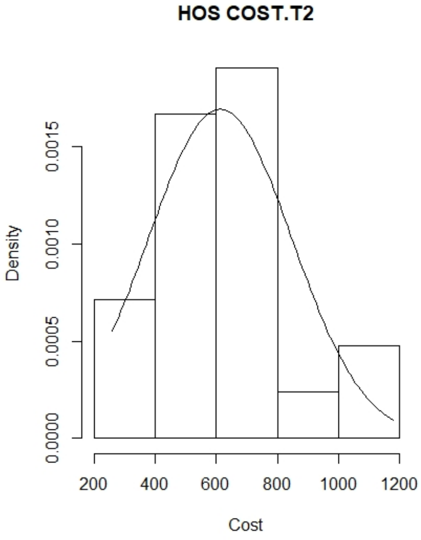


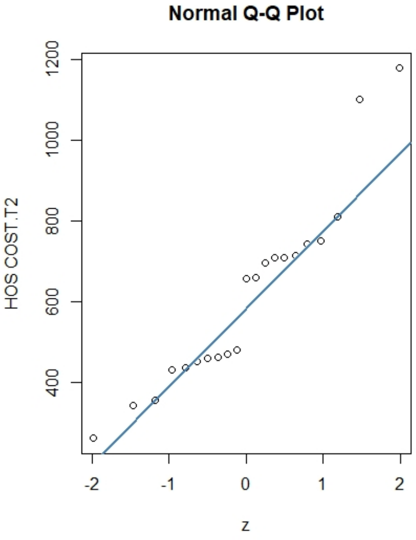
The T1 costs for hostel accomodation do not appear to follow a normal distribution. The histogram appears bi-modal, and there are systematic devations from the straight line on the normal quantile plot. Perhaps there is some other variable that has not been considered or the sampling was bias.



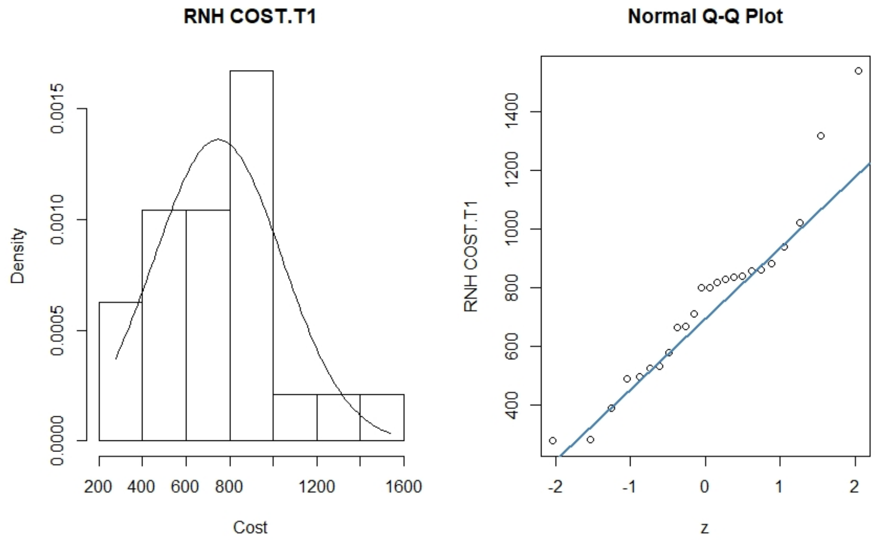


For the T2 costs for hostel accomodation we see something different. The histogram appears to be reasonably normal, however the normal quantile plot shows clustering around certain values. Because the histogram does not show the 'full picture' we can say this distribution is again not normal, because the normal quantile plots suggests it isn't.

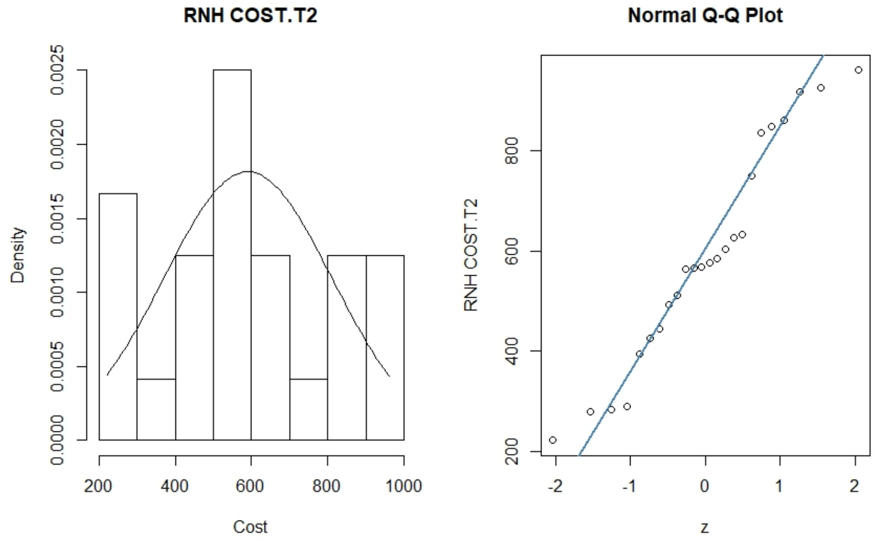




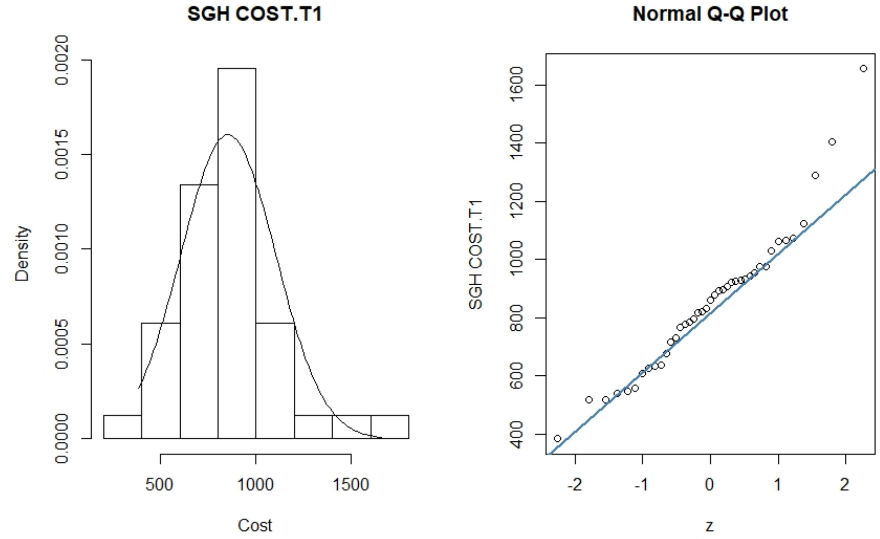
The T1 costs for residential and nursing homes appear to follow a normal distribution. The histogram is slightly right skewed but appears fairly normal, and the points in the normal quantile plot roughly follow a linear relationship without too much clustering or deviation from the straight line.



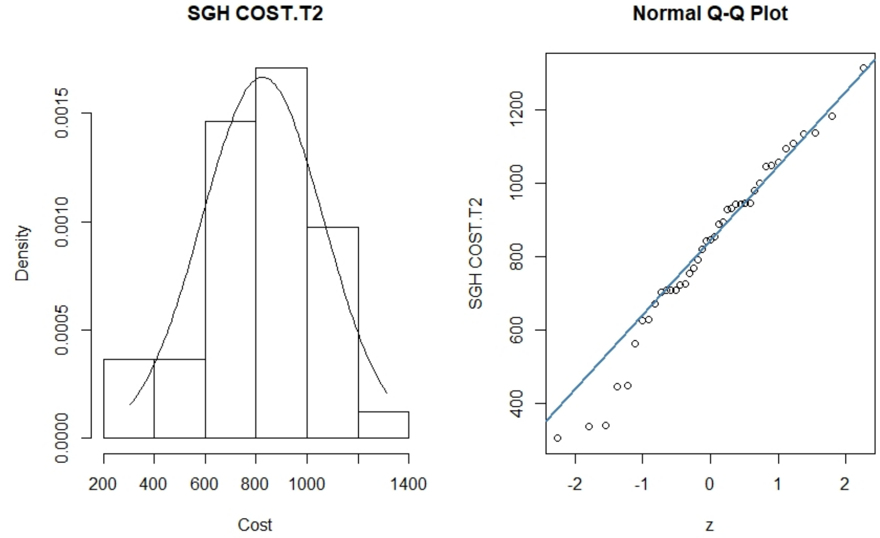
For the T2 costs for residential and nursing homes we get a strange histogram. We appear to have quite large tails but the distribution still seems to be fairly symmetric. The points on the normal quantile plot roughly follow a linear relationship with not not too much deviation from the line except at the tails. We could use a normal distribution but we would have to be careful, perhaps a t-distribution would be more suitable because of our large tails.



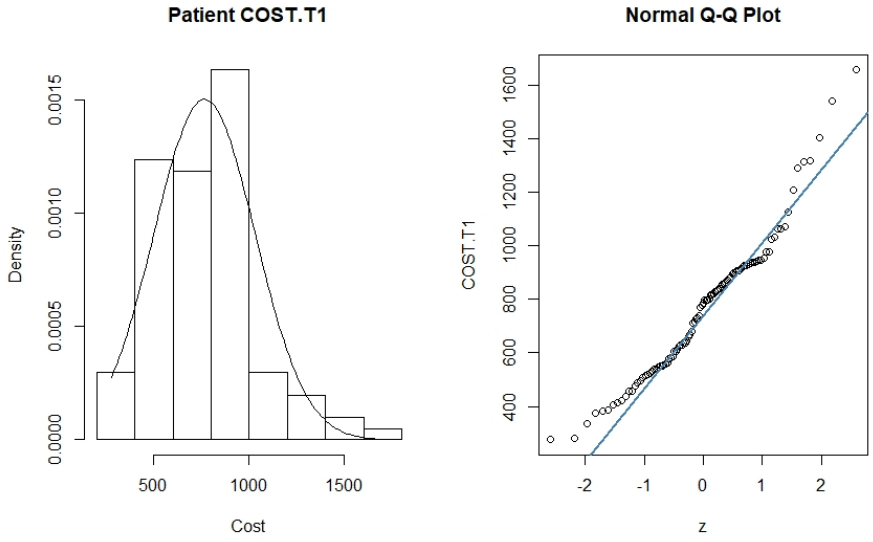
The T1 costs for professionally staffed homes appear to follow a normal distribution quite closely. The histogram looks very normal (only slightly right skewed), and the points on the normal quantile plot closely follow a linear relationship apart from the few outliers. This could because we have a relatively large n for professionally staffed homes.



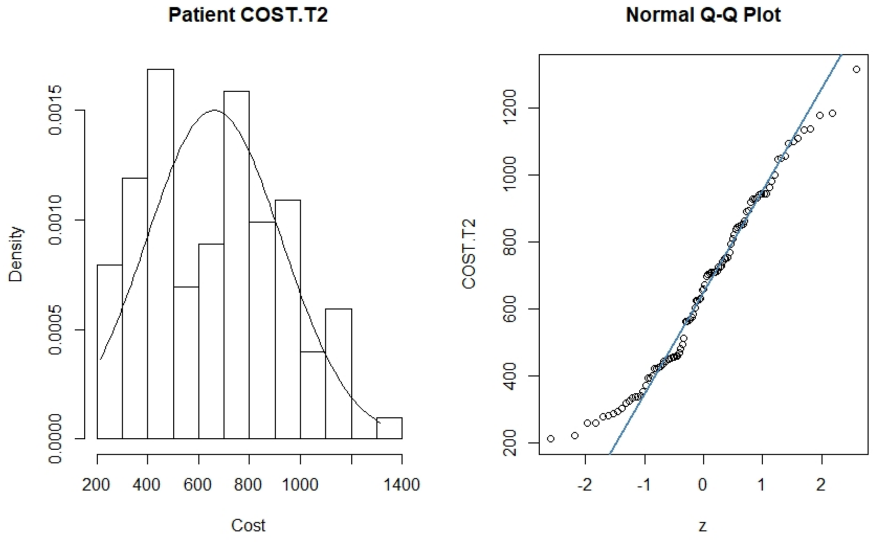
We get a similar situation for the T2 costs for professionally staffed homes. The histogram again looks quite normal, however it is left skewed in this case. The points on the normal quantile plot again closely follow a linear relationship apart from some of the smaller values.



The plots below are for the T1 costs of all patients regardless of the accomodation type. We can see some signs of normality; on the normal quantile plot a large bulk of the data follows some linear trend, but we have some systematic deviations at the tails. Overall the histogram shows some normality but the distribution appears to be right skewed. We would need to be careful if we use a normal distribution to model this data, perhaps some transformation could help us.



Below are the plots for the T2 costs of all the patients and accomodation types. Our histogram appears bimodal so a normal model would be unwise. However, our normal quantile plot seems to suggest reasonable normality for the data. Most of the points follow the same linear trend with the tails being exceptions, there seems to be less systematic deviation in the normal quantile plot, but it would still be risky to use a normal distribution to model this data. The bimodality of the histogram suggests there is anither underlying variable that has not been considered, a transformation could also be used.



4 Associations

4.1 Linear Models

Below is the R script that creates linear models for predicting costs at T2 from costs at T1 for every accomodation type and for the overall data set.

```
lmdom <- lm(domdis$COSTS.T2~domdis$COSTS.T1)
lmhos <- lm(hosdis$COSTS.T2~hosdis$COSTS.T1)
lmrnh <- lm(rnhdis$COSTS.T2~rnhdis$COSTS.T1)
lmsgh <- lm(sghdis$COSTS.T2~sghdis$COSTS.T1)
lmdis <- lm(learndis$COSTS.T2~learndis$COSTS.T1)
```

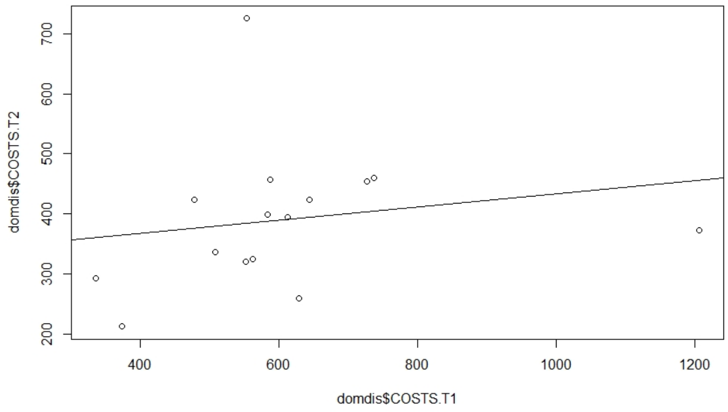
Below is the R output for the summaries of each of the linear models

```
> summary(lmdom)

Call:
lm(formula = domdis$COSTS.T2 ~ domdis$COSTS.T1)

Residuals:
    Min      1Q  Median      3Q     Max
-152.85  -66.10    3.70   48.94  340.69

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     324.4758   103.4638   3.136  0.00788 **
domdis$COSTS.T1   0.1088     0.1628   0.668  0.51555
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 121.5 on 13 degrees of freedom
Multiple R-squared:  0.03323,Adjusted R-squared:  -0.04114
F-statistic: 0.4468 on 1 and 13 DF,  p-value: 0.5155

> summary(lmhos)

Call:
lm(formula = hosdis$COSTS.T2 ~ hosdis$COSTS.T1)

Residuals:
    Min      1Q  Median      3Q     Max
-356.32 -157.87   39.95   93.46  569.67

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     627.32289  163.82292   3.829  0.00113 **
hosdis$COSTS.T1  -0.02037    0.21388  -0.095  0.92511
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 241.6 on 19 degrees of freedom
Multiple R-squared:  0.0004773,Adjusted R-squared:  -0.05213
F-statistic: 0.009073 on 1 and 19 DF,  p-value: 0.9251
```

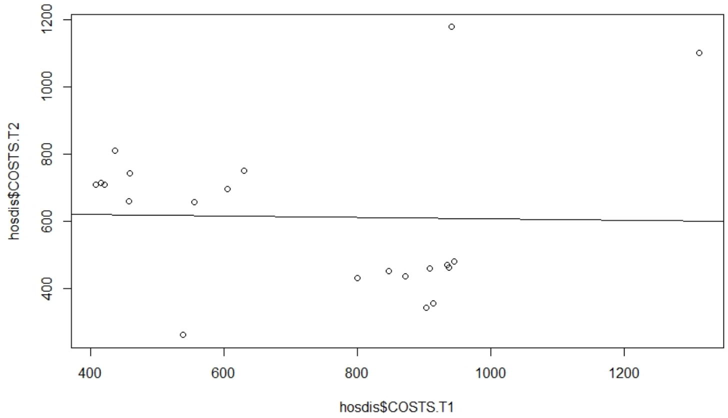
```
> summary(lmrnh)

Call:
lm(formula = rnhdis$COSTS.T2 ~ rnhdis$COSTS.T1)

Residuals:
    Min      1Q  Median      3Q     Max
-347.50 -128.13  -27.94  109.76  359.92

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     390.4512   119.4008   3.270   0.0035 **
rnhdis$COSTS.T1   0.2674     0.1491   1.793   0.0867 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 209.9 on 22 degrees of freedom
Multiple R-squared:  0.1275,Adjusted R-squared:  0.08787
F-statistic: 3.216 on 1 and 22 DF,  p-value: 0.0867

> summary(lmsgh)

Call:
lm(formula = sghdis$COSTS.T2 ~ sghdis$COSTS.T1)

Residuals:
    Min      1Q  Median      3Q     Max
-520.28 -117.25   22.53  153.95  488.12

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     831.501942 136.834644   6.077 4.05e-07 ***
sghdis$COSTS.T1  -0.007232   0.154079  -0.047    0.963
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 242.5 on 39 degrees of freedom
Multiple R-squared:  5.649e-05,Adjusted R-squared:  -0.02558
F-statistic: 0.002203 on 1 and 39 DF,  p-value: 0.9628

> summary(lmdis)

Call:
lm(formula = learndis$COSTS.T2 ~ learndis$COSTS.T1)

Residuals:
    Min      1Q  Median      3Q     Max
-413.51 -230.12   -7.11  202.78  627.52

Coefficients:
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       458.47856   78.73975   5.823 7.19e-08 ***
learndis$COSTS.T1   0.26435    0.09733   2.716   0.0078 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 258.1 on 99 degrees of freedom
Multiple R-squared:  0.06935,Adjusted R-squared:  0.05995
F-statistic: 7.377 on 1 and 99 DF,  p-value: 0.007797
```

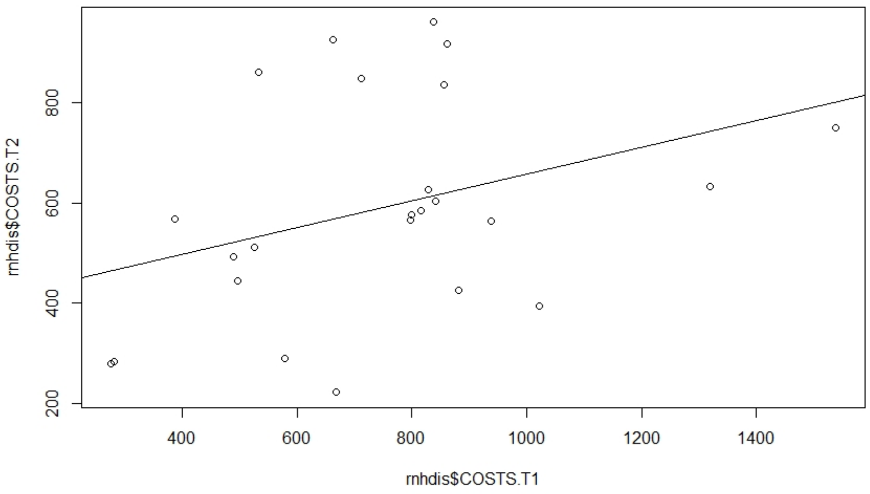
The below graph plots T2 costs against T1 costs for domestic accomodation with a regression line. There is a weak positive correlation but it is clear there is little association between the 2 varibales from the graph and notably the R value is only 0.03323



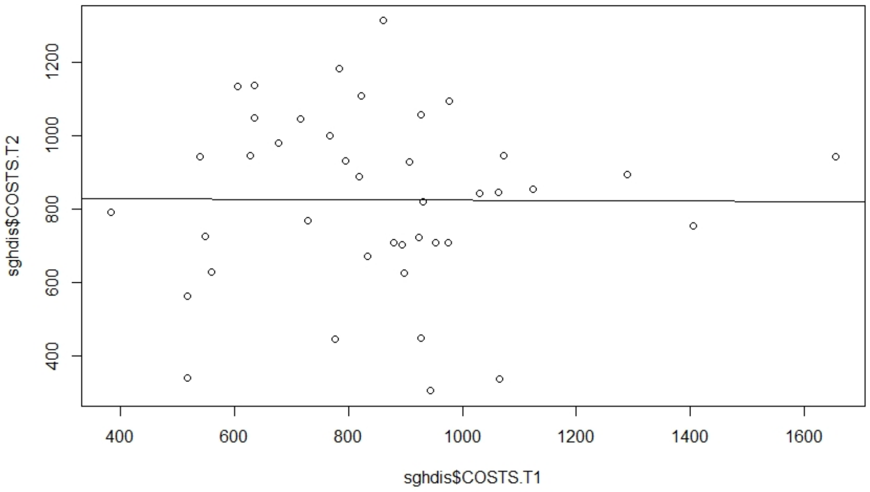
The below graph plots T2 costs against T1 costs for hostel accomodation with a regression line. There appears to be an incredibly weak negative correlation, with 2 seperate clusters of points, so again there is little association between the variables.



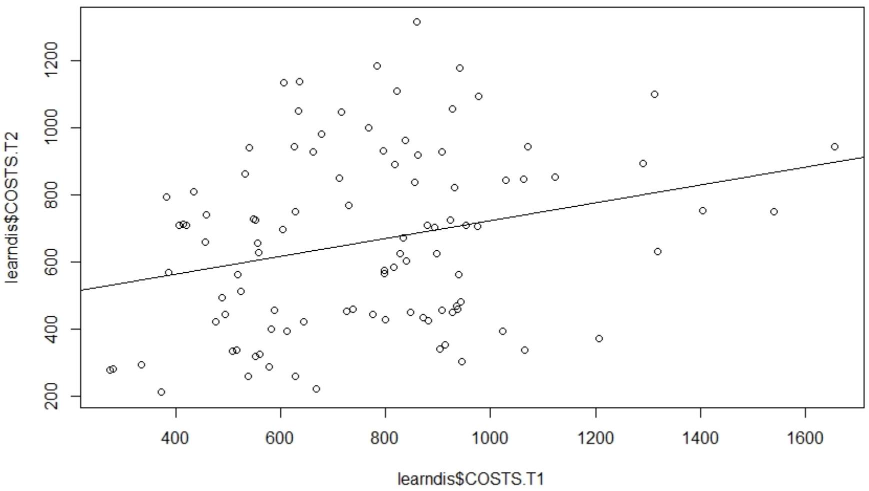
The next graph plots T2 costs against T1 costs for residential and nursing homes. There appears to be some positive correlation, but again it is weak. Infact it is the strongest linear model with an R value of 0.1275.



The next graph plots T2 costs against T1 costs for professionally staffed homes. There appears to be np correlation at all and an R value of 5.649e-05, would indeed suggest there is no association between the 2 variables.



The final graph shows all the T2 costs against all the T1 costs for all the accomodation types. There appears to be a positive correlation but it is weak with an R value of 0.06935, so there is a weak association between the variables.



5 Stability of Associations

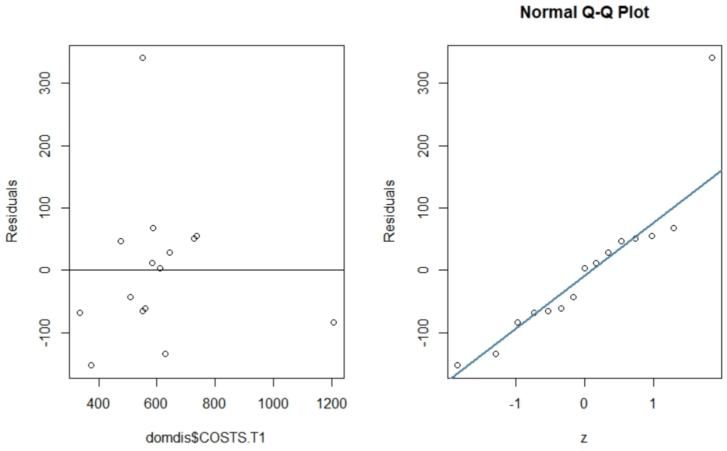
5.1 Residuals

The below R script stored the residuals of the linear models in some variables.

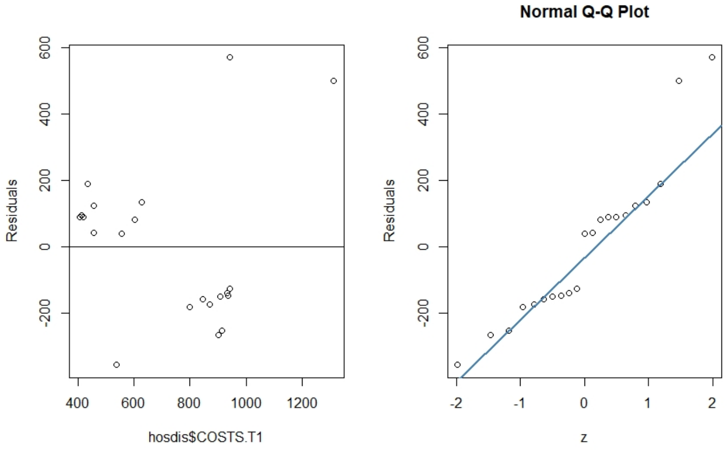
```
residdom <- resid(lmdom)
residhos <- resid(lmhos)
residrnh <- resid(lmrnh)
residsgh <- resid(lmsgh)
residdis <- resid(lmdis)
```

The following commentary will be breif because we already assessed the normality of the distributions previously, and assesing the normality of the residuals will give us the same results as the T2 costs. A residual plot is also provided as an alternate way of visually assesing the normality of the residuals for each of the T2 costs.

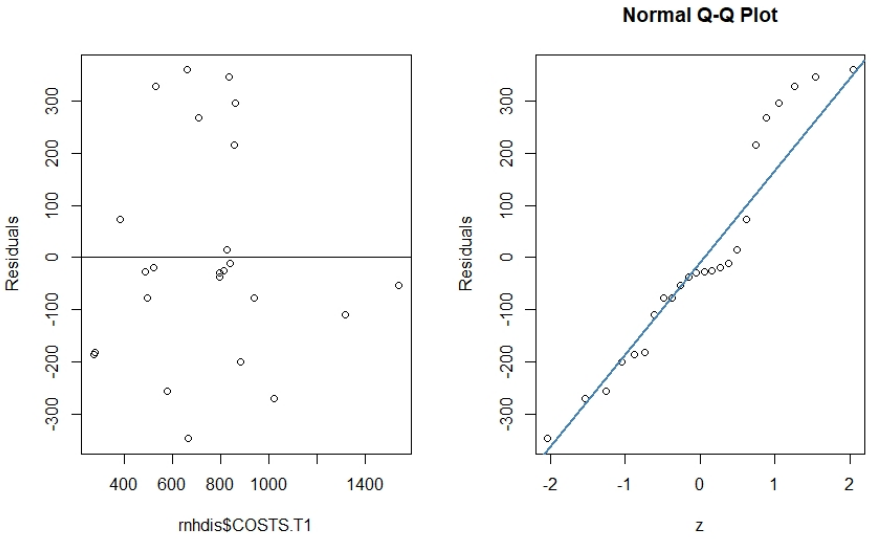
The residulas below for domestic accomodation are normal. So any probabalistic statements made are valid.



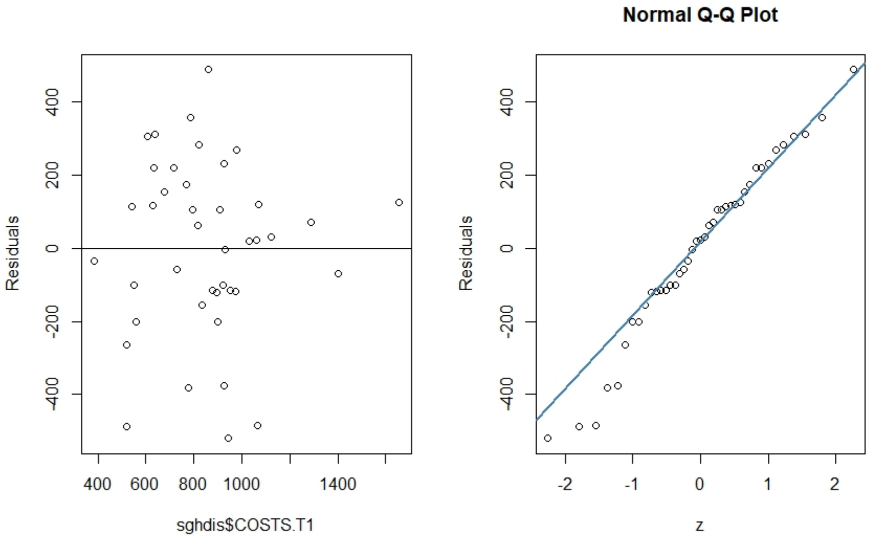
The residuals below for hostel accomodation are not normal. So any probabalistic statements made are not valid.



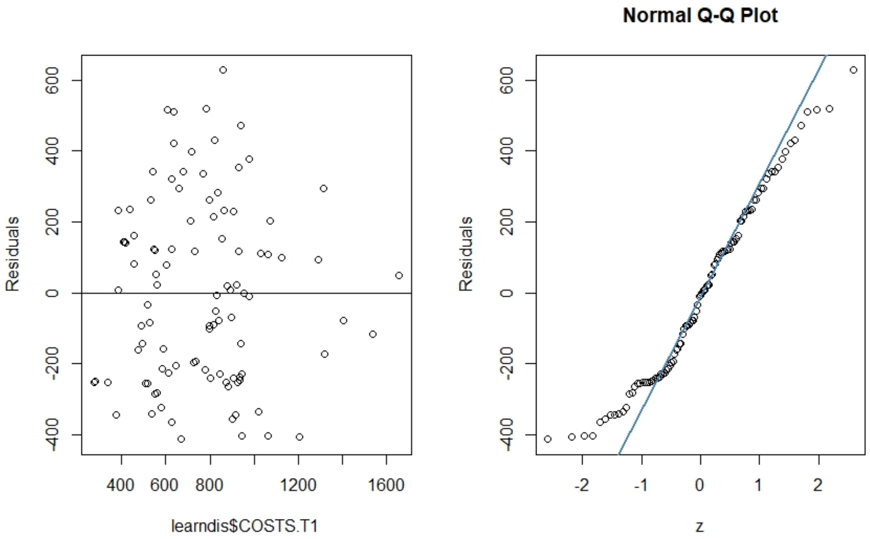
The residuals below for residential and nursing homes are not normal.



The residuals below for professionally staffed homes are normal.



The residuals below are for all T2 costs for all accomodation types. For the majority of the data they are normal, but the tails deviate slightly. We must be careful but we could say any probabalistic statements are valid.



6 Multiple Regression

Finally, we decided to use multiple regression to see whether T2 costs can be predicted from T1 costs, using using accomodation type as a second variable. The R script below does this for us, storing the linear model in a variable for later use.

```
lmdis2 <- lm(learndis$COSTS.T2~learndis$COSTS.T1+learndis$ACCOM)
```

Below is some R output that summarises the linear model, we can use this to asses how well the data fits to the model.

```
> summary(lmdis2)

Call:
lm(formula = learndis$COSTS.T2 ~ learndis$COSTS.T1 + learndis$ACCOM)

Residuals:
    Min      1Q  Median      3Q     Max
-528.84 -150.95    2.83  126.74  546.52

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       337.77748   77.92170   4.335 3.60e-05 ***
learndis$COSTS.T1   0.08686    0.08788   0.988  0.32547
learndis$ACCOMHOS 211.77891   75.25584   2.814  0.00593 **
learndis$ACCOMRNH 187.60669   73.61525   2.548  0.01241 *
learndis$ACCOMSGH 413.42936   69.98914   5.907 5.28e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 220.4 on 96 degrees of freedom
Multiple R-squared:  0.3418,Adjusted R-squared:  0.3144
F-statistic: 12.46 on 4 and 96 DF,  p-value: 3.322e-08

> coefficients(lmdis2)
      (Intercept) learndis$COSTS.T1 learndis$ACCOMHOS learndis$ACCOMRNH learndis$ACCOMSGH
     337.77747920        0.08685749      211.77891481      187.60668932      413.42936376
```

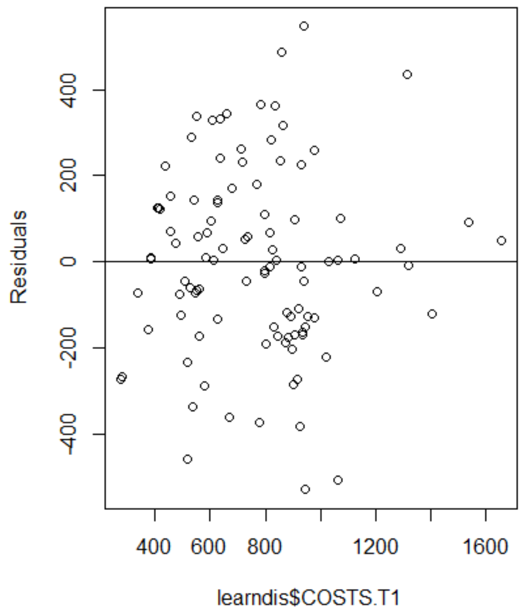
We get a formula of the form,

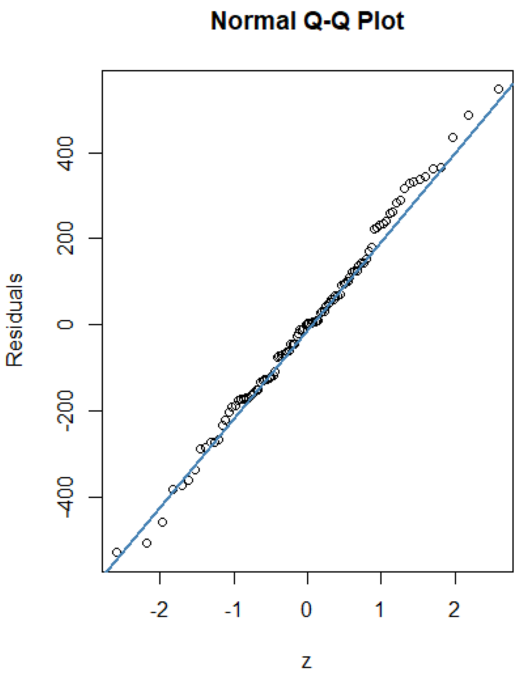
$$COSTS.T2 = 337.8 + 0.08686 \cdot COSTS.T1 + 211.8 \cdot ACCOM.HOS + 187.6 \cdot ACCOM.RNH + 413.4 \cdot ACCOM.SGH$$

where $ACCOM.HOS$, $ACCOM.RNH$ and $ACCOM.SGH$ are mutually exclusive binary variables.

We also get an R value of 0.3418, which suggests we have a weak positive correlation. Although notably this linear model is the strongest one yet.

If we begin to asses the normality of the residuals, we see below that they appear very normal. All the points closely follow a linear relationship on the normal quantile plot and the residual plot is resonably sparse. This means we can make accurate probabalistic statements about our predictions from the linear model.


Residuals
learndis$COSTS.T1


Normal Q-Q Plot
Residuals
z