Stats 762 Assignment 2

Stephen Wang 173417367

01/04/2021

Question 1

```
afghan.df = read.table("afghan.data", row.names=1, header=T)
afghan.df$pashtun<-factor(afghan.df$pashtun)
afghan.glm2nb <- glm.nb(formula = incidents ~ calories + pashtun + troops + offset(log(pop)), data = af
summary(afghan.glm2nb)
```

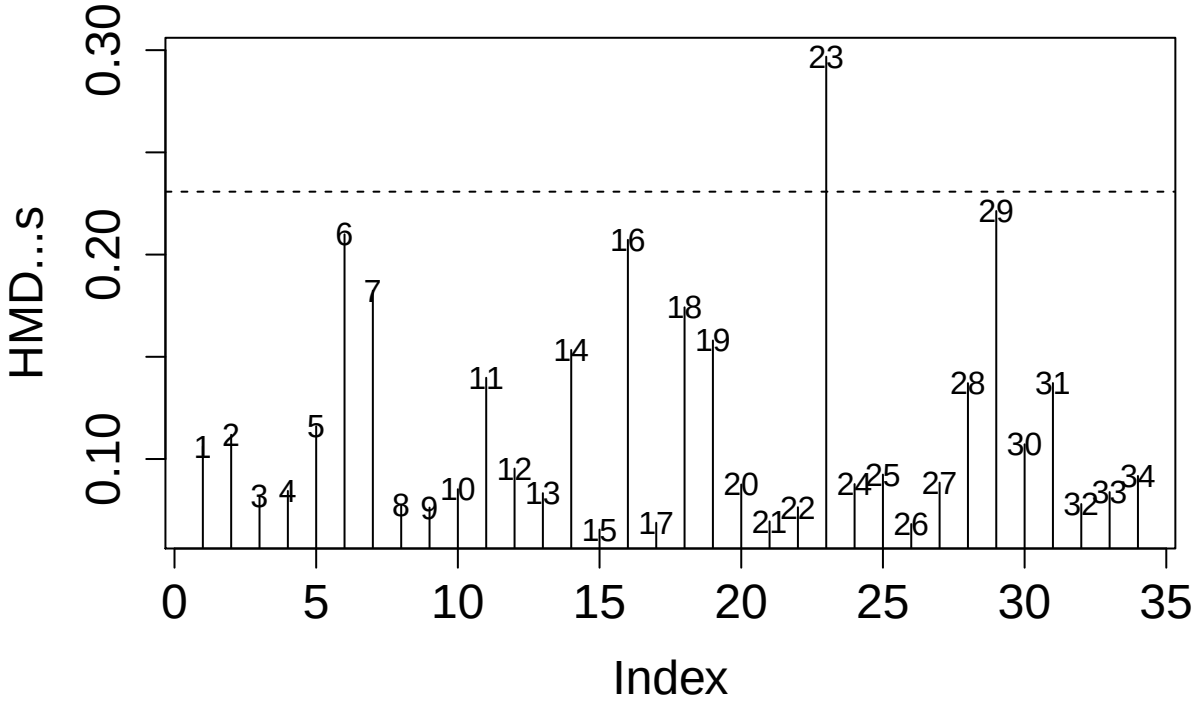
```
##
## Call:
## glm.nb(formula = incidents ~ calories + pashtun + troops + offset(log(pop)),
##     data = afghan.df, init.theta = 1.590287141, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9970  -1.0560  -0.4750   0.3728   2.6028
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.809e+00  4.272e-01 -11.256  < 2e-16 ***
## calories     2.566e-02  1.176e-02   2.181   0.0292 *
## pashtun1     1.806e+00  3.194e-01   5.656 1.55e-08 ***
## troops       8.089e-05  9.911e-05   0.816   0.4144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.5903) family taken to be 1)
##
##     Null deviance: 75.425  on 33  degrees of freedom
## Residual deviance: 36.312  on 30  degrees of freedom
## AIC: 302.88
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.590
##           Std. Err.:  0.388
##
##  2 x log-likelihood:  -292.878
```

```
HMD<-hatvalues(afghan.glm2nb)
plot(HMD,ylab="HMD's",type="h", cex=1.5,cex.axis=1.5, cex.lab=1.5)
```

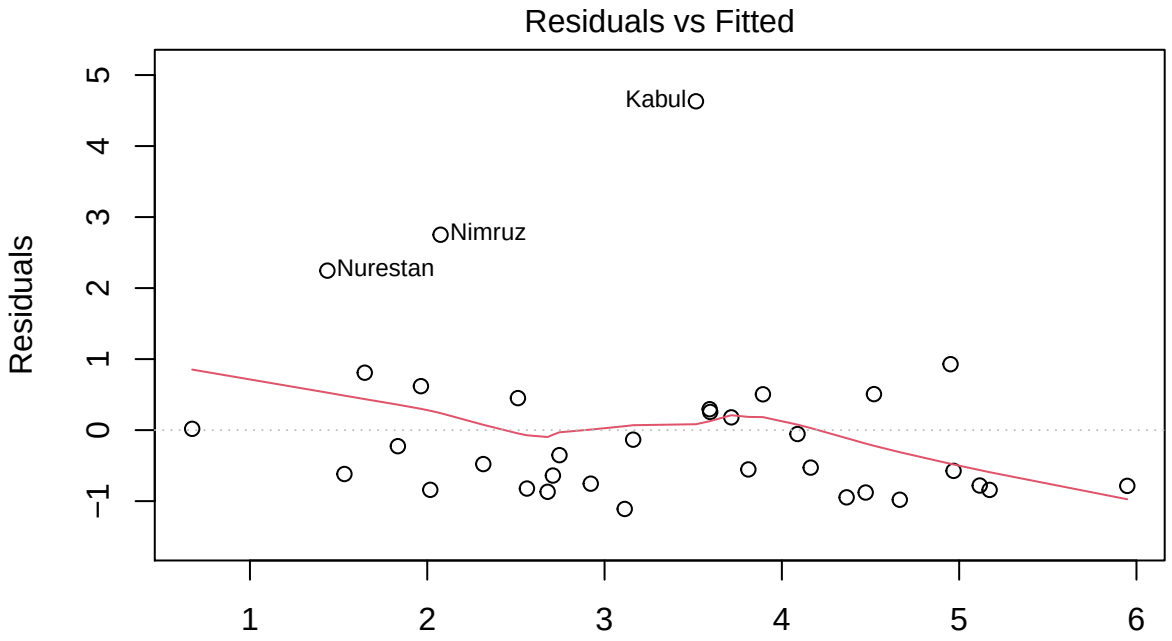
```
## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
```

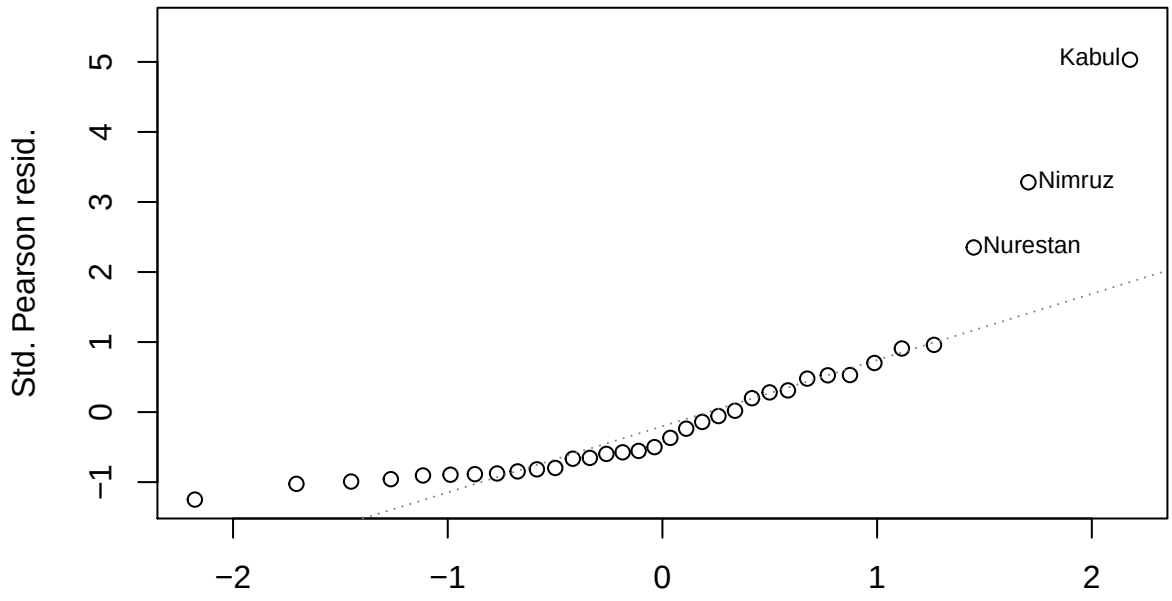
```
## substituted for <e2>

## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
## substituted for <99>
```

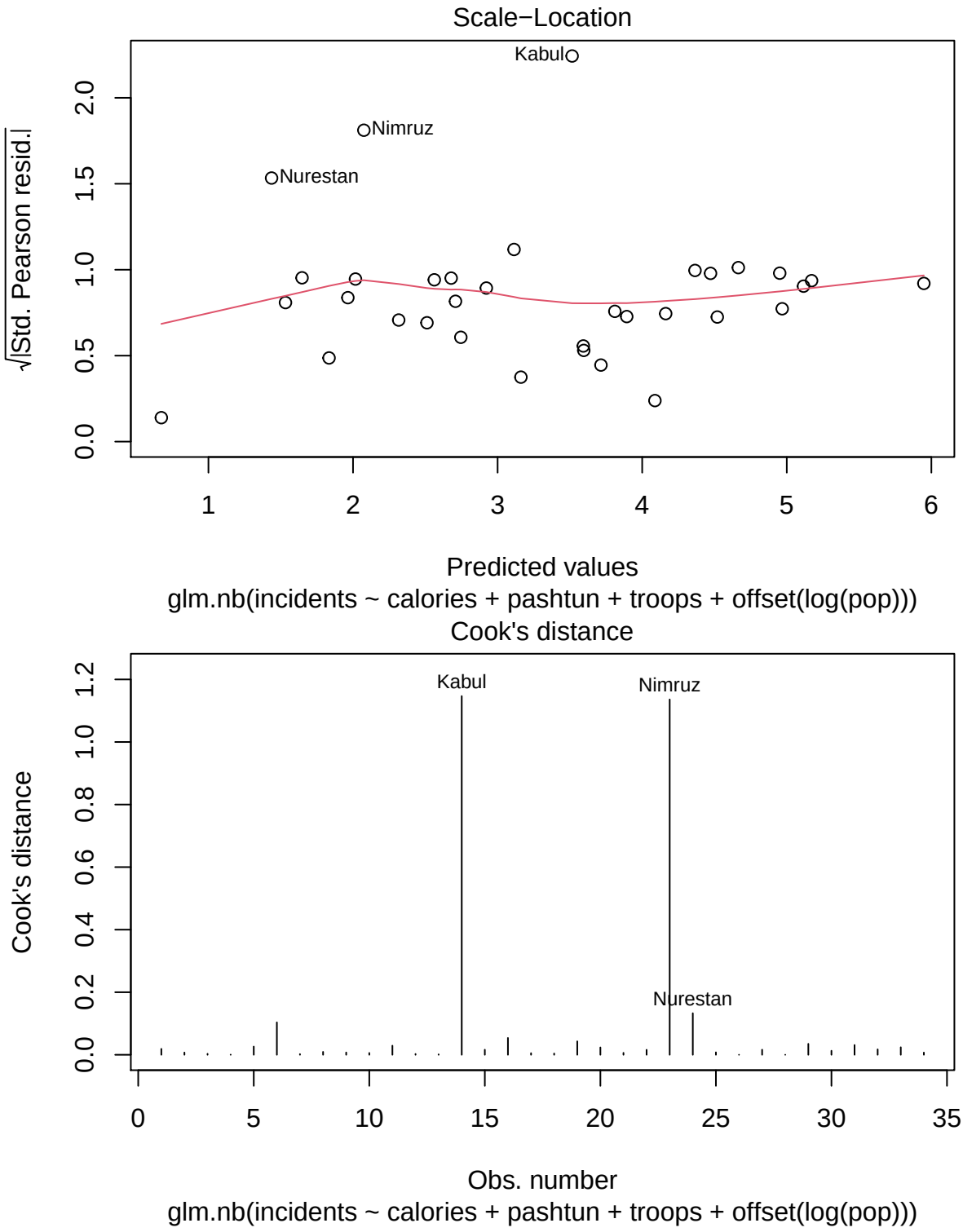
```
text(HMD)
abline(h=3*3/39, lty=2)
```



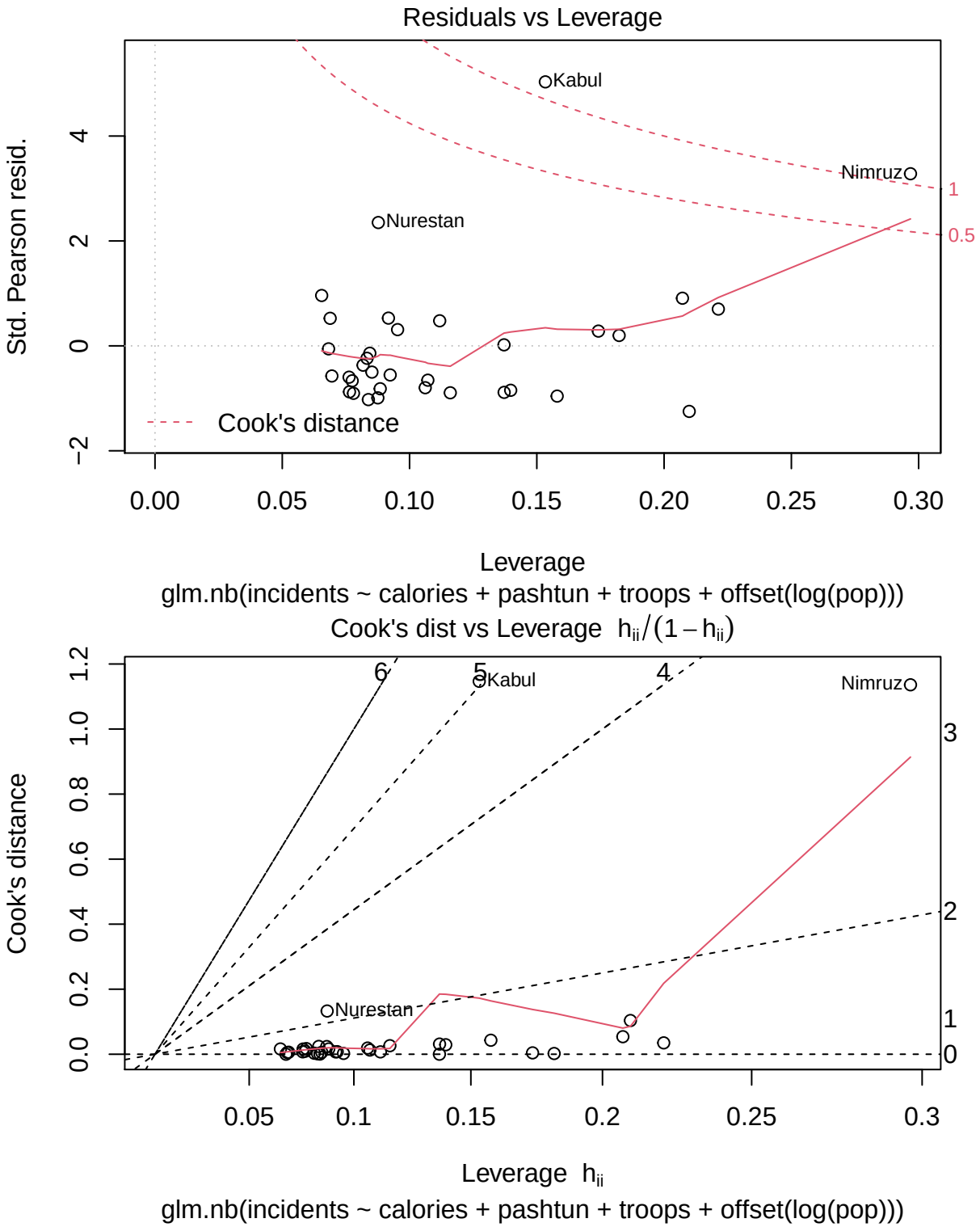

```
plot(afghan.glm2nb, which=1:6)
```


Residuals vs Fitted
Residuals
Kabul
Nimruz
Nurestan
Predicted values
glm.nb(incidents ~ calories + pashtun + troops + offset(log(pop)))


Normal Q−Q
Std. Pearson resid.
Kabul
Nimruz
Nurestan
Theoretical Quantiles
glm.nb(incidents ~ calories + pashtun + troops + offset(log(pop)))


Scale−Location
Kabul
Nimruz
Nurestan
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
6
Predicted values
glm.nb(incidents ~ calories + pashtun + troops + offset(log(pop)))

Cook's distance
Kabul
Nimruz
Nurestan
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
35
Obs. number
glm.nb(incidents ~ calories + pashtun + troops + offset(log(pop)))



Residuals vs Fitted: at the far left, the red lowess fit goes from 1 to 0, hovers around the horizontal band around residuals=0, until moving down to the far right where residual=-1. This may indicate non-constant error variance. Furthermore, there are three data points (Nurestan, Nimruz and Kabul) which stand out from the pattern of residuals.

Cook's Distance: confirms there are two data points, Kabul and Nimruz, with high levels of influence (cook's distance>1) which we need to take into consideration. Although Nurestan has been standing out from the rest

of the diagnostic plots, the Cook's Distance suggest that it is not a data point with high levels of influence.

Residuals vs Leverage: majority of data points are well situated in the "OK" region but there are two points, Kabul and Nimruz, which are high-leverage and potentially high-leverage outliers respectively. Possible indication of a few more high leverage outliers in the bottom right of the graph from 0.15-0.22.

Normal QQ plot: there is a light tail but the majority of observations fit the QQ line. There are three data points (Nurestan, Nimruz and Kabul) that are have a severe gap at the higher end of the quantile which indicates these have more extreme values than expected from a Normal distribution.

HMD: the data point 23 (Nimruz) has the largest weight of 0.30 but we only need to be concerned of any high leverage data points with weight greater than $3(k+1)/n=3(4+1)/34=0.44$ where $k=4$ and $n=34$.

Scale Location: the red line is approximately horizontal which confirms homoskedasticity.

Question 2

```
1-pchisq(36.312, 30)
```

```
## [1] 0.1980393
```

Null hypotheses: the fitted model adequately explains the response/observed data. The goodness of fit test returned a high p-value (>0.05) which supports the null hypotheses and indicates that the negative binomial model adequately explains the response.

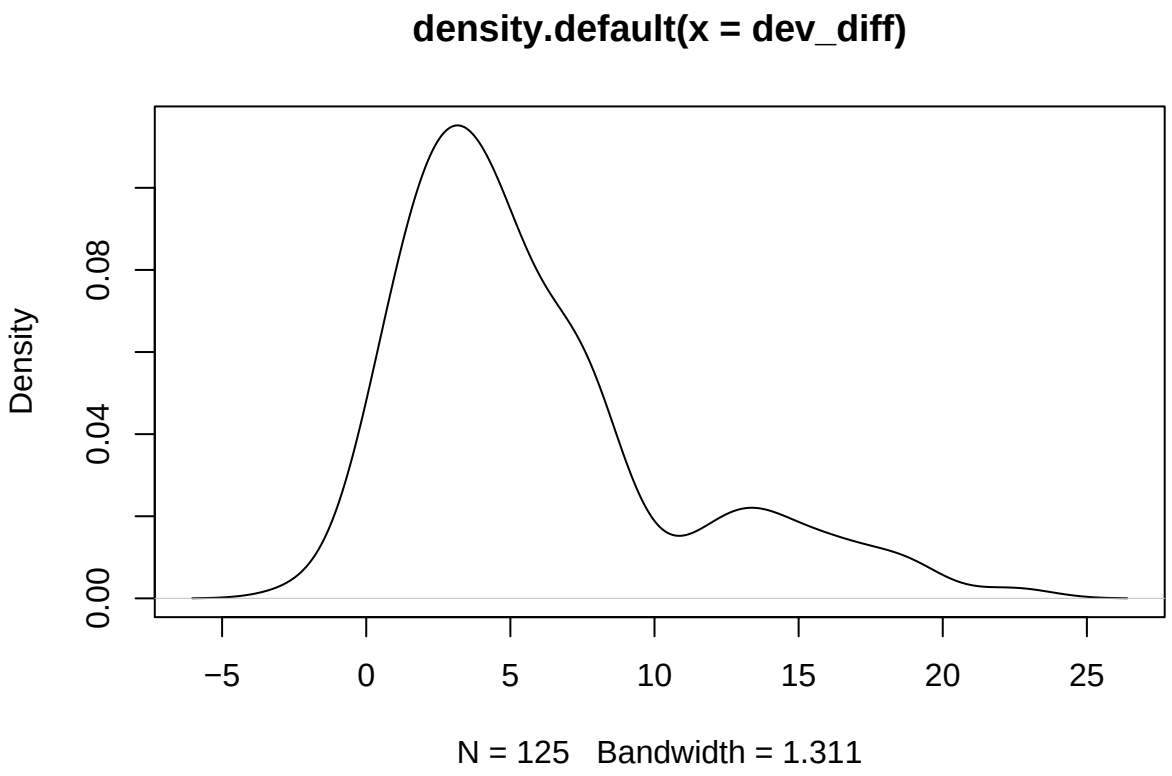
Question 3

```
set.seed(1)
bootstrap=function(){
calories.new=sample(afghan.df$calories)
pashtun.new=sample(afghan.df$pashtun)
troops.new=sample(afghan.df$troops)
pop.new=sample(afghan.df$pop)
#
new.glm<-glm.nb(afghan.df$incidents~calories.new+pashtun.new+troops.new+offset(log(pop.new)), link='log
return(new.glm$null.deviance-new.glm$deviance)
}
#
dev_diff<-replicate(n=125, bootstrap())
```

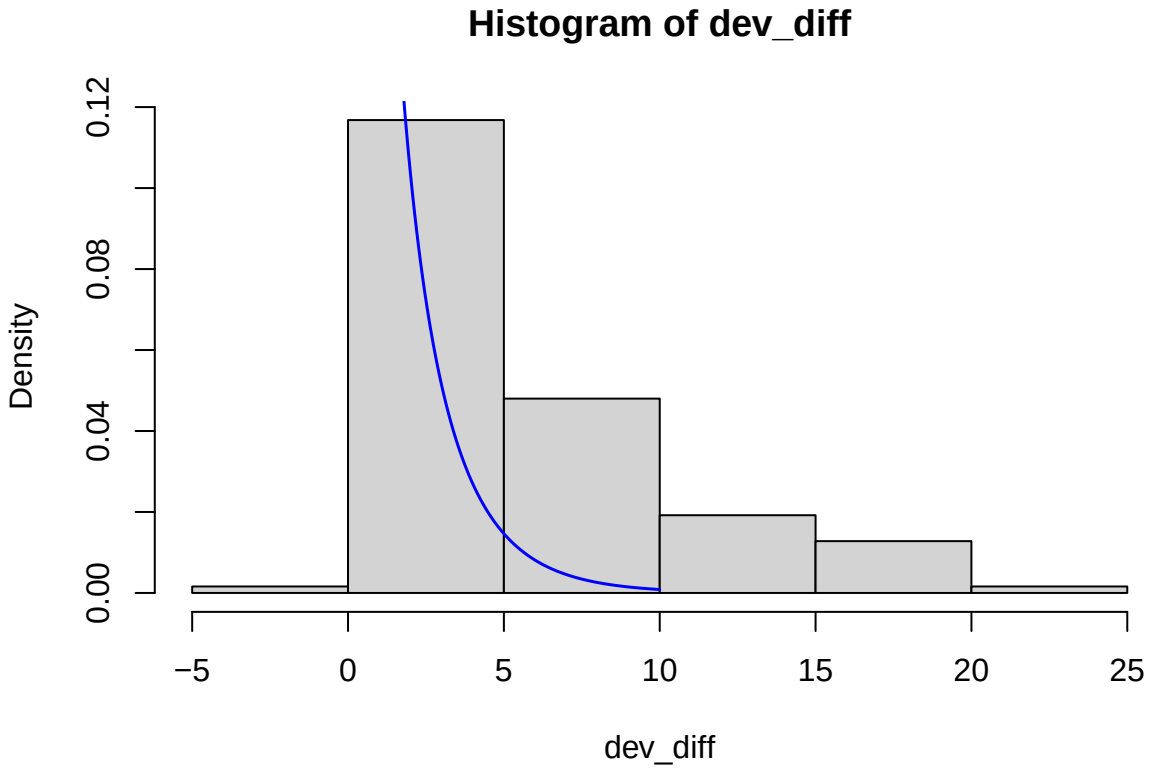
```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge
## Warning in glm.nb(afghan.df$incidents ~ calories.new + pashtun.new + troops.new
## + : alternation limit reached
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge
## Warning in glm.nb(afghan.df$incidents ~ calories.new + pashtun.new + troops.new
## + : alternation limit reached
## Warning: glm.fit: algorithm did not converge
## Warning in glm.nb(afghan.df$incidents ~ calories.new + pashtun.new + troops.new
## + : alternation limit reached
```

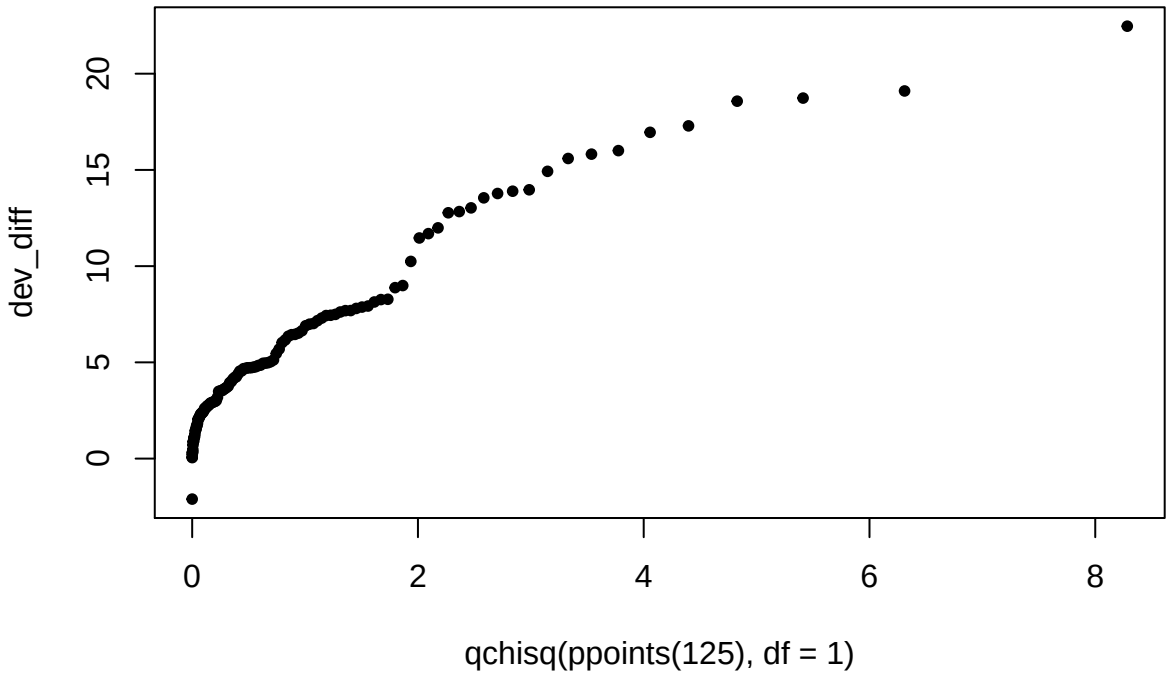
```r
xx<-seq(.01,10,length=500)
yc<-dchisq(xx,1)
plot(density(dev_diff))
```




```
hist(dev_diff,freq=FALSE)
lines(xx,yc,col="blue",lwd=1.5)
```



```
qqplot(qchisq(ppoints(125), df = 1),dev_diff,pch=20)
```

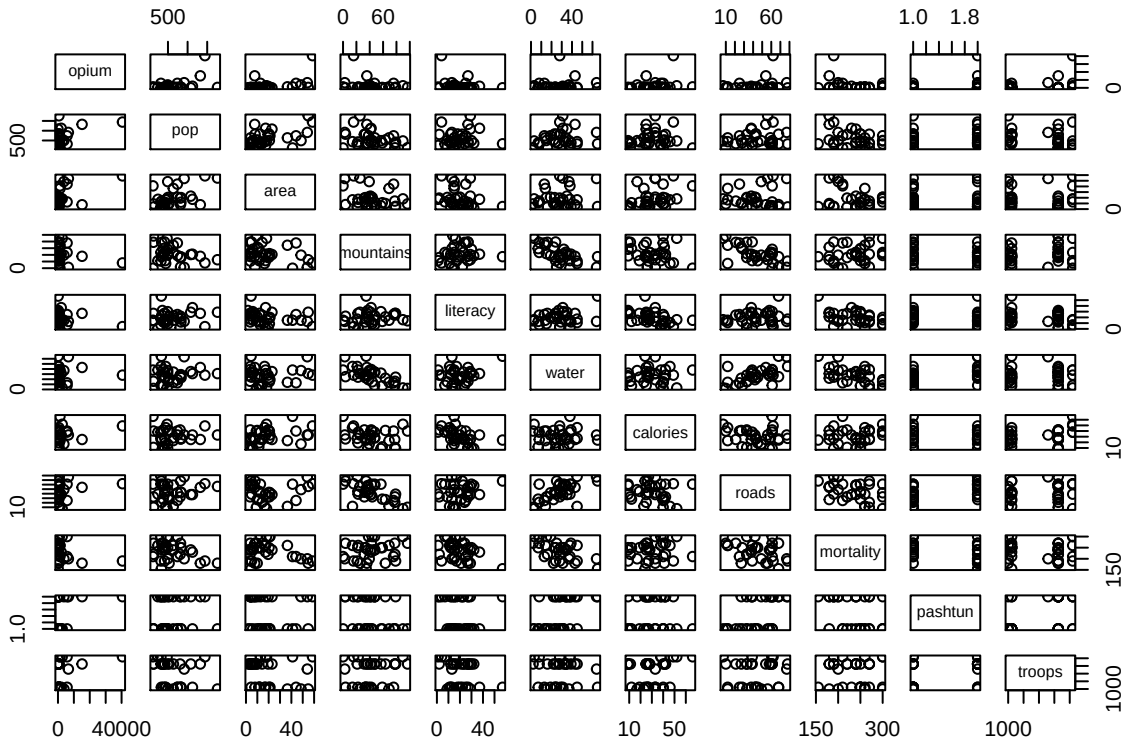


(b)

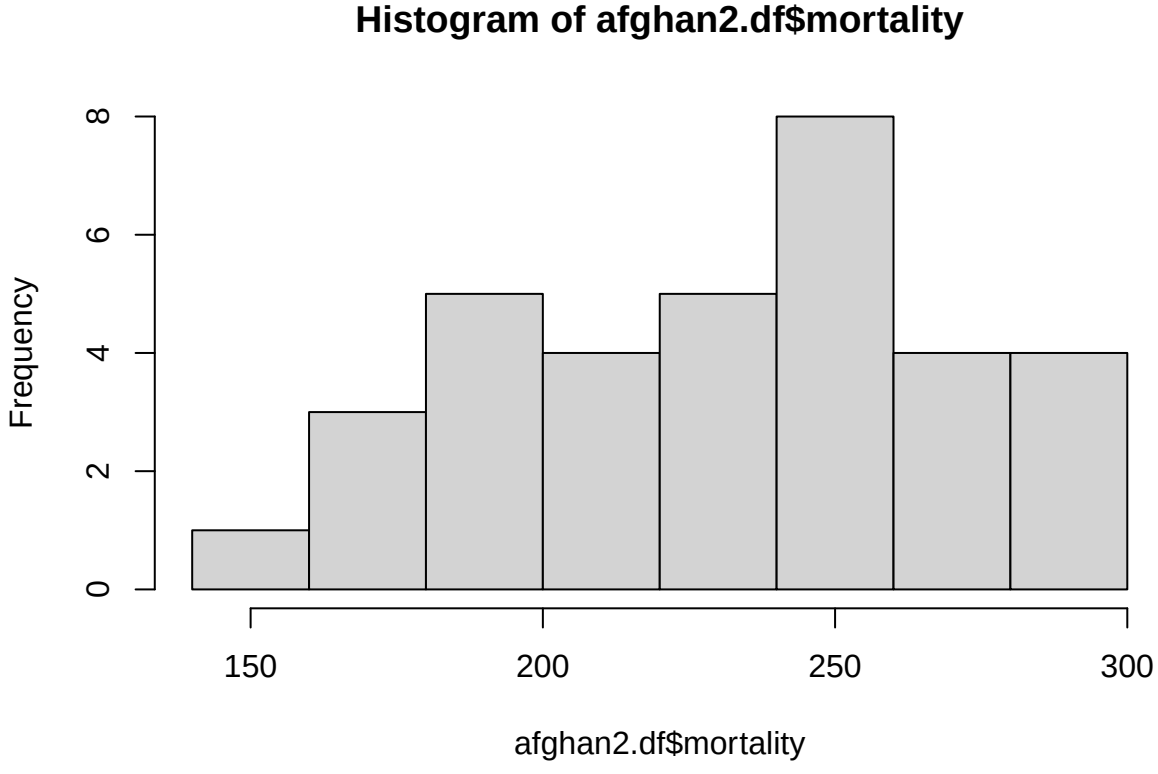
(c) The goodness of fit test determines whether a model is well fitted in regards to a specified distribution. Therefore, non-parametric bootstrap would not be an appropriate method to generate the reference distribution because it re-samples from the original dataset with no assumptions about the distribution of the data.

Question 4

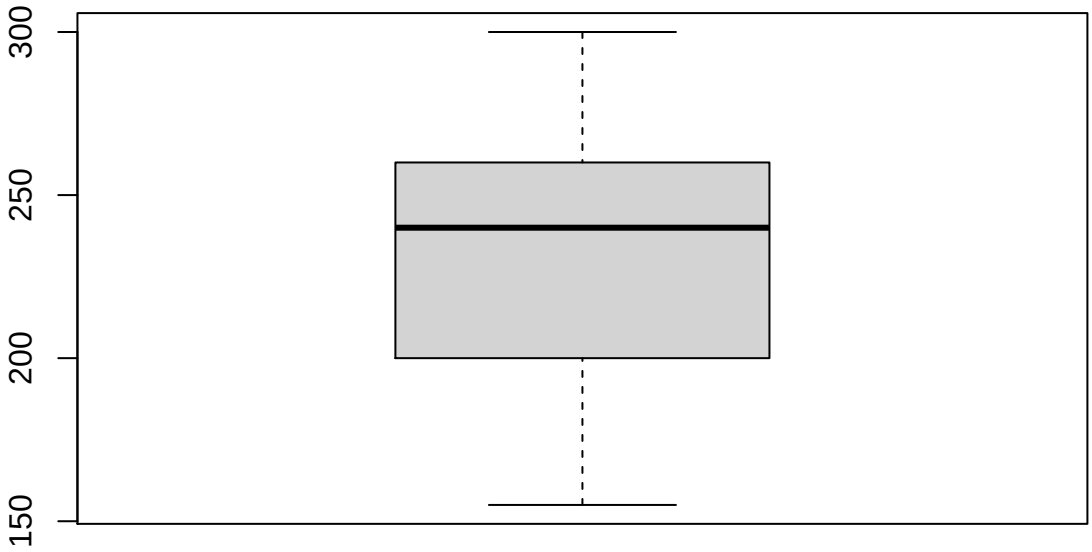
```
afghan2.df = read.table("afghan2.data", row.names=1, header=T)
afghan2.df$pashtun<-factor(afghan2.df$pashtun)
plot(afghan2.df)
```



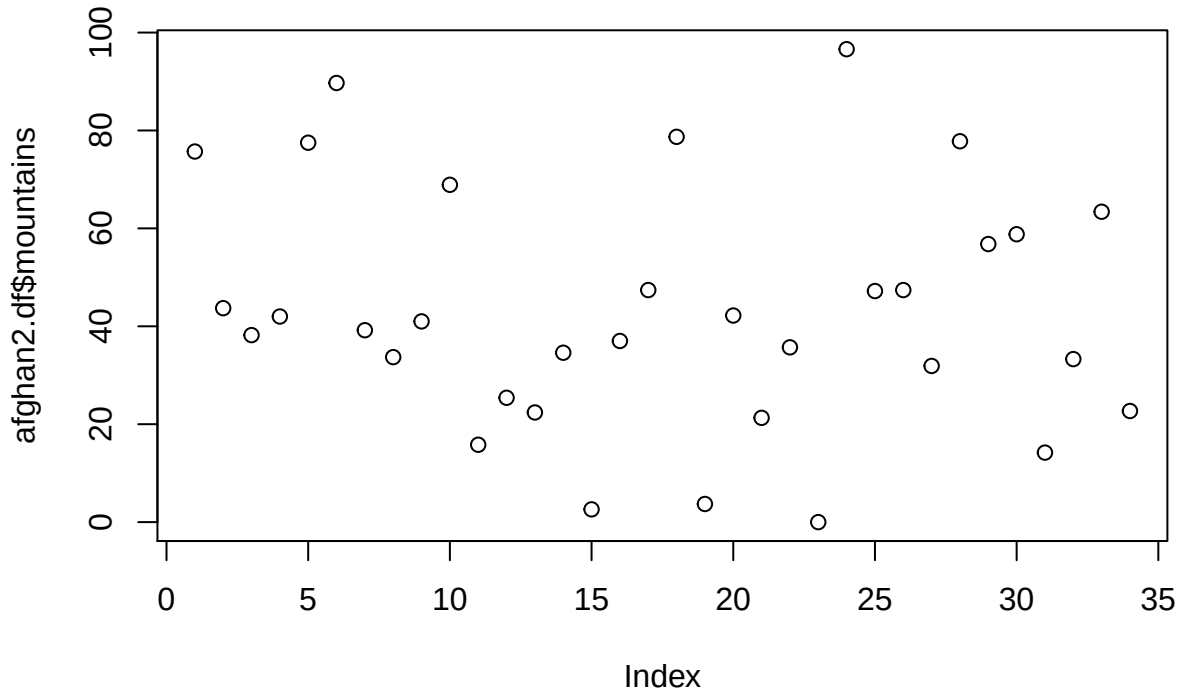

```
hist(afghan2.df$mortality)
```




```
boxplot(afghan2.df$mortality)
```




```
plot(afghan2.df$mountains)
```



From the pairs plot, we can observe the following: - Opium doesn't have any distinct relationship with mortality or any of the other regressors. - Mortality has a slight left skewed distribution with a noticeable cluster between values 225-275. - Majority of the observations of the regressor area is situated between 0-25.

Question 5

```
afghan2.lm<-lm(formula=mortality~opium+pop+area+mountains+literacy+water+calories+roads+pashtun+troops,
summary(afghan2.lm)
```

```
##
## Call:
## lm(formula = mortality ~ opium + pop + area + mountains + literacy +
##     water + calories + roads + pashtun + troops, data = afghan2.df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.149 -17.931   2.785  16.973  42.939
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.206e+02  4.486e+01   7.147 2.81e-07 ***
## opium       -8.802e-04  9.710e-04  -0.907   0.3740
## pop         -2.810e-04  1.873e-02  -0.015   0.9882
## area        -8.601e-01  4.546e-01  -1.892   0.0712 .
## mountains   -2.045e-01  4.607e-01  -0.444   0.6614
## literacy    -1.449e+00  7.267e-01  -1.994   0.0581 .
## water       -1.204e+00  6.460e-01  -1.864   0.0752 .
## calories     4.041e-01  5.231e-01   0.773   0.4477
## roads        2.122e-01  4.960e-01   0.428   0.6728
## pashtun1    -1.805e+01  1.528e+01  -1.181   0.2495
## troops      -2.604e-03  4.436e-03  -0.587   0.5629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.37 on 23 degrees of freedom
## Multiple R-squared:  0.6079, Adjusted R-squared:  0.4374
## F-statistic: 3.566 on 10 and 23 DF,  p-value: 0.005637
```

Question 6

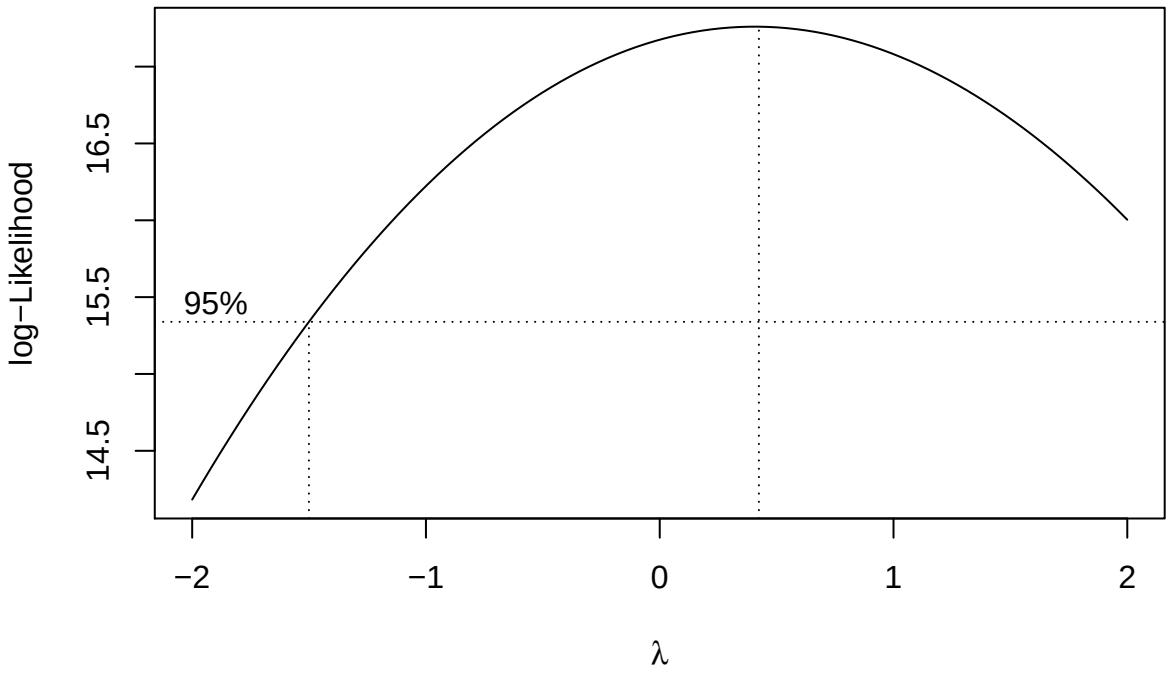
```
Xmat<-model.matrix(afghan2.lm)[,-1]
diag(solve(cor(Xmat)))
```

```
##     opium       pop      area mountains  literacy     water  calories     roads
##  1.924043  1.971880  2.156477  4.947944  2.456503  3.622145  1.833201  3.506678
##  pashtun1    troops
##  2.293227  2.137429
```

As a general rule of thumb, VIF that exceeds 5 indicate a problematic amount of collinearity among regressors. Our dataset doesn't have any VIF that is greater than 5 but we may need to be concerned with "mountains" which has a VIF of 4.95.

Question 7

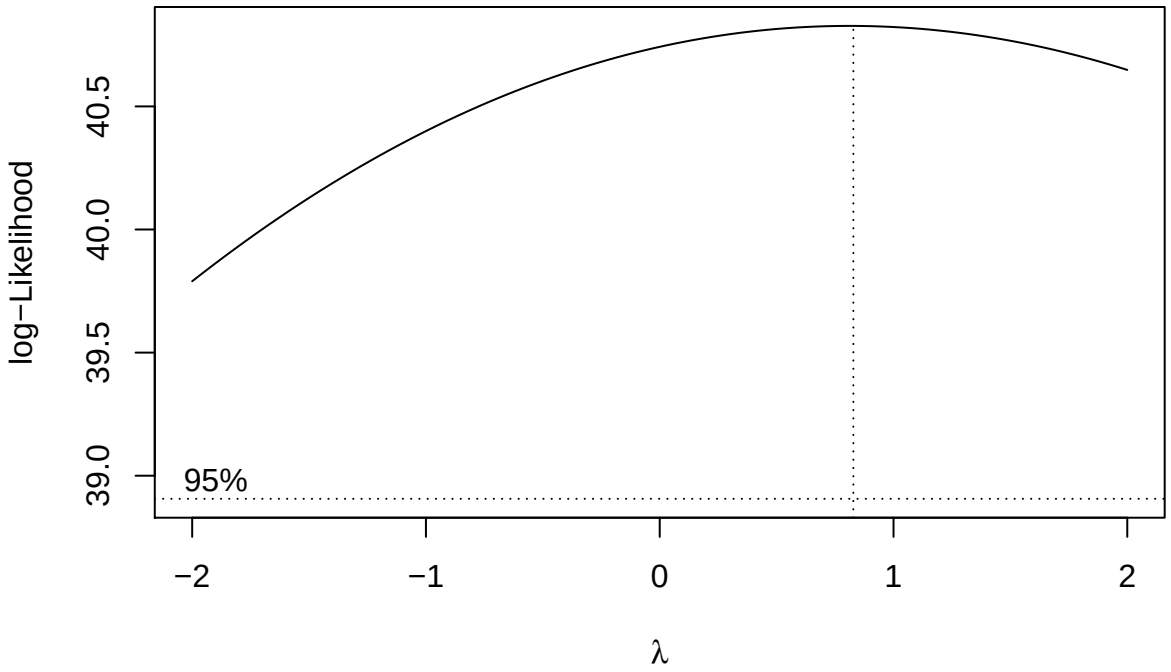
```
bc<-boxcox(lm(mortality~opium+pop+area+mountains+literacy+water+calories+roads+pashtun+troops, data=afgh
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 0.4242424
```

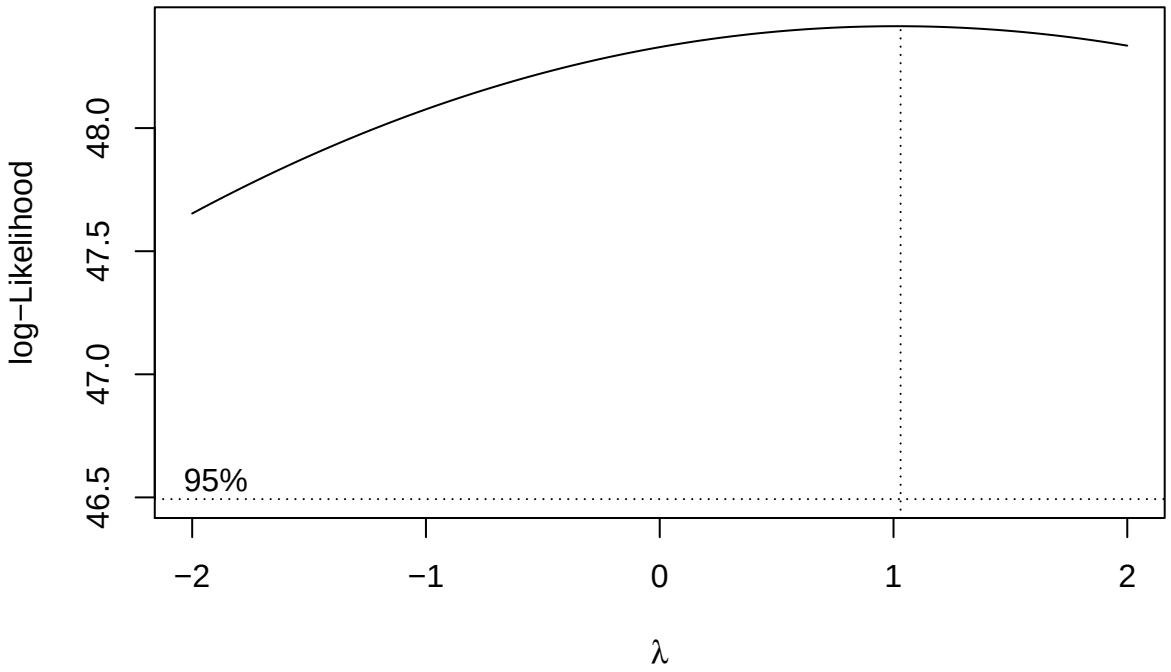
```
bc<-boxcox(lm(mortality^(1/2)~opium+pop+area+mountains+literacy+water+calories+roads+pashtun+troops, da
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 0.8282828
```

```
bc<-boxcox(lm(mortality^(2/5)~opium+pop+area+mountains+literacy+water+calories+roads+pashtun+troops, da
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 1.030303
```

```
afghan3.lm<-lm(formula=sqrt(mortality)~opium+pop+area+mountains+literacy+water+calories+roads+pashtun+t
summary(afghan3.lm)
```

```
##
## Call:
## lm(formula = sqrt(mortality) ~ opium + pop + area + mountains +
##     literacy + water + calories + roads + pashtun + troops, data = afghan2.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82311 -0.53494  0.08622  0.53824  1.44012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.824e+01  1.463e+00  12.467 1.03e-11 ***
## opium       -3.086e-05  3.166e-05  -0.975   0.3399
## pop          1.343e-05  6.107e-04   0.022   0.9826
## area        -2.772e-02  1.482e-02  -1.870   0.0743 .
## mountains   -7.650e-03  1.502e-02  -0.509   0.6155
## literacy    -4.979e-02  2.369e-02  -2.102   0.0467 *
## water       -3.906e-02  2.106e-02  -1.854   0.0765 .
## calories     1.048e-02  1.705e-02   0.614   0.5451
## roads        6.007e-03  1.617e-02   0.371   0.7137
## pashtun1    -6.338e-01  4.982e-01  -1.272   0.2160
## troops      -6.903e-05  1.446e-04  -0.477   0.6377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9575 on 23 degrees of freedom
## Multiple R-squared:  0.6169, Adjusted R-squared:  0.4504
## F-statistic: 3.704 on 10 and 23 DF,  p-value: 0.004537
```

Our boxcox plot returns lambda=0.42 which suggest we transform the response variable to achieve a linear regression for our model. Although we could transform Yˆ(2/5) to get a lambda value closer to 1, we will transform Yˆ(1/2) (which returns lambda=0.83) for the simplicity of this study.

Residual standard error have decreased from 29.37 to 0.9575 which indicates a significantly better fit.

Question 8

```
options(na.action="na.fail")
all.fits<-dredge(afghan3.lm)
```

```
## Fixed term is "(Intercept)"
```

```
head(all.fits)
```

```
## Global model call: lm(formula = sqrt(mortality) ~ opium + pop + area + mountains +
##     literacy + water + calories + roads + pashtun + troops, data = afghan2.df)
## ---
## Model selection table
##     (Intrc)     area    ltrcy    mntns      opium pshtn    water df  logLik
## 38    17.86 -0.03376 -0.06750                        +           5 -43.512
## 550   17.98 -0.02938 -0.05701                        + -0.02050  6 -42.190
## 533   17.42          -0.04369          -5.956e-05       -0.03931  5 -44.389
## 558   18.73 -0.03129 -0.04846 -0.01139                + -0.03593  7 -41.419
## 566   17.94 -0.02284 -0.05746          -2.902e-05     + -0.02274  7 -41.464
## 54    17.81 -0.02908 -0.06874          -2.239e-05     +           6 -43.106
##      AICc delta weight
## 38   99.2  0.00  0.300
## 550  99.5  0.32  0.255
## 533 100.9  1.75  0.125
## 558 101.1  1.98  0.112
## 566 101.2  2.07  0.107
## 54  101.3  2.16  0.102
## Models ranked by AICc(x)
```

The models ranked by dredge suggest that the most important factors to explain mortality are area, literacy, mountain, opium, pashtun and water. However, we need to take into consideratoin only one of the models (558) used the explanatory variable “mountain” to explain mortality.

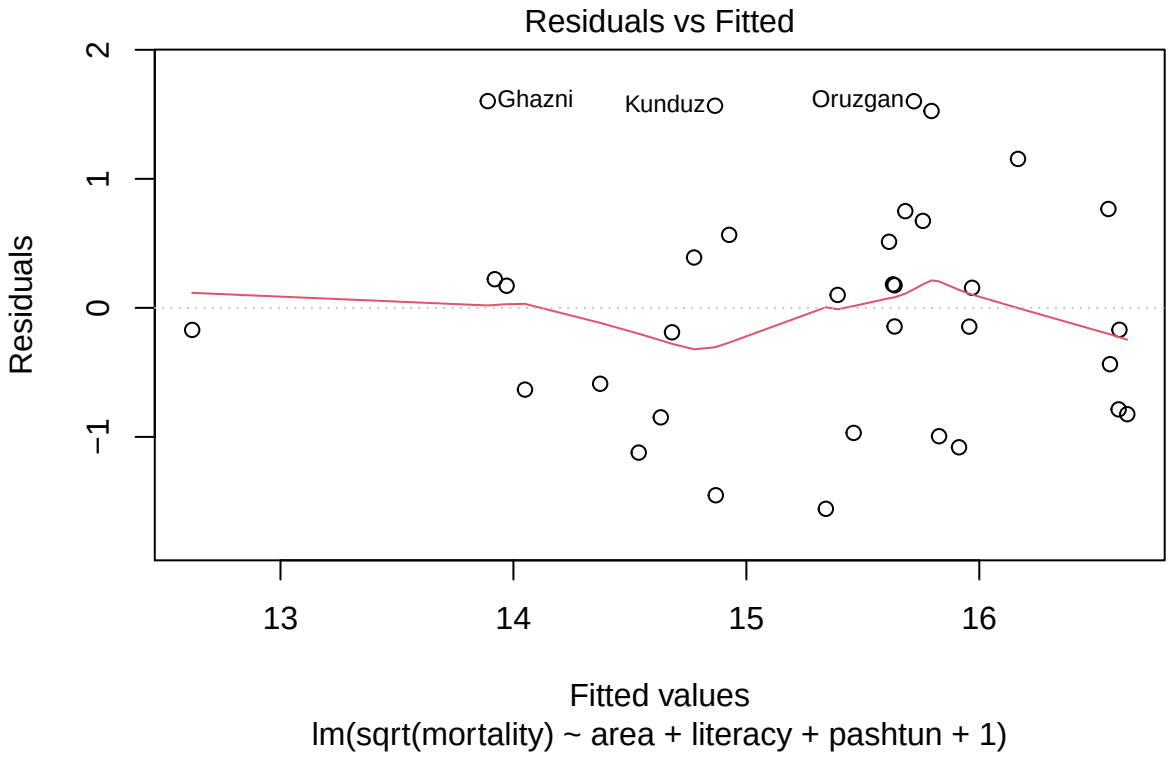
Question 9

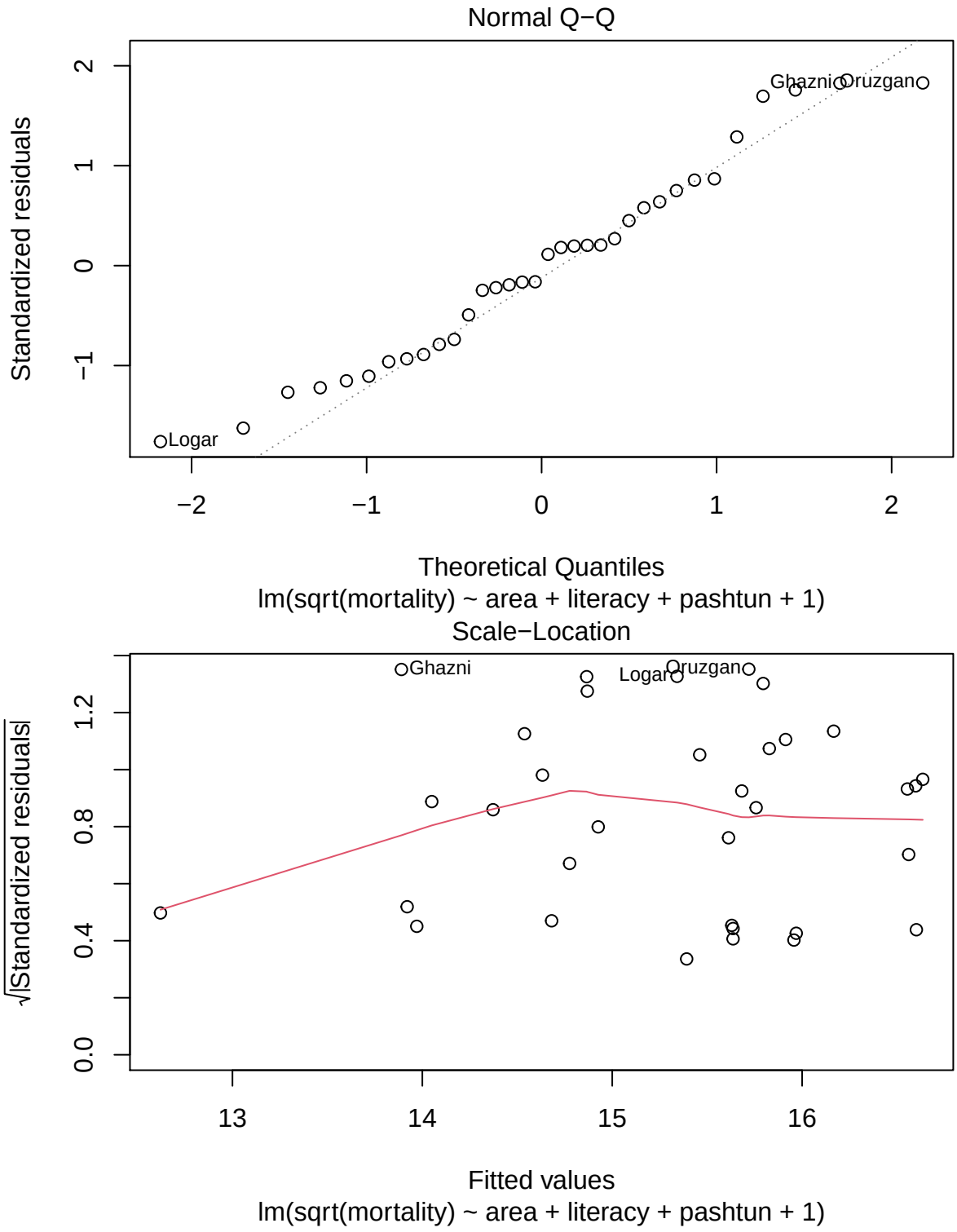
```
best.model<-get.models(all.fits, 1)[[1]]
summary(best.model)
```

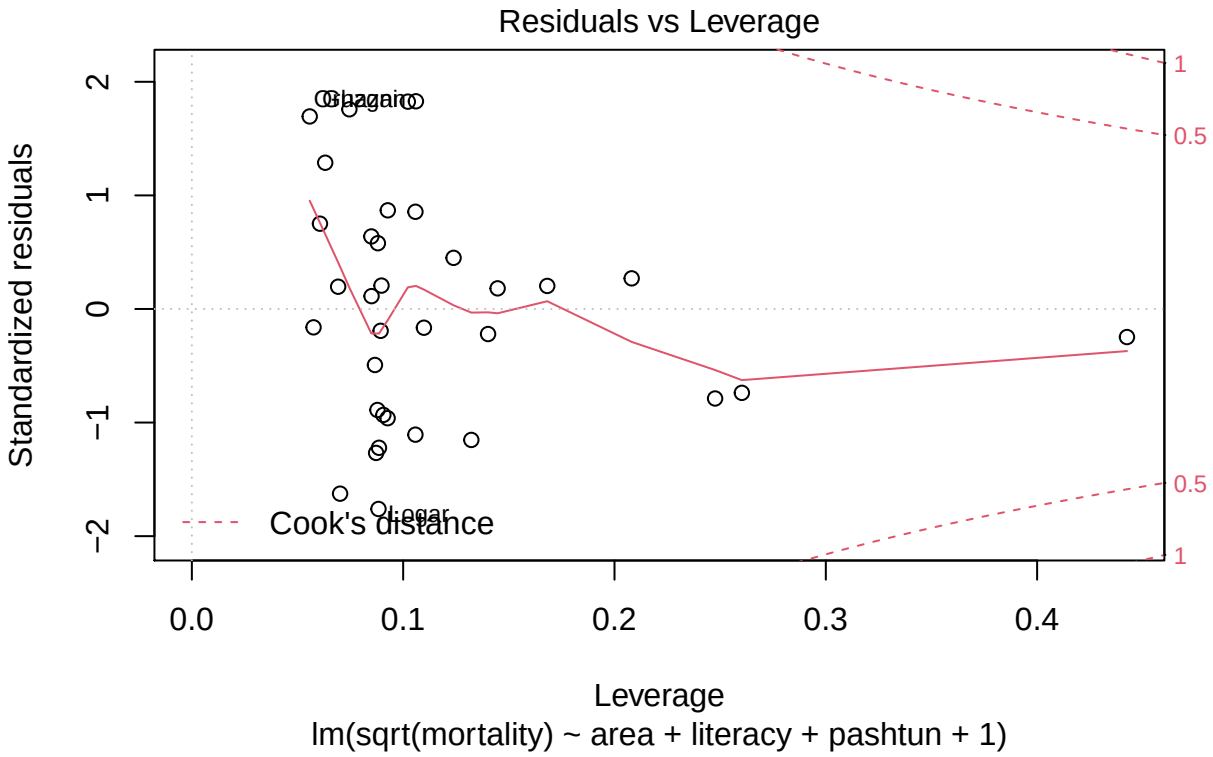
```
##
## Call:
## lm(formula = sqrt(mortality) ~ area + literacy + pashtun + 1,
##     data = afghan2.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.55726 -0.74841 -0.02218  0.55234  1.60206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.85719    0.47424  37.655  < 2e-16 ***
## area        -0.03376    0.01016  -3.322 0.002361 **
## literacy    -0.06750    0.01529  -4.415 0.000121 ***
## pashtun1    -1.23666    0.32199  -3.841 0.000590 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9263 on 30 degrees of freedom
## Multiple R-squared:  0.5324, Adjusted R-squared:  0.4857
## F-statistic: 11.39 on 3 and 30 DF,  p-value: 3.744e-05
```

```
plot(best.model)
```




Normal Q−Q
Standardized residuals
Ghazni
Oruzgan
Logar
−2
−1
0
1
2
Theoretical Quantiles
lm(sqrt(mortality) ~ area + literacy + pashtun + 1)
Scale−Location
√|Standardized residuals|
Ghazni
Logar
Oruzgan
0.0
0.4
0.8
1.2
13
14
15
16
Fitted values
lm(sqrt(mortality) ~ area + literacy + pashtun + 1)



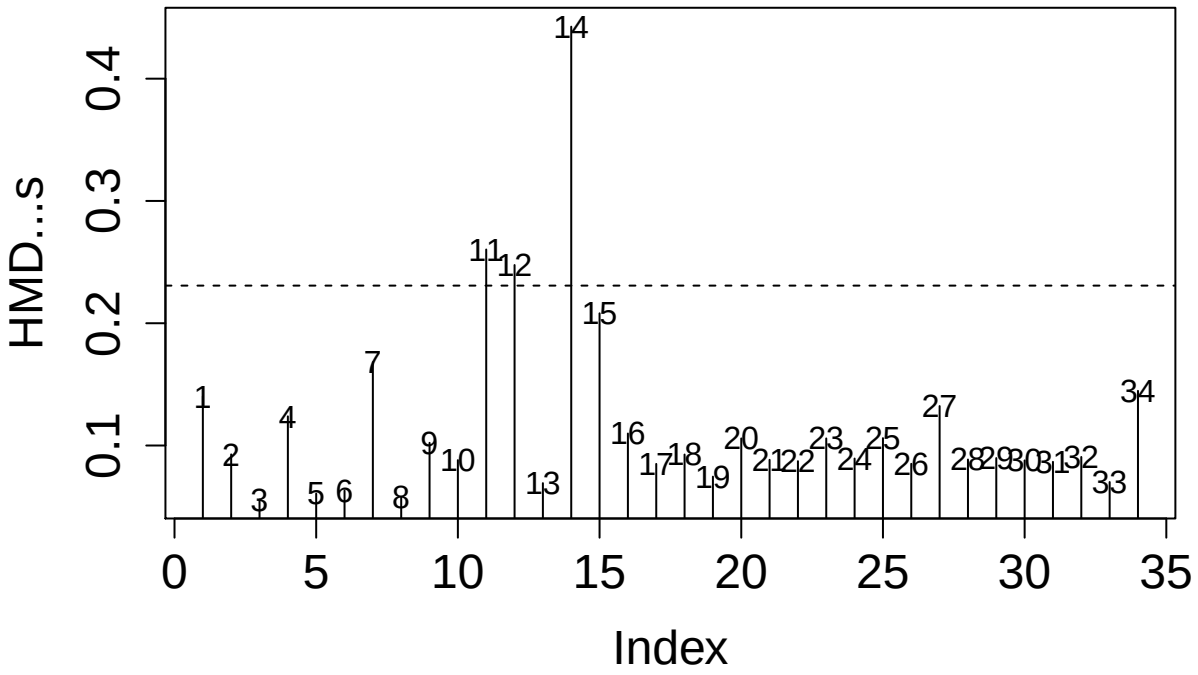

```
HMD<-hatvalues(best.model)
plot(HMD,ylab="HMD's",type="h", cex=1.5,cex.axis=1.5, cex.lab=1.5)
```

```
## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
## substituted for <e2>
```

```
## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
## substituted for <80>
```

```
## Warning in title(...): conversion failure on 'HMD's' in 'mbcsToSbcs': dot
## substituted for <99>
```

```
text(HMD)
abline(h=3*3/39, lty=2)
```



Residuals vs Fitted: observations are scattered roughly around 0 and no extreme residuals stand out from the rest of the data points.

Cook's Distance: there are two observed data points (Ghazni and Oruzgan) with higher influence than the rest of the observations but they only have a cook's distance of around 0.10 so there is no problem.

Residuals vs Leverage: we have a few observations that could be low leverage outliers and one observation that has higher leverage than the rest of the models. Most of the observation cluster can be found within reasonable standardized residuals and leverage levels.

Scale Location: the first observation on the plot has an influence on the red horizontal line. Omitting this data point, the rest of the observations look to be scattered roughly around the red line.

Normal QQ plot: there is a light tail on both ends of the quantiles but most of the observations are situated roughly around the normal distribution line.

HMD: Observation 14 (Kabul) has a value higher than $3(k+1)/n = 3(3+1)/34 = 0.35$ where $k = 3$ and $n = 34$ which indicates that it has high leverage.

Question 10

```
1-pchisq(0.9263, 30)
```

```
## [1] 1
```

The goodness of fit test returned high statistical significance (p-value=1) to suggest the model adequately explains the response variable and the diagnostic plots don't produce any extreme issues to be concerned with. Therefore, we don't need to make any more modifications to the model.

Question 11

Based on the chain rule derivative, the regression model estimates that a unit change in any of the regressors (X_i) will be associated with a change in Y of $2b_i\hat{Y} = 2b_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$.Therefore, we can provide the following findings: - the intercept of the regression model slope is 17.857 and uses pushtun=0 as the baseline - a unit change in area size is associated with a change in mortality of -0.066 times the current mortality, - a unit change in literacy is associated with a change in mortality of -0.134 times the current

mortality, - and if the region is pushtun majority (pushtun=1), there is a change in mortality of -2.472 times the current mortality.