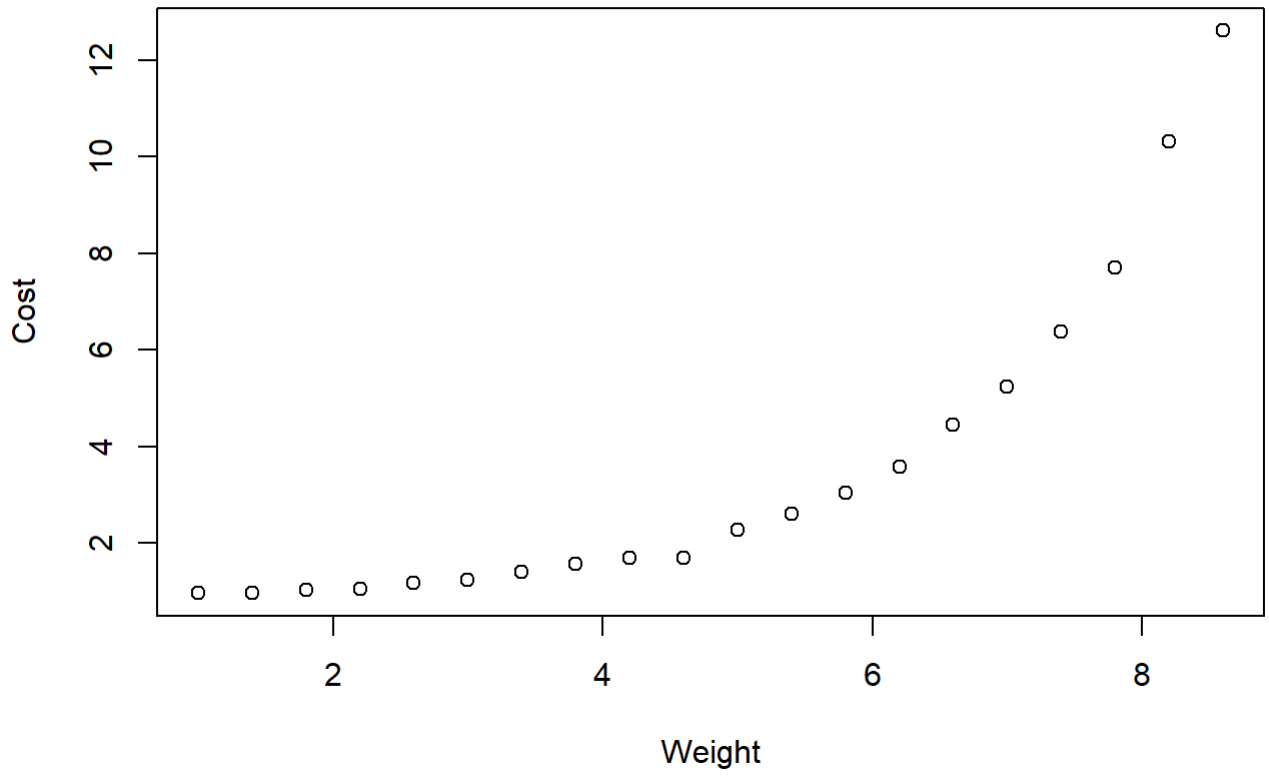
MATH40005 Coursework

Kirev, Ivan CID: 01738166

Step 1: Reading in the Data (4 marks)

Here we read in and plot the data contained in `data01738166.txt` .

```
# read in the data
df <- read.table("data01738166.txt", sep=",", header=TRUE)

# extract the columns of data to separate vectors
weight <- df$weight
cost <- df$cost

# plot a scatterplot of the data
plot(x=weight, y=cost, type='p', xlab="Weight", ylab="Cost")
```

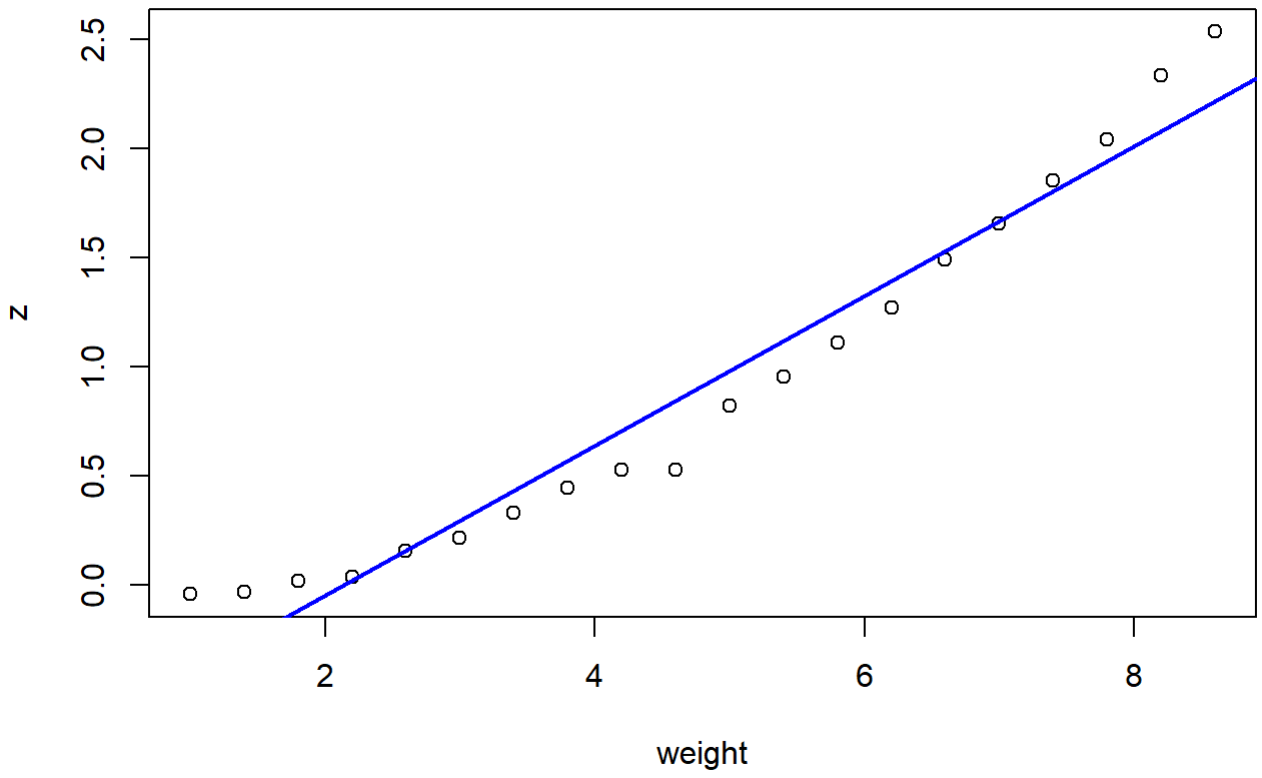


Looking at the plot, it appears that a linear relationship between the two variables is possible.

Step 2: Fitting a linear model (4 marks)

We use the `lm` function to fit a linear model, where we have transformed the cost variable by $Z= \log({\text{cost}}).$ Then we extract $\hat{\beta_0},$ and $\hat{\beta_1}$ and use them to plot the line showing the estimated linear relationship.

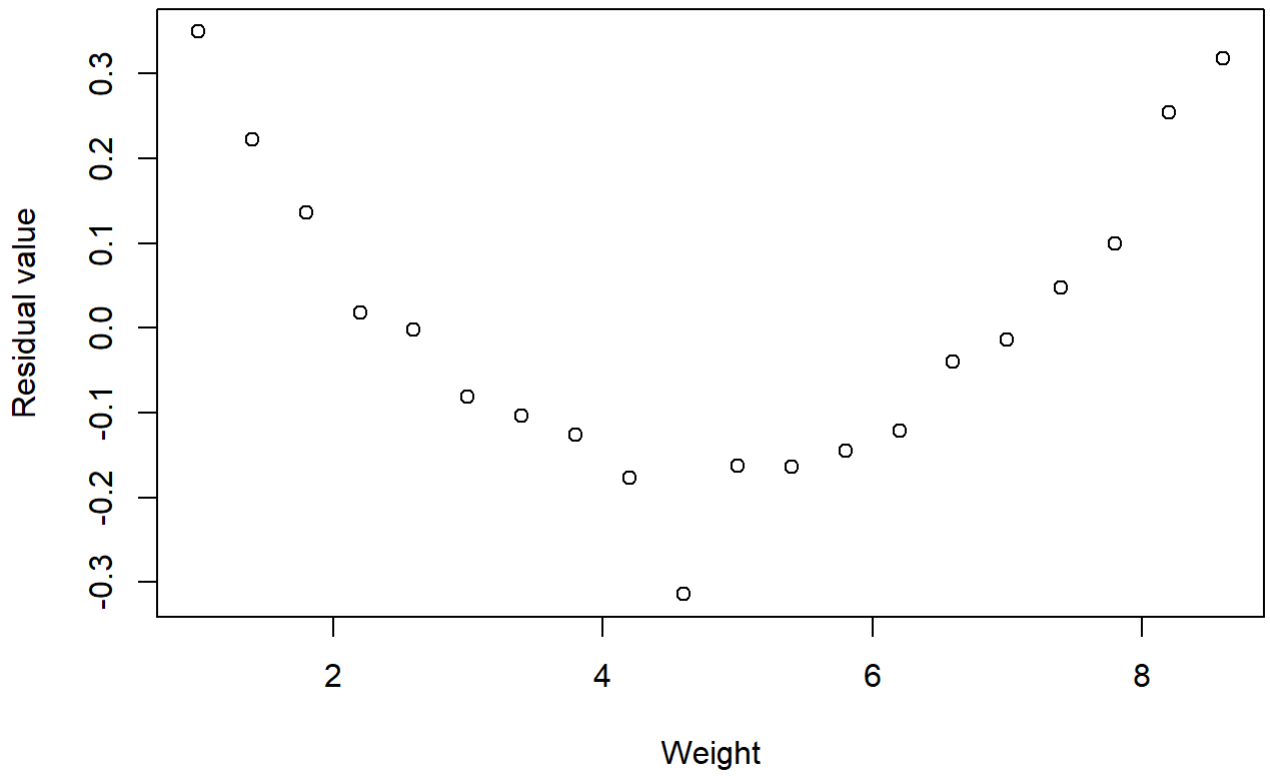
```
z <- log(cost)
model <- lm(z ~ weight)

# extract parameter coefficients from model object
beta0hat <- model$coefficients[1]
beta1hat <- model$coefficients[2]

plot(weight, z)
abline(a = beta0hat, b=beta1hat, col="blue", lwd=2)
```



Step 3: Analysis of results (4 marks)

From the `lm` function we extract the residuals and plot them agains the weight.

```
#extract the residuals from the linear model
residuals <- model$residuals

#plot the residuals agains the weight
plot(weight, residuals, xlab="Weight", ylab="Residual value")
```

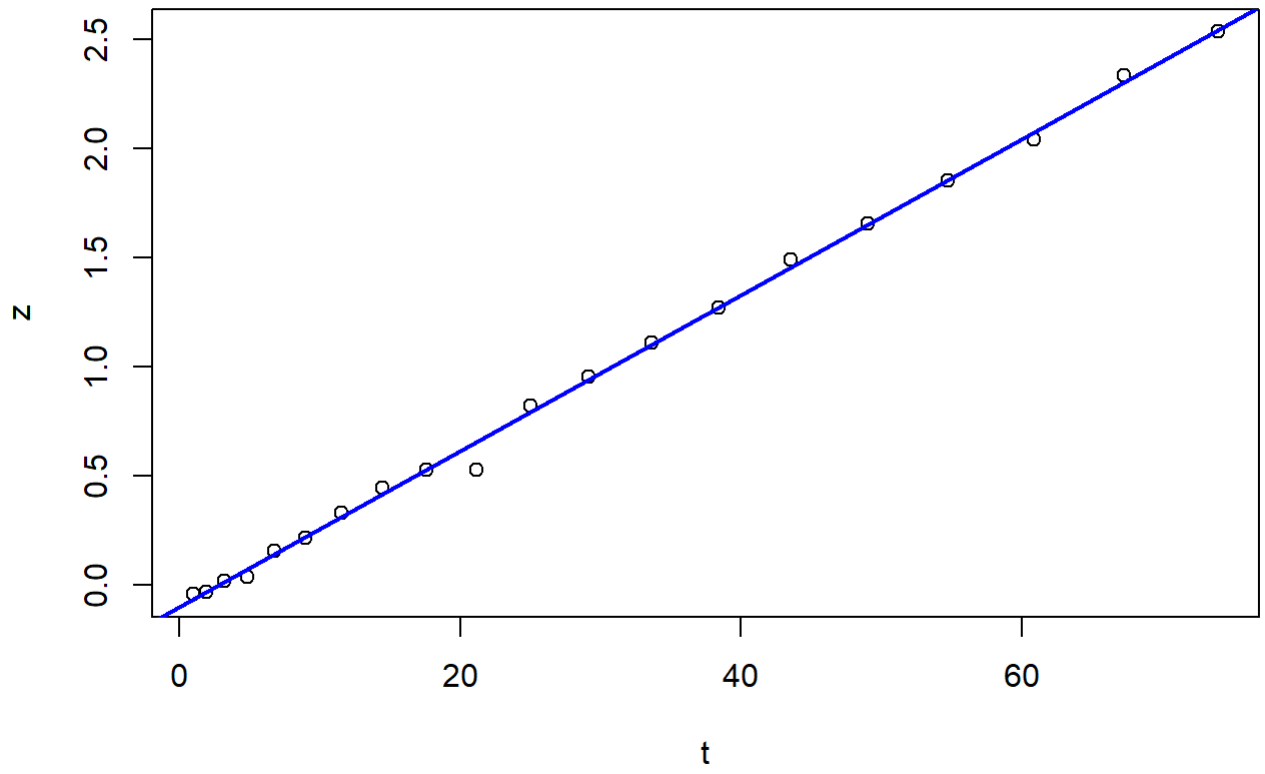


We notice that the residuals appear to follow more of a ‘U’ shape, rather than be randomly distributed around 0. This suggests that our model \[\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i\]may be incorrect.

Step 4: Fitting a second linear model (if desired) (4 marks)

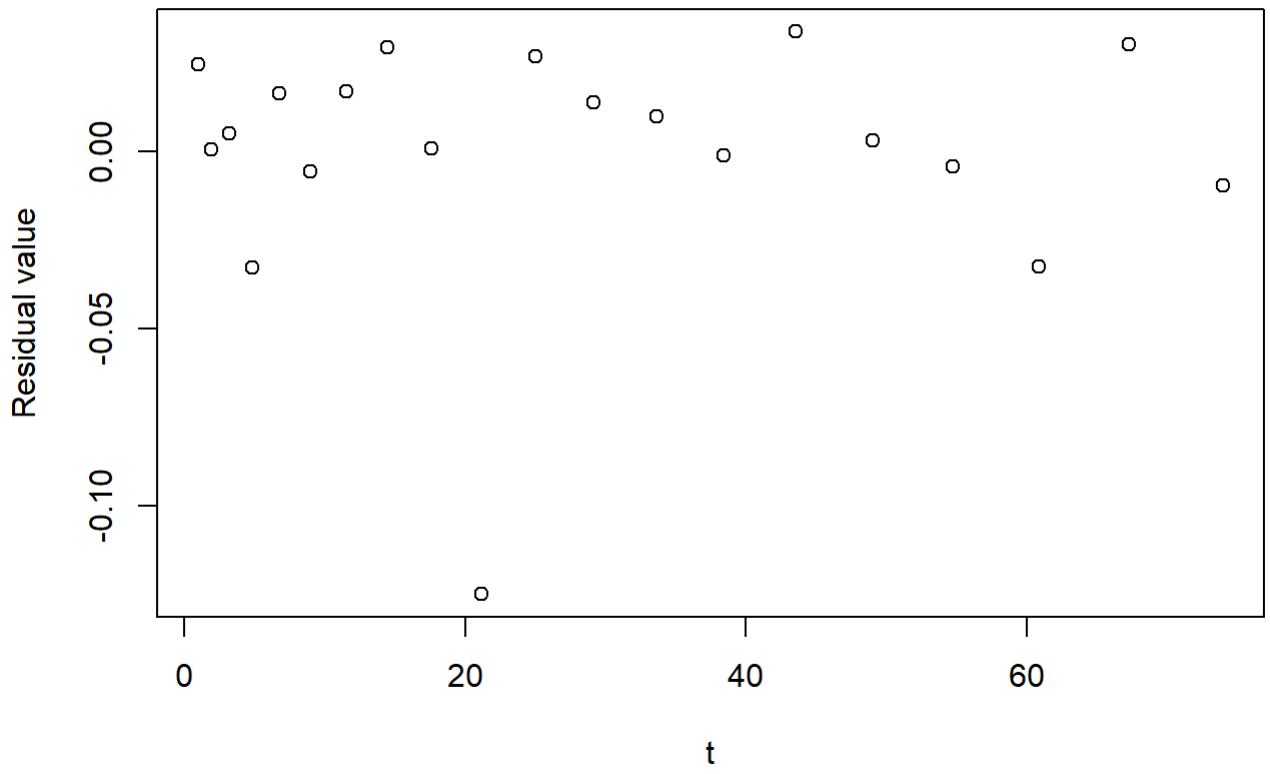
Now lets consider the transformation \(t = (\text{weight})^2\) and create a second linear model.

```
t <- (weight)^2
model_2 <- lm(z ~ t)

#extract parameter coefficients from the second linear model
beta0hat_2 <- model_2$coefficients[1]
beta1hat_2 <- model_2$coefficients[2]

plot (t, z)
abline (a = beta0hat_2, b=beta1hat_2, col="blue", lwd=2)
```



Similarly to Step 3, we extract the residuals from the second linear model and plot them against t

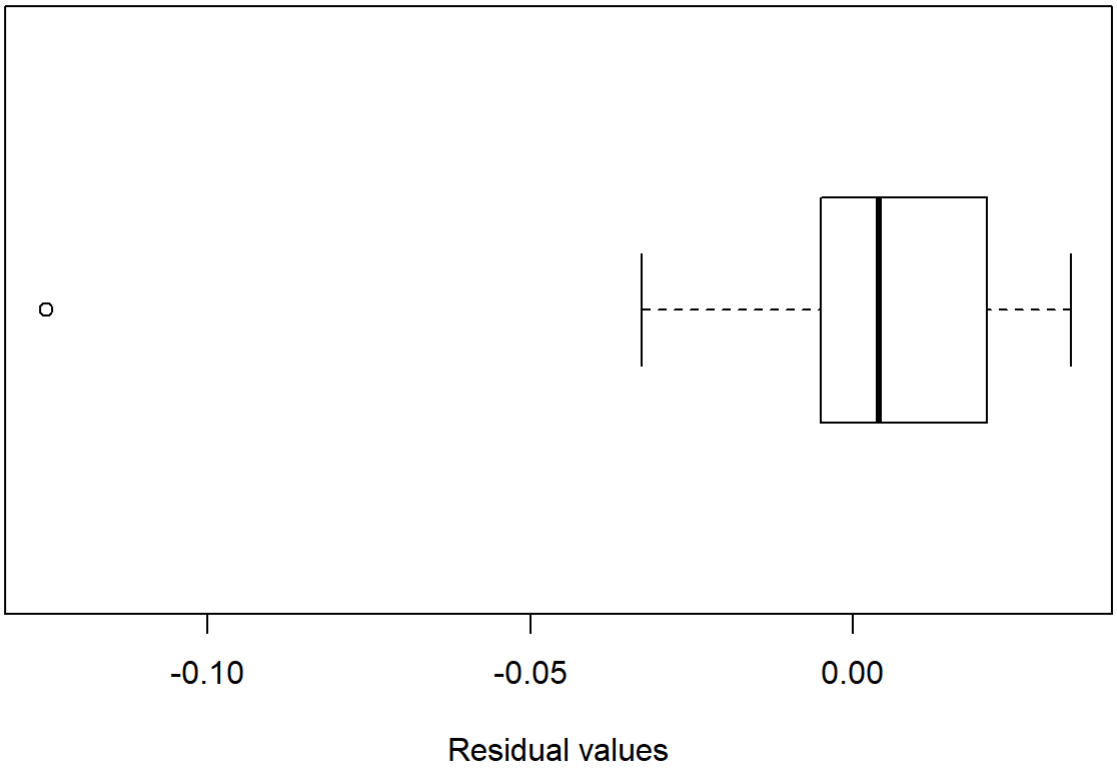
```
#extract the residuals from the second linear model
residuals_2 <- model_2$residuals

#plot the residuals agains t
plot(t, residuals_2, xlab="t", ylab="Residual value")
```



These residuals appear to be centred around 0 which suggests that this model is a better fit to the data and $\log(Y_i) = \beta_0 + \beta_1 x_i^2 + \epsilon_i$ may be a better model than $\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \text{ from step 3 }$ or than simply $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$

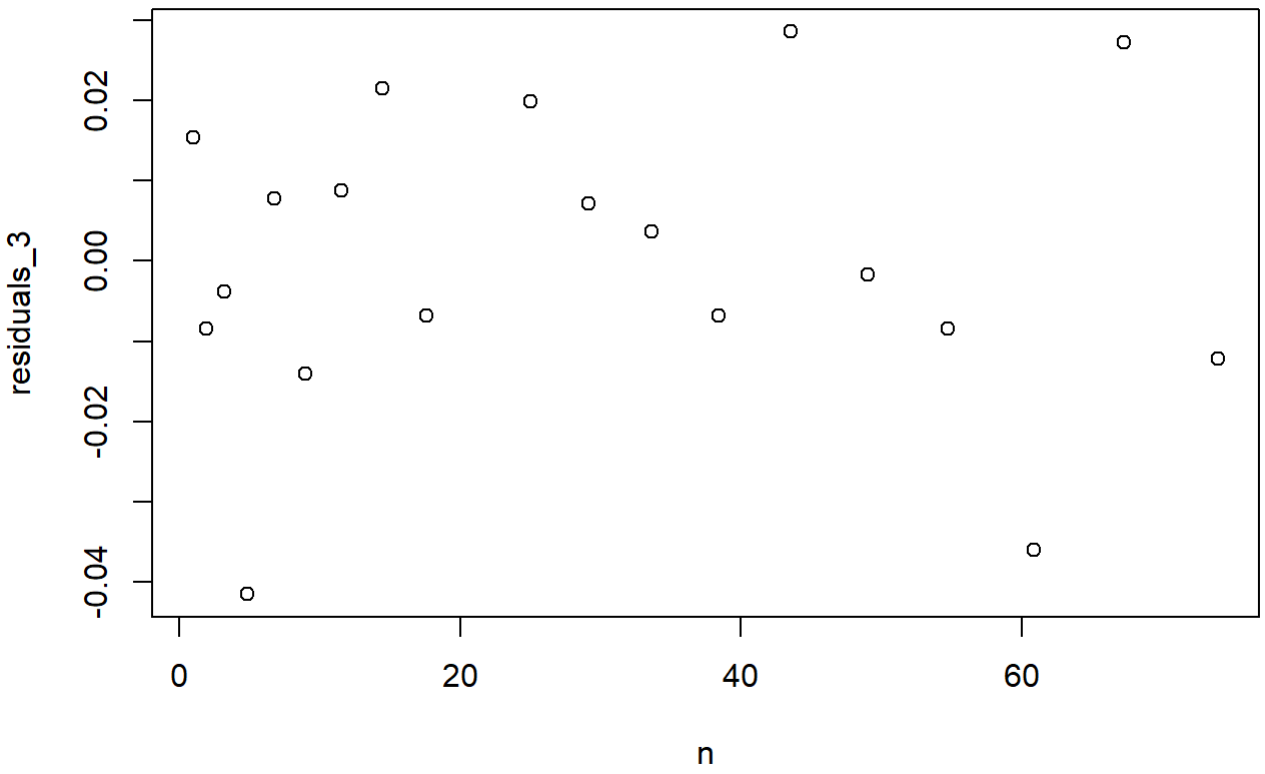
However, the 10th residual appears to be much smaller than the others. It could be that this data point was recorded incorrectly, with a larger error than the others, or it could be a true data point and our model that assumes normally-distributed errors is incorrect. If we decide that the observation is the result of a measurement error, we could see if its residual value is an outlier using a boxplot and Tukey's criterion.

```
boxplot(model_2$residuals, horizontal=TRUE, xlab="Residual values")
```



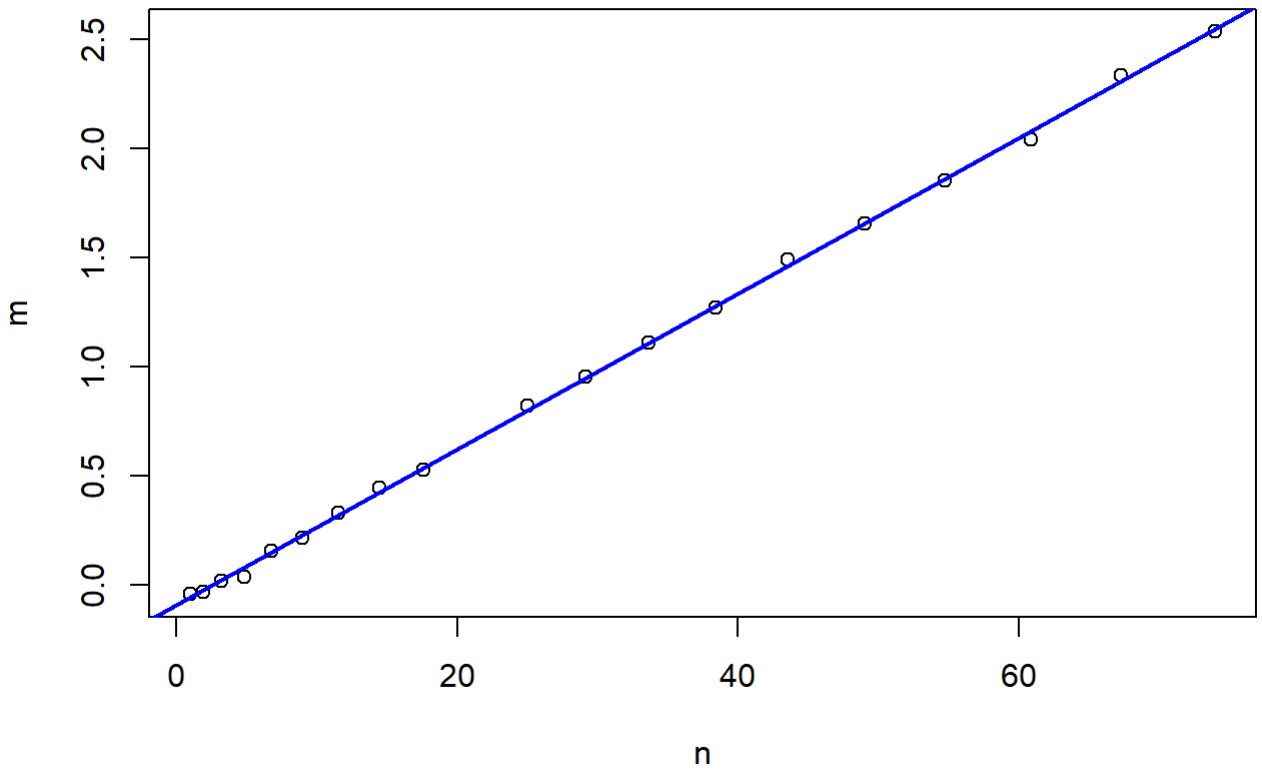
This point appears to be an outlier, and if we remove it and replot the residuals, the residual plot appears to show values centred around 0, which suggests that this model is a better fit to the data.

```
# the outlier is the 10th value, so remove this value to define the inliers
x <- df$weight[-10]
y <- df$cost[-10]
m <- log(y)
n <- x^2
model_3 <- lm(m ~ n)
residuals_3 <- model_3$residuals
plot(n, residuals_3)
```



If we finally plot the fitted regression line for this model, we see that there is a good fit, as suggested by the plot of the residuals.

```
# extract parameter coefficients from model_3 object
beta0hat_3 <- model_3$coefficients[1]
beta1hat_3 <- model_3$coefficients[2]

# plot the (transformed) values and the regression line
plot(n, m)
abline(a = beta0hat_3, b=beta1hat_3, col="blue", lwd=2)
```



Step 5: Interpreting results (4 marks)

From the last steps we derived that a good fit for the data is the relation \[\log{Y_i} = \beta_0 + \beta_1x_i^2+\epsilon_i,\] where our data sample now consists of 19 elements since we removed one in step 4.

We have assumed that the simple linear regression model has independent errors \(\epsilon_i \sim N(O,\sigma^2)\) and have computed maximum likelihood estimates \((\hat{\beta_0},\hat{\beta_1}, \hat{\sigma^2})\) for the parameters \((\beta_0,\beta_1,\sigma^2)\). It can be shown that \(\hat{\beta_0}\) and \(\hat{\beta_1}\) are unbiased estimators, while \(\hat{\sigma^2}\) is biased, but \(\frac{19}{17}\hat{\sigma^2}\) is unbiased estimator.