Homework 02

Diptanshu Singh

Septemeber 16, 2018

Introduction

In homework 2 you will fit many regression models. You are welcome to explore beyond what the question is asking you.

Please come see us we are here to help.

Data analysis

Analysis of earnings and height data

The folder `earnings` has data from the Work, Family, and Well-Being Survey (Ross, 1990). You can find the codebook at http://www.stat.columbia.edu/~gelman/arm/examples/earnings/wfwcodebook.txt

```
gelman_dir <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
heights    <- read.dta (paste0(gelman_dir,"earnings/heights.dta"))
```

Pull out the data on earnings, sex, height, and weight.

1. In R, check the dataset and clean any unusually coded data.

```
summary(heights)
```

```
##       earn           height1         height2            sex
##  Min.   :     0   Min.   :4.000   Min.   : 0.000   Min.   :1.000
##  1st Qu.:  6000   1st Qu.:5.000   1st Qu.: 3.000   1st Qu.:1.000
##  Median : 16400   Median :5.000   Median : 5.000   Median :2.000
##  Mean   : 20015   Mean   :5.122   Mean   : 5.186   Mean   :1.631
##  3rd Qu.: 28000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:2.000
##  Max.   :200000   Max.   :6.000   Max.   :98.000   Max.   :2.000
##  NA's   :650      NA's   :8       NA's   :6
##       race            hisp             ed            yearbn
##  Min.   :1.000   Min.   :1.000   Min.   : 2.00   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:12.00   1st Qu.:34.00
##  Median :1.000   Median :2.000   Median :12.00   Median :50.00
##  Mean   :1.187   Mean   :1.953   Mean   :13.31   Mean   :46.98
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:15.00   3rd Qu.:60.00
##  Max.   :9.000   Max.   :9.000   Max.   :99.00   Max.   :99.00
##
##      height
##  Min.   :57.00
##  1st Qu.:64.00
##  Median :66.00
##  Mean   :66.56
```

```
##   3rd Qu.:69.00
##   Max.   :82.00
##   NA's   :8
```

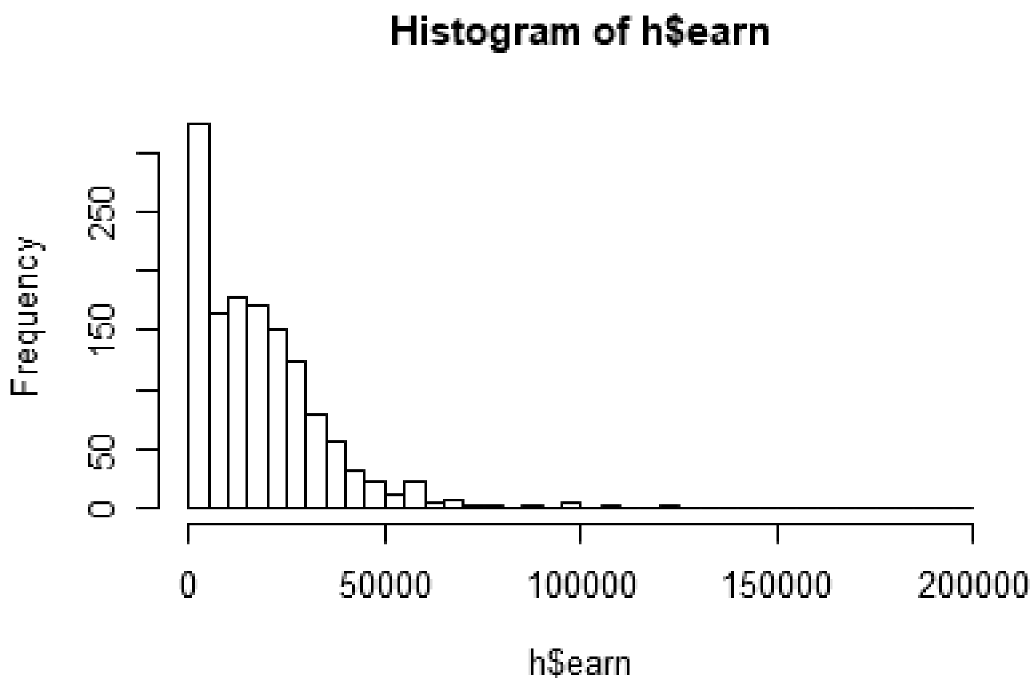
```r
#Null values in the heights variable. Removing rows with null heights
#Removing height1 and height2 columns as height = height1 + 12*height2
h_h <-  heights[!is.na(heights$height),c(-2,-3)]

count(h_h, vars = ed)
```

```
## # A tibble: 19 x 2
##     vars     n
##    <dbl> <int>
##  1     2     1
##  2     3     1
##  3     4     3
##  4     5     8
##  5     6    13
##  6     7    11
##  7     8    64
##  8     9    41
##  9    10    54
## 10    11    72
## 11    12   759
## 12    13   153
## 13    14   254
## 14    15    94
## 15    16   267
## 16    17    92
## 17    18   132
## 18    98     1
## 19    99     1
```

```r
#Removing rows with ed = 98 and 99
h_ed <- h_h[which(h_h$ed < 98),]

count(h_ed, vars = yearbn)
```

```
## # A tibble: 74 x 2
##     vars     n
##    <dbl> <int>
##  1     0     2
##  2     1     2
##  3     2     3
##  4     3     1
##  5     4     2
##  6     5     2
##  7     6     3
##  8     7     5
##  9     8    14
## 10     9     7
## # ... with 64 more rows
```

```r
#Removing rows with yearbn = 99 as the year_survey = 90
h_yr <- subset(h_ed, yearbn != 99)

#Other observations for the data
#Race,Hisp will be a categorical variable and not continuos

#Null values are present in earn (650 observations); Since this is the dependent variable; we will divi
h <- h_yr[!is.na(h_yr$earn),]
h_new <- h_yr[is.na(h_yr$earn),]
#Removing observations with height = NA

#Modelling dataset
hist(h$earn, breaks = 50)
```

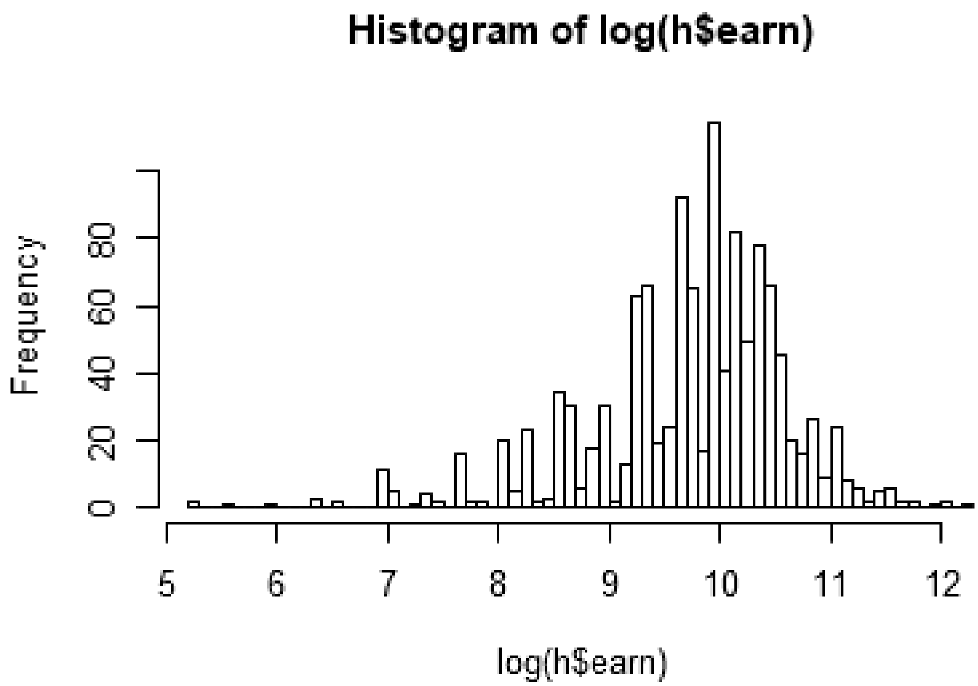



```r
count(h , earn )
```

```
## # A tibble: 134 x 2
##      earn     n
##     <dbl> <int>
##   1     0   187
##   2   200     2
##   3   265     1
##   4   400     1
##   5   600     3
##   6   700     2
##   7  1000    11
##   8  1200     5
```

```
## 9  1400     1
## 10  1500     4
## # ... with 124 more rows
```

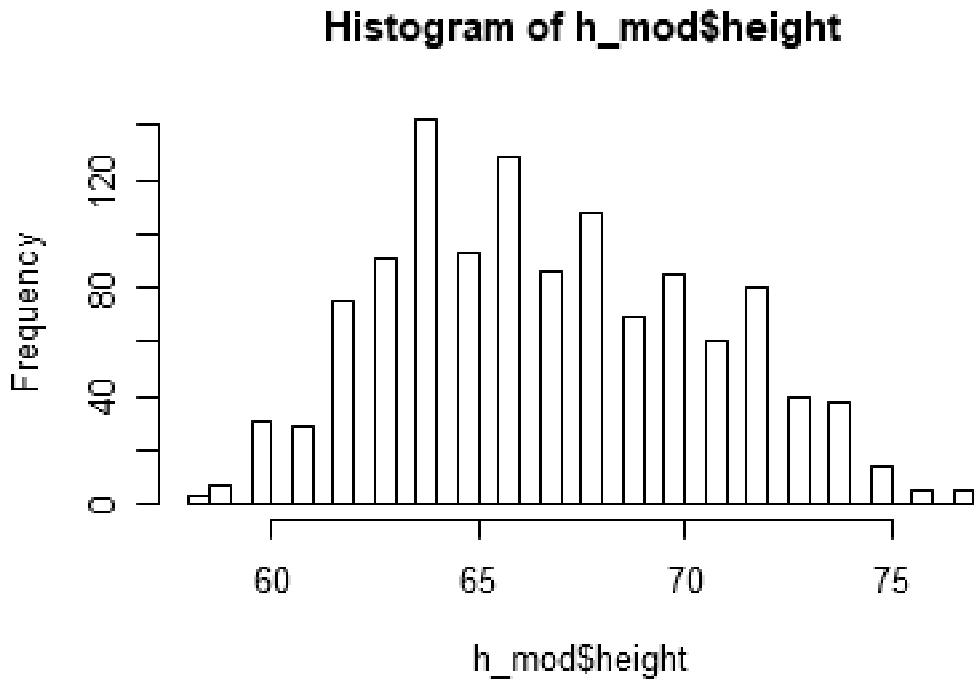
```r
#187 customers has zero income
hist(log(h$earn), breaks = 50)
```




```r
# Since income is better modelled as log; we might need to remove 0 income households and remodel.
h_mod <- h[ which(h$earn > 0 ),]
rownames(h_mod) <- 1:nrow(h_mod)
#h_log is the final dataset for modelling log earning
#h is final dataset for modelling earnings
```

2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

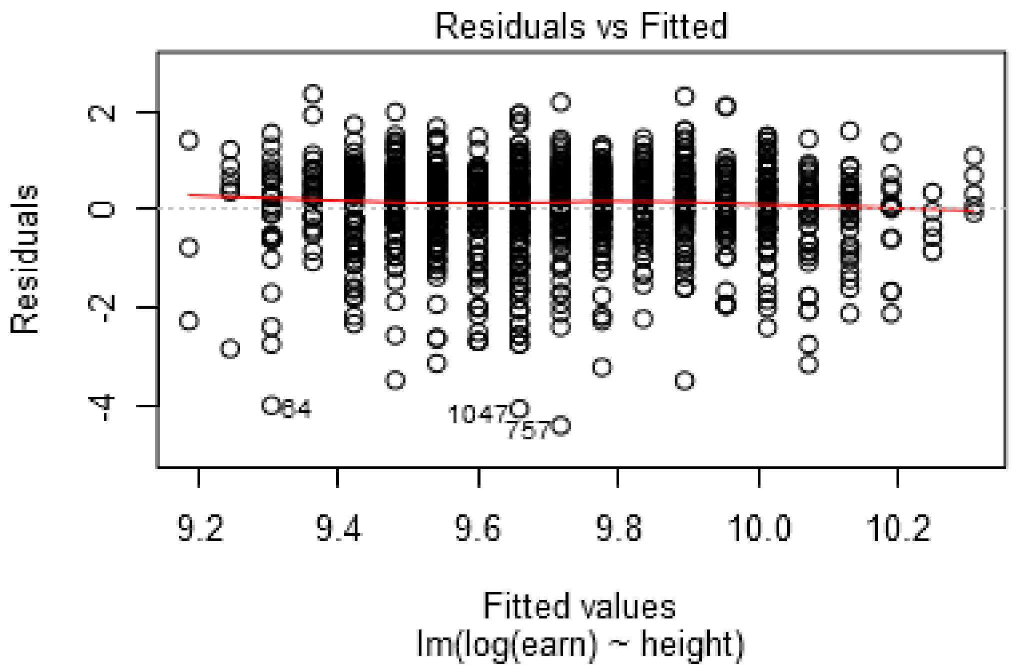
```r
#Fitting a log earning ~ height model
hist(h_mod$height, breaks = 30 )
```

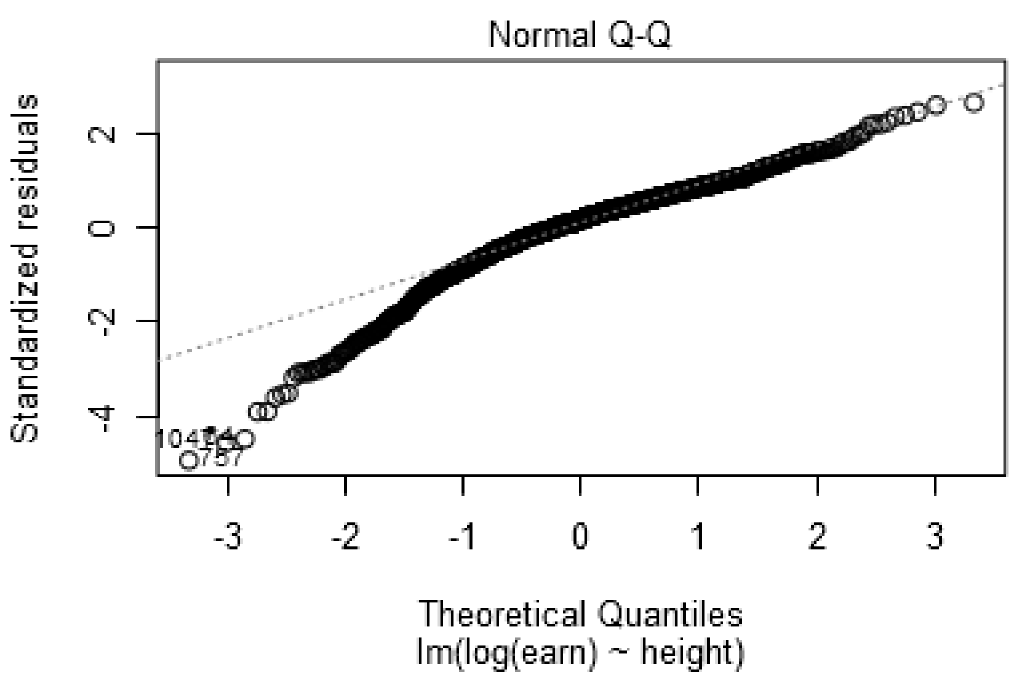


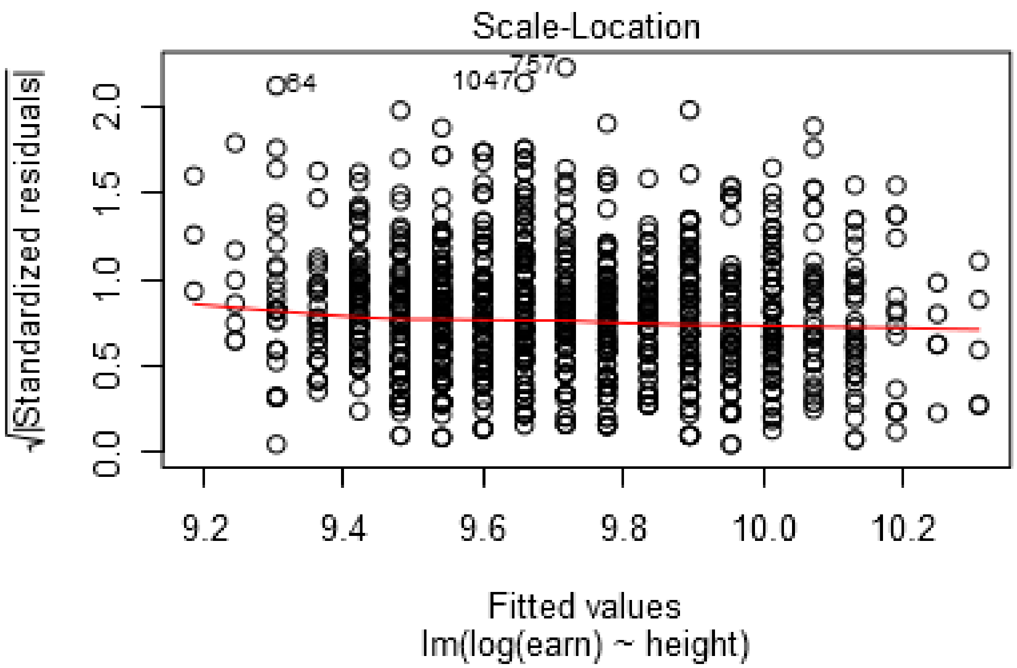

```r
# This follows a normal dist so using as it is in the model

lm_1 <- lm(log(earn) ~ height , data = h_mod)
summary(lm_1)
```

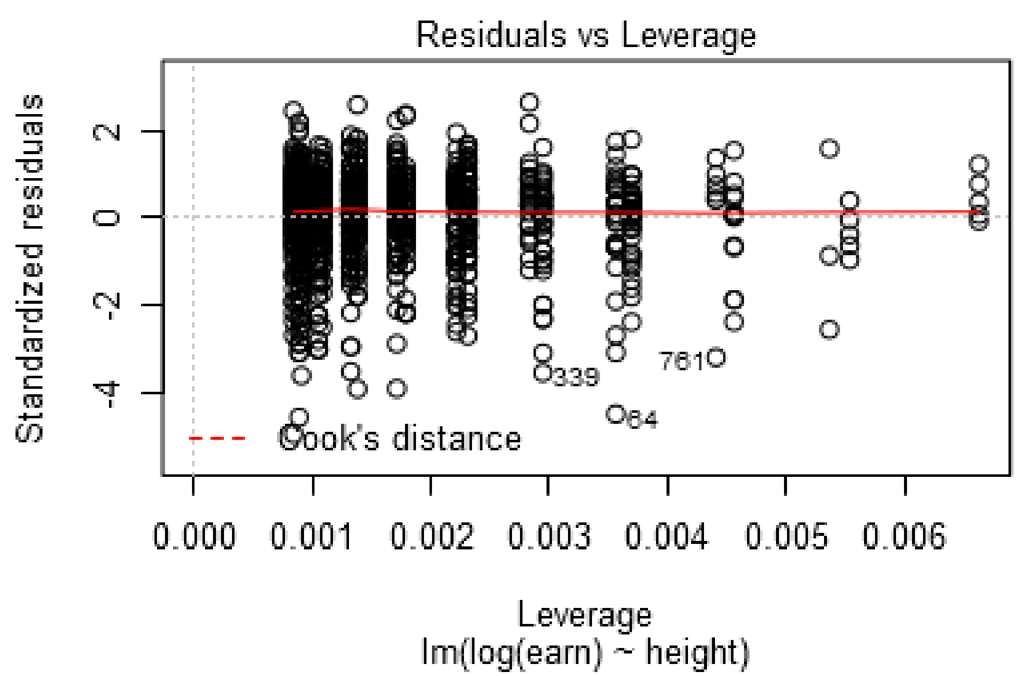
```
##
## Call:
## lm(formula = log(earn) ~ height, data = h_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4193 -0.3974  0.1416  0.5834  2.3571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.756445   0.451727  12.743   <2e-16 ***
## height      0.059122   0.006739   8.772   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8932 on 1187 degrees of freedom
## Multiple R-squared:  0.06089,    Adjusted R-squared:  0.06009
## F-statistic: 76.96 on 1 and 1187 DF,  p-value: < 2.2e-16
```

```r
plot(lm_1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(earn) ~ height)
64
1047
757


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(earn) ~ height)
757






```
#Intercept 5.75 gives the log avergae household income of population whose theoretical height is 0.
#Average earnings is (exp(lm_1$coefficients[1])) ~ 316.22$
```

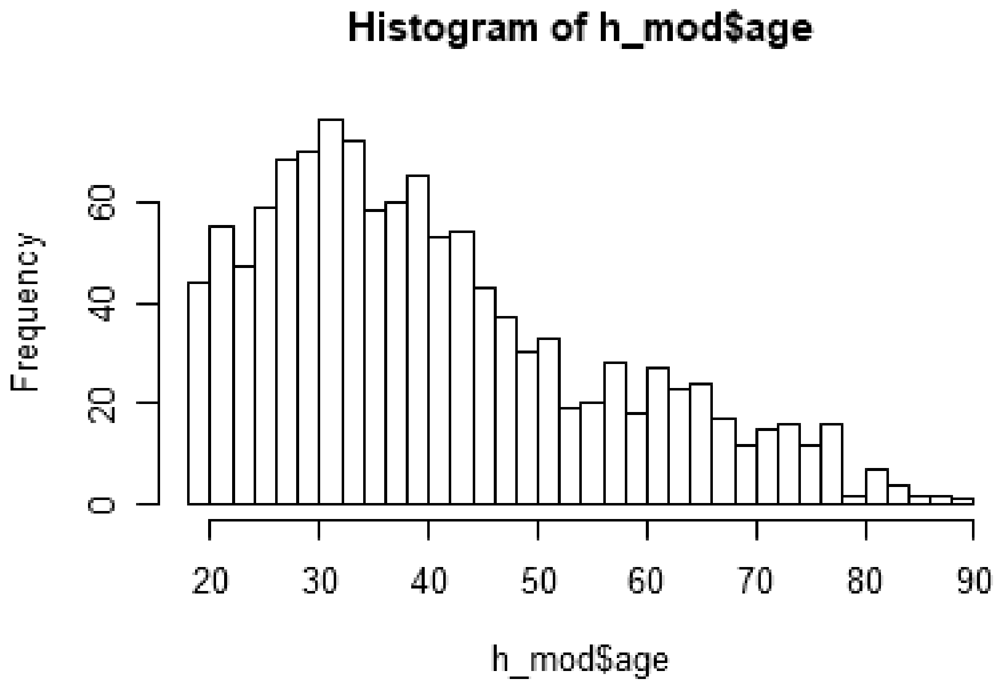
```r
#In order for the earnings to correspond to average value, we will transform the height values

h_mod$h_cent <- h_mod$height - mean(h_mod$height)
lm_2 <- lm(log(earn) ~ h_cent , data = h_mod)
summary(lm_2)
```

```
##
## Call:
## lm(formula = log(earn) ~ h_cent, data = h_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4193 -0.3974  0.1416  0.5834  2.3571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.712696   0.025905 374.937   <2e-16 ***
## h_cent      0.059122   0.006739   8.772   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8932 on 1187 degrees of freedom
## Multiple R-squared:  0.06089,    Adjusted R-squared:  0.06009
## F-statistic: 76.96 on 1 and 1187 DF,  p-value: < 2.2e-16
```

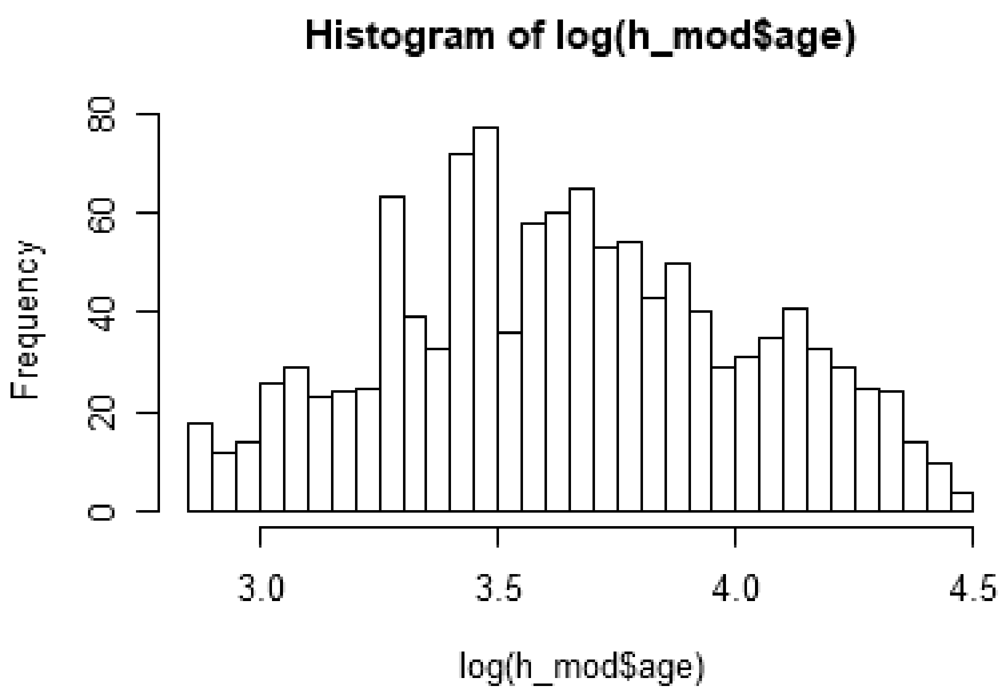
```r
#Intercept 9.71 gives the log avergae household income of population that has average height
#Average earnings is (exp(lm_1$coefficients[1])) ~ 16526$
# The amount of variance being explained by height is quite low; which makes sense as earnings has no e
```

3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and age. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

```r
#The data mentions that height and age are taken for one of the members in the household
#This will have dependency on whether the person interviewed is male or female; so for the model we are
#Since heights and age have different scales; in order for the coefficients to be comparable; we will u

h_mod$age <- 90 - h_mod$yearbn
hist(h_mod$age, breaks = 50 )
```



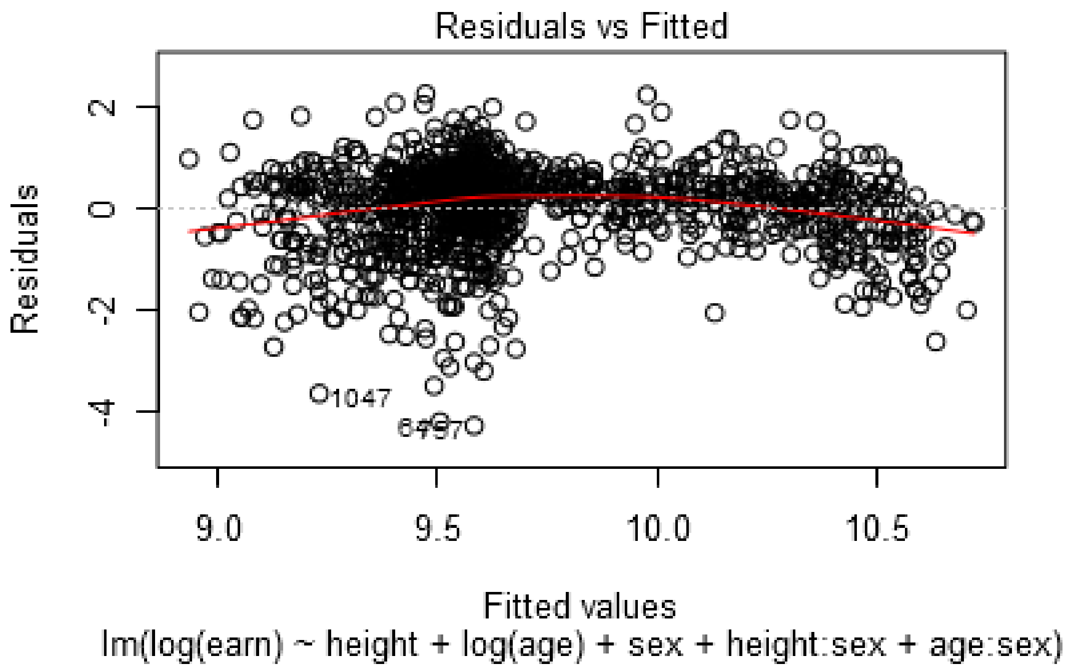

```r
#This is far from normal distribution
hist(log(h_mod$age), breaks = 50)
```

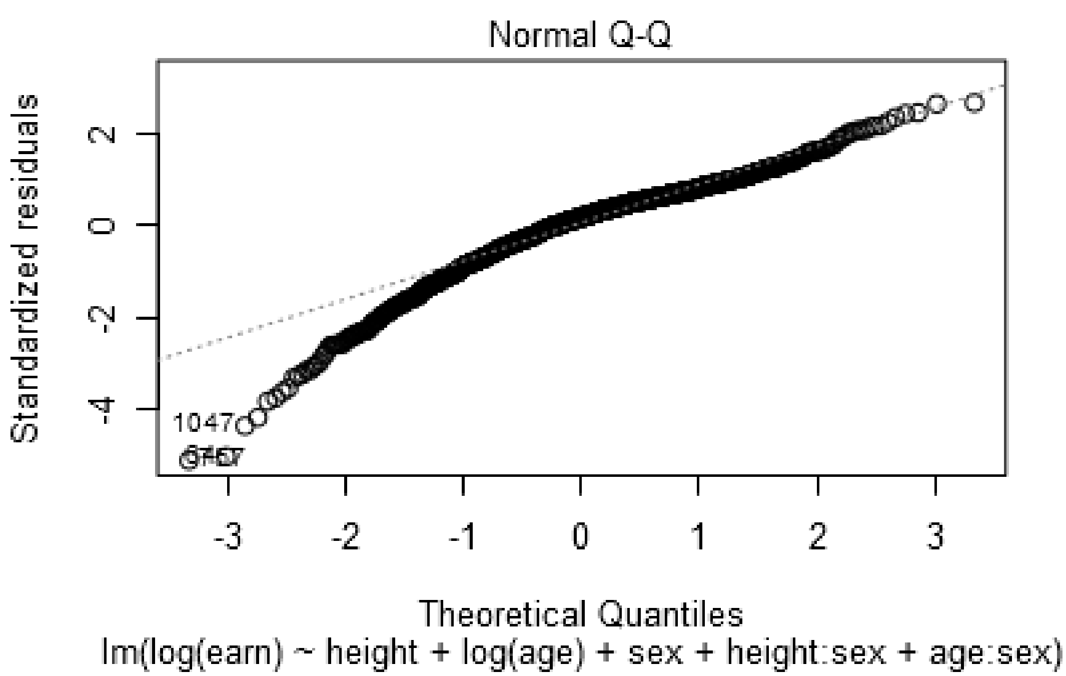


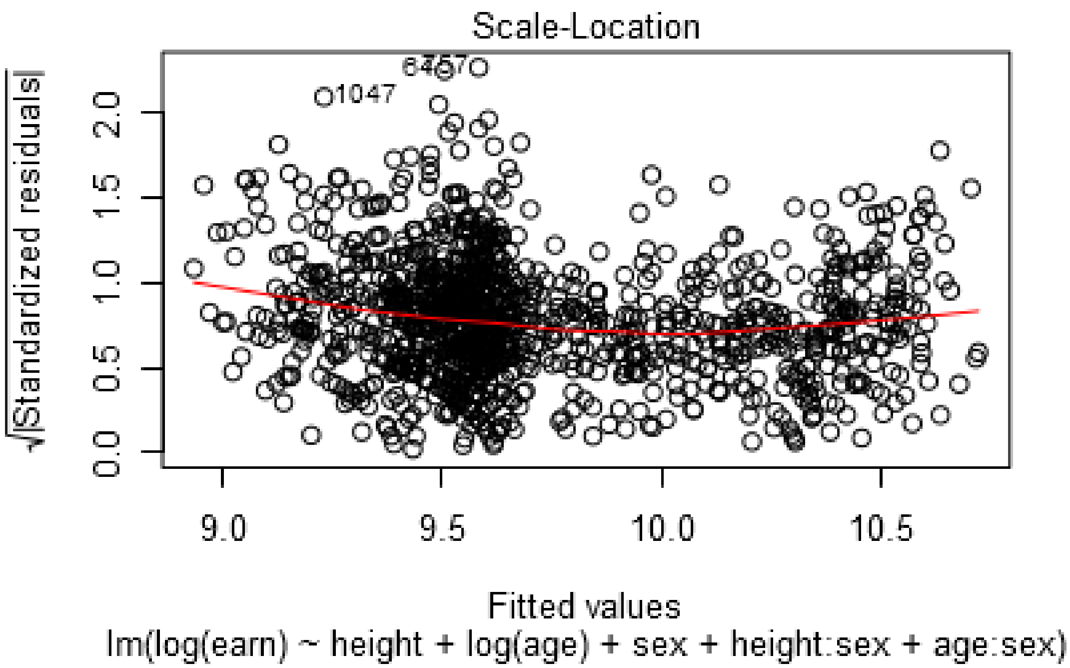
```r
#Its a symmetric distribution now so should result in better prediction

#Plotting variables one by one
lm_3 <- lm( log(earn) ~ height + log(age) + sex + height:sex + age:sex , data = h_mod)
summary(lm_3)
```

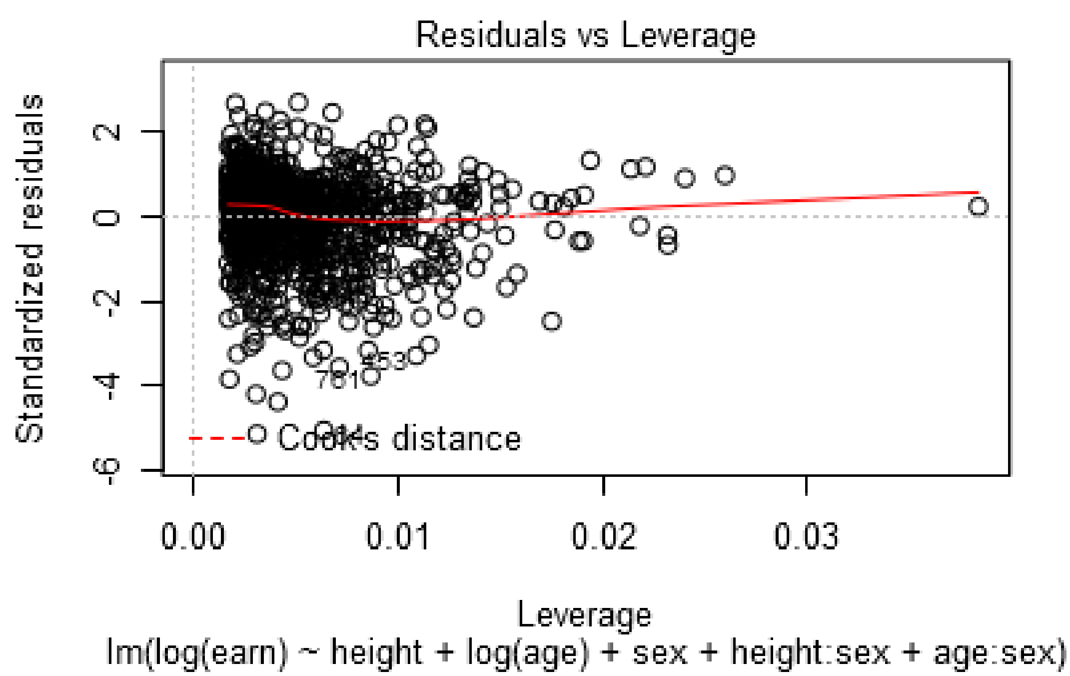
```
##
## Call:
## lm(formula = log(earn) ~ height + log(age) + sex + height:sex +
##     age:sex, data = h_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2855 -0.4086  0.1564  0.5334  2.2468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.280730   2.089605   0.134   0.8932
## height       0.046677   0.028233   1.653   0.0985 .
## log(age)     1.918726   0.177213  10.827   <2e-16 ***
## sex          1.526440   1.215770   1.256   0.2095
## height:sex  -0.015893   0.017808  -0.892   0.3723
## sex:age     -0.022066   0.002553  -8.642   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8367 on 1183 degrees of freedom
## Multiple R-squared:  0.1789, Adjusted R-squared:  0.1754
## F-statistic: 51.54 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```r
#Coefficient are explained in the next section
plot(lm_3)
```


Residuals vs Fitted
Residuals
2
0
-2
-4
1047
9.0
9.5
10.0
10.5
Fitted values
lm(log(earn) ~ height + log(age) + sex + height:sex + age:sex)


Normal Q-Q
Standardized residuals
2
0
-2
-4
1047
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(earn) ~ height + log(age) + sex + height:sex + age:sex)

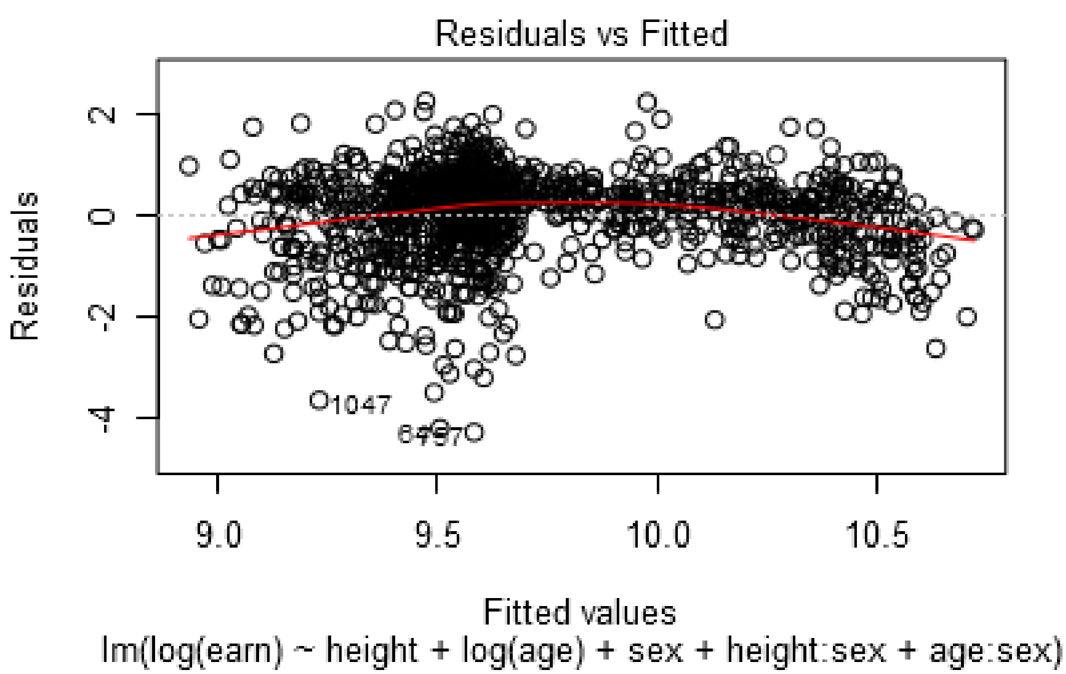


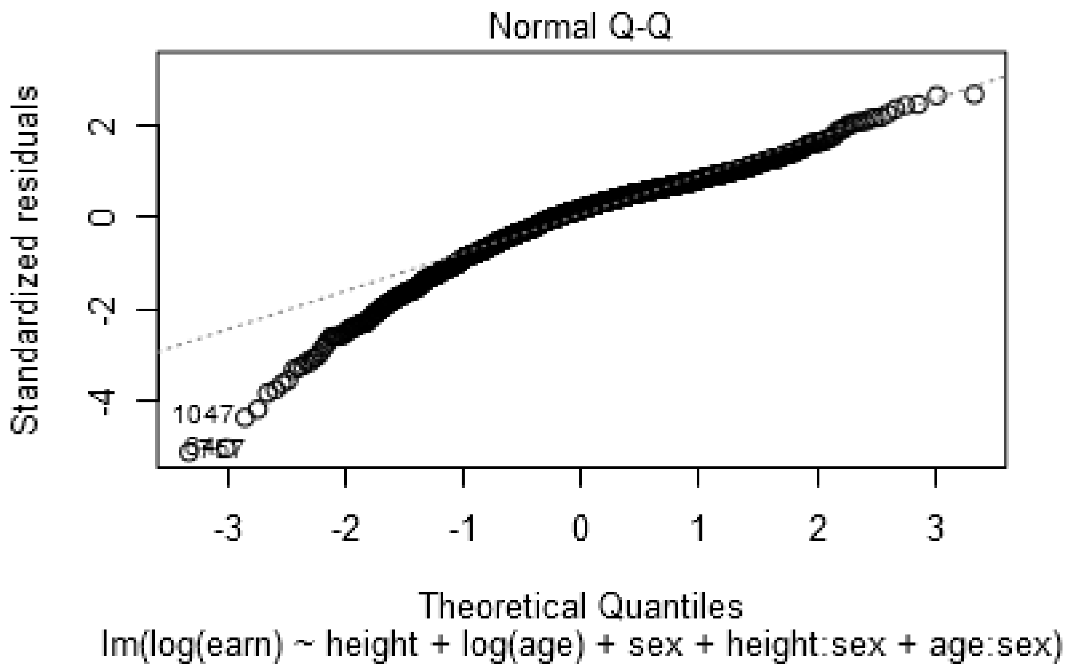


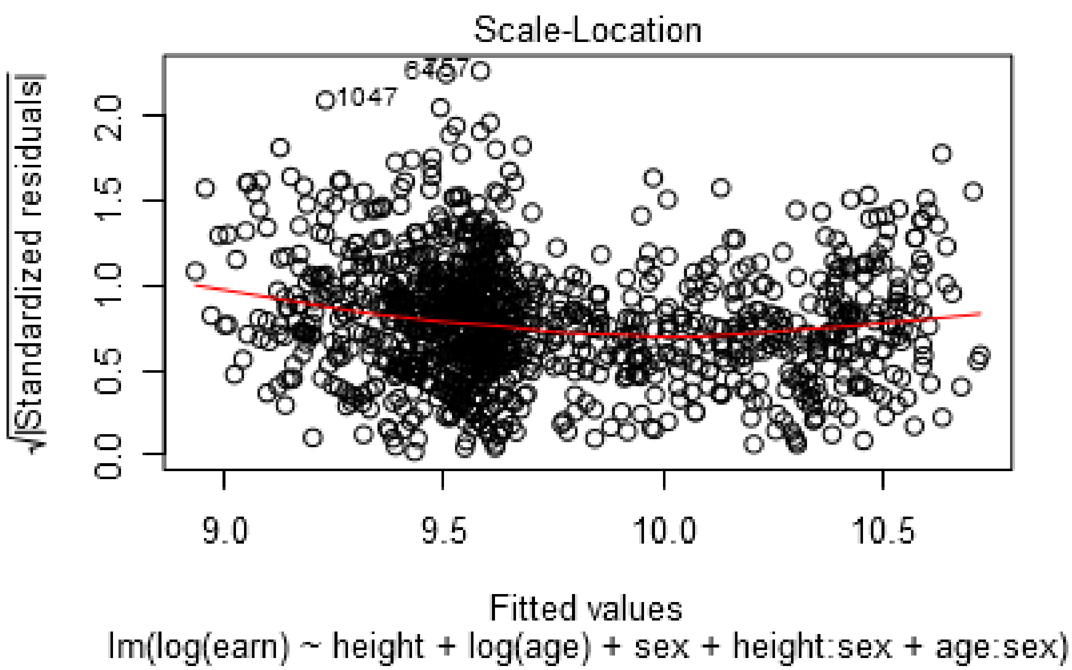

```
#Removing outliers : 1810, 116, 1296
lm_final <- lm( log(earn) ~ height + log(age) + sex + height:sex + age:sex , data = h_mod)
summary(lm_final)
```

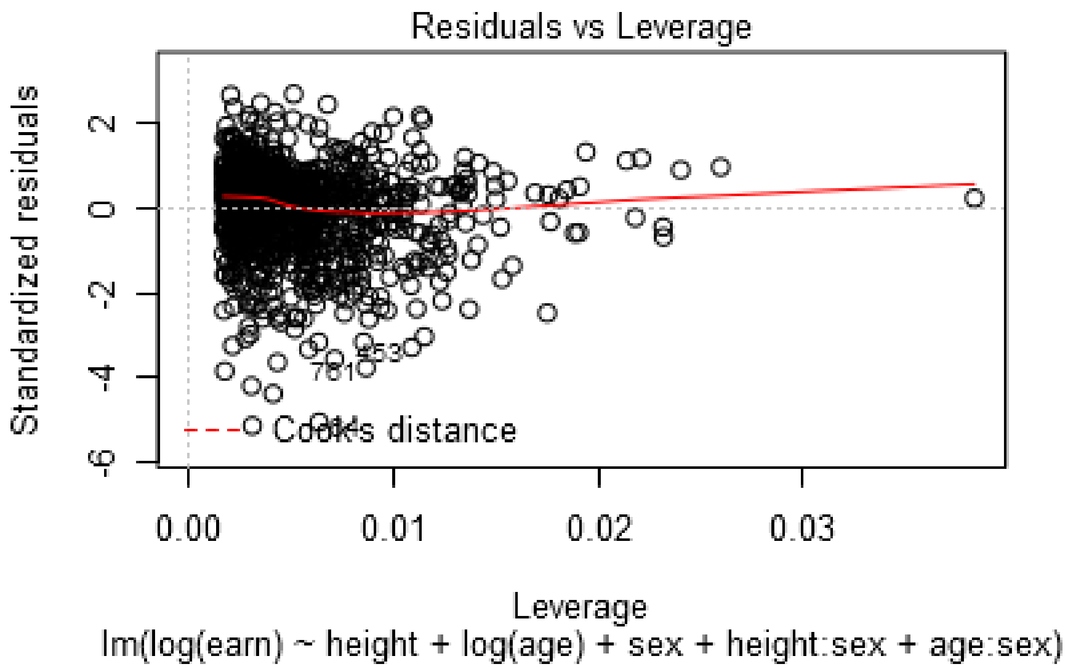
```
##
## Call:
## lm(formula = log(earn) ~ height + log(age) + sex + height:sex +
##     age:sex, data = h_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2855 -0.4086  0.1564  0.5334  2.2468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.280730   2.089605   0.134   0.8932
## height       0.046677   0.028233   1.653   0.0985 .
## log(age)     1.918726   0.177213  10.827   <2e-16 ***
## sex          1.526440   1.215770   1.256   0.2095
## height:sex  -0.015893   0.017808  -0.892   0.3723
## sex:age     -0.022066   0.002553  -8.642   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8367 on 1183 degrees of freedom
## Multiple R-squared:  0.1789, Adjusted R-squared:  0.1754
## F-statistic: 51.54 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```r
plot(lm_final)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(earn) ~ height + log(age) + sex + height:sex + age:sex)


Scale-Location
√|Standardized residuals|
Fitted values
lm(log(earn) ~ height + log(age) + sex + height:sex + age:sex)



4. Interpret all model coefficients.

```
#Explanation of coefficients :
#Log of earnings is modelled
#Intercept : Average earning of hypothetical population with zero age; zero height and male.
#height : With Every unit increase in height; earnings increase by e^0.046. or around 4.7% increment
#log(age) : WIth every increment of age ; the earnings increase in factor of e^1.91
#sex : Compared to males; if female responded then earning is around e^1.526 or around 4.7 times
#height: sex :Gives how much difference is their between males and females in correlation of height and
#age :sex :Gives how much difference is their between males and females in correlation of age and earni
```

5. Construct 95% confidence interval for all model coefficients and discuss what they mean.

```
co <- lm_final$coefficients
se <- sqrt(diag(vcov(lm_final)))
tab_final <- as.data.frame(cbind(co,se))
tab_final$t_value <- co/se
tab_final$up <- tab_final$co + 1.96*tab_final$se
tab_final$low <- tab_final$co - 1.96*tab_final$se
colnames (tab_final) <- c("Coefficient","St Error","T_Value","UpperLimit95","LowerLimit95")
tab_final
```

```
##               Coefficient    St Error    T_Value UpperLimit95 LowerLimit95
## (Intercept)    0.28072968 2.089605153  0.1343458   4.37635578 -3.814896418
## height         0.04667689 0.028233233  1.6532605   0.10201403 -0.008660247
## log(age)       1.91872562 0.177212776 10.8272421   2.26606266  1.571388582
## sex            1.52644021 1.215770272  1.2555334   3.90934994 -0.856469523
## height:sex    -0.01589283 0.017807574 -0.8924758   0.01901002 -0.050795672
## sex:age       -0.02206632 0.002553334 -8.6421579  -0.01706178 -0.027070853
```

Analysis of mortality rates and various environmental factors

The folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

Variables, in order:

- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all facilities
- DENS Population per sq. mile in urbanized areas, 1960
- NONW % non-white population in urbanized areas, 1960
- WWDRK % employed in white collar occupations
- POOR % of families with income < $3000
- HC Relative hydrocarbon pollution potential
- NOX Same for nitric oxides
- SO@ Same for sulphur dioxide
- HUMID Annual average % relative humidity at 1pm
- MORT Total age-adjusted mortality rate per 100,000

For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations in regression.

```r
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

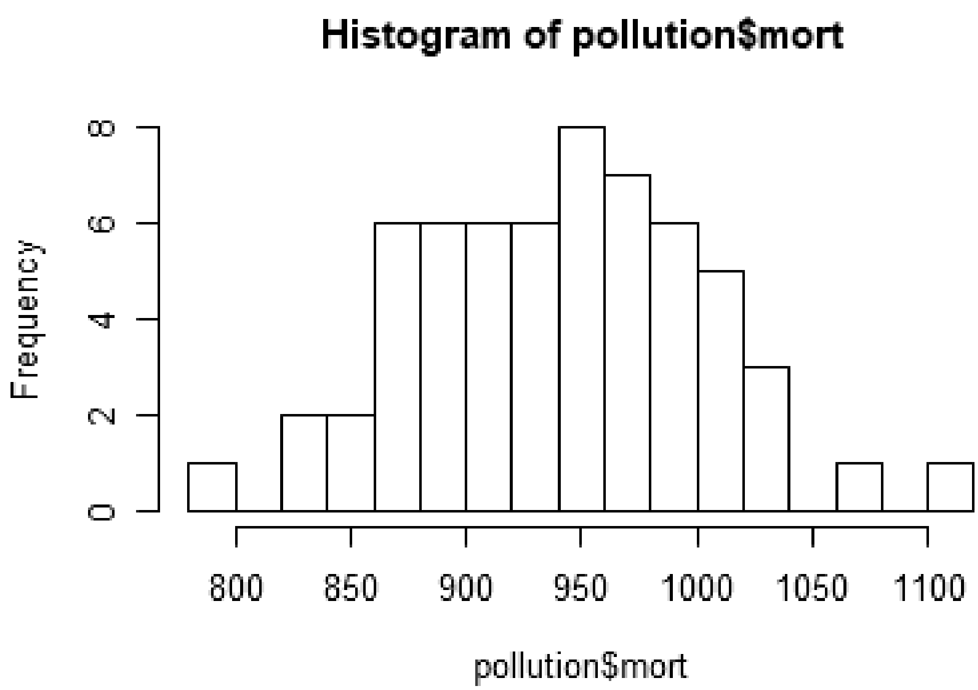
1. Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

```r
summary(pollution)
```

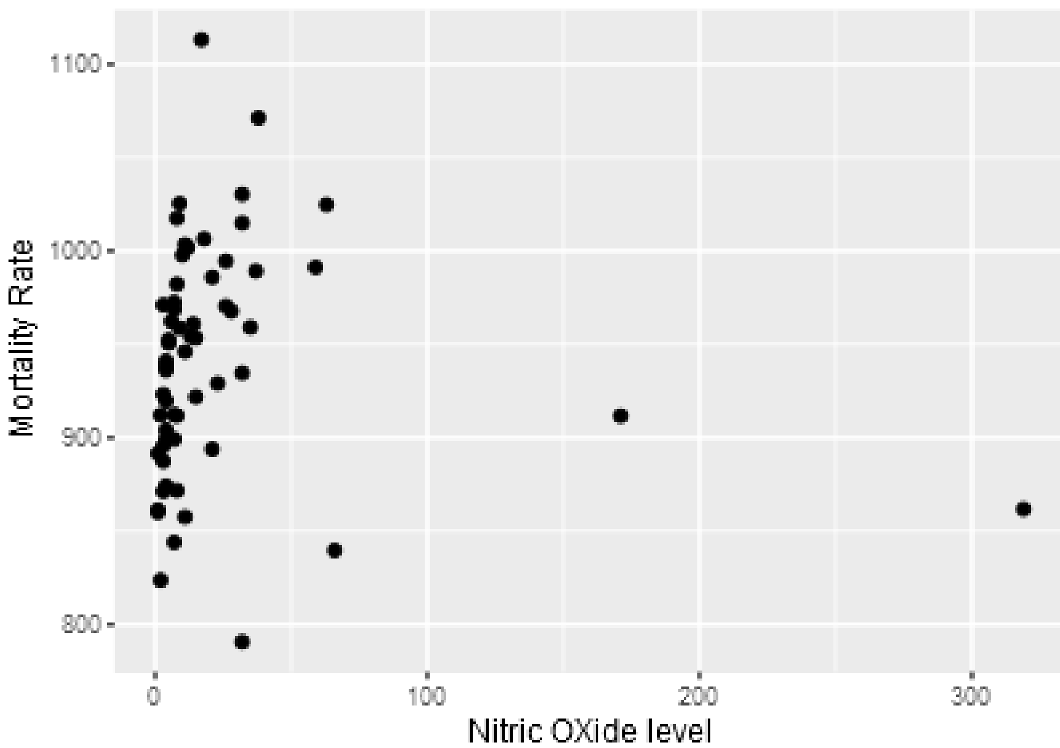
```
##       prec            jant            jult           ovr65
##  Min.   :10.00   Min.   :12.00   Min.   :63.00   Min.   : 5.600
##  1st Qu.:32.75   1st Qu.:27.00   1st Qu.:72.00   1st Qu.: 7.675
##  Median :38.00   Median :31.50   Median :74.00   Median : 9.000
##  Mean   :37.37   Mean   :33.98   Mean   :74.58   Mean   : 8.798
##  3rd Qu.:43.25   3rd Qu.:40.00   3rd Qu.:77.25   3rd Qu.: 9.700
##  Max.   :60.00   Max.   :67.00   Max.   :85.00   Max.   :11.800
##       popn            educ            hous            dens
##  Min.   :2.920   Min.   : 9.00   Min.   :66.80   Min.   :1441
##  1st Qu.:3.210   1st Qu.:10.40   1st Qu.:78.38   1st Qu.:3104
##  Median :3.265   Median :11.05   Median :81.15   Median :3567
##  Mean   :3.263   Mean   :10.97   Mean   :80.91   Mean   :3876
##  3rd Qu.:3.360   3rd Qu.:11.50   3rd Qu.:83.60   3rd Qu.:4520
##  Max.   :3.530   Max.   :12.30   Max.   :90.70   Max.   :9699
##       nonw           wwdrk            poor             hc
##  Min.   : 0.80   Min.   :33.80   Min.   : 9.40   Min.   :  1.00
```

```
##  1st Qu.: 4.95   1st Qu.:43.25   1st Qu.:12.00   1st Qu.:  7.00
##  Median :10.40   Median :45.50   Median :13.20   Median : 14.50
##  Mean   :11.87   Mean   :46.08   Mean   :14.37   Mean   : 37.85
##  3rd Qu.:15.65   3rd Qu.:49.52   3rd Qu.:15.15   3rd Qu.: 30.25
##  Max.   :38.50   Max.   :59.70   Max.   :26.40   Max.   :648.00
##       nox              so2             humid            mort
##  Min.   :  1.00   Min.   :  1.00   Min.   :38.00   Min.   : 790.7
##  1st Qu.:  4.00   1st Qu.: 11.00   1st Qu.:55.00   1st Qu.: 898.4
##  Median :  9.00   Median : 30.00   Median :57.00   Median : 943.7
##  Mean   : 22.65   Mean   : 53.77   Mean   :57.67   Mean   : 940.4
##  3rd Qu.: 23.75   3rd Qu.: 69.00   3rd Qu.:60.00   3rd Qu.: 983.2
##  Max.   :319.00   Max.   :278.00   Max.   :73.00   Max.   :1113.2
```

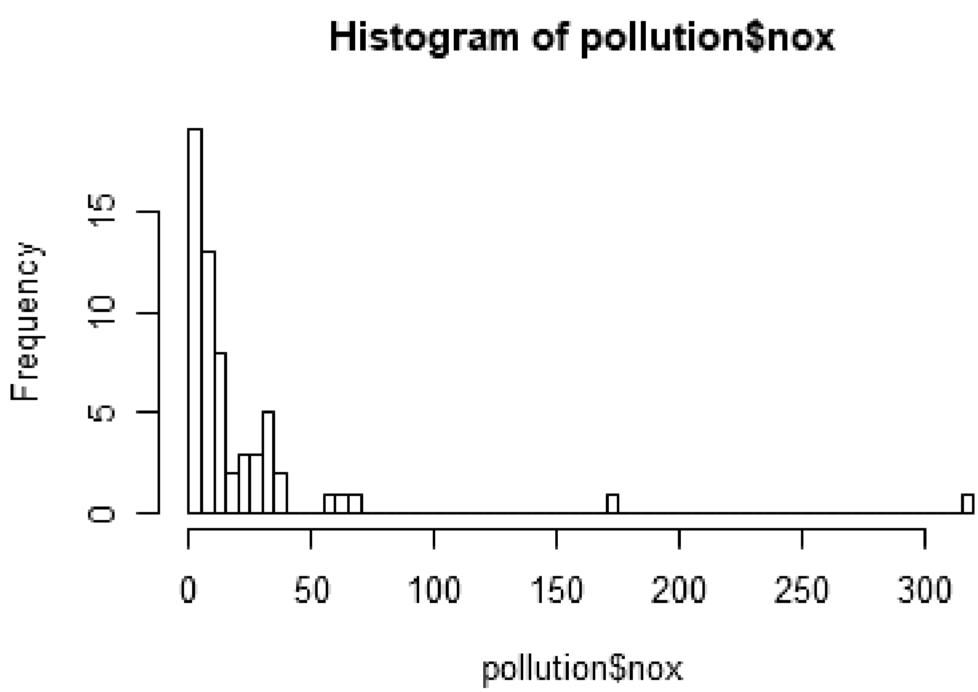
```r
#Data looks fairly well cleaned
hist(pollution$mort , breaks = 20 )
```



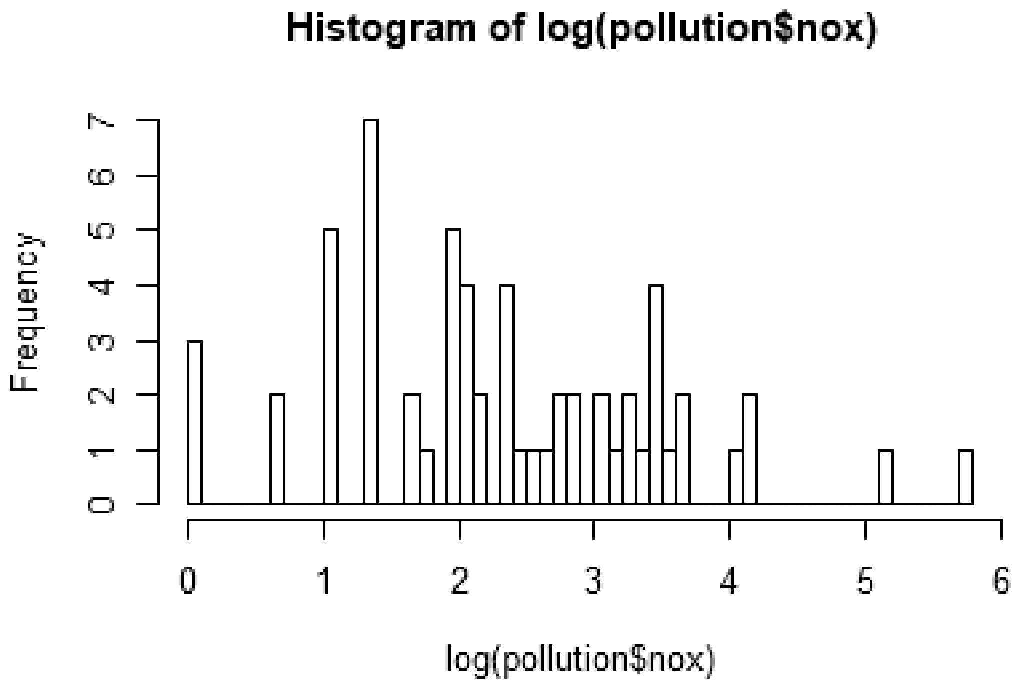

```r
#Mortality rate follows symmetric distribution
qplot(pollution$nox, pollution$mort , xlab = "Nitric OXide level" , ylab = "Mortality Rate"  )
```



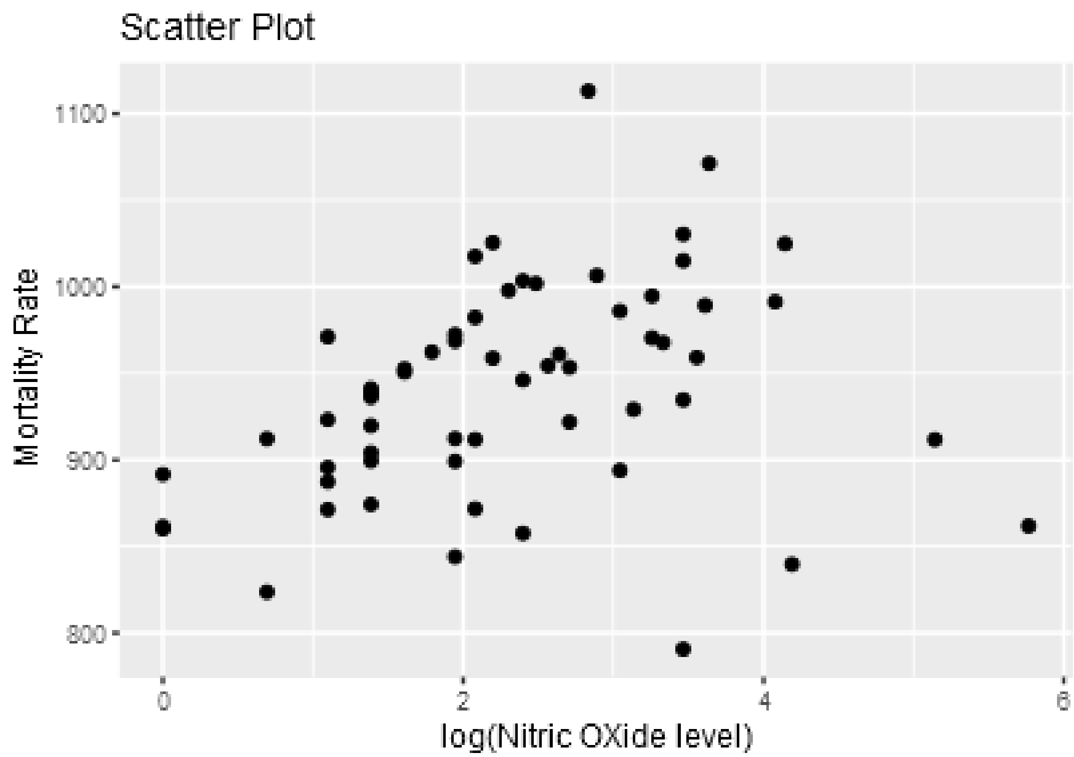

```
#These seem to be a few outliers which would skew the data. Also the nitric oxide variable is right ske
hist(pollution$nox , breaks = 50 )
```



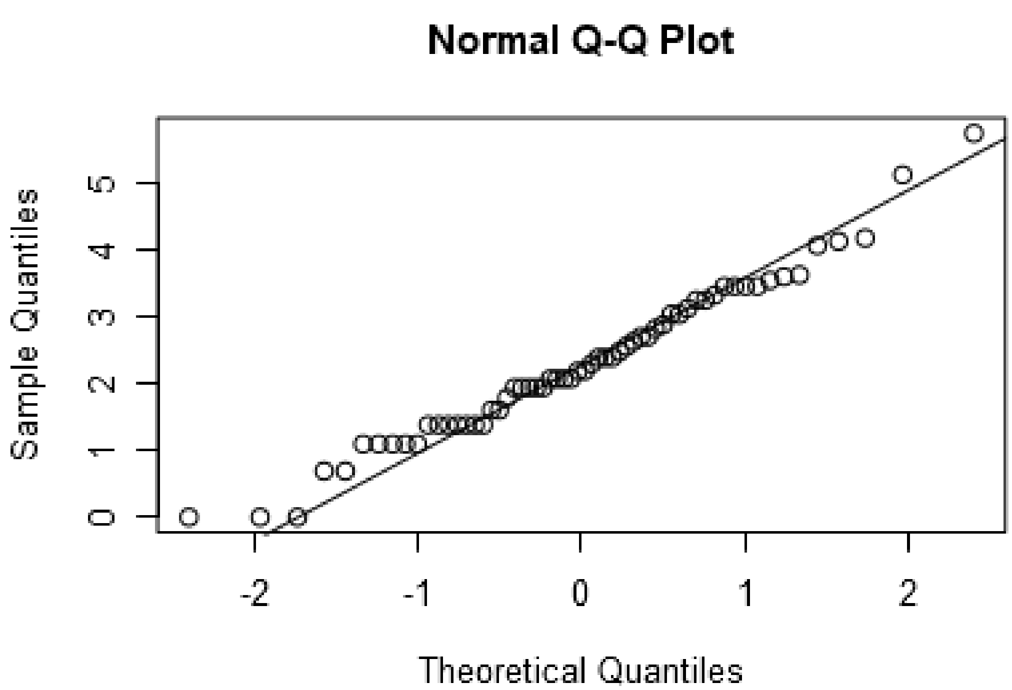
```r
#Hence converting to log transformation
hist(log(pollution$nox), breaks = 50)
```



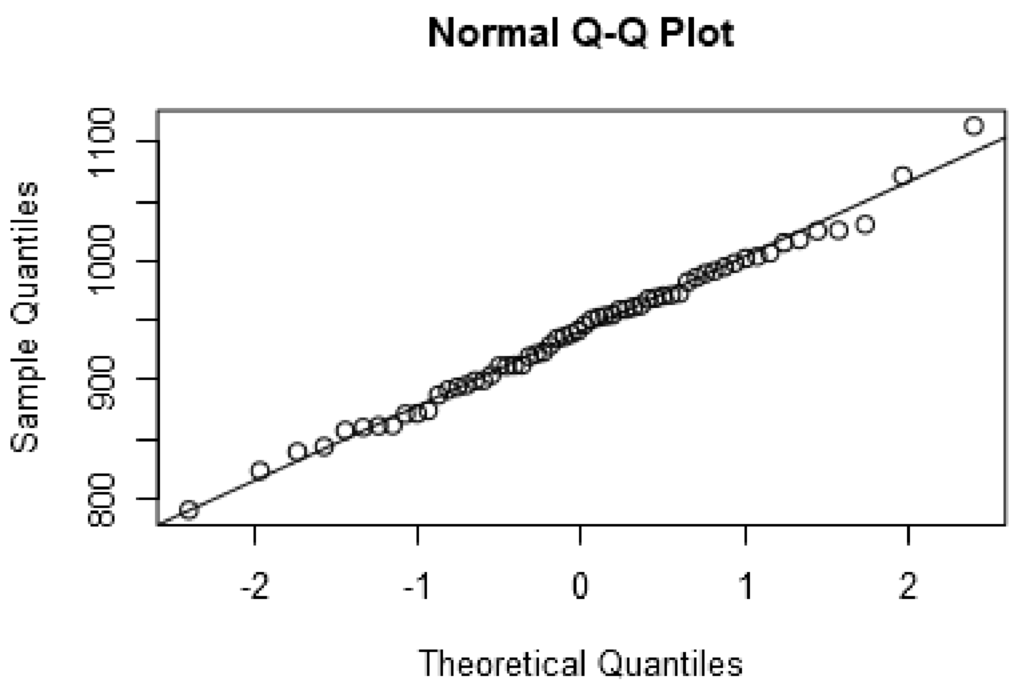

```r
#This distribution is more spread out and should lead to better prediction of Mortaility Rate
qplot(log(pollution$nox), pollution$mort , xlab = "log(Nitric OXide level)" , ylab = "Mortality Rate",
```




```
#Comparing with Normal distribution :
qqnorm( log(pollution$nox) , plot.it =  T)
qqline( log(pollution$nox) , distribution = qnorm)
```



```r
# A few outliers are still present

#Comparing with Normal distribution:
qqnorm( pollution$mort , plot.it =  T)
qqline( pollution$mort , distribution = qnorm)
```




```r
lm_4 <- lm(mort ~ nox , data = pollution)
summary(lm_4)
```

```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.654  -43.710    1.751   41.663  172.211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 942.7115     9.0034 104.706   <2e-16 ***
## nox          -0.1039     0.1758  -0.591    0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.55 on 58 degrees of freedom
## Multiple R-squared:  0.005987,   Adjusted R-squared:  -0.01115
## F-statistic: 0.3494 on 1 and 58 DF,  p-value: 0.5568
```

```r
#Explains around 5% variance in mortality rate
```

2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```r
# Log transformation of independent variable nox leads to more symmetric distribution and hence would r

lm_5 <- lm(mort ~ log(nox) , data = pollution)
summary(lm_5)
```

```
##
## Call:
## lm(formula = mort ~ log(nox), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -167.140  -28.368    8.778   35.377  164.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  904.724     17.173  52.684   <2e-16 ***
## log(nox)      15.335      6.596   2.325   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.01 on 58 degrees of freedom
## Multiple R-squared:  0.08526,    Adjusted R-squared:  0.06949
## F-statistic: 5.406 on 1 and 58 DF,  p-value: 0.02359
```

```r
# This model explains around 8,5% variance in R which is better compared to previous model
```

3. Interpret the slope coefficient from the model you chose in 2.

```r
#Mortality rate is positively correlated to Nox level.
#For every unit scale increase in logarithmic scale of NOx; the average mortality rate increases by 15%
```

4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

```r
co <- lm_2$coefficients
se <- sqrt(diag(vcov(lm_2)))
tab_2 <- as.data.frame(cbind(co,se))
tab_2$t_value <- co/se
tab_2$up <- tab_2$co + 1.96*tab_2$se
tab_2$low <- tab_2$co - 1.96*tab_2$se

#Confidence Interval for Intercept :
c(tab_2[1,"low"]   , tab_2[1,"up"])
```

```
## [1] 9.661922 9.763469
```

```r
#Confidence Interval for Nox Coefficient :
c(tab_2[2,"low"]   , tab_2[2,"up"])
```

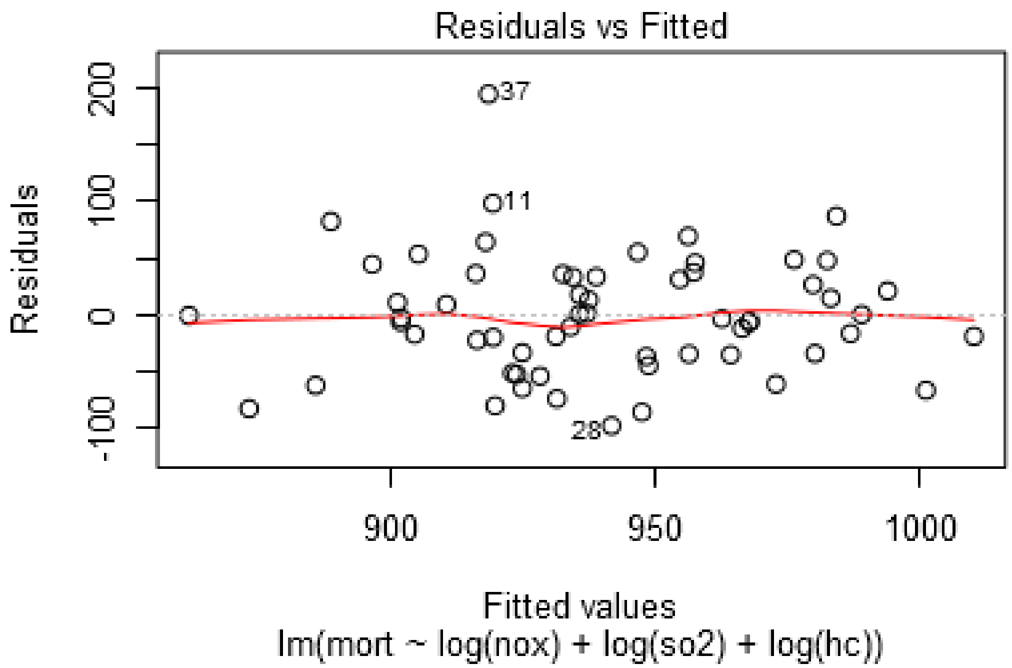
```
## [1] 0.04591262 0.07233136
```

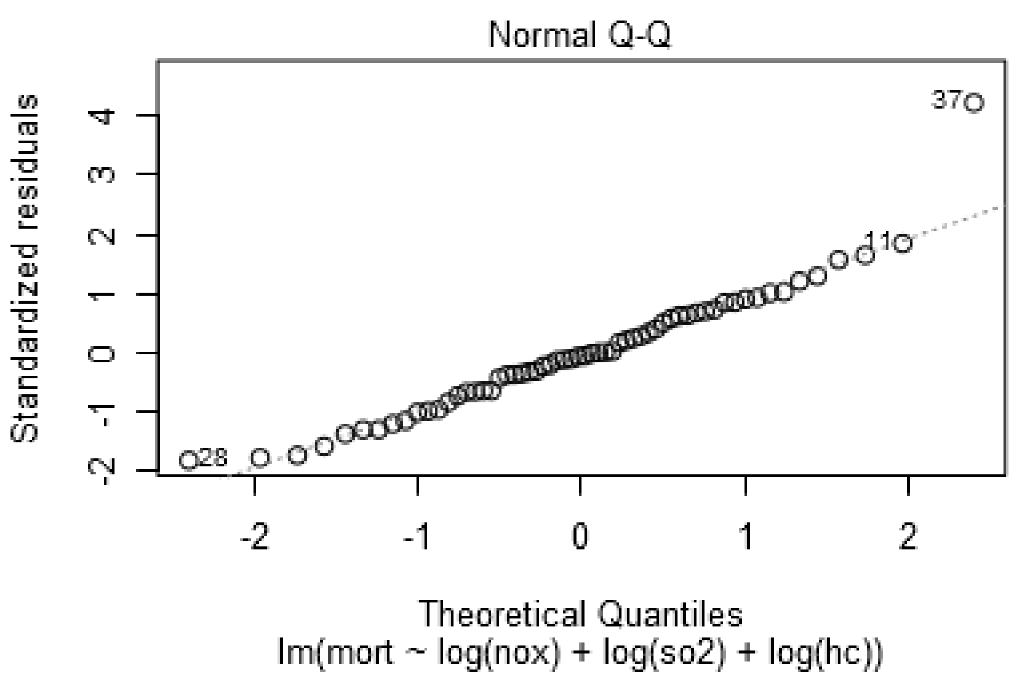
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

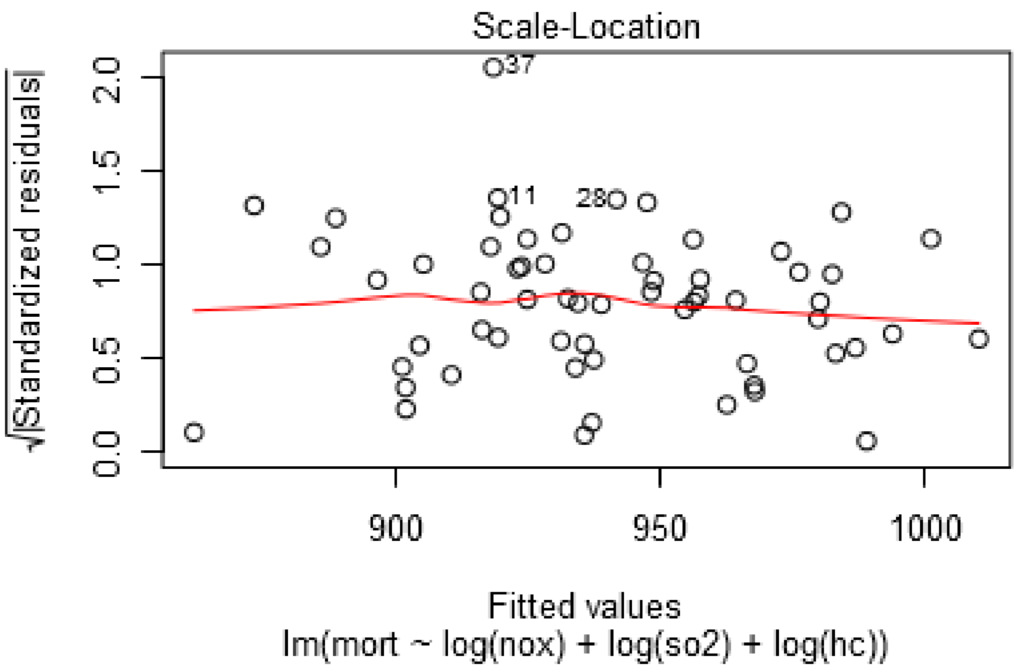
```r
lm_5 <- lm(mort ~ log(nox) + log(so2) + log(hc) , data = pollution)
summary(lm_5)
```

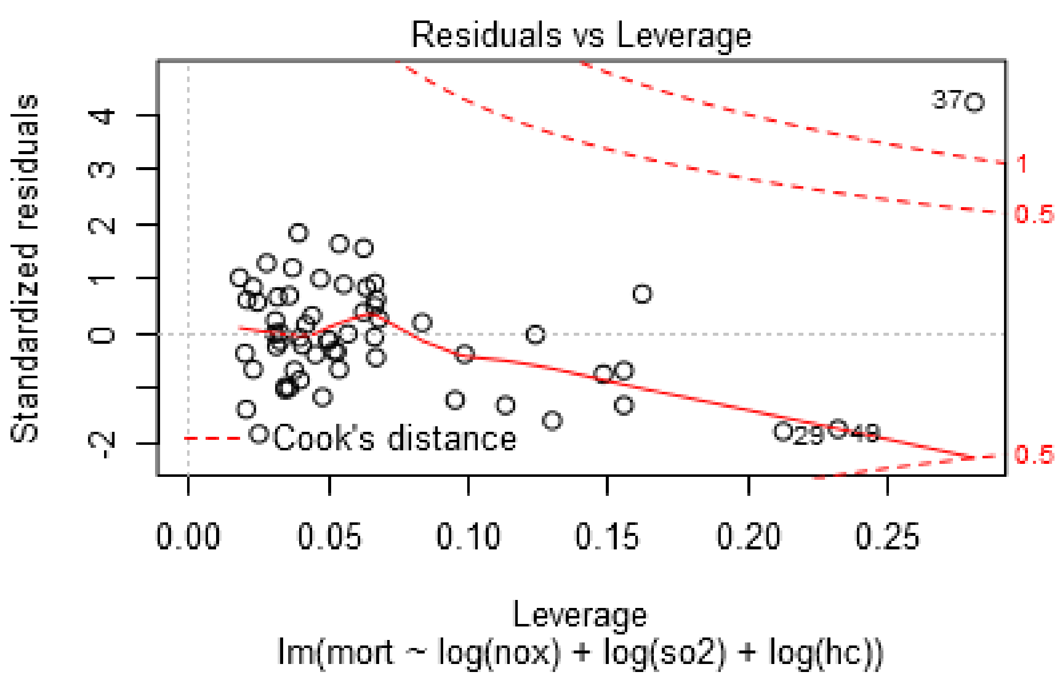
```
##
## Call:
## lm(formula = mort ~ log(nox) + log(so2) + log(hc), data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.793 -34.728  -3.118  34.148 194.567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  924.965     21.449  43.125  < 2e-16 ***
## log(nox)      58.336     21.751   2.682  0.00960 **
## log(so2)      11.762      7.165   1.642  0.10629
## log(hc)      -57.300     19.419  -2.951  0.00462 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.36 on 56 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2363
## F-statistic: 7.086 on 3 and 56 DF,  p-value: 0.0004044
```

```r
plot(lm_5)
```


Residuals vs Fitted
Residuals
200
100
0
-100
37
11
28
900
950
1000
Fitted values
lm(mort ~ log(nox) + log(so2) + log(hc))


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
37
11
28
-2
-1
0
1
2
Theoretical Quantiles
lm(mort ~ log(nox) + log(so2) + log(hc))



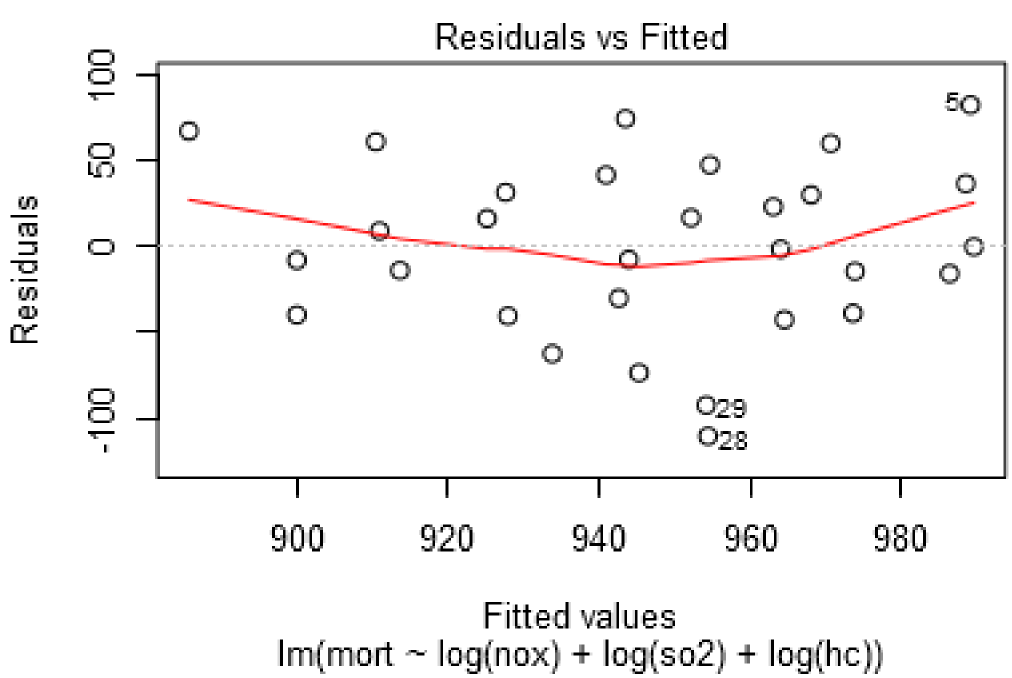


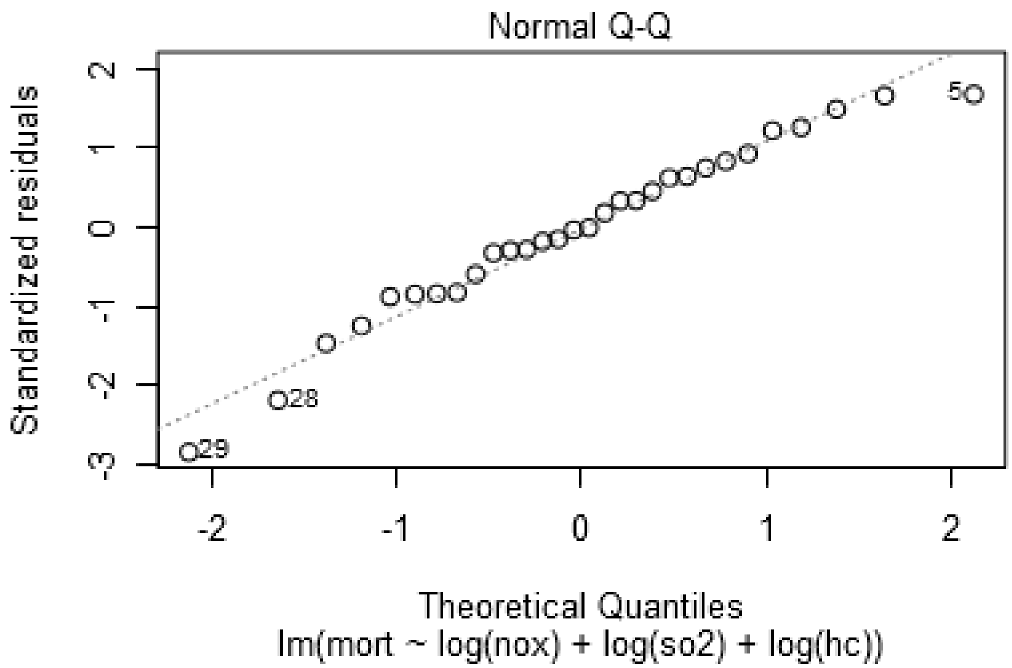
6. Cross-validate: fit the model you chose above to the first half of the data and then predict for the second half. (You used all the data to construct the model in 4, so this is not really cross-validation, but it gives a sense of how the steps of cross-validation can be implemented.)

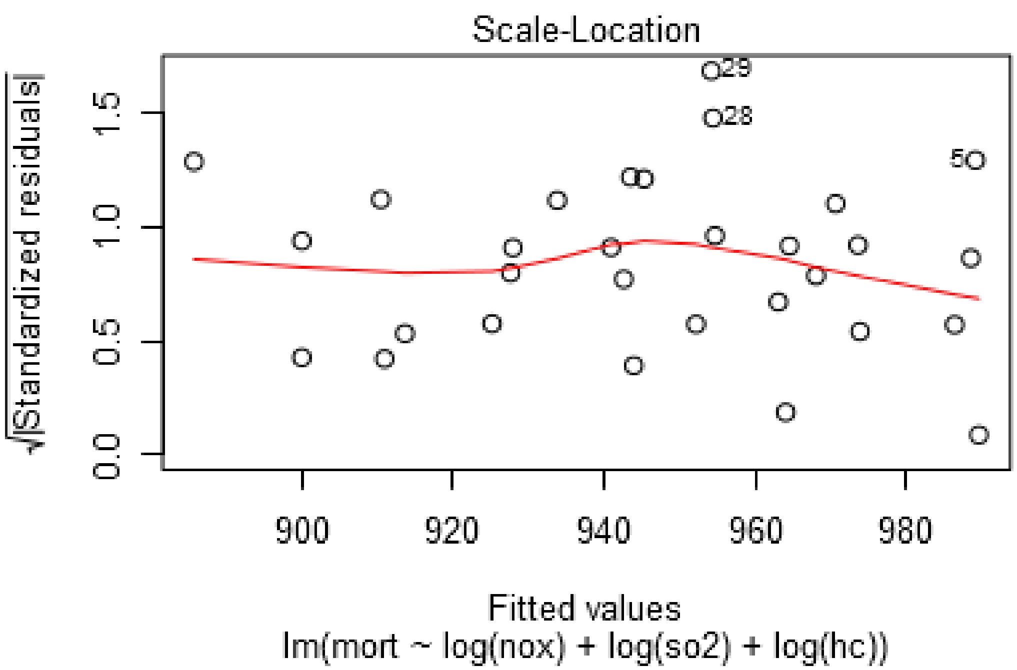
```r
p_train <- pollution[c(1:30),]
p_test <- pollution[c(31:60),]

lm_6 <- lm(mort ~ log(nox) + log(so2) + log(hc) , data = p_train)
summary(lm_6)
```

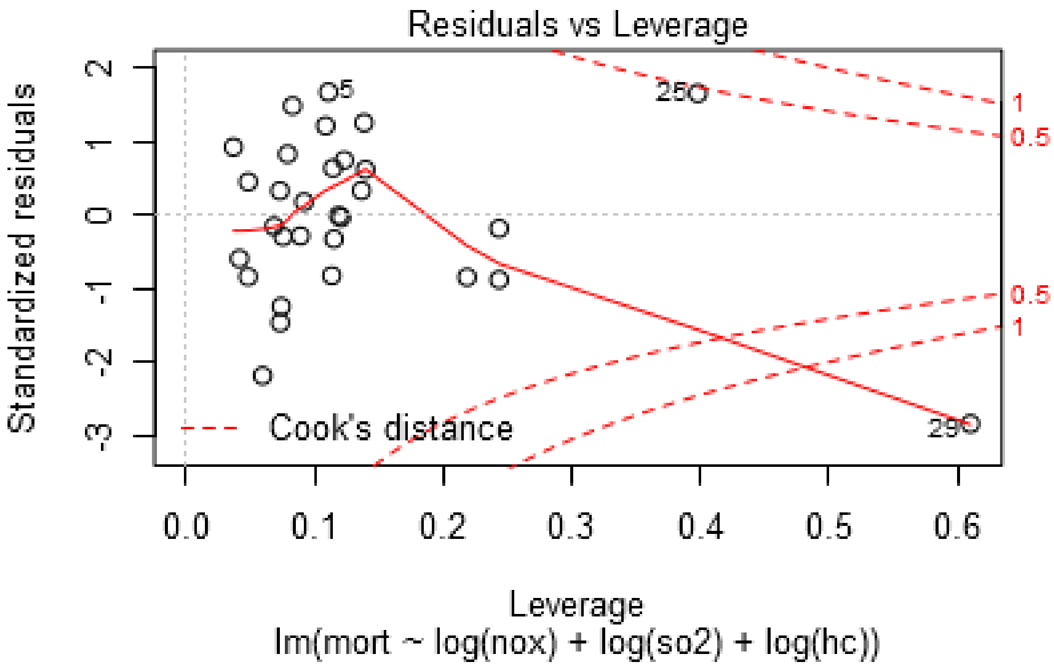
```
##
## Call:
## lm(formula = mort ~ log(nox) + log(so2) + log(hc), data = p_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.358  -36.766   -1.032   35.049   82.107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   899.97      25.71  35.009   <2e-16 ***
## log(nox)       10.57      29.59   0.357   0.7240
## log(so2)       21.87      12.32   1.774   0.0877 .
## log(hc)       -17.47      26.21  -0.667   0.5108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.07 on 26 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.1659
## F-statistic: 2.922 on 3 and 26 DF,  p-value: 0.05277
```

```r
plot(lm_6)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mort ~ log(nox) + log(so2) + log(hc))
29
28
5
-3
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
lm(mort ~ log(nox) + log(so2) + log(hc))
29
28
5
0.0
0.5
1.0
1.5
900
920
940
960
980




```r
#Predicting using training model
p_test$pred_mort <- predict(lm_6, newdata = p_test)

p_test$res <- p_test$mort - p_test$pred_mort
qplot(p_test$pred_mort, p_test$res , xlab = "predicted mortality rate", ylab = "residual mortaility rat
```

Study of teenage gambling in Britain

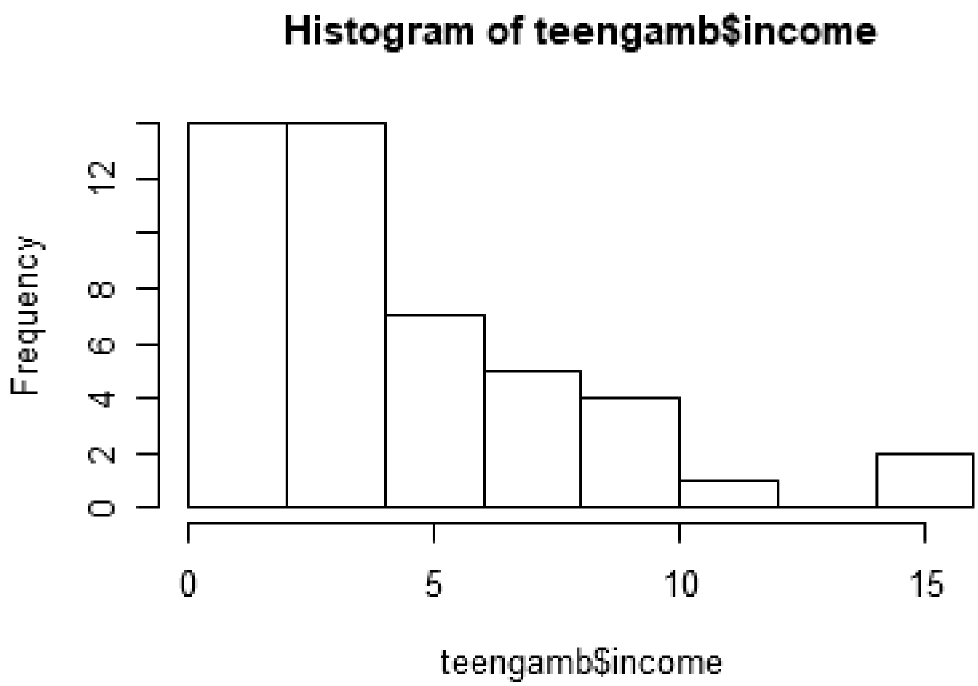
```r
data(teengamb)
```

1. Fit a linear regression model with gamble as the response and the other variables as predictors and interpret the coefficients. Make sure you rename and transform the variables to improve the interpretability of your regression model.

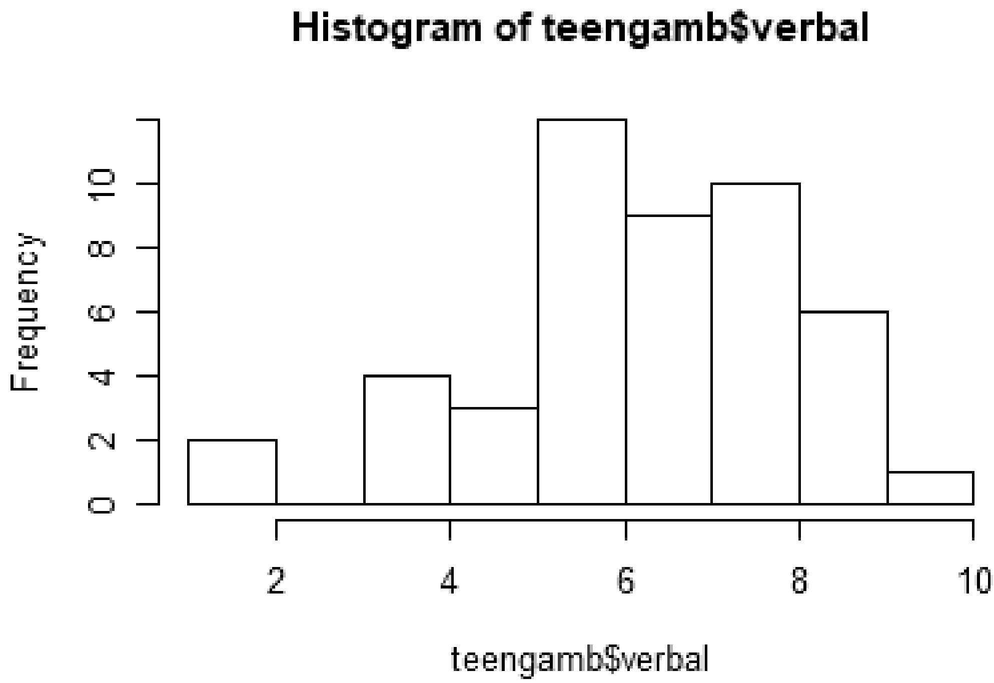
```r
summary(teengamb)
```

```
##       sex             status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

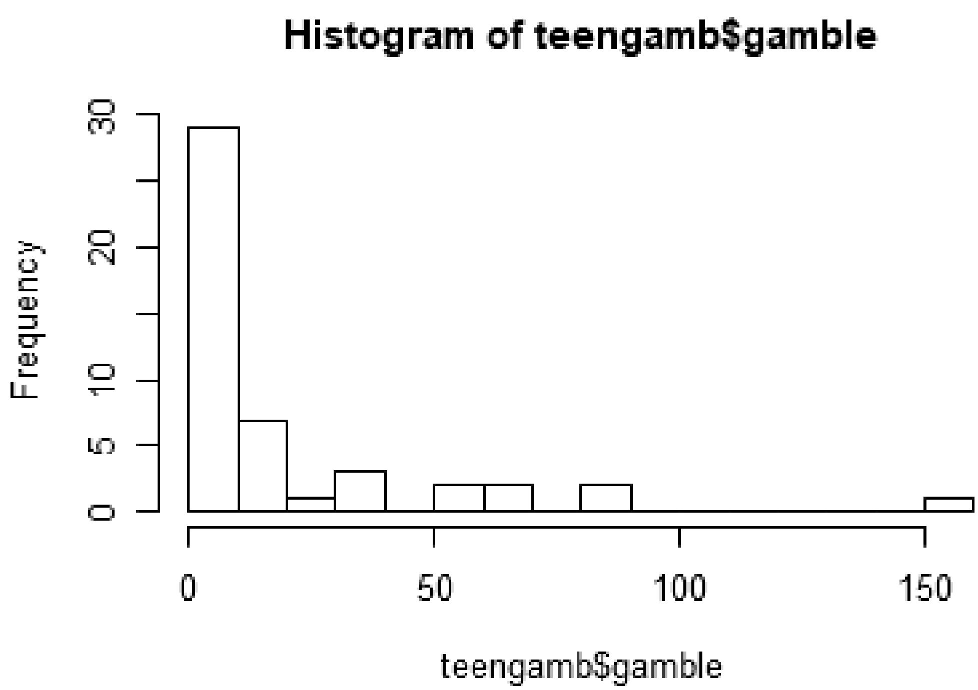
```r
#Sex is a categorical variable
#Status is assumed a linear variable
#We might need to check if correlation exists between status and income
#Verbal is also a linear scale
# Checking distribution
hist(teengamb$income)
```




```r
#Transforming income as it is right skewed
teengamb$inc_log <- log(teengamb$income)

hist(teengamb$verbal)
```



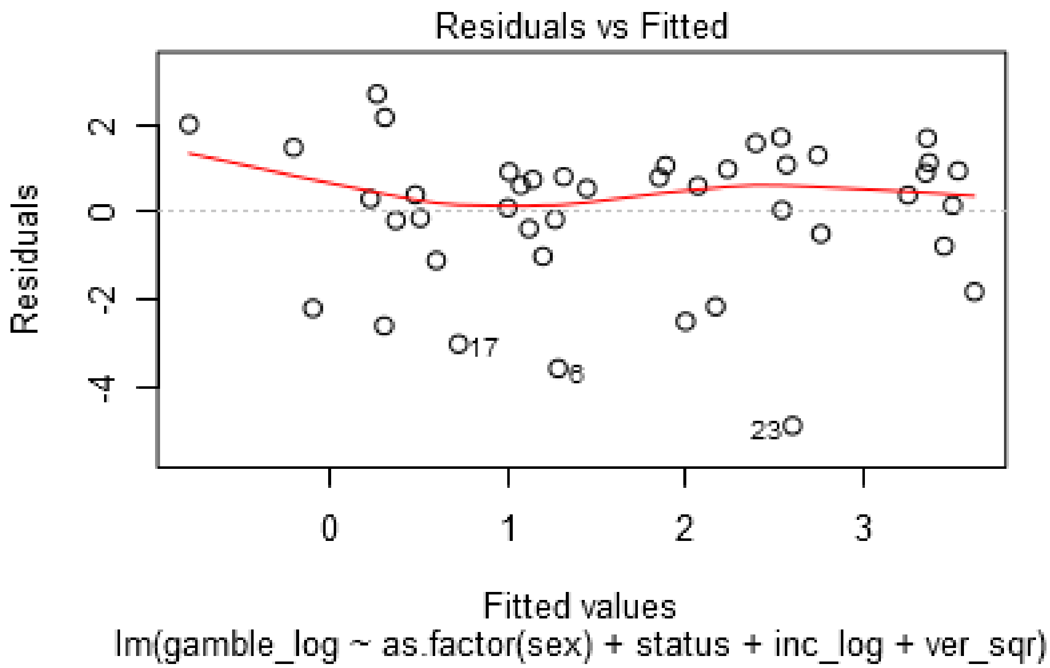

```r
#Transforming verbal as it is left skewed
teengamb$ver_sqr <- (teengamb$verbal)^2

hist(teengamb$gamble, breaks = 20)
```

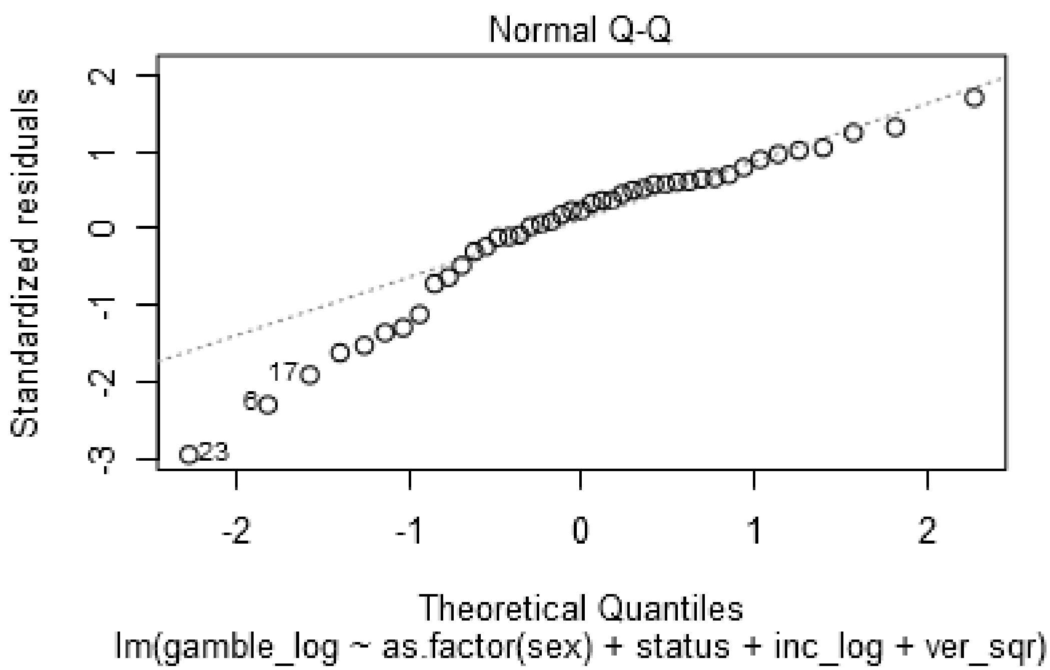


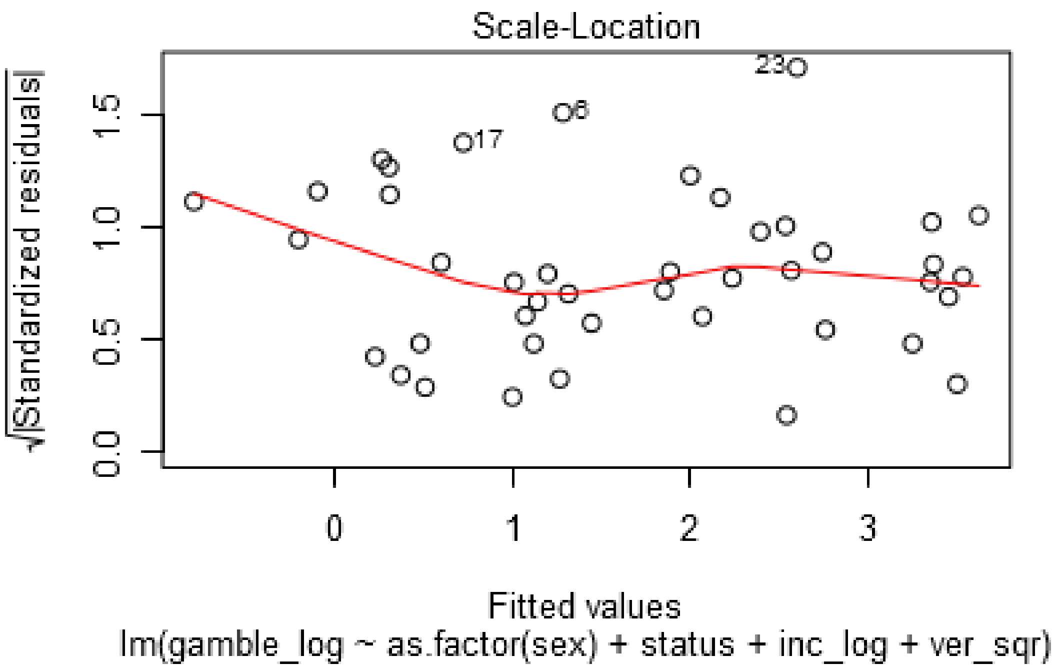
```r
#Transforming gamble as it is right skewed
teen <- teengamb[which(teengamb$gamble > 0 ),]
teen$gamble_log <- log(teen$gamble)

lm_7 <- lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr , data = teen)
summary(lm_7)
```

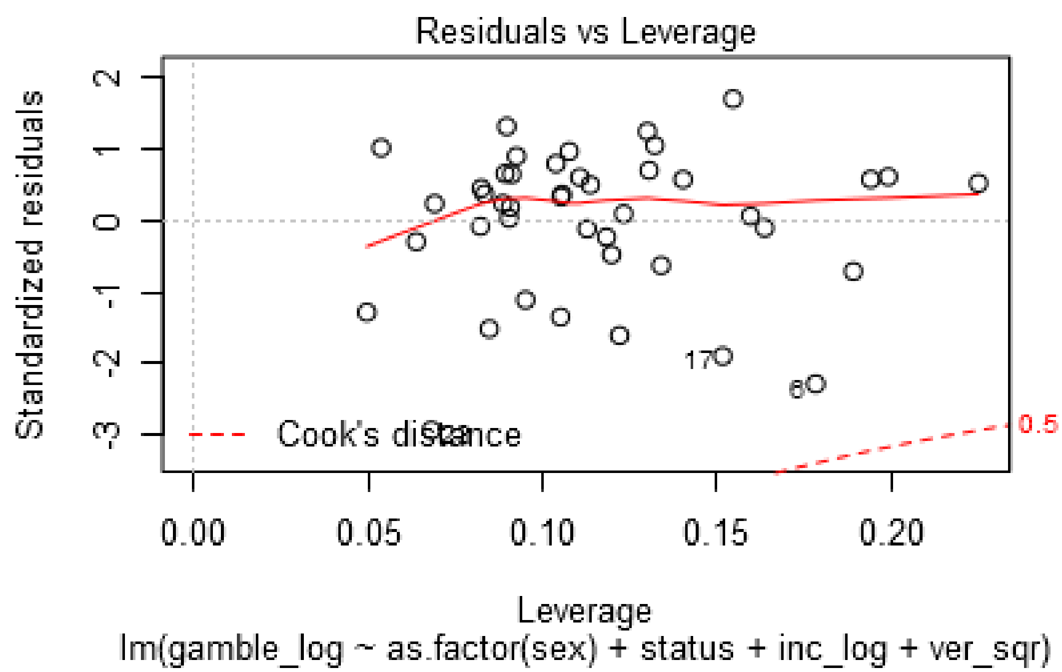
```
##
## Call:
## lm(formula = gamble_log ~ as.factor(sex) + status + inc_log +
##     ver_sqr, data = teen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9045 -0.6390  0.3925  1.0240  2.7069
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.67495    1.22652   0.550   0.5853
## as.factor(sex)1 -1.59139    0.63078  -2.523   0.0159 *
## status           0.02873    0.02121   1.355   0.1836
## inc_log          0.97574    0.37228   2.621   0.0125 *
## ver_sqr         -0.02123    0.01440  -1.475   0.1486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.733 on 38 degrees of freedom
## Multiple R-squared:  0.3489, Adjusted R-squared:  0.2803
## F-statistic:  5.09 on 4 and 38 DF,  p-value: 0.002197
```

```r
plot(lm_7)
```


Residuals vs Fitted
Residuals
Fitted values
lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr)
2
0
-2
-4
0
1
2
3
17
6
23


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr)
2
1
0
-1
-2
-3
-2
-1
0
1
2
17
6
23

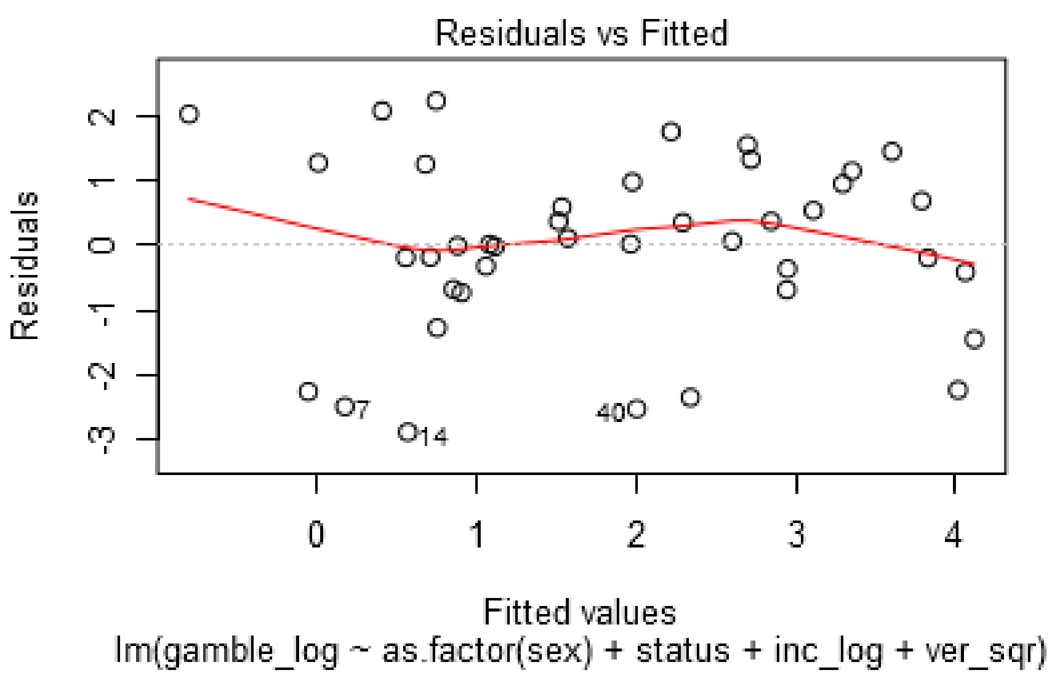


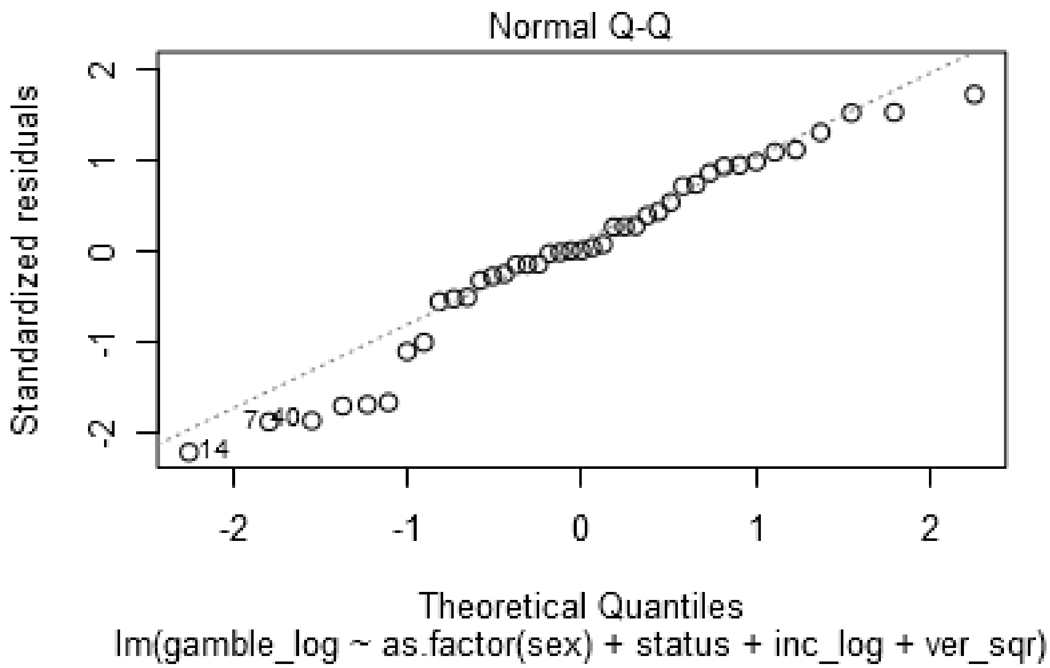


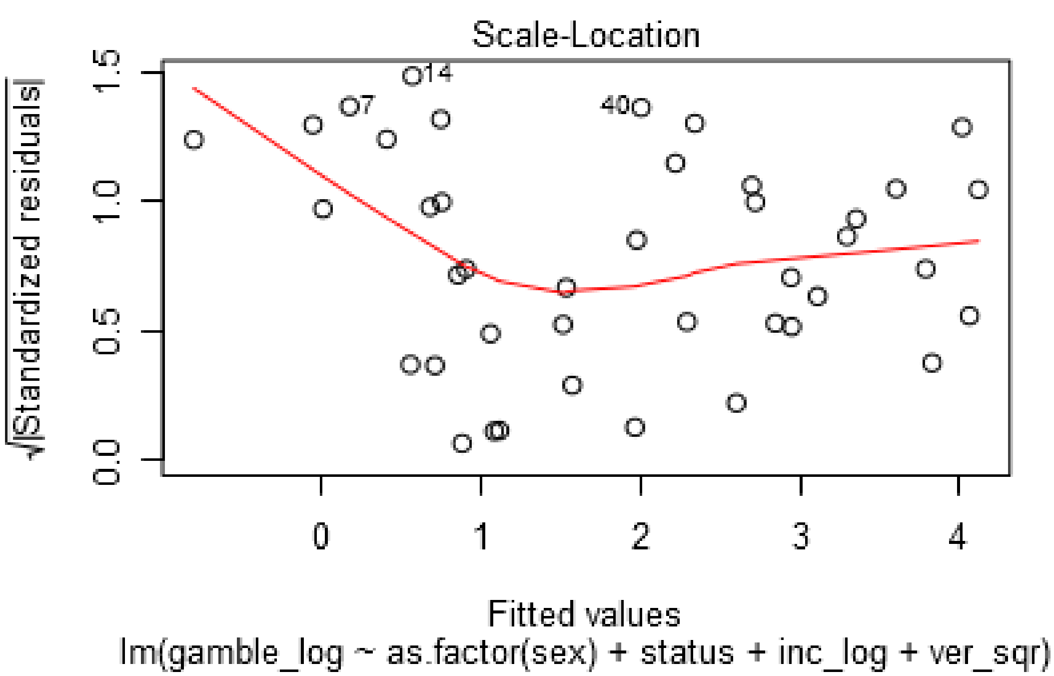

```
#Removing outlier 6 and 23
rownames(teen) <- 1:nrow(teen)
lm_8 <- lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr , data = teen[c(-19,-3),])
summary(lm_8)
```

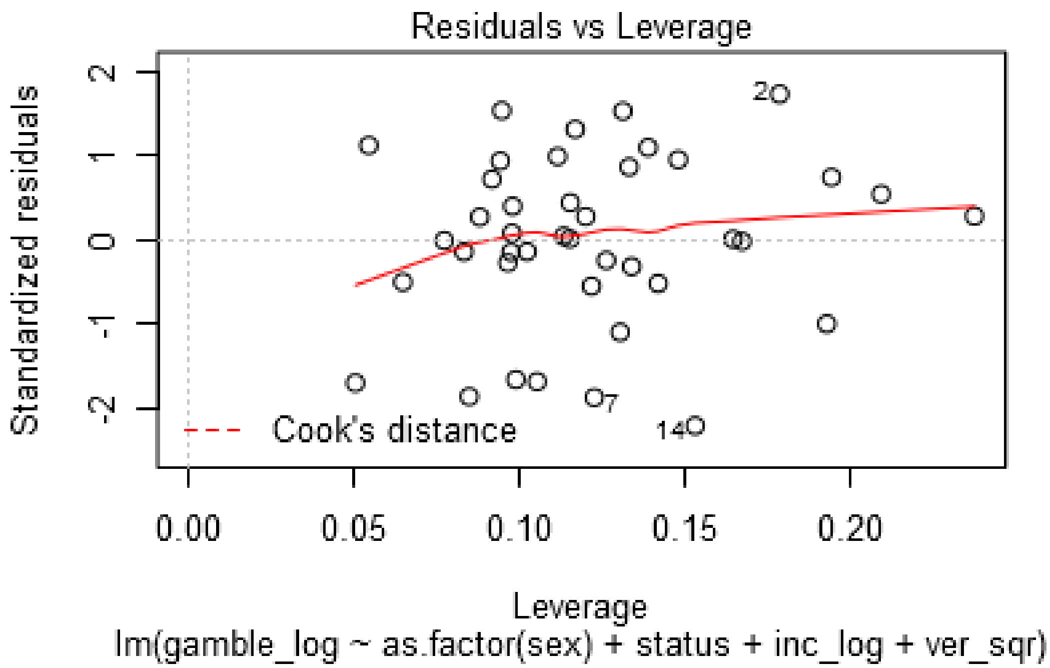
```
##
## Call:
## lm(formula = gamble_log ~ as.factor(sex) + status + inc_log +
##     ver_sqr, data = teen[c(-19, -3), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8732 -0.6734  0.0218  0.9777  2.2277
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.61481    1.04160   0.590  0.55870
## as.factor(sex)1 -1.32875    0.55262  -2.404  0.02147 *
## status           0.04673    0.01852   2.523  0.01620 *
## inc_log          0.99597    0.30559   3.259  0.00244 **
## ver_sqr         -0.03541    0.01223  -2.895  0.00640 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.419 on 36 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4333
## F-statistic: 8.645 on 4 and 36 DF,  p-value: 5.358e-05
```

```
plot(lm_8)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr)


Scale-Location
√|Standardized residuals|
Fitted values
lm(gamble_log ~ as.factor(sex) + status + inc_log + ver_sqr)



2. Create a 95% confidence interval for each of the estimated coefficients and discuss how you would interpret this uncertainty.

```
co <- lm_8$coefficients
se <- sqrt(diag(vcov(lm_8)))
tab_8 <- as.data.frame(cbind(co,se))
tab_8$t_value <- co/se
tab_8$up <- tab_8$co + 1.96*tab_2$se
```

```
## Warning in tab_8$co + 1.96 * tab_2$se: longer object length is not a
## multiple of shorter object length
```

```
tab_8$low <- tab_8$co - 1.96*tab_2$se
```

```
## Warning in tab_8$co - 1.96 * tab_2$se: longer object length is not a
## multiple of shorter object length
```

```
tab_8
```

```
##                          co         se   t_value          up         low
## (Intercept)      0.61481172 1.04159651  0.590259  0.66558521  0.56403822
## as.factor(sex)1 -1.32875242 0.55262061 -2.404457 -1.31554305 -1.34196180
## status           0.04673108 0.01852339  2.522814  0.09750457 -0.00404242
## inc_log          0.99597442 0.30558637  3.259224  1.00918379  0.98276504
## ver_sqr         -0.03540828 0.01223004 -2.895188  0.01536521 -0.08618178
```

3. Predict the amount that a male with average status, income and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values of status, income and verbal score. Which CI is wider and why is this result expected?

```r
teen_new <- teen[FALSE,]
teen_new[1,] <- sapply(teen[which(teen$sex == 0),], mean, na.rm = TRUE)
teen_new[2,] <- sapply(teen[which(teen$sex == 0),], max, na.rm = TRUE)

tab_pred <- as.data.frame(predict(lm_8, newdata = teen_new , interval = "prediction" ))
tab_pred$interval <- tab_pred$upr - tab_pred$lwr
#
```

School expenditure and test scores from USA in 1994-95

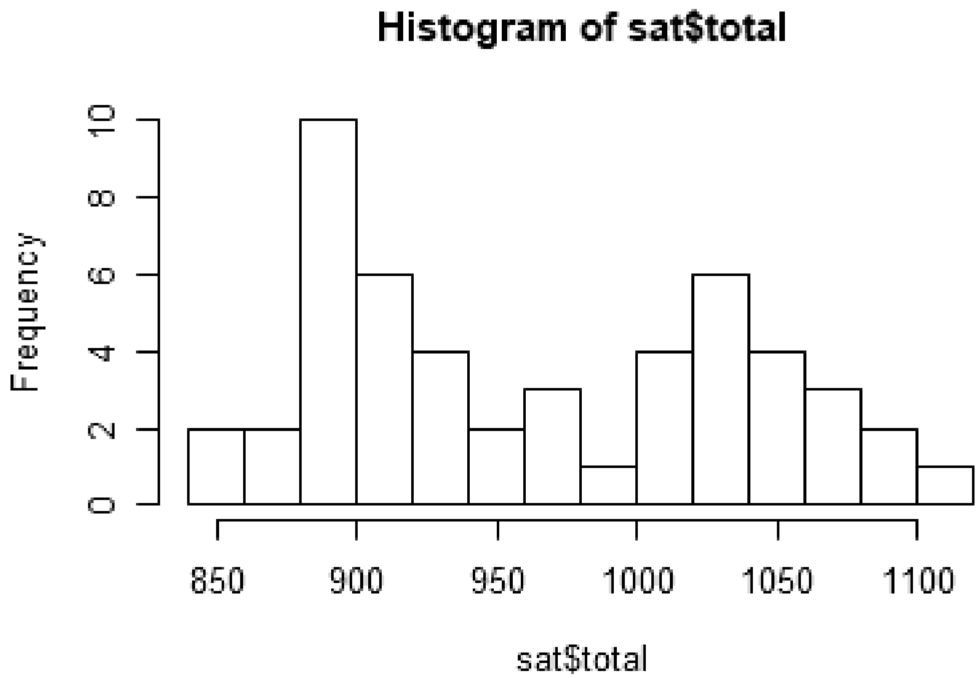
```r
data(sat)
```

1. Fit a model with total sat score as the outcome and expend, ratio and salary as predictors. Make necessary transformation in order to improve the interpretability of the model. Interpret each of the coefficient.

```r
summary(sat)
```

```
##      expend          ratio          salary          takers
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
##      verbal           math           total
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :448.0   Median :497.5   Median : 945.5
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

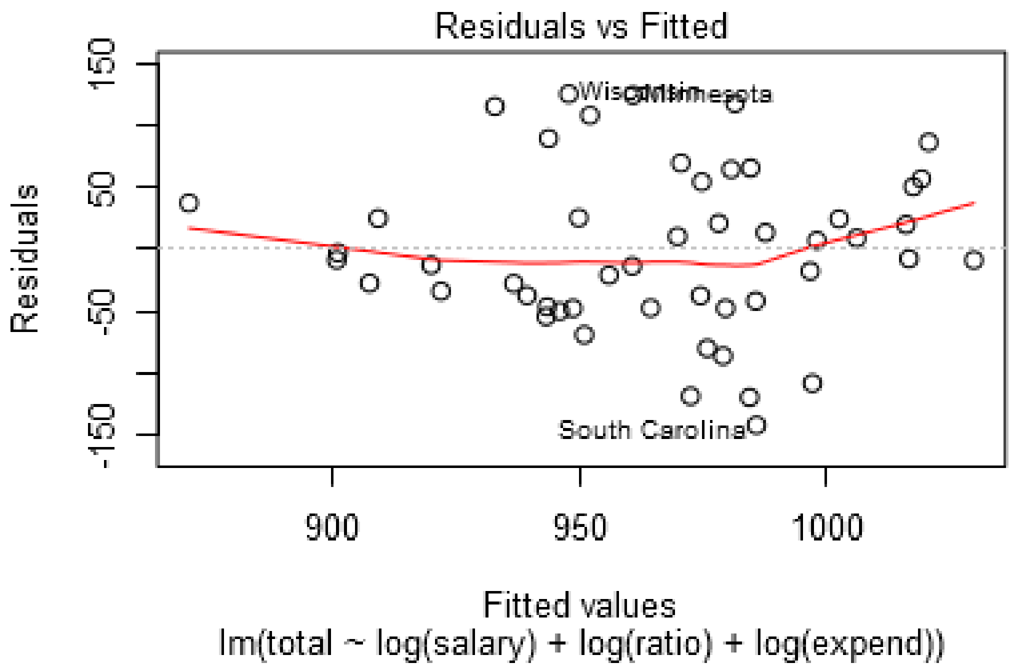
```r
#No nulls and data is empirically correct

#Most of the variables are right skewed, so taking logs for dependent varibles
hist(sat$total, breaks = 10 )
```



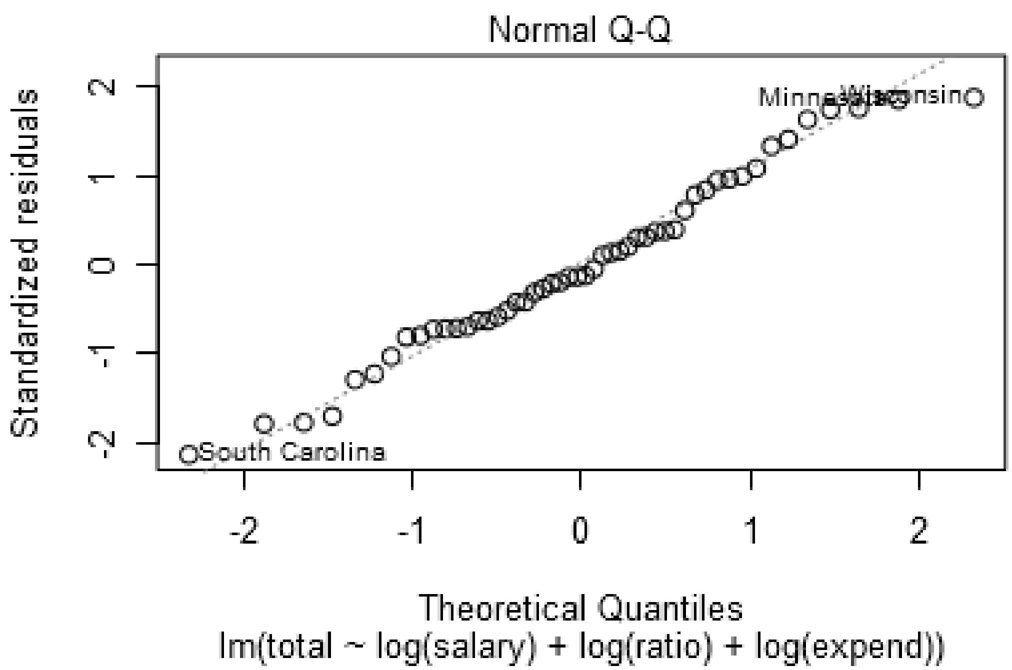

```r
#Independent variable is a bimodal distribution

lm_9 <- lm(total ~ log(salary) + log(ratio) + log(expend) , data = sat)
summary(lm_9)
```

```
##
## Call:
## lm(formula = total ~ log(salary) + log(ratio) + log(expend),
##     data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -141.883  -45.280   -8.312   47.040  125.150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1572.9      301.3   5.221 4.17e-06 ***
## log(salary)   -311.1      161.2  -1.930   0.0598 .
## log(ratio)     117.3      121.2   0.968   0.3381
## log(expend)     92.9      133.7   0.695   0.4905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.08 on 46 degrees of freedom
## Multiple R-squared:  0.2229, Adjusted R-squared:  0.1722
## F-statistic: 4.397 on 3 and 46 DF,  p-value: 0.008403
```

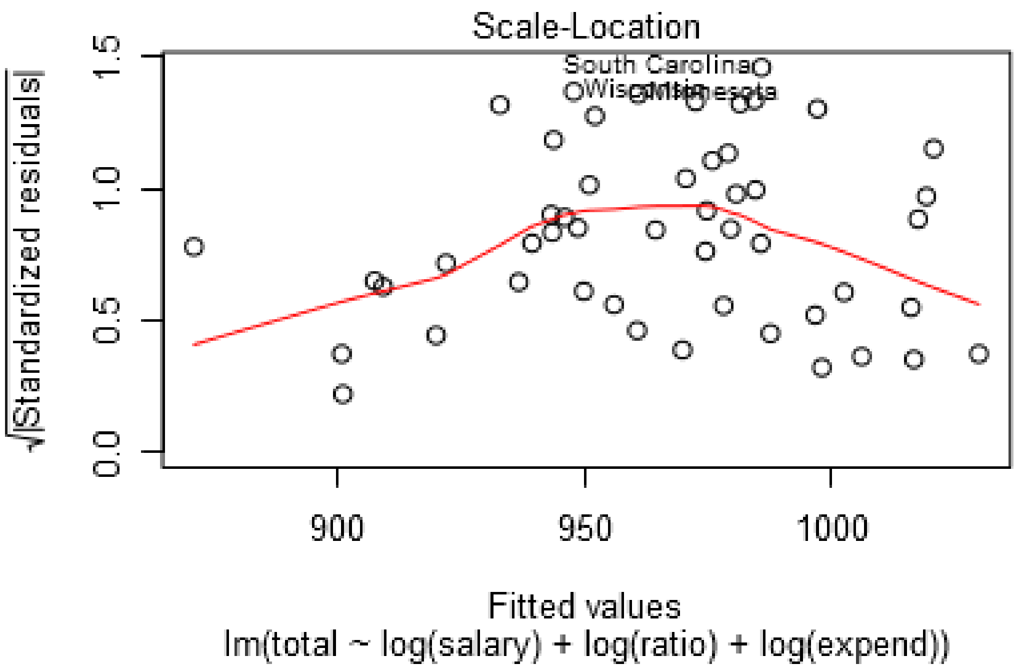
```
plot (lm_9)
```

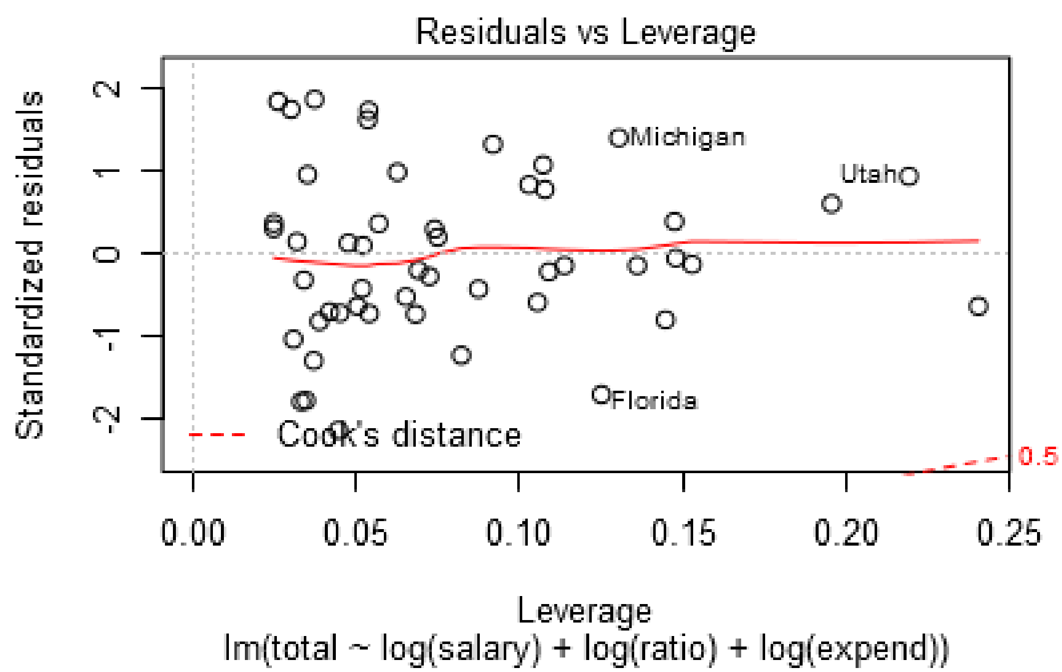


Residuals vs Fitted
Residuals
150
50
-50
-150
900
950
1000
Fitted values
lm(total ~ log(salary) + log(ratio) + log(expend))
Wisconsin
Minnesota
South Carolina



Normal Q-Q
Standardized residuals
2
1
0
-1
-2
Theoretical Quantiles
lm(total ~ log(salary) + log(ratio) + log(expend))
Minnesota
Wisconsin
South Carolina





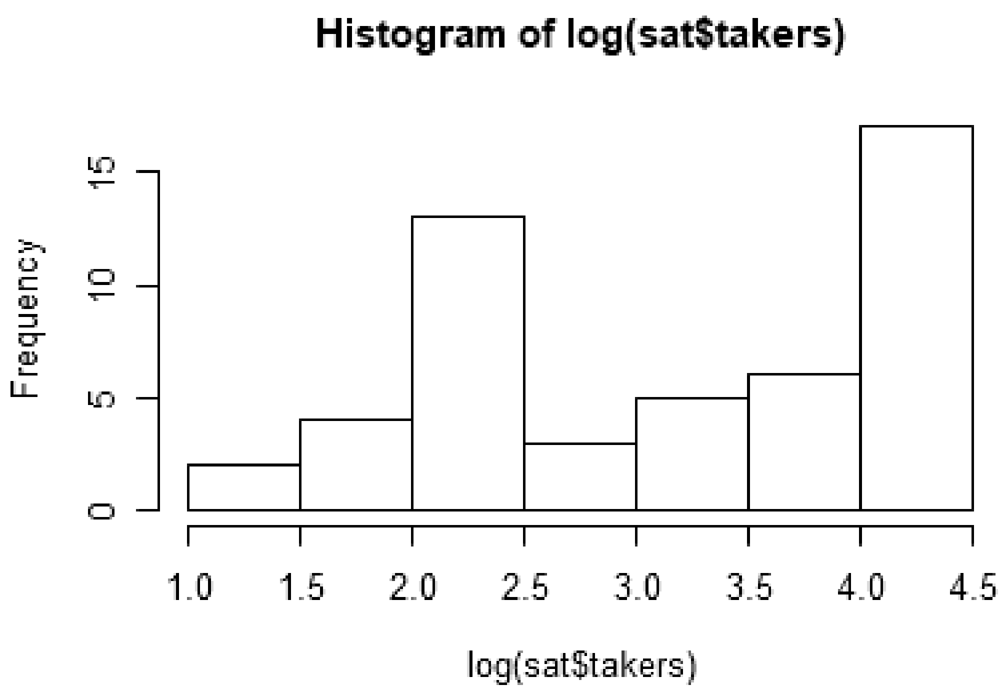

```
#We can use bootstrapping as it doesnt assume how the distribution of coeficients will be like !!
```

2. Construct 98% CI for each coefficient and discuss what you see.

```r
co <- lm_9$coefficients
se <- sqrt(diag(vcov(lm_9)))
tab_9 <- as.data.frame(cbind(co,se))
tab_9$t_value <- co/se
tab_9$up <- tab_9$co + 2.58*tab_9$se
tab_9$low <- tab_9$co - 2.58*tab_9$se
```

3. Now add takers to the model. Compare the fitted model to the previous model and discuss which of the model seem to explain the outcome better?

```r
hist(log(sat$takers))
```




```r
#Takers is also a bimodal distribution; hence it is likely that it will be able to explain the variance
lm_0 <- lm(total ~ log(salary) + log(ratio) + log(expend) + log(takers), data = sat)
summary(lm_0)
```

```
##
## Call:
## lm(formula = total ~ log(salary) + log(ratio) + log(expend) +
##     log(takers), data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.597 -14.263   0.338  15.002  56.373
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  981.203    117.961   8.318 1.19e-10 ***
## log(salary)   33.024     63.616   0.519    0.606
## log(ratio)     5.454     45.799   0.119    0.906
## log(expend)   61.583     49.995   1.232    0.224
## log(takers)  -80.872      4.797 -16.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.45 on 45 degrees of freedom
## Multiple R-squared:  0.8938, Adjusted R-squared:  0.8843
## F-statistic: 94.65 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
#This model explains around 89% of variance which is stark improvement over the previous model
```

Conceptual exercises.

Special-purpose transformations:

For a study of congressional elections, you would like a measure of the relative amount of money raised by each of the two major-party candidates in each district. Suppose that you know the amount of money raised by each candidate; label these dollar values D_i and R_i. You would like to combine these into a single variable that can be included as an input variable into a model predicting vote share for the Democrats.

Discuss the advantages and disadvantages of the following measures:

- The simple difference, $D_i - R_i$

This variable is not normalized
Advantages : Makes it easy to intepret model
Disadvantages : THe difference will depend on the total dollar value area is contributing

- The ratio, D_i/R_i

Advantages : Normalised variable so scale is easy to define and understand Disadvantages : Depends on the Will be a good variable to model if both the candidates are. If districts are polarized ie all of them has strong support for either one of the contestants If the support for one candidate is very high, this might lead to very skewed distribution

- The difference on the logarithmic scale, $logD_i - logR_i$

Advantages : This might be used when one of the candiates have substantially higher funding compared to other one. In such a case the (2) option cant be used
Disadvantages : Cant be used for bimodal

- The relative proportion, $D_i/(D_i + R_i)$.

Advantages : This is normalized variable.The chance for having bimodal distribution even in case of polarised voting area is less Its easy to intepret as well as represent Disadvantages: NA

Transformation

For observed pair of x and y, we fit a simple regression model

$$\mathrm{y} = \alpha + \beta \mathrm{x} + \epsilon$$

which results in estimates $\hat{\alpha} = 1$, $\hat{\beta} = 0.9$, $SE(\hat{\beta}) = 0.03$, $\hat{\sigma} = 2$ and $r = 0.3$.

1. Suppose that the explanatory variable values in a regression are transformed according to the $\mathrm{x}^\star = \mathrm{x} - 10$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $\hat{\alpha}^\star$, $\hat{\beta}^\star$, $\hat{\sigma}^\star$, and $r^\star$. What happens to these quantities when $\mathrm{x}^\star = 10\mathrm{x}$? When $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?
2. Now suppose that the response variable scores are transformed according to the formula $\mathrm{y}^{\star\star} = \mathrm{y} + 10$ and that $\mathrm{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $\hat{\alpha}^{\star\star}$, $\hat{\beta}^{\star\star}$, $\hat{\sigma}^{\star\star}$, and $r^{\star\star}$. What happens to these quantities when $\mathrm{y}^{\star\star} = 5\mathrm{y}$? When $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?
3. In general, how are the results of a simple regression analysis affected by linear transformations of y and x?
4. Suppose that the explanatory variable values in a regression are transformed according to the $\mathrm{x}^\star = 10(\mathrm{x} - 1)$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^\star)$ and $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star)$.
5. Now suppose that the response variable scores are transformed according to the formula $\mathrm{y}^{\star\star} = 5(\mathrm{y}+2)$ and that $\mathrm{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^{\star\star})$ and $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star})$.
6. In general, how are the hypothesis tests and confidence intervals for β affected by linear transformations of y and x?

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your opinions.