Summative Coursework 1

MTH2006 Statistical Modelling and Inference

College of Engineering, Mathematics and Physical Sciences

March 2024

1 Alfheim Rainfall Totals

The sample was obtained from Daily rainfall totals being measured in Alfheim over 10 years (in mm).

1.1 Introduction

contains the answers to Part (a)

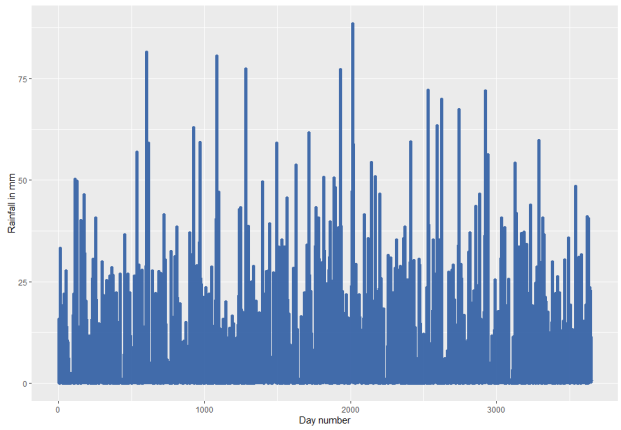

Figure 1: Raw Rainfall data

The sample had a minimum of 0.00016mm of rainfall and a maximum of 88.56mm, mean value 3.69 and first and third quartiles 0.306 and 0.939 (and hence the inter-quartile range is 0.633). Figure 1 shows each of the daily rainfall totals compared over the 3650 days, and the maximum can clearly be seen around day number 2000.

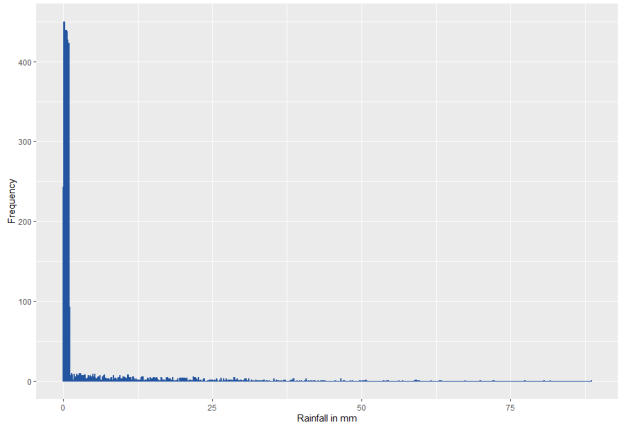

Figure 2: Rainfall Histogram

Figure 2 is a Histogram which shows the frequency of our data, which is very large close to zero, on the more dry days, but drops off rapidly as the days become what we would classify as 'wet'. for these wet days past a certain rainfall value, the frequencies taper down exponentially. The different shapes of each of the 'wet' and 'dry' ranges of the graph hint at using suitable models for each.

1.2 Proposing a Distribution

contains the answers to Part (b)

By splitting the data and the distribution between dry days, defined $<$ 1mm Rainfall, and wet days $\geq$ 1mm, it is proposed that on dry days, the rainfall total is uniformly distributed, with the probability of having such days is $1-\phi$ with $0 \leq \phi \leq 1$, and on wet days it is exponentially distributed with a probability distribution function (PDF) $f(x) = ce^{-\theta x}$, as shown in Fig. 3.

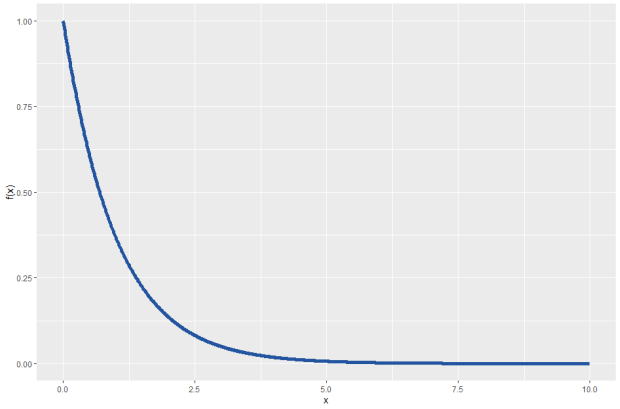

Figure 3: Exponential PDF
$(c = 1, \theta = 1)$

The value of c can then be determined by the integration of the PDF between 1 and infinity, which will be equal to the total probability over this range, ϕ:

$$\int_1^\infty ce^{-\theta x} dx = \phi$$

This integral can then be calculated as $[-\theta^{-1} ce^{-\theta x}]_1^\infty$, and provided our parameter

$\theta > 0$ (a requirement of the exponential distribution), $e^{-\infty} \to 0$ implies $-\theta^{-1}ce^{-\theta} = \phi$, which can finally be rearranged to give:

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

1.3 Likelihood of a small sample

contains the answers to Part (c)

The small sample of rainfall totals given was: $\vec{Y} = 30, 0.2, 10, 0$. The Likelihood function, denoted $L((\phi,\theta)^T;\vec{Y})$ is the Joint PDF of this sample, and **assuming independence**, will equal to the product of the PDF of each $Y_i \in \vec{Y}$.

$$L((\phi,\theta)^T;\vec{Y}) = \prod_{i=1}^{4} f_{Y_i}(Y_i)$$

And from the same logic through which c was calculated, we can find the PDF of our uniform distribution, $f_u(x) = 1 - \phi$ (the exponential PDF will now be denoted as $f_e(x)$ to avoid confusion). Hence, the Likelihood can be easily calculated:

$$\begin{aligned}\prod_{i=1}^{4} f_{Y_i}(Y_i) &= f_e(30)f_u(0.2)f_e(10)f_u(0.2)\\ &= (1-\phi)^2 ce^{-\theta(30)}ce^{-\theta(10)}\end{aligned}$$

Finally, by substituting c from 1.2 and therefore **assuming $(\theta > 0)$)**, then collecting like terms gives likelihood function for our small sample $\vec{Y}$ as:

$$L((\phi,\theta)^T;\vec{Y}) = \phi^2\theta^2(1-\theta)^2e^{-42\theta}$$

1.4 Estimating parameters

contains the answers to Part (d)

To estimate ϕ and θ, The days are partitioned into Wet and Dry as previous, and the partial derivative of the logarithm of our likelihood function with respect to our parameter equated to zero will give us the maximum likelihood estimators $\hat{\phi}$ and $\hat{\theta}$. So the first step to estimate these parameters is to calculate our likelihood function, which depends on ϕ, θ and (where D and W are the sets of wet and dry data values, m the number of wet days and n the total number of days in the sample):

$$\begin{aligned}L((\phi,\theta)^T;Y) &= \prod_{Y_i\in D} f_u(Y_i)\prod_{Y_i\in W} f_e(Y_i)\\ &= (1-\theta)^{n-m}(\phi\theta e^{\theta})^m \prod_{Y_i\in W} e^{-\theta Y_i}\end{aligned}$$

and since $a^{b_1}\times a^{b_2}\times ... \times a^{b_n} = a^{b_1+...+b_n}$, it follows that the likelihood function is:

$$L((\phi,\theta)^T;Y) = (1-\theta)^{n-m}\phi^m\theta^m e^{\theta(m-\sum_{Y_i\in W} Y_i)}$$

where $W := \{Y_i | Y_i > 1\}$ $D := \{Y_i \mid Y_i \le 1\}$,

and $m := \mid W \mid$, $n := \mid Y \mid$

Next the Log-Likelihood is calculated by taking the natural logarithm of the likelihood function (*with our likelihood function now shortened as* $L_Y := L((\phi,\theta)^T;Y)$):

$$\begin{aligned}\ell = \ln L_Y = \ln(1-\theta)^{n-m} + \ln\phi^m +\\ \ln\theta^m + \ln e^{\theta(m-\sum_{Y_i\in W} Y_i)}\end{aligned}$$

Not quite finished:

$$\ln e^{\theta(m-\sum_{Y_i\in W} Y_i)} = \ln e^{m\theta(1-\frac{1}{m}\sum_{Y_i\in W} Y_i)}$$

and since $\frac{1}{m}\sum_{Y_i\in W} Y_i = \overline{W}$ the method of moments estimator can be used to make this in terms of m, ϕ, θ:

$$\begin{aligned}\overline{W} \approx E(W) &= \int_1^{\infty} x f_e(x)dx = \int_1^{\infty} x\phi\theta e^{\theta(1-x)}dx\\ &= \frac{\phi(\theta+1)}{\theta}\end{aligned}$$

and substituting into our log-likelihood function to simplify:

$$\begin{aligned}\ell = \ln(1-\theta)^{n-m} + \ln\phi^m +\\ \ln\theta^m + m\theta - m\phi\theta + m\end{aligned}$$

we can then partially derive this and equate to zero to find our **Maximum likelihood estimates, $\hat{\phi}$ and $\hat{\theta}$**

$$\frac{d\ell}{d\phi} = 0 \text{ and } \frac{d\ell}{d\theta} = 0$$

(at $\phi = \hat{\phi}$ *and* $\theta = \hat{\theta}$*)*

Calculating these partial derivatives is straightforward and gives

$$\frac{d\ell}{d\phi} = \frac{m}{\phi} + m\theta \tag{1}$$

$$\frac{d\ell}{d\theta} = \frac{-(m-n)}{1-\theta} + \frac{m}{\theta} + m + m\phi \tag{2}$$

which gives the quadratic equation for ϕ and the relationship to calculate θ:

$$\hat{\phi}^2 + \hat{\phi} + \frac{n}{m} - 1 = 0 \text{ and } \hat{\theta} = -\hat{\phi}$$

Using our data values to estimate these parameters, with $n = 3650$, $m = 735$, gives complex results, so cannot be correct, and despite retrying the calcutions several times this was the closest attempt.

1.5 Numerical Optimization

contains the answers to Part (e)

To numerically optimimize out parameters we can define our log likelihood function in R as: *(In order to test, put* `+ (theta*(1-Yi)) #FormatErr*` *on the previous line)*

```
log_likelihood <- function(params) {
  phi <- params[0]
  theta <- params[1]
  l <- 1
  dataset = alfheim
  for (i in nrow(dataset$y)) {
    print(i)
    Yi = dataset$y[i]
    if (Yi < 1) {
      l = l + log(1-phi)
  } else if (Yi >= 1){
      l = l + log(phi) + log(theta)
      + (theta*(1-Yi)) #FormatErr*
  }
  }
  -log(l) #optim maximizes functions
  #so this will minimize our log(L)
}
```

However, when `(optim(par = c(1,-1), log_likelihood))` was used this did not adjust our initial parameter values despite lots of research, debugging and effort was not resolvable.

2 World Happiness

The World Happiness Report ranked 136 countries average happiness by responses to a survey from a representative sample of the population. This dataset can be explored to look at what variables could potentially explain happiness and why.

2.1 Linear Regression

contains the answers to Question 2 Part (a)

Using R to fit a Linear Regression model to find how Happiness relates to the natural logarithm of the Gross Domestic Product (GDP) based on the dataset gives the model:

$$H = \beta_0 \ln G + \beta_1 + \epsilon$$

Where β_0 and β_1 are constant coefficients, and ϵ is the 'error', the difference between a point and the. R estimates these coefficients to be

$$\beta_0 = 6.49 \text{ and } \beta_1 = -8.97$$

Figure 4 shows this linear regression function as the straight line of best fit in blue, with a 95% confidence band in grey around it, and a scatter plot of our data.

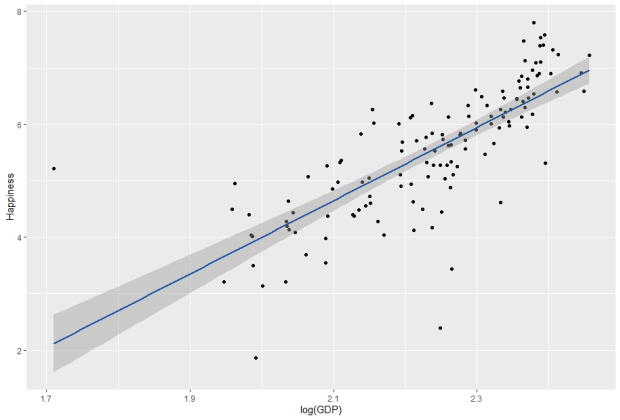

Figure 4: Linear Regression Model

2.2 Goodness of Fit

contains the answers to Part (b)

In order to quantify goodness of fit wee can do a hypothesis test with the null and alternative hypotheses:

$$H_0 : \beta_0 = 6.49, \beta_1 = -8.97$$
$$H_0 : \beta_0 \neq 6.49, \beta_1 \neq -8.97$$

R automatically works out our P-value at $< 2.2e^{-16}$, which is in the significance level (also given by R) 0.01 so for any significance level ≥ 0.01 we **Accept H_0 and reject H_1**.

2.3 Residual Analysis

contains the answers to Part (c)

In order to check the linear model, it is noted that the confidence intervals and test results for the mode rely on some key assumptions (**in bold**) about the model. We can test these assumptions by examining the residuals.

The following figure shows the residual error (in blue) between our dataset values (the natural logarithm of the GDP) and the fitted linear regression, where Each individual line represents the residual error for the point on it's end. The black line shows the line of best fit, and its clear that aside from the obvious outliers (with visibly much longer residuals) , our data follows this linear pattern and it is safe to proceed with this assumption that **there is linearity between the response (Happiness) and the explanatory (natural logarithm of GDP) variables**.

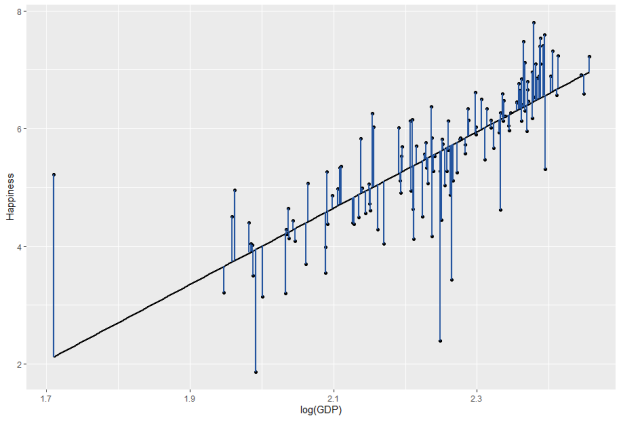

Figure 5: Residual Plot

Figure 6 shows the standardized residuals vs the dataset values, which standardizes our residuals and plots them against the fitted values. As is shown there is no particular correlation or pattern so it is safe to assume that **constant variance of residuals**

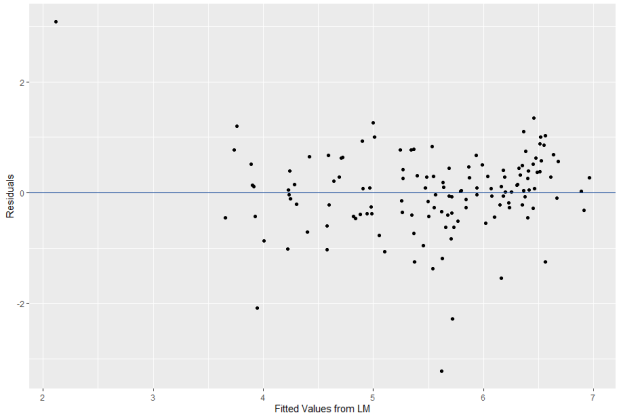

Figure 6: Standardized Residuals

Figure 7, a Q-Q plot, which compares quantiles from a standard normal distribution to our residules, and in this case shows a mostly strong linearity apart from around the first quarter of values. If this is not random then the model may not be very effective, but if it is simply a 'lucky' occurance that the sample has data causing this dip, then **normality of errors can be assumed**.

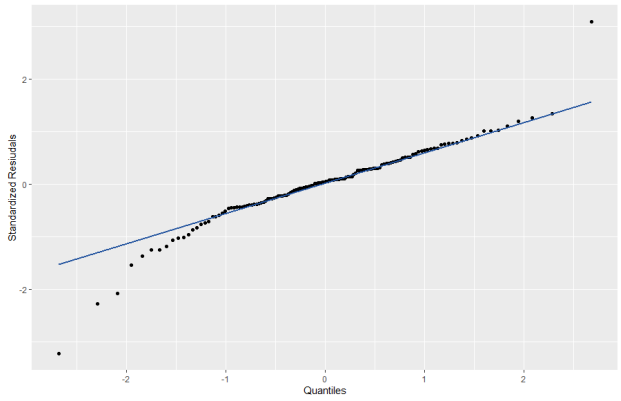

Figure 7: Q-Q plot

2.4 UK Prediction

contains the answers to Part (d)

As an example to utilize the model, assuming that the UK GDP will increase by 10%, the 95% prediction interval can be calculated by R, and in this case gave an interval of [5.54, 8.51] , centered on the new fitted value 7.02 for thee estimated happiness based on a 10% increase in GDP in our linear model.

2.5 Model Comparison

contains the answers to Part (e)

Using stepwise selection, the best subset of coefficients for the natural logarithm of each variable to explain happiness can be calculated in R, Which due to timing issues was unable to be completed, however it is predictable that The more complex model, with coefficents for each of the variables, will be more accurate. This is always true, since it involves more data and can account for these relationships with other variables, and even if it was found that no other variable affected happiness (all coefficients $\neq$ GDP are $= 0$), then our complex model would be the same as our simple linear model:

$$\text{Happiness} \approx 6.49 \ln(\text{GDP}) - 8.97$$