Math Coursework 2
Prathamesh Kulkarni (Student ID: 36004026)
Date: 16-11-2021

Q1.a)
Forward Selection
Reading the data, the given data was stored as a .csv file.

Since Elevation is a categorical variable hence we convert it to a factor.

We use forward selection, we start with the simplest model and keep on adding variables.

The aim is to see if we can predict the temperature using the two variables North latitude and elevation and see which one is more significant in predicting the temperature.

Four models are been created to test make sure we test all possible
combination for forward selection. Which are g1, g2, g3, g4
The four models are as follows.
g1 = glm(Temperature~1, data=temp_df)
g2 = glm(Temperature~North.latitude, data=temp_df)
g3 = glm(Temperature~Elevation, data=temp_df)
g4 = glm(Temperature~North.latitude + Elevation, data=temp_df)

Now that the models are created we now use add1() function to consider each model.

In this analysis we keep on adding one variable at a time. We start with the most simplest model which is g1. We can see how significant each variable is based on the p values. We now choose to add the variable that gives the most significant improvement, which is the variable with the smallest p-value.

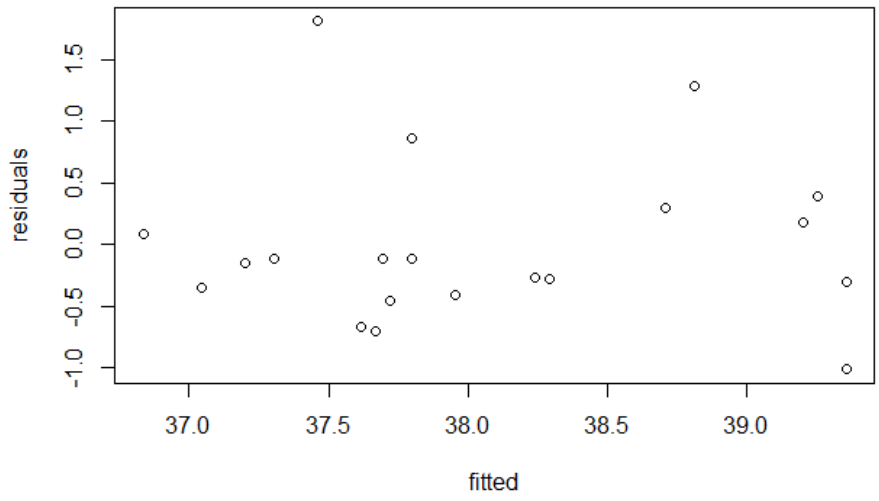
In our case this is North.latitude
Now we choose the g2 model which has the North.latitude variable.

Ideally for forward selection we keep on adding the variable with smallest p value i.e. with most significance and stop adding variables till the p values are above 0.05, hence here we can see the the elevation is giving out a p value of 0.5355 which is greater than 0.05 hence we stop our
analysis and conclude that only using North.latitude variables gives a good prediction of temperature and adding elevation variable wont make any difference as it is not a significant variable.

We will now check the validity of our selected model. which is g2. We check the validity of the model. Since p-value is high the model is valid. Hence, g2 is our final model.

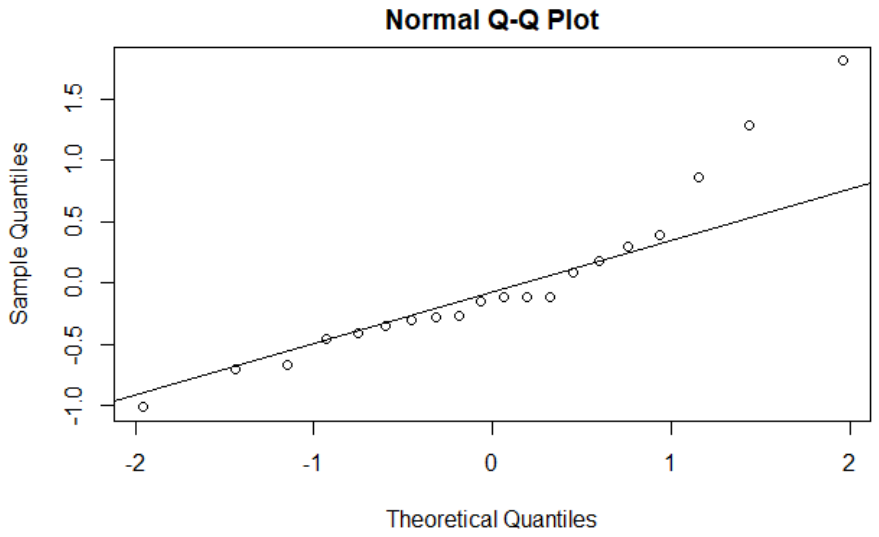
1-pchisq(8.7086,18)
0.9661156

Diagnostics
Constant variance assumption
We look for increasing or decreasing spread of points as fitted values increase. No real evidence of this.



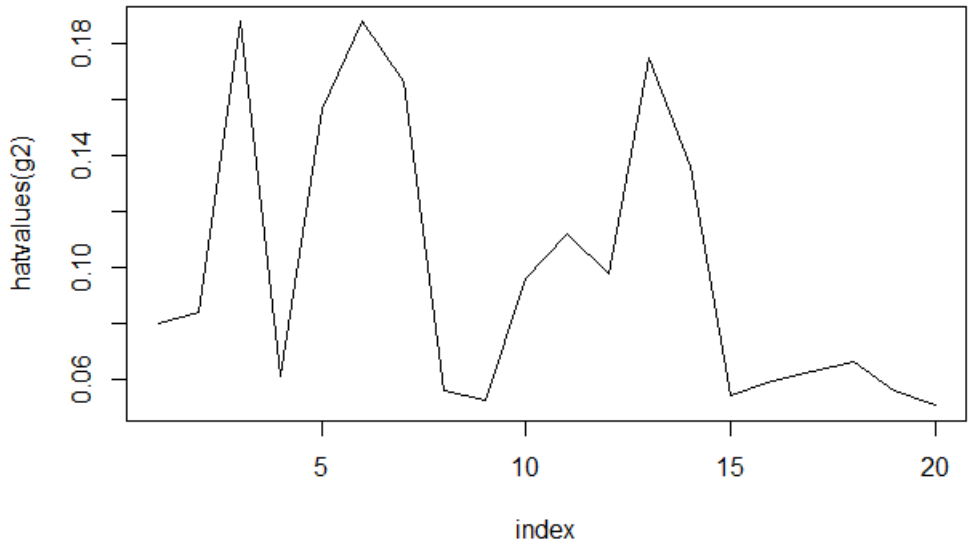
Checking the normality of the residuals

We now check the normality of residuals and can say that normality seems plausible.
From the QQplot we can see that there might be some outliers after point 1.



Leverages

We can see that there are three points around 3, 6, 14 which seem to be leverage values indicating points with too much influence.



Q1.b)Interpretation

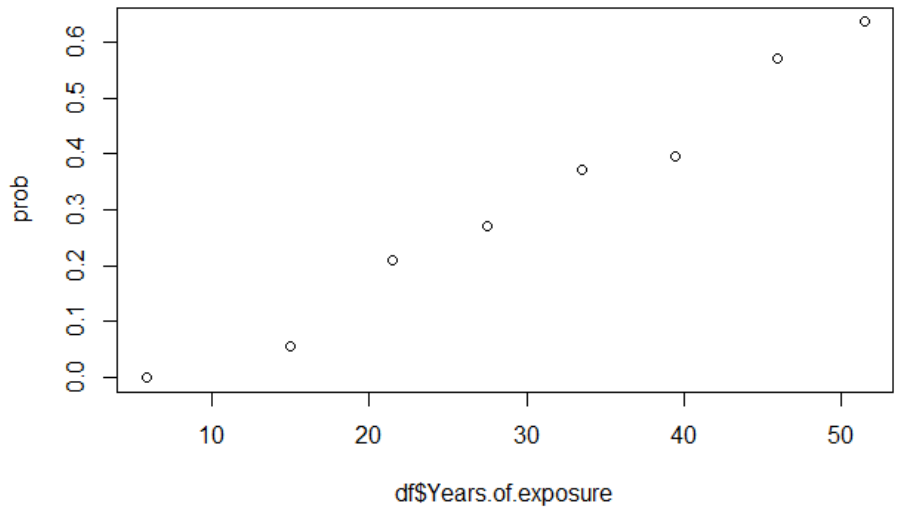
We compare if both the variables are significant hence we run the anova test to see which variables are significant. We use the g4 model since it has both the variables.

We can see that Elevation has p value of 0.566 hence we can remove elevation from the model and that model g2 which we found out is valid.

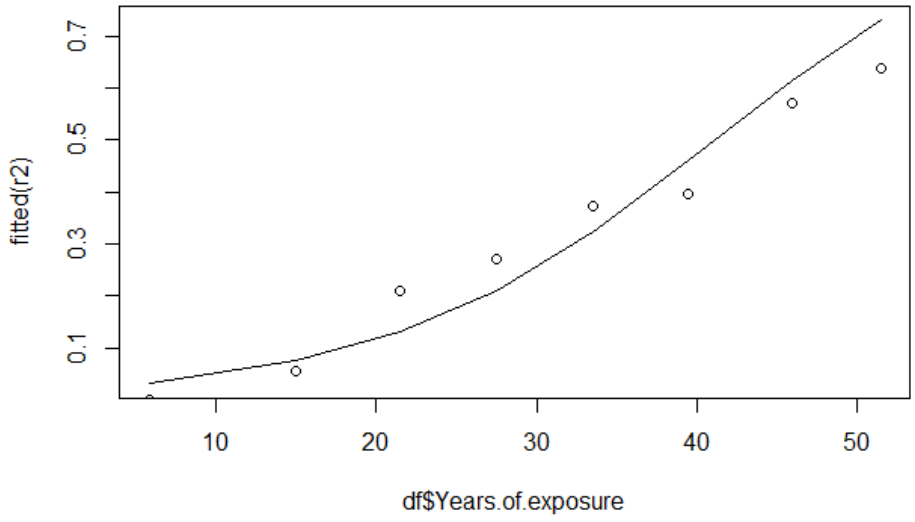
Q2.

a) The two link functions can be logit and probit from the binomial family since it is a proportion data

```
prob = df$No..affected.by.pneumoconiosis/df$No..of.coalminers.exposed
plot(df$Years.of.exposure, prob)
```



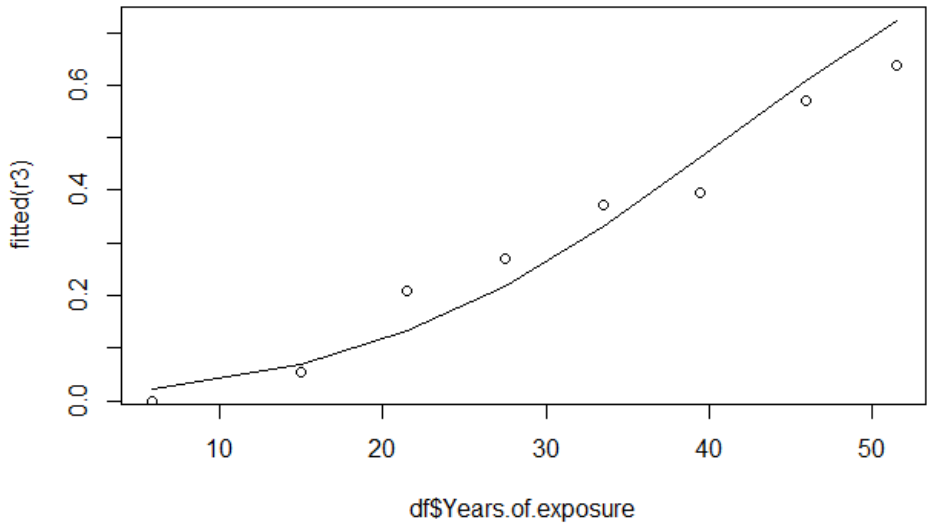

```
r2=glm(prob~df$Years.of.exposure,family=binomial(link="logit"),weights=df$No..of.coalminers.exposed, data = df)
plot(df$Years.of.exposure, fitted(r2),type="l")
points(df$Years.of.exposure, prob)
```



Checking the validity of logit model

```
1-pchisq(11.566,6)
0.07238138
```

```
r3=glm(prob~df$Years.of.exposure,family=binomial(link="probit"), weights=df$No..of.coalminers.exposed, data = df)
plot(df$Years.of.exposure, fitted(r3),type="l")
points(df$Years.of.exposure, prob)
```



Checking the validity of probit model

```
1-pchisq(8.8616,6)
0.1815125
```

Q2.b) since the p value of the probit is larger than the logit model we can say that the probit model fits better.

Q2.c) For the prediction of the proportion affected with an exposure for 25 years we use predict or predict.glm with the model and new data.

5.1

```
newdata = data.frame(Years.of.exposure = 25)
predict.glm(r2, newdata, type="response")
```

Note: R markdown is not provided since the content was not concise.

Feedback comments

Qn 1: A-

Qn 2: 1+2+0=3

Index of comments