MTH2006 Summative Coursework for Year 2, Term 1

Section 1

Introduction

Taking the daily rainfall totals over 10 years for the village of Alfheim, I will discuss the distribution of the data and find a suitable expression for the constant c as provided. Using this, I will find the likelihood function for a small given sample of data. Finally, I will partition the data to produce maximum likelihood estimators for the distribution parameters and compare them to the results of numerical optimisation.

Features of the Distribution of the Overall Data

A density histogram can be considered a rough emulation of the pdf across the range space, in this case it shows the frequency over 10 years of days occurring with certain rainfall totals in Alfheim. $Figure\ 1$ shows a majority of days fall into the lowest bin of 0-3mm of rainfall, and the remaining data demonstrate a long tail with relatively very low frequency of days, especially with more than 50 mm of rainfall.

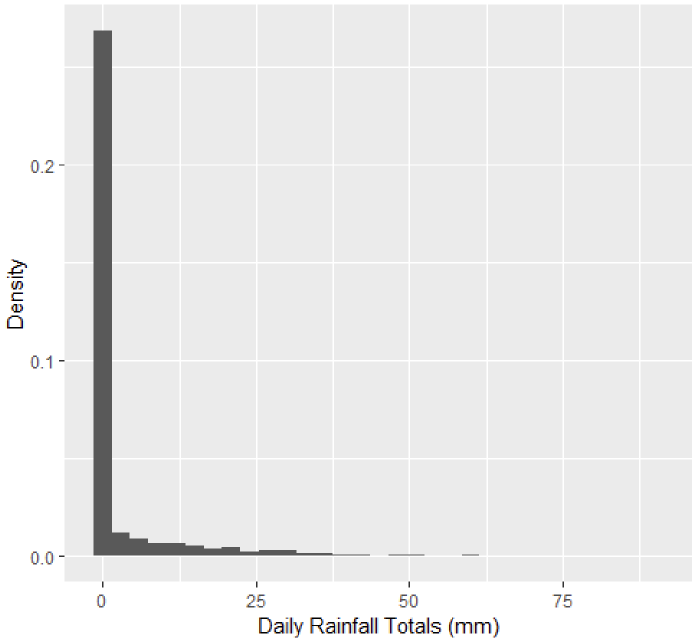


Figure 1: Histogram plot showing the distribution of rainfall totals in the full dataset

This is demonstrated similarly by the scatter plot in $Figure\ 2$ showing a very strong positive skew across the data that is consistent throughout the 10 years of data collection and seems independent of the day of the data being recorded ie. a constant variance. The scatter plot also suggests a consistent concentration ie. constant mean near to 0.

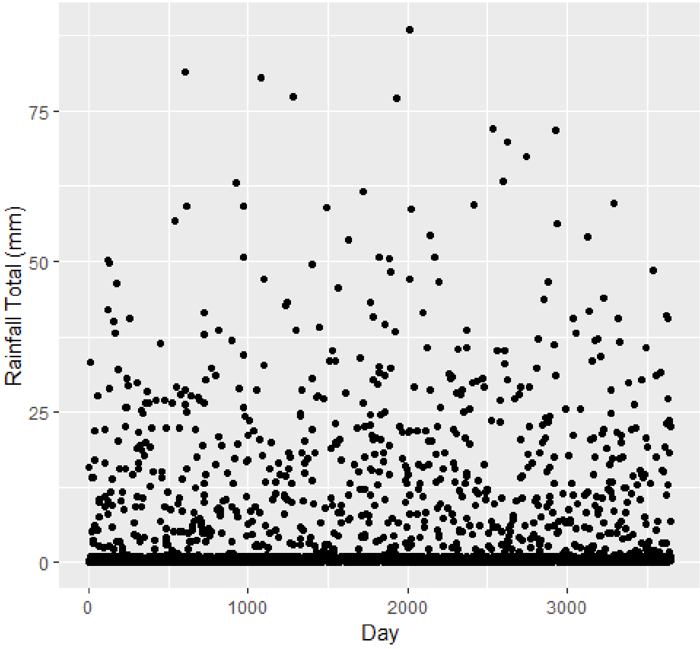


Figure 2: Scatter plot showing rainfall totals in the full dataset

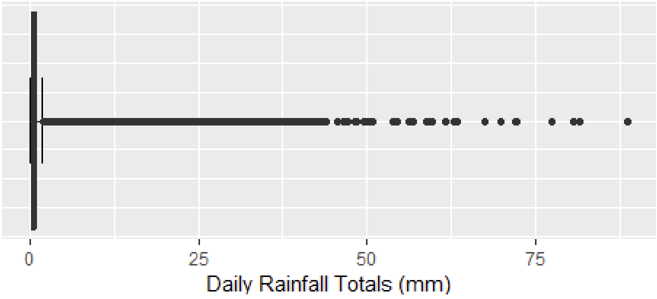


Figure 3: Boxplot showing the distribution of rainfall totals in the full dataset

It seems unreasonable to use $Figure\ 3$ above to describe the data as the datapoints are so highly concentrated near to 0 with such a small IQR that majority of the data is plotted as outliers, and it is difficult to infer information about the quartiles or shape of the distribution.

Instead the following boxplots and histograms have been plotted to show the distribution of rainfall totals separately for $y \leq 1$ and for $y > 1$.

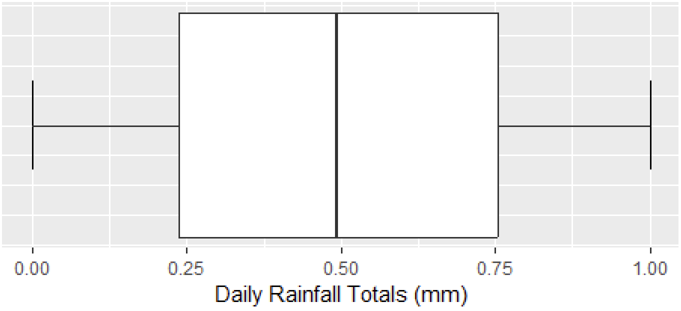


Figure 4: Boxplot showing distribution of rainfall totals where y ≤ 1

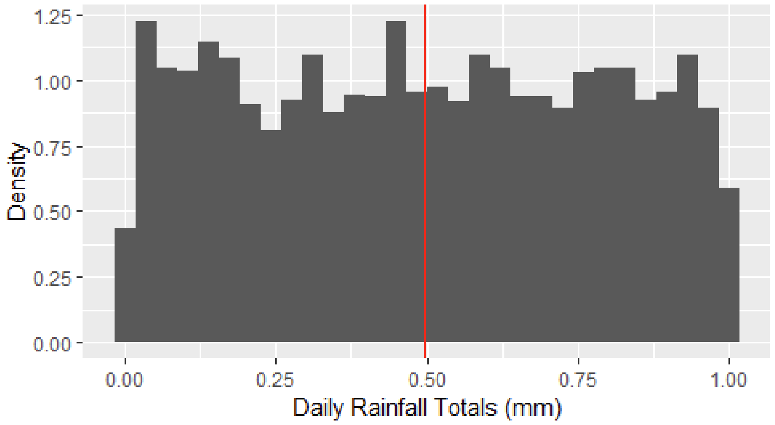


Figure 5: Histogram plot showing distribution of rainfall totals where y ≤ 1, red line is the mean for y ≤ 1

The rainfall totals where $0 < y \leq 1$ can be seen to reasonably follow a uniform distribution due to a mean and median almost at $y = 0.5mm$ and equal spread across the range $[0, 1]$.

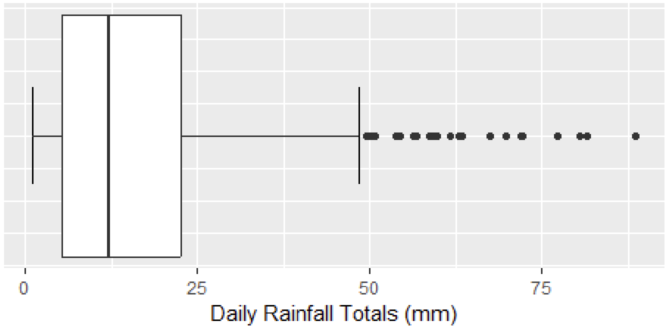


Figure 6: Boxplot showing distribution of rainfall totals where y > 1

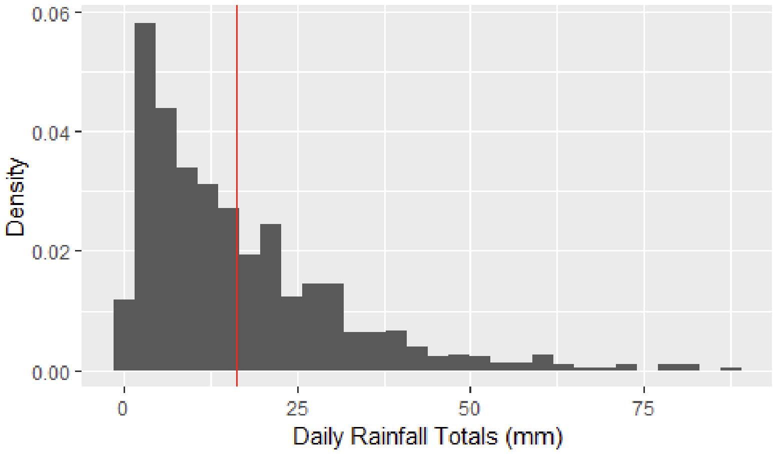


Figure 7: Histogram plot showing distribution of rainfall totals where y > 1, red line is the mean for y > 1

The rainfall totals where $y > 1$ show a strong positive skew, such that >75% of the datapoints lie below 25mm of rainfall. The boxplot demonstrates a long, right-hand tail and the histogram reinforces this with a mean (≈15mm) to the right of the median (≈12mm). Given that the data is measured as a positive continuous variable, the exponential distribution is a reasonable shape to represent the data where $y > 1$.

Finding a Suitable Expression for Constant “c”

Given the following distributions for the partitioned data:

$$where\ y \leq 1, Y\ follows\ a\ Uniform\ Dist. such\ that\ pdf: 1 - \phi$$

$$where\ y > 1, Y\ follows\ an\ Exponential\ Dist. such\ that\ pdf: ce^{-\theta y}$$

I will use the integration that the sum of the pdfs over the range space must sum to 1, therefore:

$$1 = \int_0^1 f_U(y).dy + \int_1^\infty f_E(y).dy$$

$$= [(1-\phi)y]_0^1 + c\left[-\frac{1}{\theta}e^{-\theta y}\right]_1^\infty$$

$$\lim_{y\to\infty}(-\frac{1}{\theta}e^{-\theta y}) \to 0, therefore:$$

$$= 1 - \phi + \frac{c}{\theta}e^{-\theta}$$

$$so:\ c = \phi\theta e^{\theta}$$

Finding the Likelihood Function for a given Sample

$$for\ the\ sample: x = \{30, 0.2, 10, 0\}, where\ m = no. of\ x_i \leq 1\ and\ n = no. of\ x_i > 1$$

$$L(\phi, \theta; x) = \prod_{i=1}^{m}(1-\phi)\prod_{j=1}^{n}\left(ce^{-\theta x_j}\right)$$

$$= (1-\phi)^m(\phi\theta)^n\prod_{j=1}^{n}(e^{\theta(1-x_j)})$$

$$(1-\phi)^m(\phi\theta)^n \exp\left[\theta\sum_{j=1}^{n}\left(1-x_j\right)\right]$$

$$since: \sum_{j=1}^{n}\left(1-x_j\right) = n - n\bar{x}_{x>1}, therefore:$$

$$L(\phi, \theta; x) = (1-\phi)^2(\phi\theta)^2\exp[-38\theta]$$

This likelihood function is only applicable to the small sample x assuming that it is rainfall totals measured in Alfheim or elsewhere but can still be determined to follow the same distribution shown above for both $x \leq 1$ and $x > 1$.

Additionally, this likelihood function is assuming that all the rainfall measurements are daily measurements which are independent from each other.

Deriving MLEs for ϕ and θ by applying the Likelihood Function to partitioned data Y

Using the partitions of the original dataset as described previously, for n = 3650:

$$L(\phi, \theta; y) = (1-\phi)^{n-m}(\phi\theta)^{n} \exp\left[m\theta\left(1-\bar{y}_{y>1}\right)\right]$$

$$where\ m = no.of\ y_i > 1$$

Now to find the Log-Likelihood and MLEs:

$$l(\phi, \theta; y) = \log L(\phi, \theta; y)$$

$$= (n-m)\log(1-\phi) + n\log(\phi) + n\log(\theta) + m\theta\left(1-\bar{y}_{y>1}\right)$$

Partial Differentiation can be used to find the MLEs since there is more than one parameter:

$$\frac{\partial l}{\partial \phi} = \frac{m-n}{1-\phi} + \frac{n}{\phi}$$

$$\frac{\partial l}{\partial \phi} = 0, therefore:\ \ \frac{\phi - 1}{\phi} = \frac{m-n}{n}\ and\ so:\ \ \phi = 2 - \frac{m}{n}$$

$$\frac{\partial^2 l}{\partial \phi^2} = \frac{m-n}{(1-\phi)^2} - \frac{n}{\phi^2}, m \leq 3560\ therefore\ \frac{\partial^2 l}{\partial \phi^2} < 0\ so\ \hat{\phi} = \ 2 - \frac{m}{n}$$

$$\frac{\partial l}{\partial \theta} = \frac{n}{\theta} + m\left(1-\bar{y}_{y>1}\right)$$

$$\frac{\partial l}{\partial \theta} = \ 0, therefore:\ \ \frac{n}{m} = \ -m\left(1-\bar{y}_{y>1}\right)\ and\ so:\ \ \theta = \frac{n}{m\left(\bar{y}_{y>1}-1\right)}$$

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{n}{\theta^2}, \theta^2 > 0\ therefore\ \frac{\partial^2 l}{\partial \theta^2} < 0\ so\ \hat{\theta} = \frac{n}{m\left(\bar{y}_{y>1}-1\right)}$$

Calculated for the Alfheim dataset Y: $\hat{\phi} = \ 1.799\ (4sf)$ and $\hat{\theta} = \ 0.3236\ (4sf)$.

An assessment of the correlation between these MLEs can be made by reviewing the covariance matrix of their joint distribution which can be done by calculating the expected information:

$$-\frac{\partial^2 l}{\partial \phi^2} = \frac{n}{\phi^2} - \frac{m\ -n}{(1-\phi)^2}\ and - \frac{\partial^2 l}{\partial \theta^2} = \frac{n}{\theta^2}$$

$$also:\ -\frac{\partial^2 l}{\partial \theta \partial \phi} = -\frac{\partial^2 l}{\partial \phi \partial \theta} = 0$$

$$therefore:\ \ J(\phi, \theta; Y) = \begin{bmatrix} \frac{n}{\phi^2} - \frac{m\ -n}{(1-\phi)^2} & 0 \\ 0 & \frac{n}{\theta^2} \end{bmatrix}$$

$$and:\ Cov\left(\hat{\phi}, \hat{\theta}\right) = \ I(\phi, \theta)^{-1} = E(J)^{-1} = c\begin{bmatrix} \left(\frac{n}{\phi^2} - \frac{m\ -n}{(1-\phi)^2}\right)^{-1} & 0 \\ 0 & \frac{\theta^2}{n} \end{bmatrix}$$

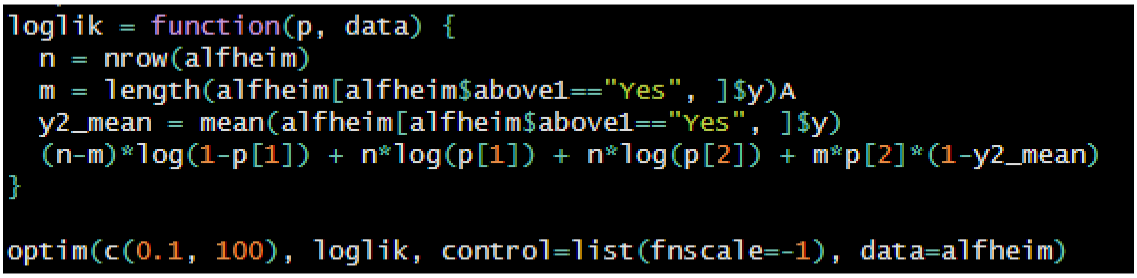
Given that the distribution parameters do not affect the range space of Y, in the asymptotic limit as $n \to \infty$ (ie. for a large enough sample size n) then the joint distribution of the MLEs can be assumed to be:

$$(\hat{\phi}, \hat{\theta}) \sim BVN\big((\phi, \theta), I(\phi, \theta)^{-1}\big)$$

Therefore, seeing as the off-diagonal entries of their covariance matrix are 0, it can be inferred that there is no correlation expected between the MLEs.

Finding the Estimates for ϕ and θ using Numerical Optimisation

Using the following code I was able to estimate the parameters digitally, this is using the Golden Section Method built into the R function "optim", using for optimising more than one parameter:



```
loglik = function(p, data) {
  n = nrow(alfheim)
  m = length(alfheim[alfheim$above1=="Yes", ]$y)A
  y2_mean = mean(alfheim[alfheim$above1=="Yes", ]$y)
  (n-m)*log(1-p[1]) + n*log(p[1]) + n*log(p[2]) + m*p[2]*(1-y2_mean)
}

optim(c(0.1, 100), loglik, control=list(fnscale=-1), data=alfheim)
```

Starting at the values: $\phi = 0.1 \; and \; \theta = 100$, this function results in maximising the Log-Likelihood function to find the best fitting parameters for the given data resulting in:

$$\hat{\phi} = 0.5559 \; (4sf) \; and \; \hat{\theta} = \; 0.3241 \; (4sf)$$

These are reasonable results in comparison to the MLEs derived by hand. The MLE for θ is accurate since it is similar to the digital result. However the MLE for ϕ must be biased as it did not give a sensible result considering that $0 < \phi < 1$ was an original constraint on the parameter and is not fulfilled by the MLE method but is fulfilled digitally.

Summary

I have investigated the shape and features of the data and its distribution overall, and used this to justify the given distributions. After finding a suitable expression for constant "c", I was able to derive MLEs for ϕ and θ distribution parameters, followed by an evaluation against digital results via numerical optimisation.

Section 2

Introduction

By investigation the results of the World Happiness Report data for 2023, I will fit and assess the simple linear model of GDP on Happiness and find any unusual outliers/ overly influential observations which may be affecting its reliability. I will make an out-of-sample prediction using my model and finally determine the significance of a stepwise selected optimal model compared to the simple linear model used.

Fitting a Simple Linear Model and Assessing its Fit

The scatter plot in $Figure\ 8$ demonstrates a simple linear regression model fitting GDP onto Happiness as a single explanatory variable. The blue line is the line of best fit and shows a grey 95% confidence band.

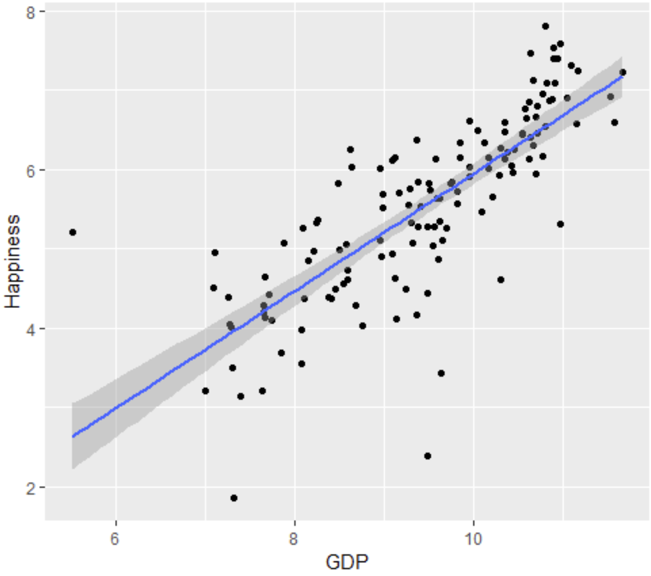


Figure 8: Scatter plot showing the simple linear model Happiness ~ GDP applied to the data

The line of best fit demonstrates the best estimates for the intercept and slope coefficients of the linear model which are:

$$\hat{\beta}_0 = -1.455\ (4sf)$$

$$\hat{\beta}_1 = 0.7403\ (4sf)$$

Given that $\hat{\beta}_1$in this model has $std.error = 0.0507\ (4sf)$ and is $t \approx 14.61\ (4sf)$ std. deviations away from 0, via t-distribution: $P(T \geq t) < 2 * 10^{-16}$ (ie. a very small number) we can say that this model produces a LOBF with a slope that is significantly different from 0. In other words there is a meaningful relationship between GDP and Happiness.

Simliarly, the goodness of fit can be evaluated: I will evaluate the goodness of fit under the following hypothesis at the 95% significance level:

$$H_0: R^2 = 0$$

$$H_1: R^2 > 0$$

The model shows the coefficient of determination to be $R^2 = 0.6144$ which indicates a reasonably strong positive association between GDP and Happiness. The f-test performed on this value returns:

$$pvalue < 2.2 * 10^{-16}$$

(ie. a very small number) so I can reject the null hypothesis even at the 95% significance level chosen as this result indicates that this R^2 value is unlikely to have occurred by chance and is sufficient evidence to suggest that GDP is a significant cause of variation in Happiness.

Additionally, we can compare the AIC for the simple linear model and the null model to evaluate whether it is worth using GDP as an explanatory variable. AIC for the null model ≈ 425 vs. the AIC when using GDP ≈ 298. This indicates that that the simple linear model involving GDP is better than the null model considering there are more parameters in the model which are penalised in AIC in order to reduce unnecessary complexities. In other words, using GDP as an explanatory variable is a necessary complexity to have a better fitting model for this data.

Using Standardised Residuals to find Outliers/ Overly Influential Observations

The assumptions about the true residuals of a linear model are as follows:

1) They are normally distributed
2) They have 0 expectation
3) They have constant variance
4) They are all independent from each other

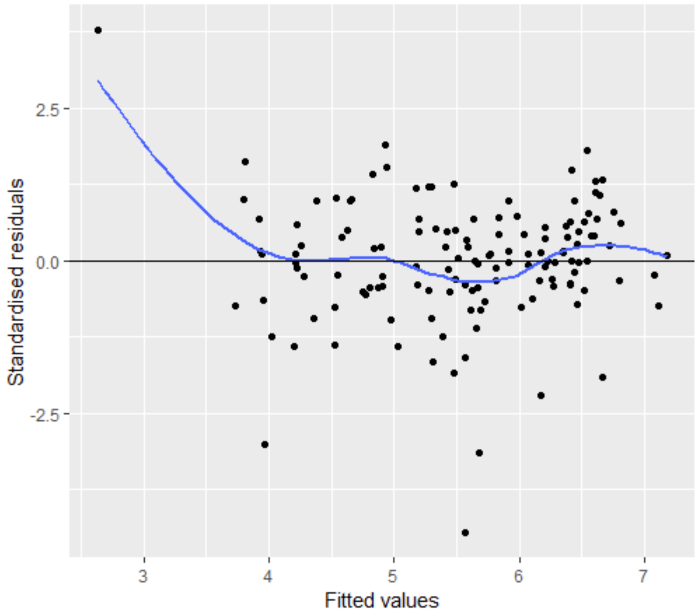


Figure 9: scatter plot demonstrating standardised residuals of the linear model vs. the fitted values

At the lower end, the residuals are higher than they should be as shown in $Figure\ 9$ - ie. the smoothing curve is far above the assumed mean of 0. As the happiness score increases we can see a tendency for the residuals to concentrate at 0 – this suggests some outlier(s) on the lower end which have led to an unusually high Happiness Score for their GDP (as expected by the linear model). Overall, this shows that the mean differs from 0 depending on the fitted values and is not constant. However the variation seems constant across Happiness scores - ie. doesn't seem to depend on the fitted values. It is more difficult to tell at the extremes due to less frequent data points, but in general there is little evidence to suggest a changing variance/ dependency on the fitted values.

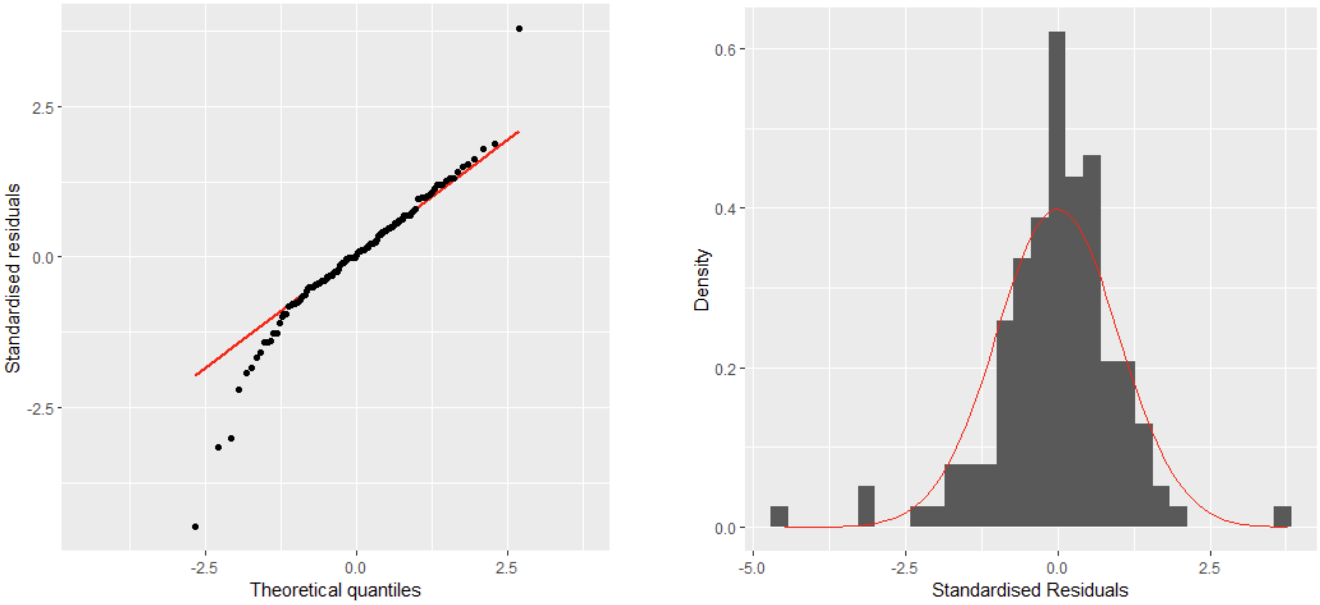


Figure 11+12: q-q and histogram plots to compare the std. residuals distribution to the expected $N(0,1)$ *shown in red*

To justify the suggestion of outliers we can look further into the distribution shape as shown in $Figure$ 11 above. The q-q plot indicates a broader distribution of the standardised residuals than expected such that it is heavier tailed than a standard normal distribution (more extreme values than expected).

Also the standardised residuals are slightly negatively skewed due to how the left-tail is heavier than the right-tail in $Figure$ 12 - ie. "the distribution leans to the right". This is due to several occurrences of negative residual values (fitted points that are far above the observed ie. low outliers in the collected data) and less but still more than expected number of positive residual values (fitted points that are far below the observed ie. high outliers in the collected data).

We can also evaluate overly influential observations which are negatively impacting the reliability of our model. This can be done by comparing each residual to its leverage (distance from the mean of the explanatory variable – indicating its capacity to affect the model) as shown in $Figure$ 13 below.

Datapoints with greater leverage have less freedom for variation that is causing a large residual and influencing the linear model coefficients. This is shown by the inner “red zone” between the red bounds in $Figure$ 13. These red bounds show the limit for a residual as leverage increases to not be considered an overly influential datapoint using Cooke’s Distance with $level = \frac{4}{n}$.

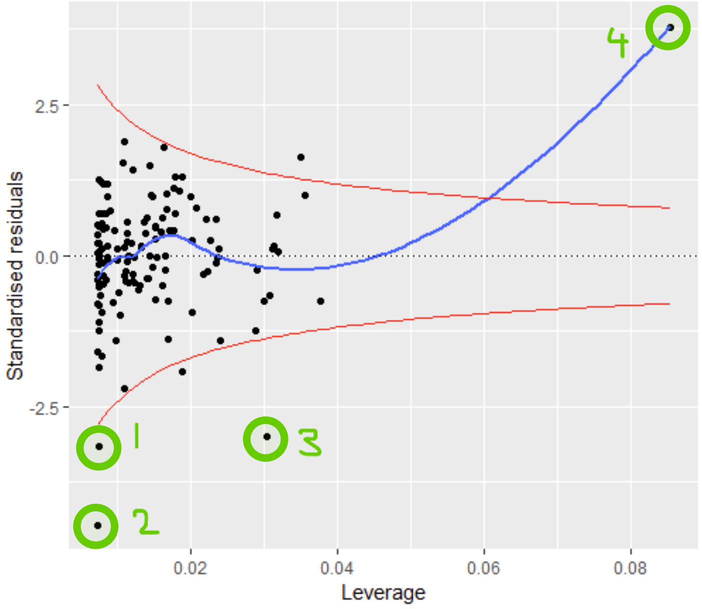


Figure 13: Residuals vs. Leverage plot showcasing the overly influential points based on Cooke's Distance, level $= \frac{4}{n}$

In this case, one standout data point with very high leverage and very high residual (fitted values are far below the observed) has too much influence and a few points not very far from the mean GDP actually have too much influence due to overwhelmingly low residual (fitted value far above the observed). These make a significant effect on the model due to low sample size: $n = 136$ countries in the full dataset.

The main overly influential points are the following countries as numbered in $Figure\ 13$:

1) Botswana
2) Lebanon
3) Afghanistan
4) Venezuela

These results could be attributed to outliers to the trend, that higher GDP leads to higher perception of Happiness. In particular country 1 – Botswana – is one of the most sparsely populated countries in the world, mostly home to tribal communities which probably effects the nature and reliability of the data. However, countries 2-4 have common features of political instability, corruption and war which would reduce the average Happiness Score and make it possibly subject to tampering etc.

Out-of-Sample Prediction

If the UK's GDP rises by 10% in the future then the model tells us:

$$Happiness\ Score = 7.261$$
$$with\ 95\%\ certainty\ in\ the\ interval: (5.828, 8.694)$$

Stepwise Optimal Model Selection and Evaluation

Using a stepwise optimisation method built into R, starting with the null model the optimal model returned is:

$$Happiness \sim SocialSupport + Corruption + Freedom + GDP + LifeExpectancy$$

This has $AIC \approx -192$ which is far better the the simple linear model that gave us $AIC \approx 297$ despite the added complexities.

Using an Analysis of Variation Table (ANOVA) method to test the significance of using the model over the simple linear model results in $f = 40.255$. The f-test performed on this value returns:

$$pvalue < 2.2 * 10^{-16}$$

(ie. a very small number) indicating that the explanation of the variation in the data using a linear model with 5 explanatory variables (stepwise result) is significantly different to the simple linear model. In other words, it is making a meaning difference to the model to add these additional explanatory variables despite the complexities introduced.

Summary

Having fit the simple linear model Happiness ~ GDP, I was able to measure the goodness of fit of this model, determine outliers and find the overly influential points using by evaluating the standardised residuals distribution. After making an out-of-sample prediction for future UK GDP data with a 95% prediction interval, I evaluated the model used to a more complex 5 variable model determined to be a significantly better fit found using numerical optimisation.