# MTH2006: Statistical Modelling and Inference

## Coursework

## Ex1 a)

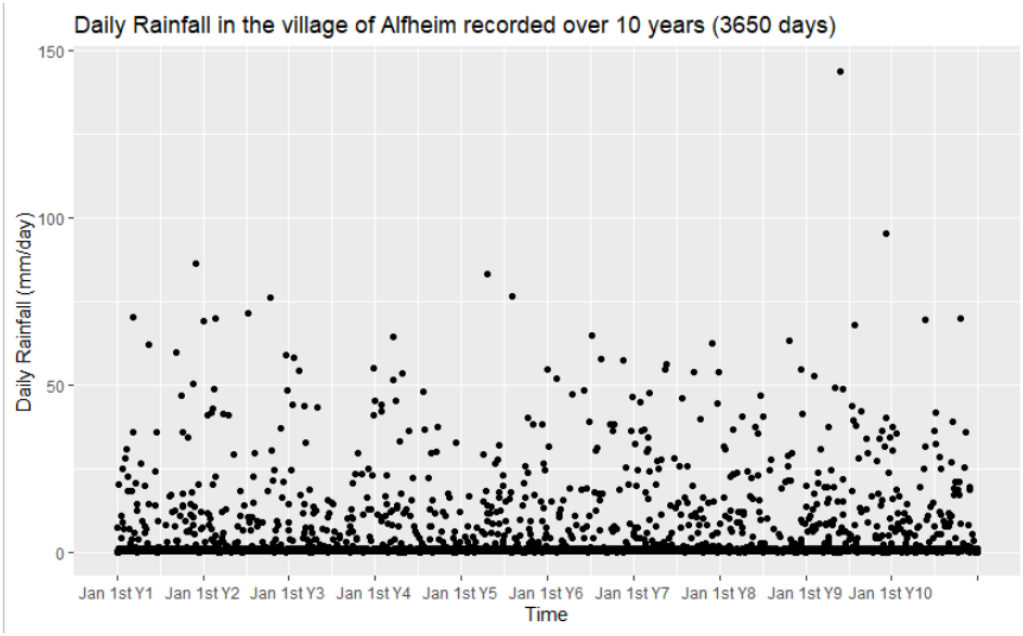



*Figure 1: Daily Rainfall in the village of Alfheim*

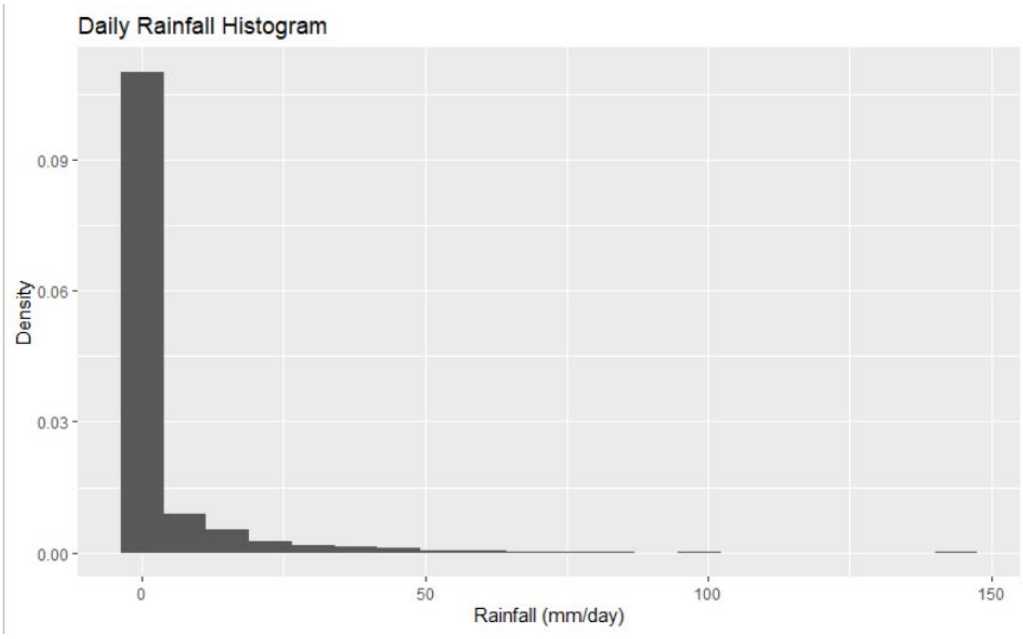



*Figure 2: Daily Rainfall Histogram*

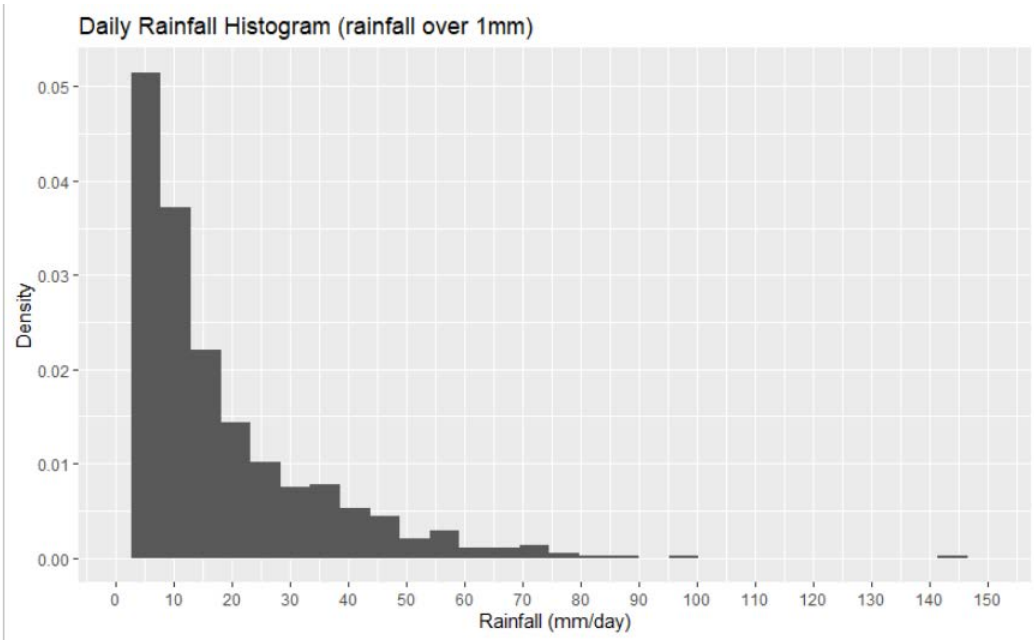



*Figure 3: Daily Rainfall Histogram (rainfall > 1mm/day)*

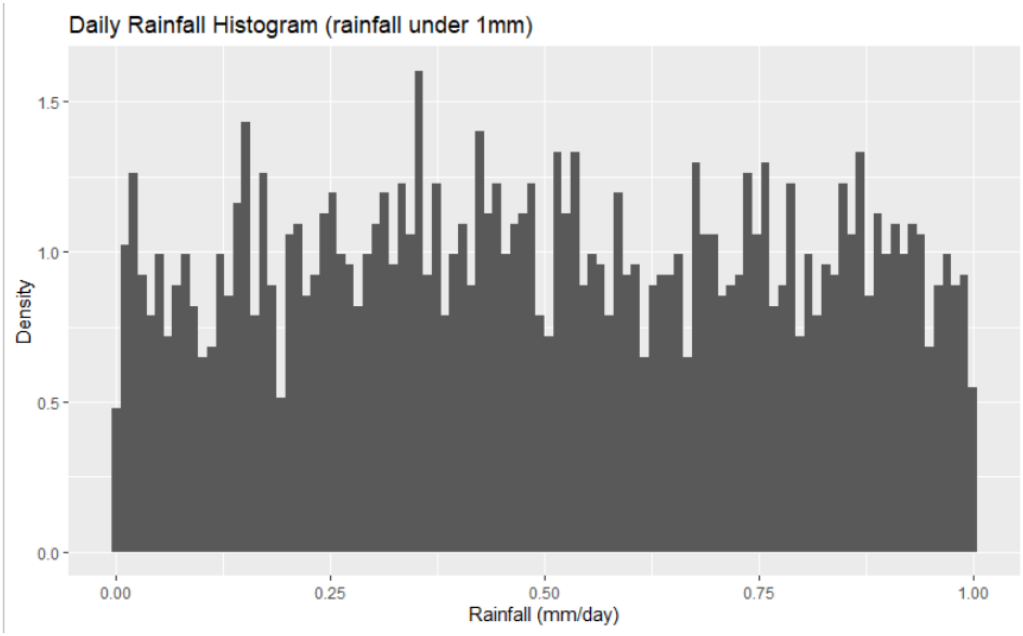



*Figure 4: Daily Rainfall Histogram (rainfall <= 1mm/day)*

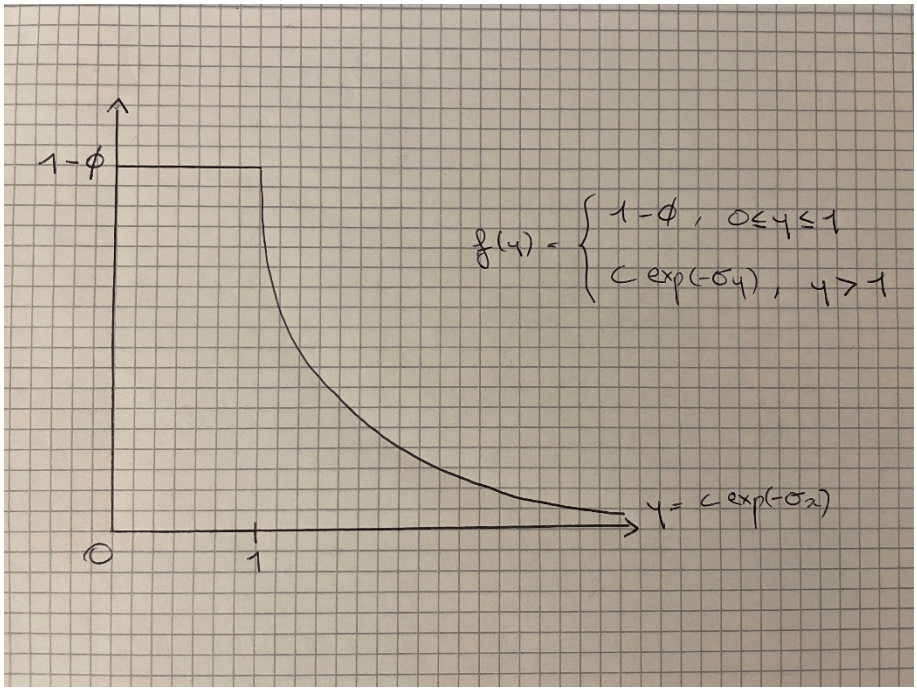
In figure 1, we have the daily recorded rainfall in Alfheim over a period of 3650 days (10 years). There does not appear to be much monthly variation – some years it rains less during the summer months, others it rains more. We can however see a large accumulation of points close to 0 whereas values above 1 seem to become rarer.

In figure 2, we notice a large density seems to be close to 0, then drops significantly as the amount of rain (mm/day) becomes larger. This could suggest we should look more in detail at values less than 1 and values above 1. Figure 3 seems to suggest that daily rainfall values above 1mm/day

follow an exponential distribution and figure 4 seems to suggest that daily rainfall values under 1mm/day follow a uniform distribution.

## Ex1 b)

Sketch of the p.d.f of $f(y), y > 0$




To find c :

$$\int_{-\infty}^{+\infty} f(y)dy = 1 = \int_{0}^{1} (1 - \Phi)\, dy + \int_{0}^{\infty} c * e^{-\theta * y} dy$$

$$1 * (1 - \Phi) - 0 * (1 - \Phi) + \lim_{y \to \infty} -\frac{e^{-\theta * y}}{\theta} + \frac{c}{\theta} = 1$$

$$c = \theta * \Phi$$

And so, $f(y) = \theta * \Phi * e^{-\theta y}, y > 1$

## Ex 1 c)

$$L(\theta, \Phi, \mathrm{y}) = \prod_{i=1}^{n} (1 - \Phi) * \prod_{i=1}^{m} (\theta * \Phi) * e^{-\theta y}$$

Where n is the number of values <=1 and m is the number of values >1

Here we have values : $\{0, 10, 20, 0.1, 0.9\}$, so n=3 and m=2

$$L(\theta, \Phi, \mathrm{y}) = (1 - \Phi)^3 * (\theta * \Phi)^2 * e^{-\theta * 30}$$

Ex 1 d)

Consider m=$\sum_{yi>1} 1$ and n=$\sum_{yi\leq 1} 1$, and $\bar{y} = \frac{1}{m} * \sum_{i=1}^{m} yi\,, yi > 1$

We know that

$L(\theta, \Phi, \mathrm{y}) = \prod_{i=1}^{n}(1-\Phi) * \prod_{i=1}^{m}(\theta * \Phi) * e^{-\theta yi} = (1-\Phi)^{n} * (\theta * \Phi)^{\mathrm{m}} * e^{-\theta * m * \bar{y}}$

And so, the log-likelihood is :

$\mathrm{l}(\theta, \Phi, \mathrm{y}) = \mathrm{n} * \log(1-\Phi) + \mathrm{m} * \log(\theta * \Phi) - \theta * \mathrm{m} * \bar{y}$

The partial derivatives give us :

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

$$\frac{\partial l}{\partial \theta} = \frac{m}{\theta} - m * \bar{y}$$

Setting these derivatives to 0 and solving for $\Phi$ and $\theta$, yields:

$$\hat{\Phi} = \frac{m}{m+n} = \frac{m}{3650}$$

$$\hat{\theta} = \frac{1}{\bar{y}}$$

Using data from Alfheim, we can estimate $\hat{\Phi} = 0.2049315$ and $\hat{\theta} = 8.285548e-05$

The 2nd derivatives of the log-likelihood function are :

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

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{m}{\theta^2}$$

$$\frac{\partial^2 l}{\partial \Phi \partial \theta} = 0$$

$$J(\Phi, \theta) = \begin{bmatrix} \frac{m}{\theta^2} & 0 \\ 0 & \frac{n}{(1-\Phi)^2} + \frac{m}{\Phi^2} \end{bmatrix}$$

And so Expected Information is :

$$I(\Phi, \theta) = \begin{bmatrix} \frac{m}{\theta^2} & 0 \\ 0 & \frac{n}{(1-\Phi)^2} + \frac{m}{\Phi^2} \end{bmatrix}$$

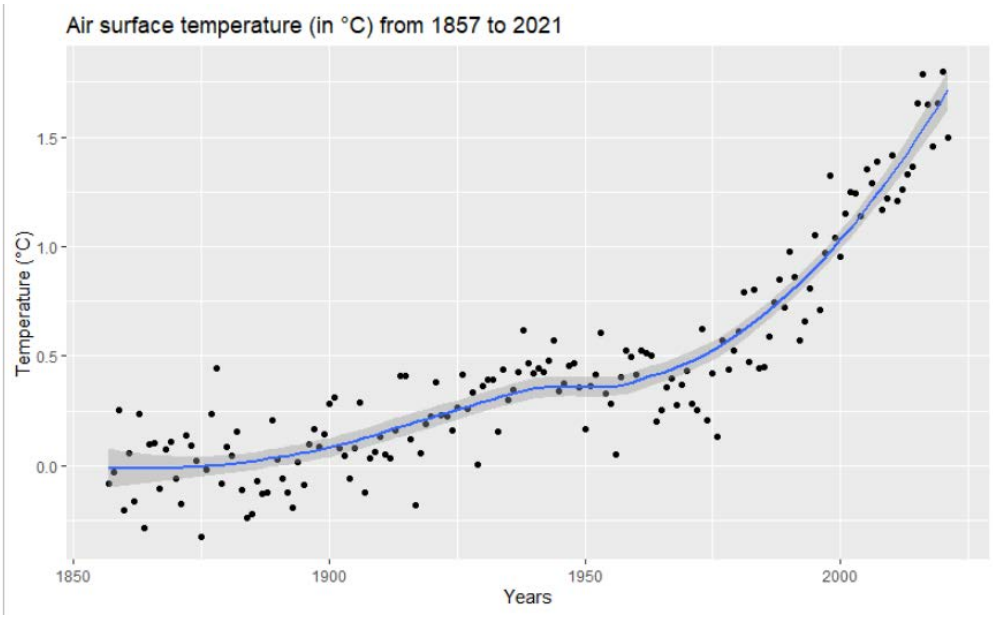
Ex 1 e)

By numerical optimization, we obtain :

$\widehat{\Phi} = 2.048813e-01$ and $\hat{\theta} = 8.285726e-05$

```
49
50
51  y1 <- alfheim %>% select(y) %>% filter(y>1)
52  y2 <- alfheim %>% select(y) %>% filter(y<=1)
53
54  loglik <- function(p, y1, y2) {
55    m <- length(y1[,1])
56    n <- length(y2[,1])
57    if(p[1] <= 0) return(-1e20)
58    if(p[2] <= 0) return(-1e20)
59    n*log(1-p[2]) + m*log(p[1]*p[2]) - p[1]*m*sum(y1)
60  }
61  optim(c(0.00009, 0.3),control = list(fnscale = -1), loglik, y1 = y1, y2 = y2)
62
63
```

Ex 2 a)




There appears to be little relationship between the years and the air surface temperature from 1857 to 1899. From 1900 to 1975, the air temperature increases almost linearly and then from 1975 to 2021, the air temperature seems to increase exponentially.

Ex 2 b)

From our linear model fit, we obtain a rate of change of temperature per year $\hat{\beta}_1 = 0.008485$ with intercept $\hat{\beta}_0 = -16.027835$, and a 95% confidence interval of (0.007693459, 0.009276968)

We test the null hypothesis $H_0 : \beta_1 = 0$ at the 1% significance level.

The test statistic :

$$\frac{\hat{\beta}_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} = \frac{0.008485}{0.000401} = 21.16 \gg t_{163,0.995} = 2.606328$$

The test statistic exceeds the critical value, so we reject the null hypothesis $H_0 : \beta_1 = 0$ at the 1% level. We obtain our p-value : $p = 6.40152e - 49$

Ex 2 c)

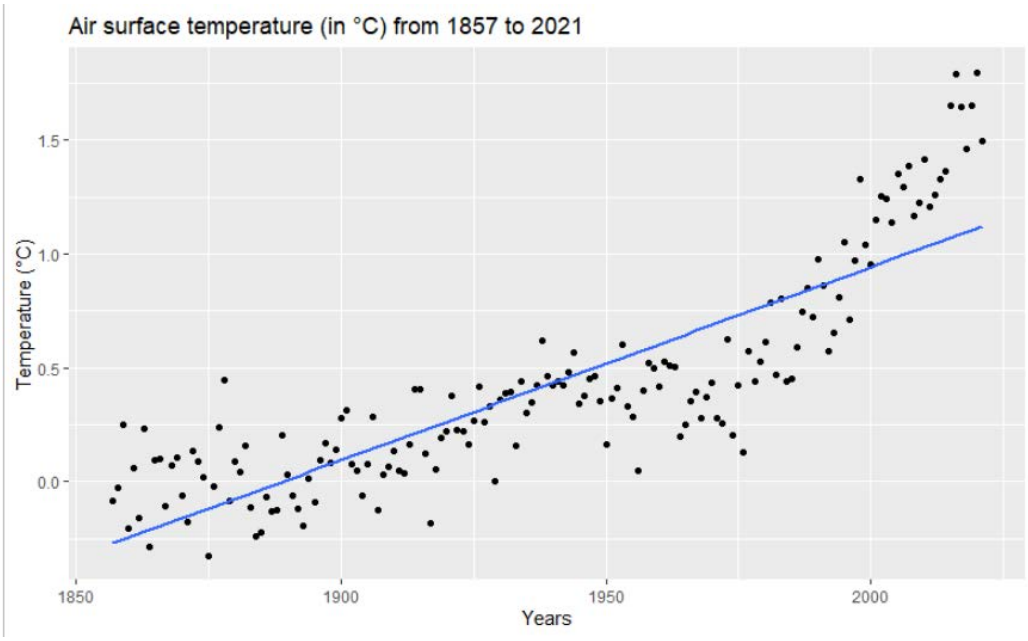



*Figure 5*

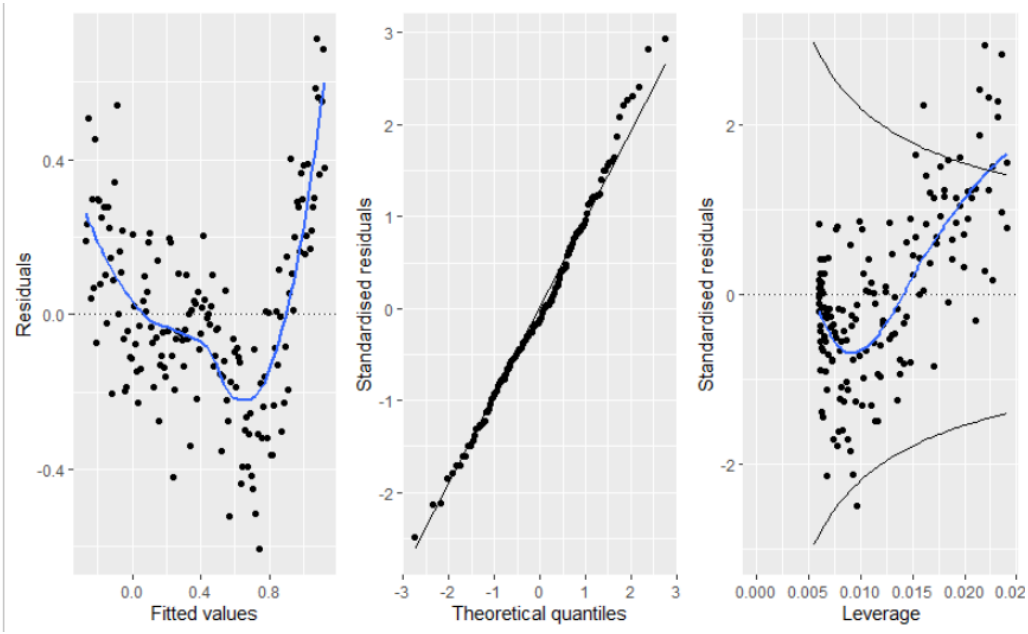


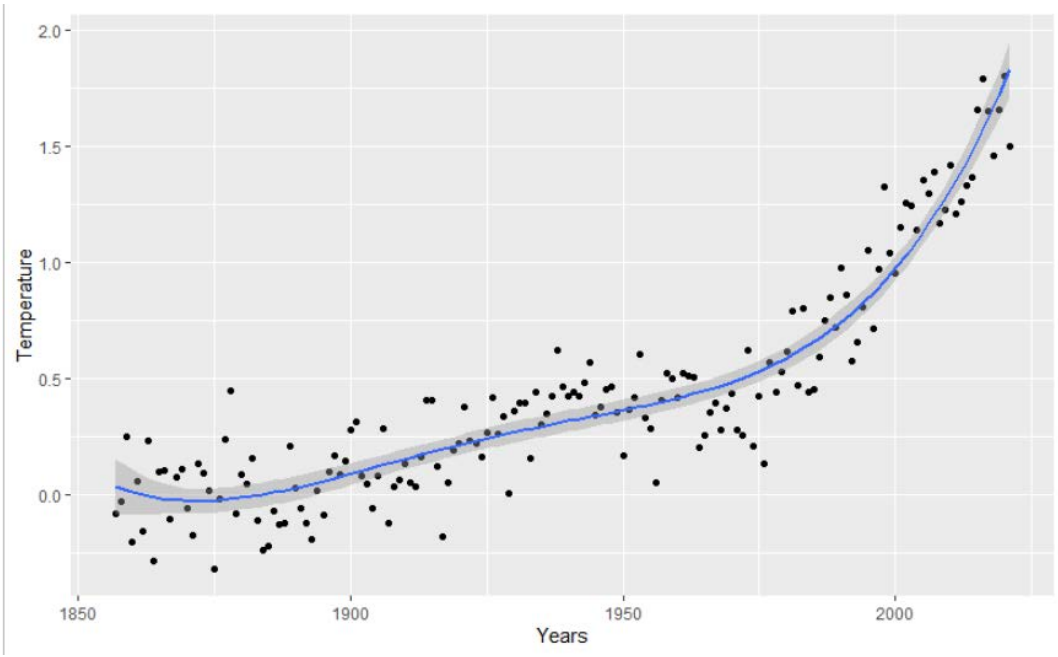

*Figure 6*

From figure 5, we see that there are many points below the line with none above, between 1950 and 2000, and then a large amount above with none under between, 2000 and 2021. We see some strong evidence of funnelling in our residual's vs fitted values plot (Figure 6), which could indicate non-constant variance. The distribution seems to depend on our fitted values and the assumption that the sample mean of residuals is 0 is not valid. Our quantile-quantile plot shows some that most of our data points are along the line, but we do see some positive skew at the top. We also see high leverage/Cook's distance for quite a few points. This suggests our model is not well specified.

## Ex 2 d)

Using stepwise regression starting with the full model, we find the best model is :

$Y_i \sim N(\beta_0 + \beta_1 * x + \beta_2 * x^2 + \beta_3 * x^3 + \beta_4 * x^4, \sigma^2)$ i.e. k=4




$$F = \frac{(RSS_{M1} - RSS_{M2}) * (p_2 - p_1)}{\frac{RSS_{M2}}{(n - p_2 - 1)}} = \frac{(9.8097 - 4.102) * (4 - 1)}{\frac{4.102}{(165 - 4 - 1)}} = 667.92 \gg F_{1,4;0.95} = 7.7$$

So, our best model is significantly better at 5% significance.

## Ex 2 e)

Using our model, we can predict the observed temperature in 2040 will be ~3.25◦C, with 95% C.I $[2.8542134167, 3.65268649]$

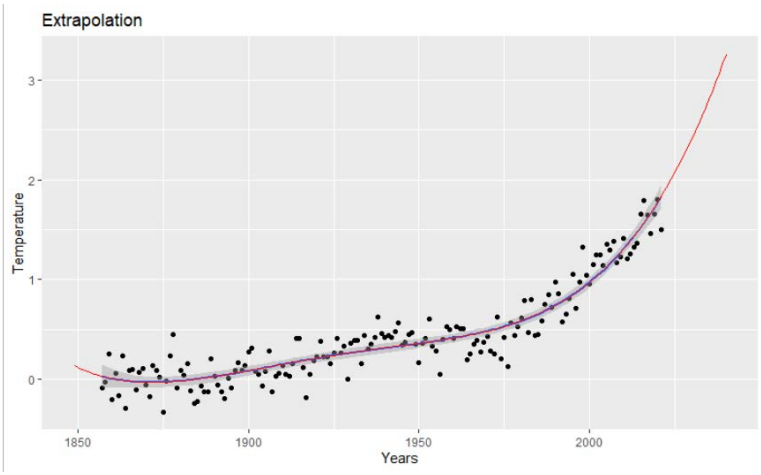



*Figure 7*