MTH2006 - Coursework 1

1 Rainfall in Alfheim

1.1 Introduction

Here we are looking at rainfall data from Alfheim that has been recorded daily of a 10-year period. In Figure 1, we can see box plots of the data by year. The upper plot shows the entire data set, which clearly makes the plots hard to interpret. However, it does show that there is significant number of outliers in every year. The bottom plot shows a significantly tighter set of values, with an upper limit of 2.5mm. The most noticeable part of this plots is that there is little change year to year of the distribution of the data, even if our previous plot has shown that the highest values can vary year to year. This is evident in the consistency of the medians which only small variation. It also shows that across the entire time period, at least 75% of the data comes in at under 1mm, suggesting the data is heavily skewed towards these lower values.

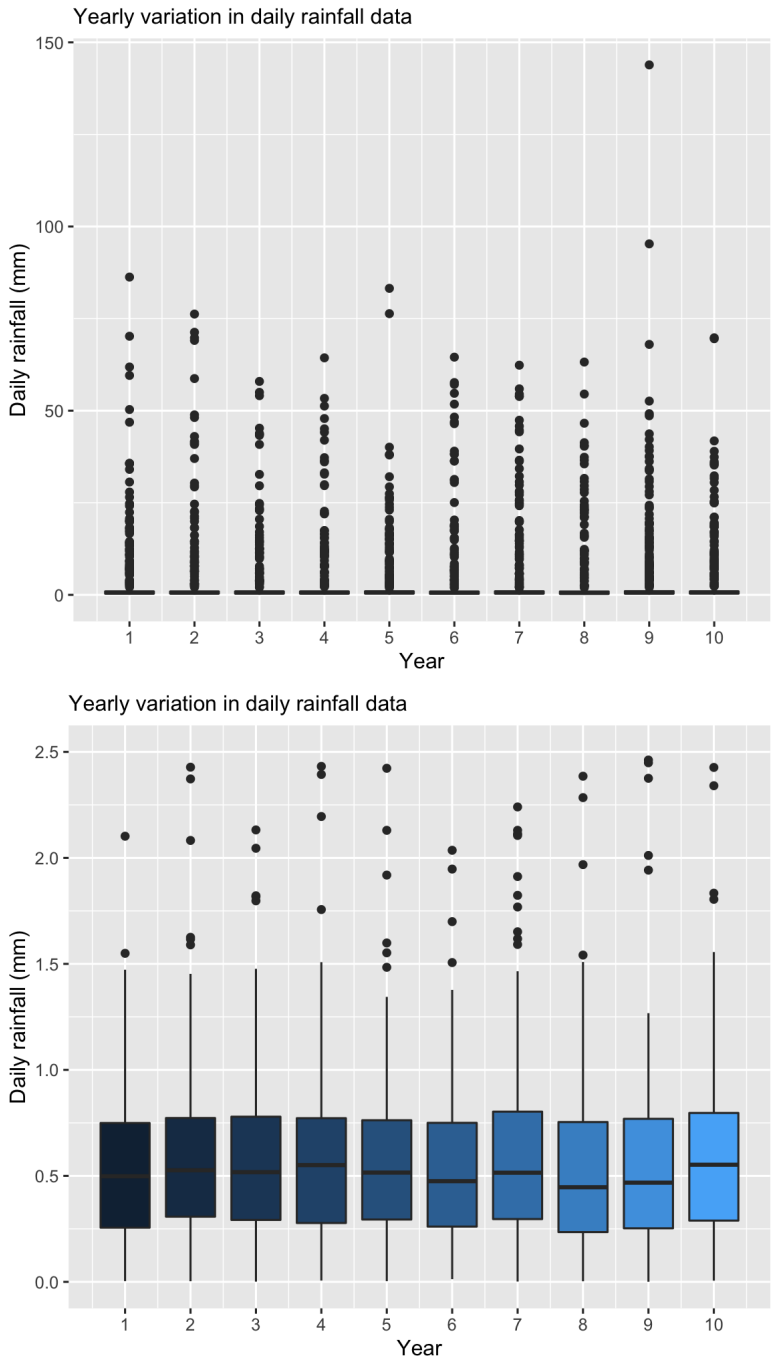

Figure 1: Yearly data of daily rainfall totals

1.2 Model fitting

We suppose that rainfall totals can be distributed uniformly for rainfall totals less than $1mm$ and the probability of having such days is $1-\phi$ with $0 \geq \phi \leq 1$. On wetter days, with rainfall total greater than 1mm, we assume that this is exponentially distributed, i.e. the p.d.f is $f(y) = c\exp -\theta y$. Hence, overall we have the p.d.f

$$f(y) = \begin{cases} 1-\phi & 0 \leq \phi \leq 1 \quad y \leq 1 \\ c\exp(-\theta y) & y > 1 \end{cases}.$$

.

We can calculate c using the property that p.d.fs integrate to 1 over their domain. Hence,

$$\begin{aligned} \int_{y\in D} f(y)\,dy = 1 = \int_0^1 1-\phi\,dy + \int_1^\infty c\exp -\theta y \\ = [y(1-\phi)]_0^1 + [-c\exp -\theta y \cdot \frac{1}{\theta}]_1^\infty \\ = 1-\phi+\exp -\theta\frac{1}{\theta} = 1 \\ \implies c = \exp(\theta)\phi\theta \end{aligned}$$

So we have,

$$f(y) = \begin{cases} 1-\phi & 0 \leq \phi \leq 1 \quad y \leq 1 \\ \phi\theta\cdot\exp(\theta(1-y)) & y > 1 \end{cases}.$$

. Figure 2 is a sketch of the p.d.f of $f(y)$.

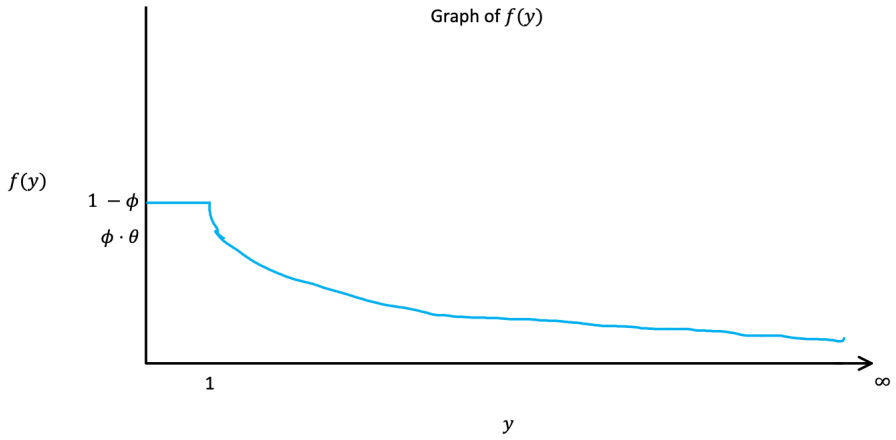

Figure 2

1.3 Likelihood Estimates

We can now use our data in conjunction with our model to form estimates for our parameters.

1.3.1 A small sample likelihood function

We let $\mathbf{y} = \{y_1, ..., y_5\}$ represent the sample {0, 10, 20, 0.1, 0.9}. Each point, y_i, has an equivalent random variable Y_i in the vector $\mathbf{Y}$. As per our model in section 1.2, values less than 1 will be distributed uniformly and values greater than 1 will be distributed exponentially. Therefore, we can form the likelihood function

$$\begin{aligned} L(\mathbf{y}; \phi, \theta) &= \prod_{i=1}^{5} f_{Y_i}(y_i; \phi, \theta) \\ &= (1-\phi)^3 \cdot \exp\theta(1-y_2) \cdot \phi\theta \cdot \exp\theta(1-y_3) \cdot \phi\theta \\ &= (1-\phi)^3 \cdot (\phi\theta)^2 \cdot \exp(-28\theta) \end{aligned}$$

for this small sample.

1.3.2 General Likelihood function

Taking inspiration from our small sample example, we can form the likelihood function for our model. Let y_i be an individual value in the data set and the data be the vector of such values, y. As before, each point, y_i, has equivalent random variable Y_i in Y. We let n be the total number of values in our data (this is a known quantity but we use n for brevity). We also let $m = \sum_{y_i>1} 1$, i.e. the total number of days were rainfall is greater than 1mm. Hence, we have the likelihood function

$$\begin{aligned} L(y; \phi, \theta) &= \prod_{i=1}^{n} f_{Y_i}(y_i; \phi, \theta) \\ &= (1-\phi)^{n-m} \cdot (\phi\theta)^m \cdot \exp\left(\theta(m - \sum_{y_i>1} y_i)\right). \end{aligned}$$

By taking the natural log of the likelihood, we get the log-likelihood which can be easier to work with.

$$l(y; \phi, \theta) = (n-m)\log(1-\phi) + (m)\log(\phi) + (m)\log(\theta) + \theta\left(m - \sum_{y_i>1} y_i\right)$$

We can then estimate the parameters as the partial derivatives w.r.t to each parameter will be equal to 0.

$$\begin{aligned} \frac{\partial l}{\partial \phi} &= \frac{m-n}{1-\phi} + \frac{m}{\phi} = 0 \\ &\implies \hat{\phi} = \frac{m}{n} \\ \frac{\partial l}{\partial \theta} &= \frac{m}{\theta} + m - \sum_{y_i>1} y_i \\ &\implies \hat{\theta} = \frac{m}{\sum_{y_i>1} y_i - m} \end{aligned}$$

To prove that the the estimations are maximised we can examine the second partial derivatives

$$\frac{\partial^2 l}{\partial \phi^2} = \frac{m-n}{(1-\phi)^2} - \frac{m}{\phi^2} < 0$$
$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{m}{\theta^2} < 0.$$

Hence, we see that these are maximums. We can then calculate the expected information which is simply

$$E\left[-\frac{\partial^2 l}{\partial\theta\,\partial\phi}\right] = E\left[-\frac{\partial}{\partial\phi}\left(\frac{m}{\theta} + m - \sum_{y_i>1} y_i\right)\right] = 0,$$
$$E\left[-\frac{\partial^2 l}{\partial\phi\,\partial\theta}\right] = E\left[-\frac{\partial}{\partial\theta}\left(\frac{m-n}{1-\phi} + \frac{m}{\phi}\right)\right] = 0.$$

Therefore, we expect there to be no correlation between the estimates for ϕ and θ and for them to be independent of each other.

1.3.3 Optimisation

We can also use R to optimise our likelihood function to find the maximum likelihood value. Below is the code used to this.

```
# --- Question 1(e) ---


# p[1] = phi, p[2] = theta
# Creating log-likelihood function
loglik <- function(p){
  if(p[1] <= 0) {
    return(-1e20)
  }
  if(p[2] <= 0) {
    return(-1e20)
  }

  # Setting values from data
  n = nrow(alfheim)
  m = nrow(alfheim_wet) # alfheim_wet - all vals s.t y > 1
  sum_y = sum(alfheim_wet[,2])

  # Calculating log-likelihood
  (n-m) * log(1-p[1]) + m * ( log(p[2]) + log(p[1]) ) - p[2] * (sum_y - m)
}

# Optimising for parameters
phi_opt <- optim(c(0.5, 0.5), loglik, control = list(fnscale = -1))$par[1]
theta_opt <- optim(c(0.5, 0.5), loglik, control = list(fnscale = -1))$par[2]

# Creating data frame of results
df <- data.frame("Optimized MLE" = c(phi_opt, theta_opt),
                 'Diff MLE' = c(phi_hat, theta_hat))
df
```

Table 1 shows use the output from this code, and compares our optimisation with our method from section 1.3.2 which used partial differentiation. We see that there is some difference in the estimators, but not a hugely significant one, with an error of less than 1% in both cases.

Parameter	Optimisation	Diff.	Difference
$\hat{\phi}$	0.20478567	0.20465753	0.0001281403
$\hat{\theta}$	0.06680815	0.06682572	0.0000175758

Table 1

2 Global Temperature

2.1 Introduction

We're examining data that contains observations of the annual mean surface air temperature over land from 1857 - 2021. The temperatures are made relative to the time period 1857 - 1899, by adding a constant such that the sample mean is 0.

Figure 3 shows a scatter plot of the data, with a smoothing line to demonstrate the general trend of the data. Between 1850 and 1875, the temperature is relatively constant and the data is relatively spread. From 1875 to 1900 we see an upturn in temperature, at a slow rate. This trend continues at a faster rate up until around 1940. Here, we see a roughly 10-year plateau, where the trend is very constant. The data tends to be more closely grouped in this area also. After 1950, we see very drastic rise in mean temperature, the trend is very clear here data is quite closely grouped and rising an ever increasing rate. Temperature are almost exclusively above 0.5°C after and the trend shows the data may stay above 1.5°C, as it has in recent years.

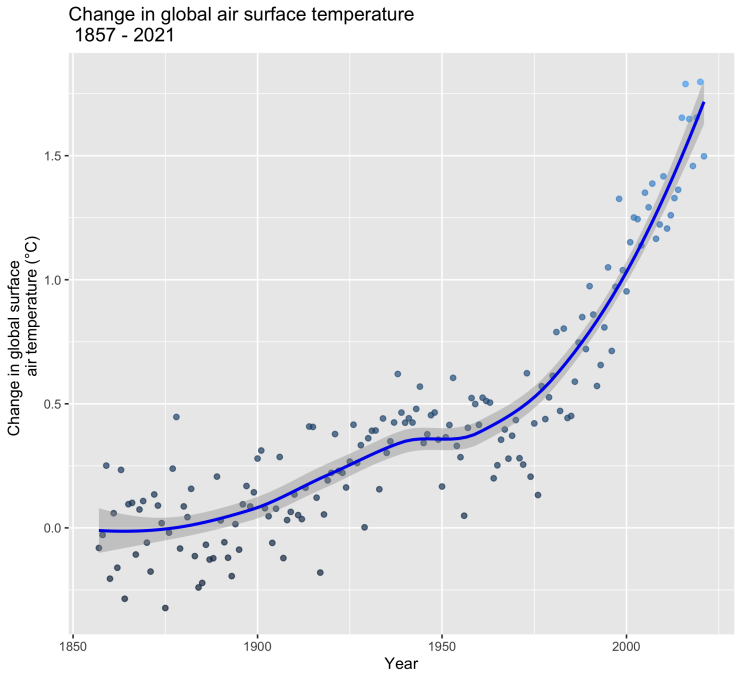

Figure 3: Scatter plot of global air temperature change

2.2 Model Fitting

We will now fit normal linear model to our data with year as covariate. The code used to fit this linear model can be seen in Appendix **??**. This model provides a point estimate of $\hat{r} = 0.00849$ (3 s.f.), for the rate off change of temperature per year i.e. on average temperature increases by

around 0.0849°C per year in the model. It has a confidence interval of

$$\hat{r} \pm t_{n-p-1;1-\alpha/2}\sqrt{\hat{\sigma}^2 \cdot c} \tag{1}$$

where t is derived from the t distribution, n (3650) is the size of the data set, p (1) is the number of parameters, $\sigma^2 \cdot c$ is the standard error of $\hat{r}$ and $100(1-\alpha)\%$ is the confidence interval. Hence we can find a 95% confidence interval of $(0.00768, 0.00929)$ (3 s.f.).

We can test if the rate of change ($\hat{r}$) is significantly different from 0 by using the a hypothesis test. We take the null hypothesis to be $r = 0$ and the alternate hypothesis to be $r \neq 0$. We will be testing at the 1% significance level. We have a test statistic that calculated by

$$T = \frac{r \hat{-} 0}{\sqrt{\sigma^2 \cdot c}}$$

Our summary of the model provides this test statistic to be 21.162 (5 s.f.) for rate and a p-value of $1.265e-48$. Since this value is significantly less than 0.01, we reject the null hypothesis and can say that the rate of change is significantly different from 0.

Figure 4 shows our linear model being fitted to our scatter plot in Figure 3. We see that it correctly models the positive relationship of year and temperature change and for a period, 1900 to 1930, it follows the path of our smoothing line relatively well. However, significantly over estimates the temperature change in the period 1950 to 1970. It then it does model the significant upturn in the later stages of the 20th century and early 21st century. Overall, this suggests the model is not particularly well-specified as, aside from the general positive relationship, it gets significant parts of the data wrong.

We can further establish how well-specified but examining the residual plots in Figure 5. We see that in Graph 1, which plots fitted values against the residuals. We see some sort of quadratic relationship here, which suggests that the model does not account for some non-linear relationship. Looking at quantile-quantile plot (Graph 2) , the data seems to fit relatively until we reach the upper quantiles where they do not line up. Graph 3, which should show no pattern at all if the model is well-specified, shows a distinct pattern of decline followed by an upturn, which shows the residuals are not equally spread along the range of predictors. Graph 4, which shows how much leverage the residuals have i.e. how much they affect the model, does redeem the linear model slightly as there are no identified influential cases.

Overall, from both Figure 4 and Figure 5 we have seen that is not a particularly well-specified as it has a few fundamental flaws.

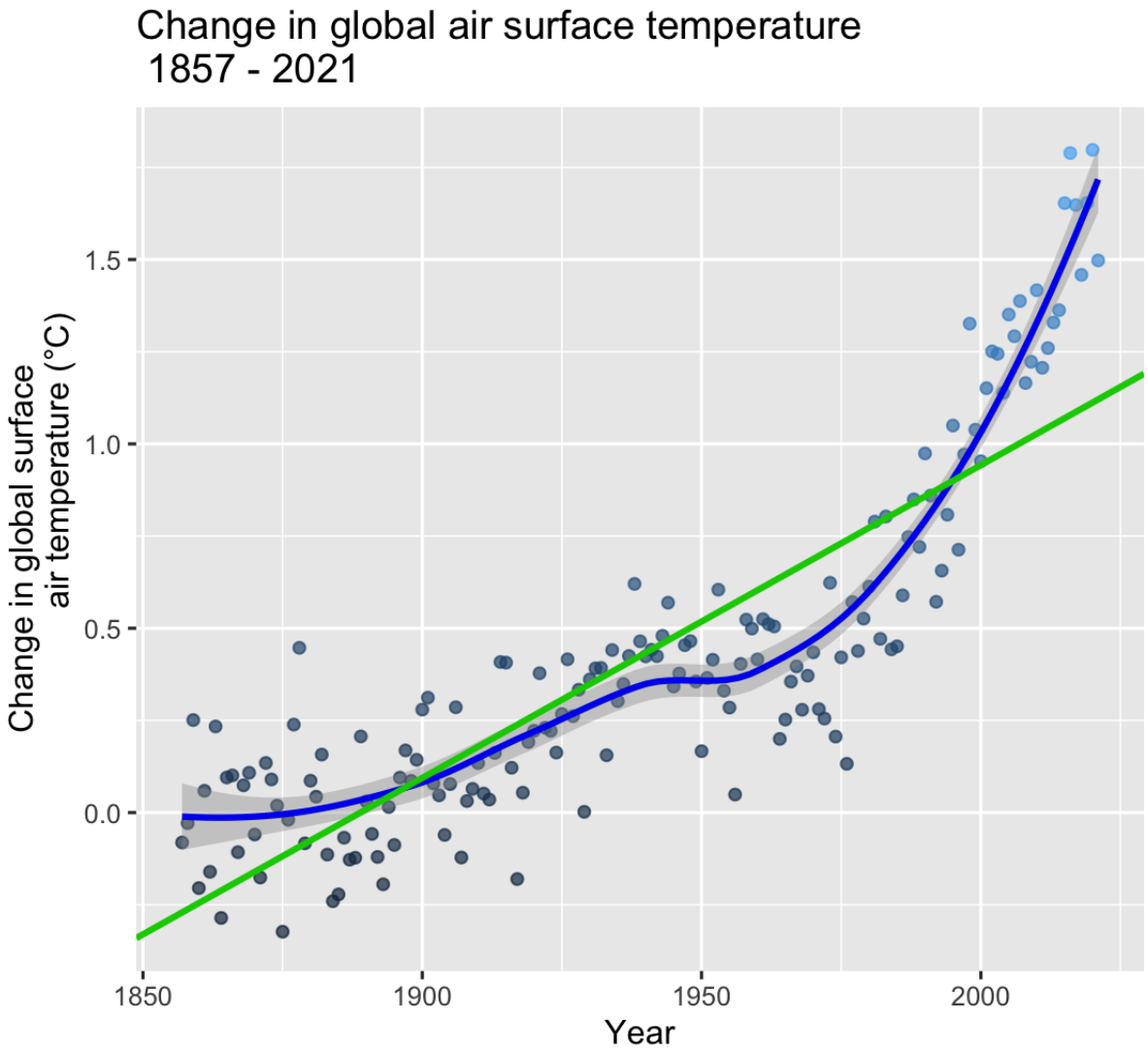

Figure 4: This is Figure 3 now including the fitted line from our linear model

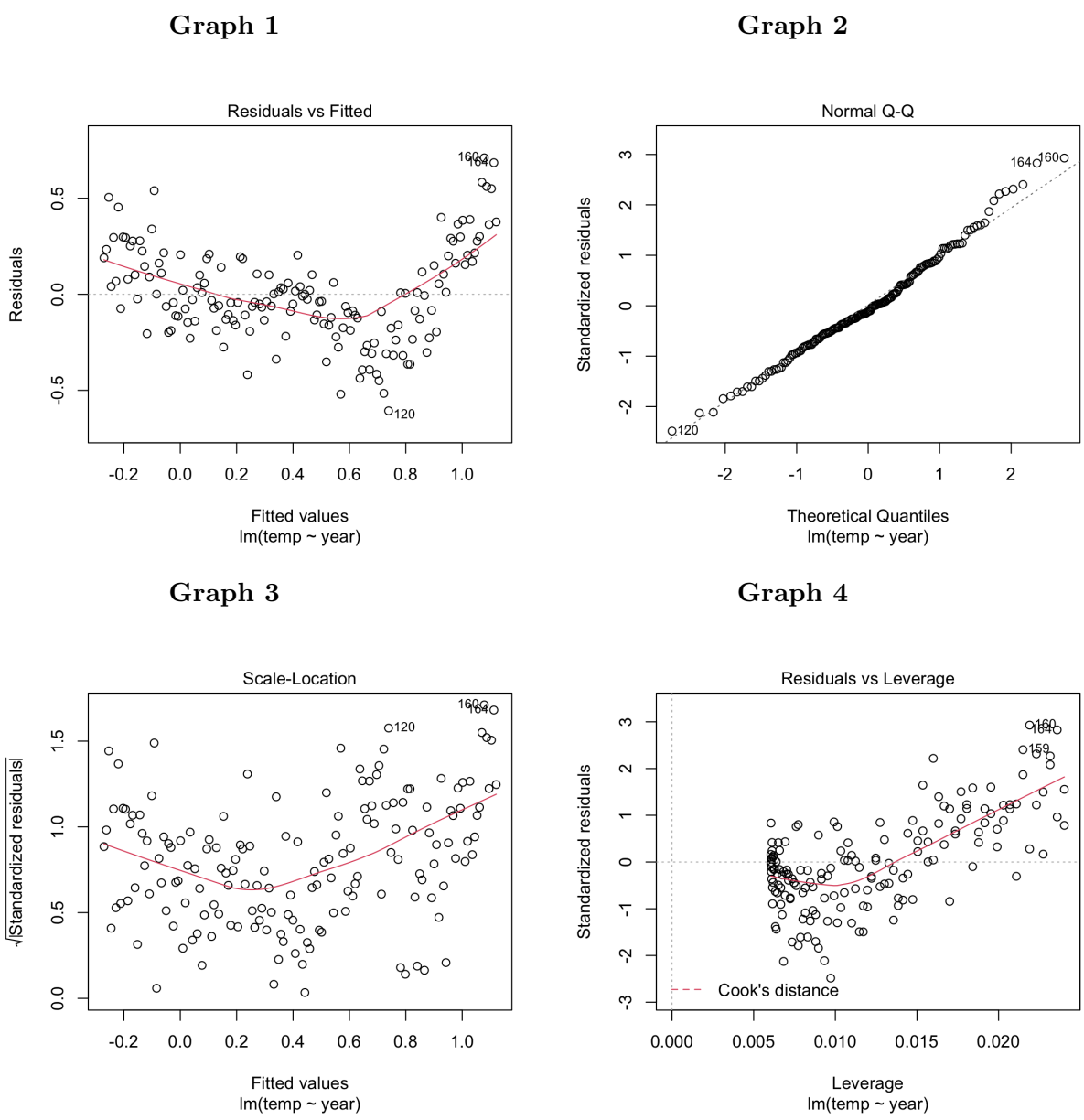

Figure 5: Residual plots of the linear model

2.3 Improvements on the Model

We will now consider modelling the data with the year as the covariate for a polynomial regression model, $Y_i \sim (\beta_0 + \beta_1 x + ... + \beta_k x^k, \sigma^2)$, for $k = \{1, 2, 3, 4\}$. We use R to compute these models. We will use the adjusted r-squared measure to compare these models, as this has the advantage of penalising models for unnecessary greater complexity, unlike r-squared which would simply assign the highest value to the most complex model.

k	1	2	3	4
R_a^2	0.732	0.844	0.872	0.886

Table 2: Adjusted r-squared values for the k-polynomial model

The table above clearly shows the best model by this measure is $k = 4$. We can test whether this is significantly better fit than the linear model, $k = 1$. We can find the test-statistic and p-value from R. We say that H_0 is that $k = 4$ is not a significantly better fit than $k = 1$ and H_1 is that $k = 4$ is a significantly better fit than $k = 1$. Since the p-value, $4.63e-75$, is significantly less than 0.05 we can say the $k = 4$ model is a significantly better fit than the linear trend model.

Using this improved model, we can now make a reasonable prediction of the observed temperature in 2040. Using R, we find that the value will be 3.253°C (4 s.f.) with a 95% confidence interval of (2.854, 3.653) (4 s.f.). Hence, the model predicts it the temperature will be 3.253°C higher than the average temperature in the pre-19th century period.

A Rainfall in Alfheim

Below is the code that has been use to plot graphs and generate values in Section 1.

```
library(tidyverse)
library(ggplot2)
load("~/OneDrive - University of Exeter/R Directory /aut2021.RData")
# ------ Question 1(a) ------

# Examining and filtering data
alfheim_dry <- filter(alfheim, y <= 1 )
alfheim_wet <- filter(alfheim, y > 1)
alfheim_wet <- filter(alfheim_wet, y < 100)
head(alfheim_wet)
head(alfheim_dry)

# Adding years column to table
year <- c(rep(1,365),rep(2,365),rep(3,365),rep(4,365),rep(5,365),
          rep(6,365),rep(7,365),rep(8,365),rep(9,365),rep(10,365))

alfheim <- mutate(alfheim, year = year)

# Plotting rainfall data
ggplot(alfheim) +
  geom_boxplot(aes(group = year, x = year, y = y,
                   fill = year)
  ) +
  ylim(0,2.5) +
  theme(legend.position="none",
        plot.title = element_text(size=11)) +
  ggtitle("Yearly variation in daily rainfall data") +
  scale_x_continuous('Year', seq(1,10,1)) +
  ylab('Daily rainfall (mm)')

# --- Question 1(d) ---

# getting information from data
m = nrow(alfheim_wet)
n = nrow(alfheim)
sum_y_wet = sum(alfheim_wet[,2])

# estimating params using formulas
phi_hat =   m / n
theta_hat = m / ((sum_y_wet) - m)

print("The estimate for phi is:")
print(phi_hat)
print("The estimate for theta is: " )
print(theta_hat)

# --- Question 1(e) ---


# p[1] = phi, p[2] = theta
# Creating log-likelihood function
loglik <- function(p){
  if(p[1] <= 0) {
    return(-1e20)
  }
  if(p[2] <= 0) {
    return(-1e20)
```

```
  }

  # Setting values from data
  n = nrow(alfheim)
  m = nrow(alfheim_wet) # alfheim_wet - all vals s.t y > 1
  sum_y = sum(alfheim_wet[,2])

  # Calculating log-likelihood
  (n-m) * log(1-p[1]) + m * ( log(p[2]) + log(p[1]) ) - p[2] * (sum_y - m)
}

# Optimising for parameters
phi_opt <- optim(c(0.5, 0.5), loglik, control = list(fnscale = -1))$par[1]
theta_opt <- optim(c(0.5, 0.5), loglik, control = list(fnscale = -1))$par[2]

# Creating data frame of results
df <- data.frame("Optimized MLE" = c(phi_opt, theta_opt),
                 'Diff MLE' = c(phi_hat, theta_hat))
df
```

B Global Temperature

Below is the code that has been use to plot graphs, perform optimisations and generate values in Section 2.

```
# ---- Question 2 ----

library(tidyverse)
library(ggplot2)
load("~/OneDrive - University of Exeter/R Directory /aut2021.RData")

# ---- Question 2(a) ----

# creating a plot of data
tempplot <- ggplot(global, aes(year, temp)) + geom_point( aes(x = year, y =
    temp,
              colour = temp), alpha = 0.7) +
              geom_smooth( aes(x = year, y = temp), colour = 'blue2') +
              labs(x = "Year", y = 'Change in global surface
                   air temperature ( C )') +
              ggtitle('Change in global air surface temperature \n 1857 - 2021')
    +
              theme(legend.position = "none")
tempplot
#ggsave(plot = tempplot, filename ='temp_plot.png')

# ---- Question 2(b) ----

# generating linear model
fit.lm <- lm(temp ~ year, data = global)
summary(fit.lm)

# finding rate from summary
rate <-  summary(fit.lm)$coefficients[2,1]
std_err <- summary(fit.lm)$coefficients[2,2]

# find confidence intervala of rate
low_CI <- rate - 2 * std_err
hi_CI <- rate + 2* std_err
```

```
print('The confidence interval is')
print((low_CI))
print((hi_CI))

# Conducting hypothesis test
p_value <- summary(fit.lm)$coefficients[2,4]
sig_level <- 0.01
test_stat <- summary(fit.lm)$coefficients[2,3]
print("The null hypothesis is that r = 0 (rate of change),
       the alternate hypothesis is r != 0")
if (p_value < sig_level){
  print("The p-value is below the significance level hence
        we reject the null hypothesis, the rate of change
        is significantly different from 0")
}

# ---- Question 2(c) ----

# plotting linear model on graph
m <- fit.lm$coefficients[1]
temp_plot_fit <-  + geom_abline(intercept = m, slope = rate,
                                      colour = 'green3', size = 1)
temp_plot_fit
#ggsave(plot = temp_plot_fit, filename ='temp_fit_plot.png')

#plot(fit.lm)


# ---- Question 2(d) ----

# Generating models and find p-value and adj.r-squared values
lm.poly1 <- lm(temp ~ poly(year,1), data = global)
ra_1 <- summary(lm.poly1)$adj.r.squared
p_1 <- anova(lm.poly1)$'Pr(>F)'[1]

lm.poly2 <- lm(temp ~ poly(year, 2), data = global)
ra_2 <-summary(lm.poly2)$adj.r.squared
p_2 <- anova(lm.poly2)$'Pr(>F)'[1]

lm.poly3 <- lm(temp ~ poly(year, 3), data = global)
ra_3 <- summary(lm.poly3)$adj.r.squared
p_3 <- anova(lm.poly3)$'Pr(>F)'[1]

lm.poly4 <- lm(temp ~ poly(year, 4), data = global)
ra_4 <- summary(lm.poly4)$adj.r.squared
p_4 <-  anova(lm.poly4)$'Pr(>F)'[1]

modelcheck <- tibble('PolyNumber' = 1:4,
                     "Adj.R-Values" = c(ra_1, ra_2, ra_3, ra_4),
                     "P-Values" = c(p_1, p_2, p_3, p_4) )


# Conducting hypothesis test
if (p_4 < 0.05) {
  print("Reject H_0")
}

# ---- Question 2(e) ----

# Predicting value for year 2040
predict = data.frame(year = 2040)
```

```
predict_val <- predict(lm.poly4, newdata = predict, interval = 'confidence')
```