HERIOT-WATT UNIVERSITY

F79MB Statistical Model B

Name: Cheong Chi King Student ID: H00301314

Assessment Project 2

1a. The summary of the linear regression model with Benefits as the response variable and Assets as the explanatory variable is

```
Residuals:
     Min      1Q  Median      3Q     Max
-57123.  -1558.   6297.   7247.  33864.

Coefficients:
             Estimate      Std. Error    t value      Pr(>|t|)
(Intercept)  -7.572e+03    3.790e+03     -1.998       0.0572 .
Ast           8.219e-02    1.180e-03     69.656       <2e-16 ***
---
Residual standard error: 18650 on 24 degrees of freedom
```

(10 observations deleted due to missingness)

```
Multiple R-squared:  0.9951,   Adjusted R-squared:  0.9949
F-statistic:  4852 on 1 and 24 DF,  p-value: < 2.2e-16
```

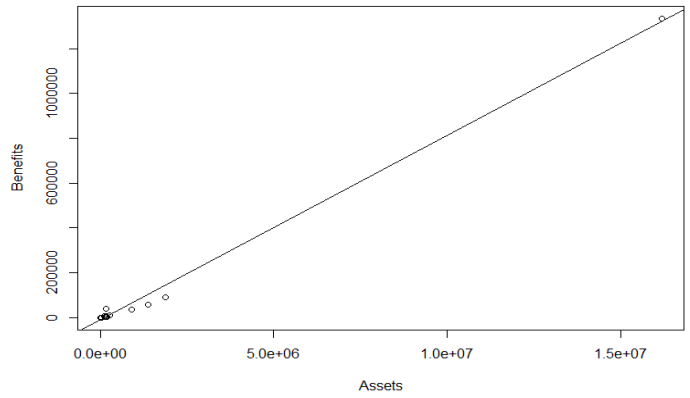

Figure 1: Benefits vs Assets and fitted line

Comment: The summary shows the gap between the minimum and maximum of the residual is large and 1st quantile value is much further to the median than the 3rd quantile value to median. The value of the slope is **8.219×10^{-02}** implies that the relationship between benefits and assets is positive. At 5% significant level, the slope is highly significant but the intercept is not significant. The residual standard error is extremely large which means a high deviation of the model from the true regression line. The estimation of the regression parameters is **R^2=99.51%** which is close to 100% implies that it is almost all the Benefits values that are predictable from the Assets values and indicates a perfect fit but R^2 cannot determine whether the model provides a adequate fit to the data, which later will check with residual plots. However, the p-value of F statistic is very low implies that there is a potential relationship between the benefits and the assets.

From Figure one, the distribution is right-skewed heavily and only one observation is left out on the right-hand end. The left out observation may be a influential point and will be verify later. Heavy positive skewness of the distributions of both variables suggests log transformation for both to make data suitable for linear relationship.

From Residual plots (figure 2), Cook's distance plot shows there is an excessively influential point in the data which is United State (observation no.26).

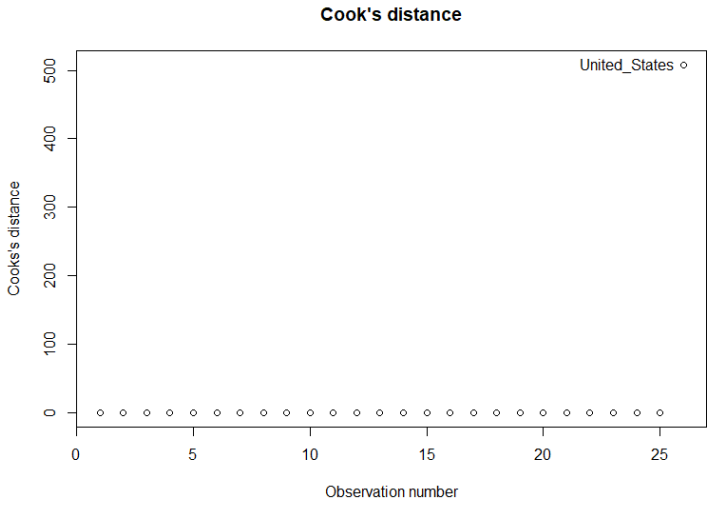

Figure 2: Cook's distance plot

Now a new linear model (Model 2) will be fitted without United States (observation number 26). The new model summary is:

```
Residuals:
                 Min       1Q     Median       3Q      Max
                -5835    -2209     -623     -428    32167

Coefficients:
                Estimate      Std. Error      t value         Pr(>|t|)
(Intercept)    4.334e+02       1.608e+03      0.269           0.79
Ast[-26]       4.533e-02       3.158e-03      14.354          5.74e-13 ***
---
Residual standard error: 7174 on 23 degrees of freedom
Multiple R-squared:  0.8996,   Adjusted R-squared:  0.8952
F-statistic:   206 on 1 and 23 DF,  p-value: 5.735e-13
```

Comments: The gap between the residuals has become smaller. The residual standard errors are decreased a lot in the new model (Model 2). The p-values for F-statistic and slope has increased, but it still remains significant. The coefficient of determination has decreased from **R^2 = 99.51% to R^2= 89.96%.**

From residual plots (figure 3), the line in residuals vs fitted plot is close to 0 but it has a clear pattern and show no linearity. The scale-location plot shows no equal variance assumption as the residuals is not spread equally. Normality of the residuals seems well, although with some deviation in both tails. Cook's distance has reduced and less than 1 implies that there is no more influential point.

Model 2 seems better than the old model. Although the coefficient of determination has decreased and p-value has increased, but it will not make too much different from old model.

Therefore, the preferred fitted model (model 2) is

$$\widehat{Benefit} = (4.334 \times 10^{2}) + (4.533 \times 10^{-02}) \times \text{Assets}$$

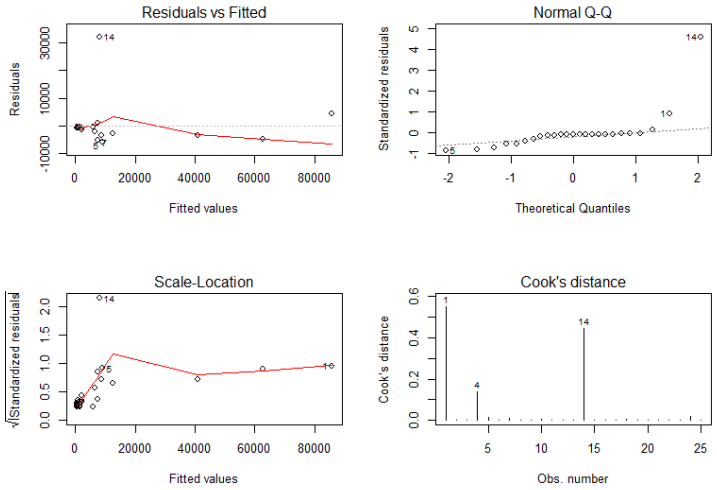

Figure 3: Plots of residuals for the regression of Benefits on Assets without observation number 26

1b. From part (a), figure 1 shows that the data is heavily right skewed implies that log-transformation for the data is needed. The summary of the linear regression model with ln(benefits) as the response variable and ln(assets) as the explanatory variable is

```
Coefficients:
                Estimate     Std. Error     t value         Pr(>|t|)
     (Intercept) -4.43978      0.72191       -6.15          2.36e-06 ***
     log(Ast)     1.09607      0.06486       16.90          7.88e-15 ***
   ---
Residual standard error: 0.7666 on 24 degrees of freedom
        (10 observations deleted due to missingness)
Multiple R-squared:  0.9225,   Adjusted R-squared:  0.9192
F-statistic: 285.6 on 1 and 24 DF,  p-value: 7.879e-15
```

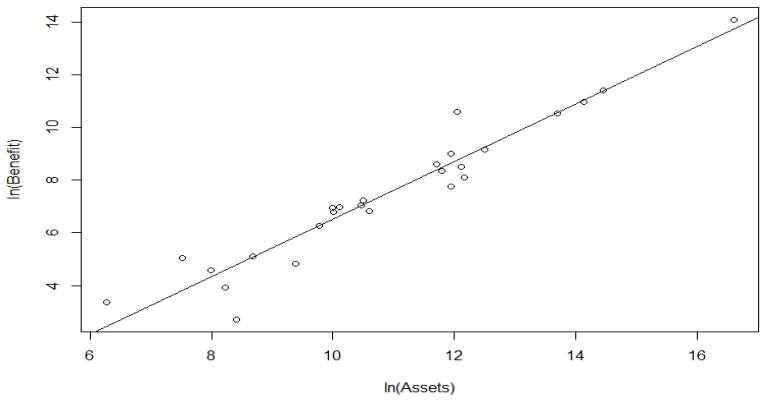

Figure 4: ln (Benefits) vs ln (Assets) and fitted line

Comment: The explanatory variable is highly significant (as in the intercept term). The relationship between ln (benefits) and ln (assets) is positive since the slope of the line is **1.09607**. The coefficient of determination is **R^2=92.5%**. The residual standard error is severely decreased in this model after the transformation which is **0.7666 on 24 degrees of freedom**. The estimation of the regression parameters seems to be less uncertain due to smaller residual standard error. The p-value of F statistic for the model is less than the 5% significance level (**$7.879*10^{-15}$**<p= 0.05). This linear model is known to be highly statistically significant which means ln(assets) is useful in explaining variable is ln(benefits).

From figure 4, the line seems reasonably good fit to the data. There is no influential point in the plot. The observations are equally spread and follow a linear pattern. Also, they show a positive relationship between ln(Benefits) and ln(Assets).

From Residual plots (figure 4), the Residuals vs Fitted plot and Scale-location plot show that the variance may be decreasing as fitted value increase which means there exists heteroscedasticity in the data and has no equal variance assumption from Scale-location plot. From the Normal Q-Q plot, the residuals follow a normal distribution, although with some deviation from normality in both tails. Cook's distance plot shows no excessively influential points.

The fitted model is

$$Ln\ (\widehat{Benefit}) = -4.43978 + 1.09607 \times \ln(\text{Assets})$$

Re-expressed as:

$$\widehat{\boldsymbol{Benefit}} = \mathbf{e}^{-4.43978} \mathbf{x}\ \mathbf{Assets}^{1.09607}$$

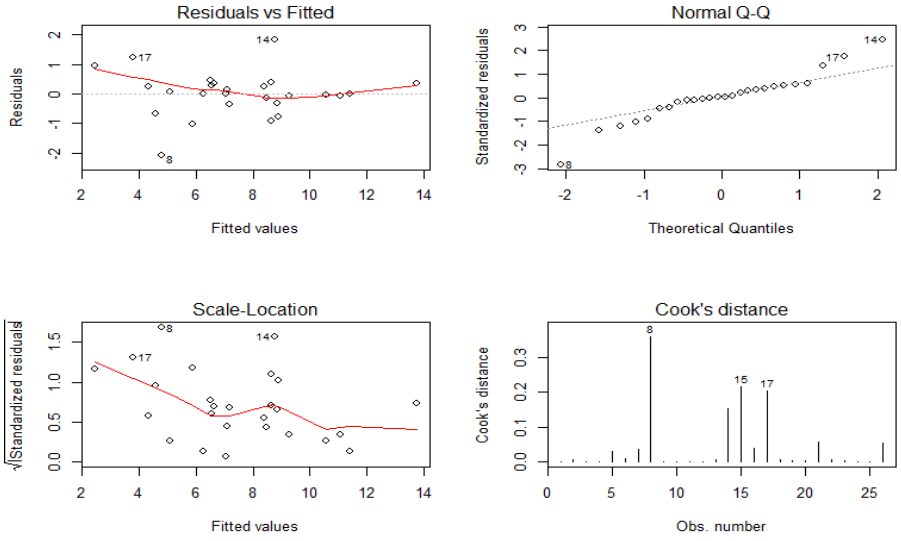

Figure 5: Plots of residuals for the regression of ln(Benefits) on ln(Assets)

1c. For both models in part (a) and part (b) which are

Part (a) model after omitted influential point: $\widehat{\boldsymbol{Benefit}}$**= (4.334 x 10^{2}) + (4.533 x 10^{-02}) x Assets**

Part (b) model: $\widehat{\boldsymbol{Benefit}} = \mathbf{e}^{-4.43978} \mathbf{x}\ \mathbf{Assets}^{1.09607}$

Both of the summaries are compared with each other:

	Model (a)	Model (b)
Coefficient of determination, R^2	0.8996	0.9225
Residual standard errors	7174 on 23 degrees of freedom	0.7666 on 24 degrees of freedom
Differences between residuals	Big	Small
P-value for F-statistic, p	< $2.2*10^{-16}$	$7.879*10^{-15}$
P-values for the intercept and slope	For intercept, p=0.79 For x, $p=5.74*10^{-13}$	For y -intercept, p=2.36e-06 For x, $p=7.88*10^{-15}$

Observation omitted	1	0

Table 1: The comparison for summaries of model (a) and model (b)

Comment: After all the comparison, model B is more preferred to be used for the data. The coefficient of determination and residual standard errors for model (b) are better than model (a) implies that the data fits better and can be explained more in model (b). Also, model (b) has smaller difference between residuals. Although the p-value for F- statistic for model (b) is lower, it is still low enough to make the model highly significant. P-values for the intercept and explanatory variable for model (b) is lower and more significant. In residual vs fitted plots and scale-location plots (figure 3 and figure 5), the linear relationship for model (b) is more reasonable than model (a) although there exists heteroscedasticity in model (b). Also, the model (b) has complete data observation while model (a) has one observation omitted implies that we get to know more information from model (b).

The preferred fitted model (Model (b)) is

$$\widehat{\boldsymbol{Benefit}} = \mathbf{e}^{-4.43978}\mathbf{x}\ \mathbf{Assets}^{1.09607}$$

Conclusion: The above fitted model seems a reasonable description of the data. Before log-transformation, the data are heavily right skewed implies that a log-transformation for the data is needed. After the transformation, the data become more reasonable in the model. The total amount paid out in benefits (millions of US dollars) increases with total amount held assets (millions of US dollars). One unit increase in ln(assets) will increase There is some evidence of heterocedasticity, which perhaps could be addressed using Weighted Least Square regression, but it doesn't look like too big a problem. The regression analysis is based on only 26 countries and there are 10 countries have missing data for benefit which means we have lost some information. The missing data is needed if we want to know more information from the model.

1d. H_0: log-transformed assets values (with all 36 values) can be model by normal distribution

H_1: log-transformed assets values (with all 36 values) don't follow normal distribution

This test is based on x^2 statistifc given as

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Where f_i and e_i are observed and expected frequencies, k is the number of cells.

Assume that log-transformed Assets values can be model by normal distribution,

$$N(\hat{\mu}, \hat{\sigma}^2), with\ \hat{\mu} = 11.077210\ and\ \hat{\sigma} = 2.335419$$

Under H_0, using 6 equal-probability cells and we have two parameters to estimate,

$$X^2 \sim X^2_{k-d-1} = X^2_{6-2-1} = X^2_3$$

where *d* is the number of model parameters estimated

With f_i= [7,6,6,6,4,7] and $e_i = \frac{36}{6} = 6$

The 6 equal-probability intervals for the normal distribution is
(-∞ ,-8.8129), (-8.8129, 10.0663), (10.0663, 11.0722), (11.0722, 12.0781), (12.0781, 13.33156), (13.33156, ∞)

Then we obtain p-value **0.801252**. For 0.05 significant level, we do not have evidence to reject H_0.

Under same hypothesis, but changing the equal-probability cells from 6 to 5

Under H_0, using 5 equal-probability cells and we have two parameters to estimate,

$$X^2 \sim X_2^2$$

With f_i= [7,8,4,10,7] and e_i=$\frac{36}{5}$=7.2
The 5 equal-probability intervals for the normal distribution is

(-∞, 9.1067), (9.1067, 10.4805), (10.4805, 11.6639), (11.6639, 13.0378), (13.0378, ∞)

Then we get p-value= **0.2710219**. For 0.05 significant level, we do not have evidence to reject H_0.

Conclusion: For 6 equal-probability cells, giving p-value, **p=0.801252**. For 5 equal-probability cells, giving p-value, **p= 0.271022**. From both tests, we do not have evidence to reject H_0. Therefore, we know that the log-transformed Assets value can be modelled by a normal distribution.

1e. There is no relationship between the conclusion from part (d) and the conclusion from part (a-c).

The linear model makes no assumption about the distribution of the explanatory variable nor response variable. The model only needs the assumption about the normality of residuals to check the model. Therefore, the normal distribution of ln (assets) (explanatory variable) is not related to the linear model.

2. Fitting the proposed model with ln(Medinc) as the response variable, and ln(Unemp), Metro (Metro with 0 = non-metropolitan, 1 = metropolitan) as explanatory variables, including interaction terms, the analysis of variance result as follow:

```
                       Df Sum Sq Mean Sq F value   Pr(>F)
log(Unemp)              1  39.62   39.62  879.81  < 2e-16 ***
Metrofactor             1   1.16    1.16   25.68 4.26e-07 ***
log(Unemp):Metrofactor  1   0.83    0.83   18.49 1.76e-05 ***
Residuals            3135 141.19    0.05
```

Both explanatory variables are significant in this model and two-way interaction between them is highly significant too. When the interaction term is highly significant, interaction effect should be considered in the model. Although the R-squared for the interaction term is low implies that the interaction term only explained a small proportion of the variability in ln(Medinc), but it doesn't imply that the interaction term is useless in the model.

To investigate more, fitting proposed model, without interaction term (Model 1) and with interaction term (Model 2), giving summaries as follow:

Model 1 (without interaction term) summary:

```
Coefficients:
             Estimate       Std. Error     t value        Pr(>|t|)
(Intercept)  11.311628      0.016596       681.599        < 2e-16 ***
log(Unemp)   -0.352968      0.011818       -29.868        < 2e-16 ***
Metrofactor1  0.039838      0.007883        5.054         4.58e-07 ***
---
Residual standard error: 0.2128 on 3136 degrees of freedom
Multiple R-squared:  0.2231,  Adjusted R-squared:  0.2226
F-statistic: 450.2 on 2 and 3136 DF, p-value: < 2.2e-16
```

Model 2 (with interaction term) summary:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              11.26561    0.01971 571.627  < 2e-16 ***
log(Unemp)               -0.31875    0.01422 -22.414  < 2e-16 ***
Metrofactor1              0.19036    0.03588   5.306 1.20e-07 ***
log(Unemp):Metrofactor1 -0.10925    0.02541  -4.300 1.76e-05 ***
---
Residual standard error:       0.2122 on 3135 degrees of freedom
Multiple R-squared:  0.2276,   Adjusted R-squared:  0.2269
F-statistic:   308 on 3 and 3135 DF, p-value: < 2.2e-16
```

The difference between Model 1 and Model 2 is not much as the coefficient of determination, R^2 are close to each other **(Model 1: R^2 =0.2231, Model 2: R^2 =0.2276)** and the p-values for F-statistic are same (highly significant). By adding interaction term in the model can obtain more information of the relationships among the variables. Therefore, we retain both explanatory variables and their interaction term in the model.

Comment: The summary (Model 2) shows that ln (Medinc) and ln (Unemp) for type 0 and type 1 is negative related. The slope for log (Unemp) for type 0 is **-0.31875** while the slope for log (Unemp) for type 1 is **-0.428 (-0.31875-0.10925).** Both explanatory variables and their interaction are highly significant (as in the intercept term). The coefficient of determination is low as R^2=**22.76%.** Residual error in this model is low. This model is highly significant as the p-value for F-statistic is low **($p< 2.2 \times 10^{-6}$)** implies that ln (Unemp), Metro and their interaction term together are useful in explaining variation in ln (Medinc).

From figure 6, the observations follow a linear pattern and shows negative relationship between ln (Medinc) and ln (Unemp).

From residual plots (figure 7), residuals vs fitted plot and Scale-location plots shows that there exists homoscedasticity as the residuals spread randomly. Residuals vs fitted plot also shows there the linearity assumption is met. Normality of the residuals seems fairly OK, although with some deviation at both tails. Plot of Cook's distance shows no overly influential points. All the assumptions of linear regression are met implies that the data fits perfectly in the model.

The fitted model is

E[ln(Medinc)] =11.26561 – [0.31875 × ln(Unemp)]+ [0.19036 x I(Type=1)]– [0.10925 × ln(Unemp) × I(Type=1)]

Conclusion: The above fitted model seems a reasonable description of the data. The interaction between ln(unemp) and metro is significant implies metro here is playing a significant role as different metro will have different effect on ln(medinc). It seems that that median household income decreases with unemployment rate interact with whether is metropolitan or not. One unit increase in ln(unemp) with non-metropolitan will have decrease of **0.31875** in ln(medinc). Also, one unit increase in ln(unemp) with non-metropolitan will have decrease of **0.428** in ln(medinc).

However, the coefficient of determination, **R^2 =0.2122** is low in this model implies that the model only explains **21.22%** of the variation in the data but it doesn't determine whether the regression model is suitable so it is not a big issue. From residual plots (figure 7), the residuals are randomly spread in residuals vs fitted plot and scale-location plot implies that the least-squares assumption is correct.

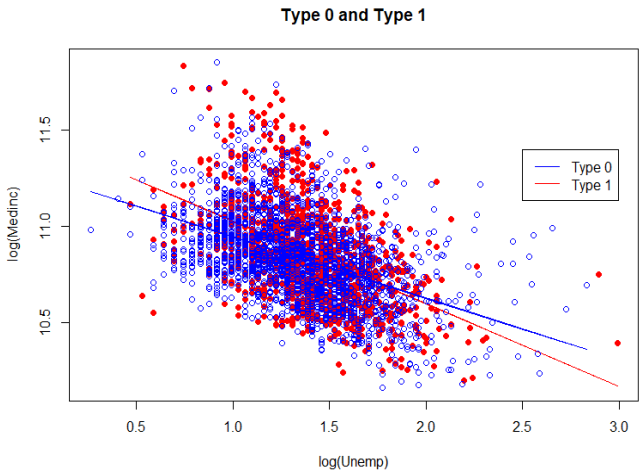

Figure 6: ln(Medinc) vs ln(Unemp) and fitted lines for type 0 (blue) and type 1 (red)

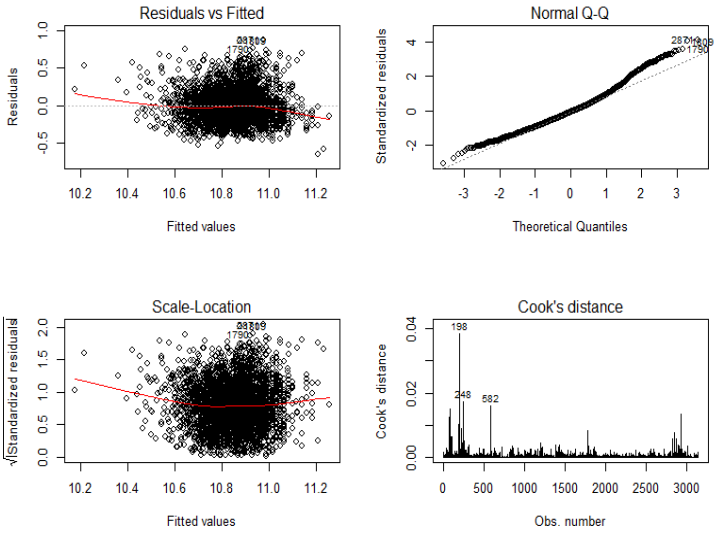

Figure 7: Plots of residuals for the regression of ln(Medinc) on ln(Unemp), Metro and their interaction term

3. For a Poisson response (Eruptions), the default link is log. Fitting the Poisson generalized linear model with log link, 2 explanatory variables (Elevation and type), and no interaction terms (Model 1) gives the following summary output.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.6824027  0.4248692  -1.606    0.108
Elevation                0.0001407  0.0002024   0.695    0.487
TypefactorStratovolcano  0.3631592  0.4411470   0.823    0.410

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.858  on 29  degrees of freedom
Residual deviance: 27.295  on 27  degrees of freedom
AIC: 71.788
```

The individual hypothesis tests suggest that both of the explanatory variables are not significant as p-value for elevation is **0.487** and p-value for type is **0.410**. The different between null deviance and residual deviance is small may implies that the response is not related to the covariate but we will check about it later.

A new model (Model 2) will be fitted with Type only.

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.5108     0.3333  -1.532    0.125
TypefactorStratovolcano  0.4418     0.4272   1.034    0.301

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.858  on 29  degrees of freedom
Residual deviance: 27.762  on 28  degrees of freedom
AIC: 70.255
```

Change in deviance is $D_{28} - D_{27} = 27.762 - 27.295 = \mathbf{0.467}$, to be compare with X_1^2, giving p-value, $p = \Pr(X_1^2 \geq 0.467) = \mathbf{0.494}$ which is not significant at 5% level implies that Elevation does not improve fit.

Another new model (Model 3) will be fitted with Elevation only.

```
Coefficients:
             Estimate       Std. Error     z value        Pr(>|z|)
(Intercept)  -0.5434214     0.3750230      -1.449         0.147
Elevation    0.0001818      0.0001910      0.952          0.341

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.858  on 29  degrees of freedom
Residual deviance: 27.987  on 28  degrees of freedom
AIC: 70.48
```

Change in deviance is $D_{28} - D_{27} = 27.987 - 27.295 = \mathbf{0.692}$, to be compare with X_1^2, giving p-value, $p = \Pr(X_1^2 \geq 0.692) = \mathbf{0.4055}$ which is not significant at 5% level implies that Elevation does not improve fit.

Comparing AIC values, Model 2 has the lowest AIC value (**70.255**), indicating it is the best fit to data among the models.

Comparing model 2 with the null model, have null deviance $D_{29} = \mathbf{28.858}$ and residual deviance $D_{28} = \mathbf{27.762}$. Change in deviance is $D_{29} - D_{28} = 28.858 - 27.762 = \mathbf{1.096}$, to be compare with X_1^2, giving p-value $p = \Pr(X_1^2 \geq 1.096) = \mathbf{0.2951}$. For 5% significant level, there is no evidence that Model 2 fits better than null model.

Model 2 is better than Model 1 and Model 3. However, null model is better than Model 2. Null model is preferred for the data.

Fitting intercept only into a new model (null model), giving following summary output.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2657     0.2085  -1.274    0.203

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.858  on 29  degrees of freedom
Residual deviance: 28.858  on 29  degrees of freedom
AIC: 69.351

Number of Fisher Scoring iterations: 5
```

In null model, null deviance is same with residual deviance, D_{29}= 28.858.

Comparing all models' AIC values, null model has the lowest AIC value (**69.351**), indicating it is the best model.

The preferred fitted model (null model) for λ is

$$Y \sim Poisson(\lambda)$$

$$\boldsymbol{\lambda = \exp\left(-0.2657\right)}$$

Conclusion: There is no relationship between eruptions and elevation+type nor elevation and type individually. In all models, they show that the coefficient of intercept and explanatory variables are not significant.

Appendix

Rcode

```
 #Q1 A
Pension=na.omit(read.delim("OECD_Pensions.txt")) #read the data without missing data
 Bnf=Pension$Benefits #response variable
 Ast=Pension$Assets  #explanatory variable

 example1.a=lm(Bnf ~ Ast) #fit in linear regression model
 summary(example1.a) #summary of the model
 plot(Ast,Bnf,ylab="Benefits",xlab="Assets") # create plot of Benefits against Assets
 abline(example1.a)       #add regression line

 plot(cooks.distance(example1.a),ylab="Cooks's distance",xlab="Observation number",main="Cook'
s distance") #make a cook's distance plot
 text(26,max(cooks.distance(example1.a)),labels=Pension$Country[26],pos=2) #label the influential
point

 example1.a.new=lm(Bnf[-26] ~ Ast[-26]) #fit in linear regression model without observation 26
 summary(example1.a.new)   #summary of the model
 par(mfrow=c(2,2))
 plot(example1.a.new,which=1:4)  #create residual plots

 #Q1 B
 example1.b=lm(log(Bnf) ~ log(Ast))  #fit in linear regression model
 summary(example1.b)     #summary of the model
 plot(log(Ast),log(Bnf),xlab="ln(Assets)",ylab="ln(Benefit)") # plot of ln(Benefits) against ln(Assets)
 abline(example1.b)  #add regression line
 par(mfrow=c(2,2))
 plot(example1.b,which=1:4)  #create residual plots

 #Q1 D
 Pension.new=read.delim("OECD_Pensions.txt") #read the data
 Assets=Pension.new$Assets
 n=length(log(Assets))
 muhat = mean(log(Assets))   #mean of the data
 sigmahat = sd(log(Assets))  #standard deviation of the data
 OECD.a= qnorm(seq(0,1,by=1/6), mean=muhat, sd=sigmahat) #equal probability intervals
 expected.a=n/6  #expected frequency
 observed.a = hist(log(Assets), breaks=OECD.a, plot=F, right=F)$counts   #observed frequecy
 Xsq.a = sum((observed.a-expected.a)^2/expected.a); Xsq.a
 pval.a = 1 - pchisq(Xsq.a,6-2-1); pval.a    #p-value

 OECD.b= qnorm(seq(0,1,by=1/5), mean=muhat, sd=sigmahat)   #equal probability intervals
 expected.b=n/5    #expected frequency
 observed.b = hist(log(Assets), breaks=OECD.b, plot=F, right=F)$counts   #equal probability intervals
 Xsq.b = sum((observed.b-expected.b)^2/expected.b); Xsq.b
 pval.b = 1 - pchisq(Xsq.b,5-2-1); pval.b    #p-value

#####################################################################
##Q2 A
Counties=read.delim("USA_counties.txt") #read the data
```

```
Medinc=Counties$Medinc #response varible
Unemp=Counties$Unemp #explanatory variable
Metro=Counties$Metro #explanatory variable
Metrofactor=factor(Metro)   #creata a factor for metro
example2.1=lm(log(Medinc) ~ log(Unemp) + Metrofactor)  #fit in linear regression model without int
eraction term
example2.2=lm(log(Medinc) ~ log(Unemp) * Metrofactor)   #fit in linear regression model with intera
ction term
summary.aov(example2.2) #produce analysis of variance summary
summary(example2.1)     #summmary of the model without interaction term
summary(example2.2)     #summmary of the model with interaction term

par(mfrow=c(1,1))
plot(log(Unemp) ,log(Medinc), main="Type 0 and Type 1",type="n") #plot of ln(Unemp) against ln(M
edinc)
points(log(Medinc)[Metro=="0"]~log(Unemp)[Metro=="0"],col="blue")   #type 0 points
lines(fitted(example2)[Metro=="0"]~log(Unemp)[Metro=="0"],col="blue") #regression line for type 0
points(log(Medinc)[Metro=="1"]~log(Unemp)[Metro=="1"],col="red",pch=19) #type 1 points
lines(fitted(example2)[Metro=="1"]~log(Unemp)[Metro=="1"],col="red")  #regression line for type 1
legend(2.5,11.4,legend=c("Type 0","Type 1"),col=c("blue","red"),lty=1:1)

par(mfrow=c(2,2))
plot(example2.2,which=1:4) #residual plots for model 2

#####################################################################
##Q3
Volcanoes=read.delim("volcanoes.txt") #read the data
Eruptions=Volcanoes$Eruptions #Poisson response variable
Elevation=Volcanoes$Elevation  #explanatory variable
Type=Volcanoes$Type #explanatory variable
Typefactor=factor(Type) #create a factor for type
example3.a=glm(Eruptions~Elevation+Typefactor,family=poisson) #fit in generalized linear model wi
th elevation+type
example3.b=glm(Eruptions~Typefactor,family=poisson)   #fit in generalized linear model with type
example3.c=glm(Eruptions~Elevation,family=poisson)    #fit in generalized linear model with elevatio
n
example3.null=glm(Eruptions~1,family=poisson)        #fit in generalized linear model with intercept o
nly
summary(example3.a) #summary of the model with elevation+type
summary(example3.b) #summary of the model with type
summary(example3.c) #summary of the model with elevation
summary(example3.null)  #summary of the model with intercept only
```

Reference

Selva Prabhakaran, Linear Regression, http://r-statistics.co/Linear-Regression.html

Kumar Rohit Malhotra,27 September 2018, Linear regression: Modeling and Assumptions

,https://towardsdatascience.com/linear-regression-modeling-and-assumptions-dcd7a201502a

Jim Frost, How To Interpret R-squared in Regression Analysis,
https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Jim Frost, Understanding Interaction Effects in Statistics,
https://statisticsbyjim.com/regression/interaction-effects/

Matthias Döring, 09 November 2018, Interpreting Generalized Linear Models

, https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/

KAREN GRACE-MARTIN, The Distribution of Independent Variables in Regression Models,
https://www.theanalysisfactor.com/the-distribution-of-independent-variables-in-regression-models/

Gaurav Bansal, What are the four assumptions of linear regression?,
https://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/