Homework 02

Tingrui Huang

Septemeber 20, 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.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
library(carData)
library(arm)
library(faraway)
#Look at the dataset.
summary(heights)
```

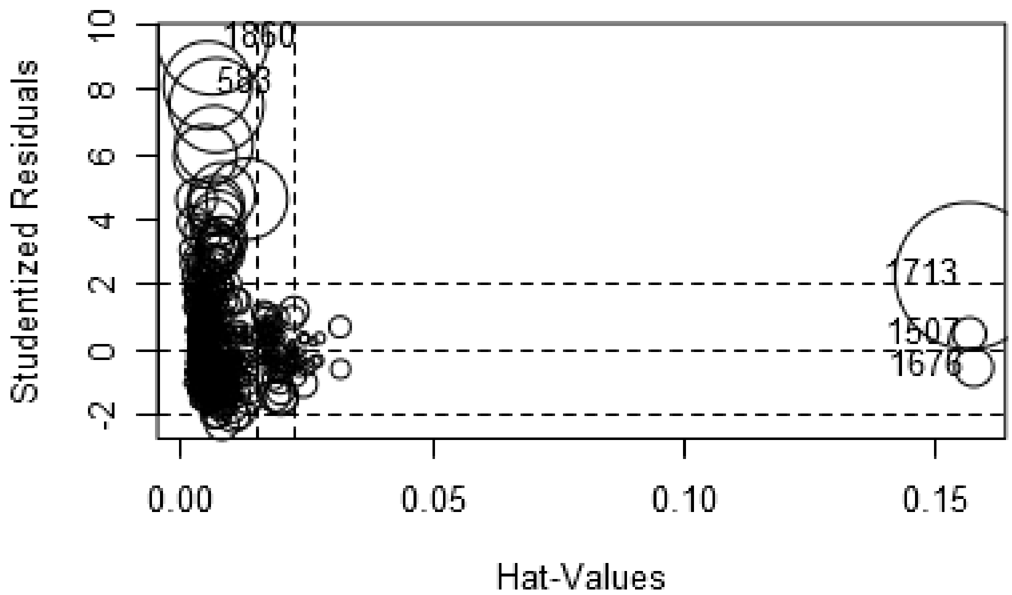
```
##       earn           height1         height2            sex
##  Min.   :     0   Min.   :4.000   Min.   : 0.000   Min.   :1.000
##  1st Qu.:  6000   1st Qu.:5.000   1st Qu.: 3.000   1st Qu.:1.000
##  Median : 16400   Median :5.000   Median : 5.000   Median :2.000
##  Mean   : 20015   Mean   :5.122   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
```

```
# In the dataset we can find that, the survey was conducted in 1990, and many respondents had age young
# which is younger than the legal age for working. Therefore, we need to remove these records.
heights$yearbn[heights$yearbn > 73] <- NA

#There are a lot of NA inputs in the dataset, so we are going to remove these NA values.
na <- which(!complete.cases(heights))
heights_clean_1 <- heights[-na,]

# Since we are going to study the relation between earn and other variables, therefore, if a person's e
# then we need to remove it from the dataset
no_income <- which(heights_clean_1$earn==0)
heights_clean <- heights_clean_1[-no_income,]

#Discover outliers by using Bonferroni outlier test
regall <- lm(earn~height1+height2+sex+race+hisp+ed+yearbn+height, data = heights_clean)
outlierTest(regall)
```

```
##      rstudent unadjusted p-value Bonferonni p
## 1860 9.571983         5.9081e-21   7.0247e-18
## 583  8.143793         9.6660e-16   1.1493e-12
## 351  7.525467         1.0388e-13   1.2351e-10
## 618  6.312884         3.8693e-10   4.6005e-07
## 1277 5.951937         3.4897e-09   4.1492e-06
## 314  4.743934         2.3521e-06   2.7967e-03
## 2020 4.636947         3.9299e-06   4.6727e-03
## 1419 4.625059         4.1579e-06   4.9438e-03
## 967  4.476911         8.3064e-06   9.8763e-03
## 1428 4.351768         1.4673e-05   1.7446e-02
```

```
influencePlot(regall)
```

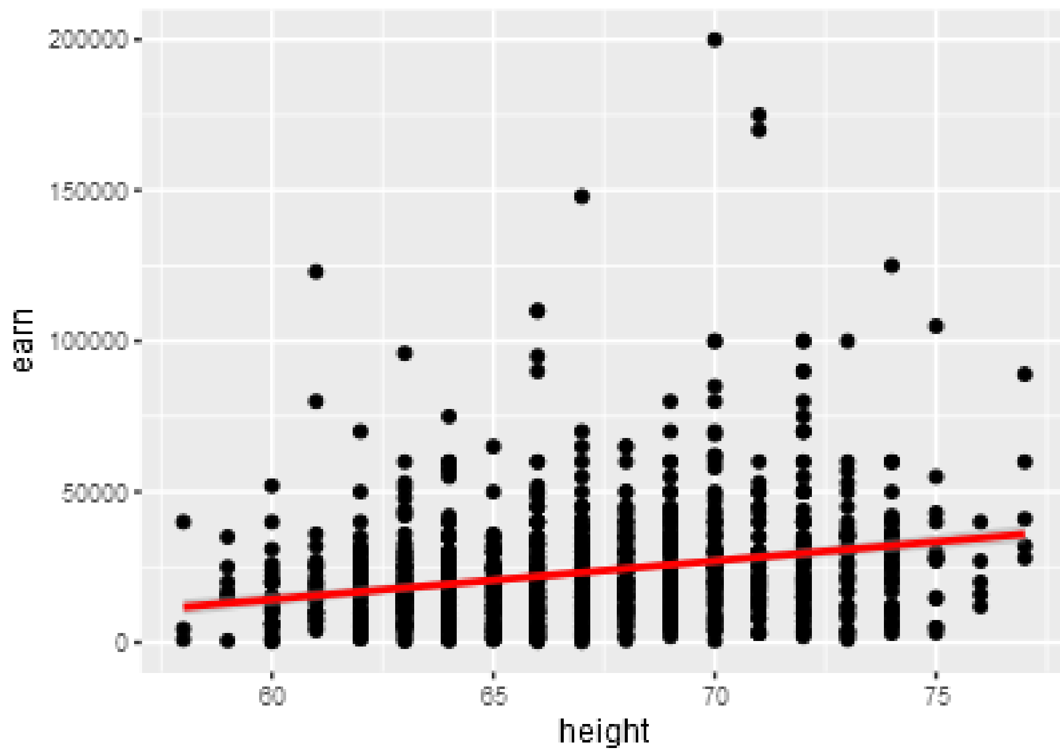



```
##         StudRes         Hat       CookD
## 583   8.1437932 0.005285456 0.041741433
## 1507  0.4764611 0.156663217 0.005274913
## 1676 -0.5598991 0.157535395 0.007331765
## 1713  2.2680271 0.156492210 0.118874562
## 1860  9.5719826 0.006303674 0.067475395
```

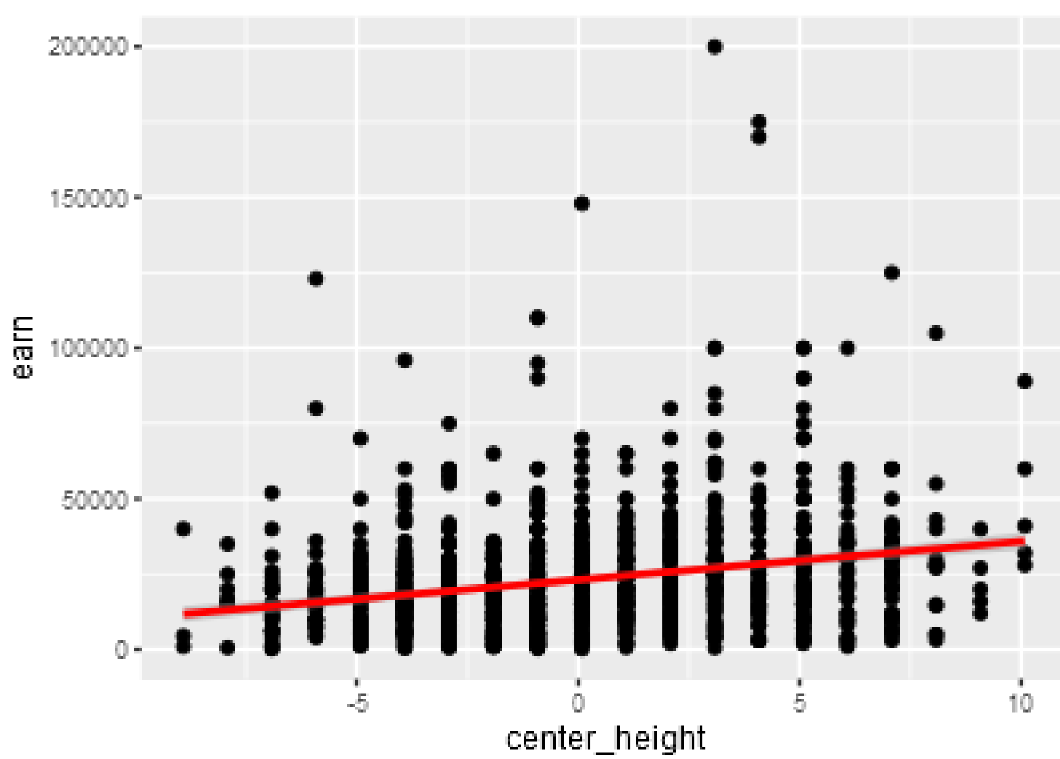
```
#Convert sex into 0 for Men and 1 for Women
heights_clean$sex <- heights_clean$sex - 1
View(heights_clean)
```

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?

```
#Regress "earn" onto "height"
reg_h_1 <- lm(earn~height, data = heights_clean)
ggplot(reg_h_1)+aes(height,earn)+geom_point()+stat_smooth(method='lm',col='red')
```




```r
#Since thereis no one's height is zero, therefore, I would center the height to its mean
center_height <- heights_clean$height - mean(heights_clean$height)
reg_h_2 <- lm(earn~center_height, data = heights_clean)
ggplot(reg_h_2)+aes(center_height,earn)+geom_point()+stat_smooth(method='lm',col='red')
```



```r
summary(reg_h_2)
```

```
##
## Call:
## lm(formula = earn ~ center_height, data = heights_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30211 -11318  -3403   6579 172953
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23128.3      547.1   42.27   <2e-16 ***
## center_height   1271.1      142.3    8.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1187 degrees of freedom
## Multiple R-squared:  0.06295,    Adjusted R-squared:  0.06216
## F-statistic: 79.74 on 1 and 1187 DF,  p-value: < 2.2e-16
```

```r
#Interpretation: in the refined model, for a person with average height (center_height=0) has a income
#And each unit increase in heights, will be resulted in 1271.1 more income.
```

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

```r
# Test 1. Put everything in, since all of the three variables could influence earning.
reg_t1 <- lm(earn~height+sex+race+ed+yearbn, data = heights_clean)
summary(reg_t1)
```

```
##
## Call:
## lm(formula = earn ~ height + sex + race + ed + yearbn, data = heights_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39218  -9692  -2217   6017 159011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18362.5    13070.0  -1.405    0.160
## height         292.7      185.9   1.575    0.116
## sex          -9876.1     1426.4  -6.924 7.19e-12 ***
## race          -793.0      840.0  -0.944    0.345
## ed            2773.3      209.9  13.214  < 2e-16 ***
## yearbn        -183.3       32.5  -5.640 2.13e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17270 on 1183 degrees of freedom
## Multiple R-squared:  0.2173, Adjusted R-squared:  0.214
## F-statistic:  65.7 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```r
# Test 2. Still put everything in but we are going to assume there are interactions between each of the
reg_t2 <- lm(earn~height*sex*race*ed*yearbn, data=heights_clean)
summary(reg_t2)
```

```
##
## Call:
## lm(formula = earn ~ height * sex * race * ed * yearbn, data = heights_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47035  -9663  -2355   6244 157840
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 596532.77 1007593.70   0.592    0.554
## height                       -9181.37   14631.38  -0.628    0.530
## sex                        -841069.40 1178035.20  -0.714    0.475
## race                       -645090.79  723224.62  -0.892    0.373
## ed                          -68052.92   78054.15  -0.872    0.383
## yearbn                      -11877.27   19889.27  -0.597    0.551
## height:sex                   13019.63   17615.52   0.739    0.460
## height:race                   9630.26   10538.30   0.914    0.361
## sex:race                   1070100.71  889459.06   1.203    0.229
## height:ed                     1086.85    1136.27   0.957    0.339
## sex:ed                       87660.37   90235.55   0.971    0.332
## race:ed                      64030.81   58292.21   1.098    0.272
## height:yearbn                  179.64     289.84   0.620    0.536
## sex:yearbn                   13643.34   24821.25   0.550    0.583
## race:yearbn                  12279.00   14993.41   0.819    0.413
## ed:yearbn                     1255.39    1503.16   0.835    0.404
## height:sex:race             -16463.03   13429.01  -1.226    0.220
## height:sex:ed                -1374.00    1349.98  -1.018    0.309
## height:race:ed                -963.90     853.03  -1.130    0.259
## sex:race:ed                 -99643.74   69481.93  -1.434    0.152
## height:sex:yearbn             -207.05     373.67  -0.554    0.580
## height:race:yearbn            -184.20     219.59  -0.839    0.402
## sex:race:yearbn             -17831.75   19588.30  -0.910    0.363
## height:ed:yearbn               -19.21      21.97  -0.875    0.382
## sex:ed:yearbn                -1347.75    1852.97  -0.727    0.467
## race:ed:yearbn               -1158.25    1153.03  -1.005    0.315
## height:sex:race:ed            1540.22    1048.63   1.469    0.142
## height:sex:race:yearbn         271.04     297.39   0.911    0.362
## height:sex:ed:yearbn            20.63      27.94   0.738    0.460
## height:race:ed:yearbn           17.43      16.96   1.028    0.304
## sex:race:ed:yearbn            1617.51    1468.61   1.101    0.271
## height:sex:race:ed:yearbn      -24.70      22.34  -1.106    0.269
##
## Residual standard error: 17170 on 1157 degrees of freedom
## Multiple R-squared:  0.2431, Adjusted R-squared:  0.2228
## F-statistic: 11.99 on 31 and 1157 DF,  p-value: < 2.2e-16
```

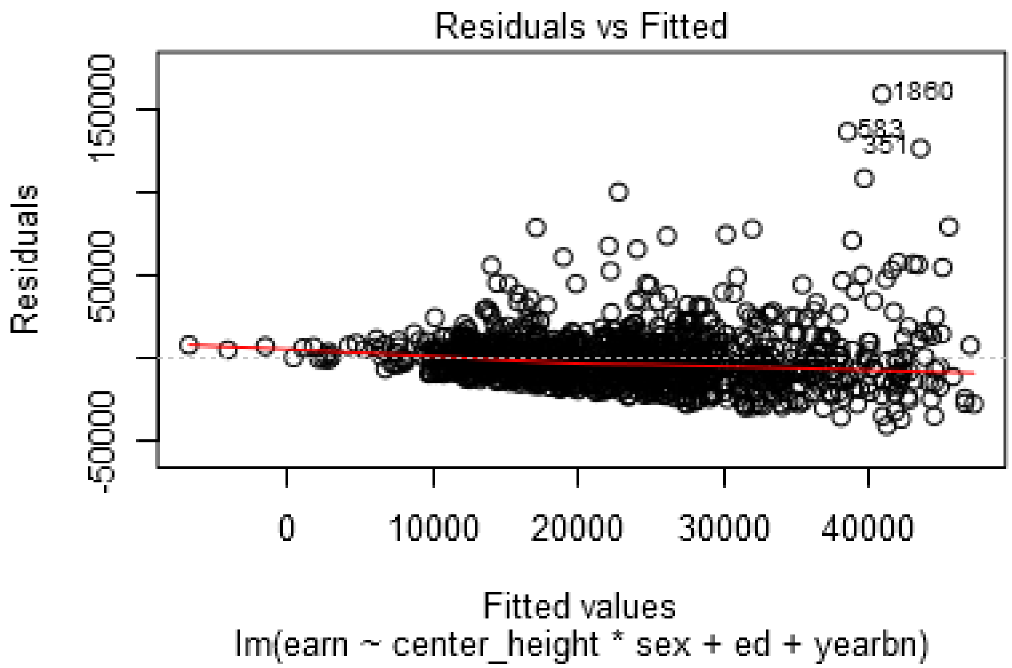
```r
# Test 3. Based on test 1 and 2, I select "sex", "ed" and "yearbn" plus centered "height"
reg_t3 <- lm(earn~center_height+sex+ed+yearbn, data = heights_clean)
summary(reg_t3)
```

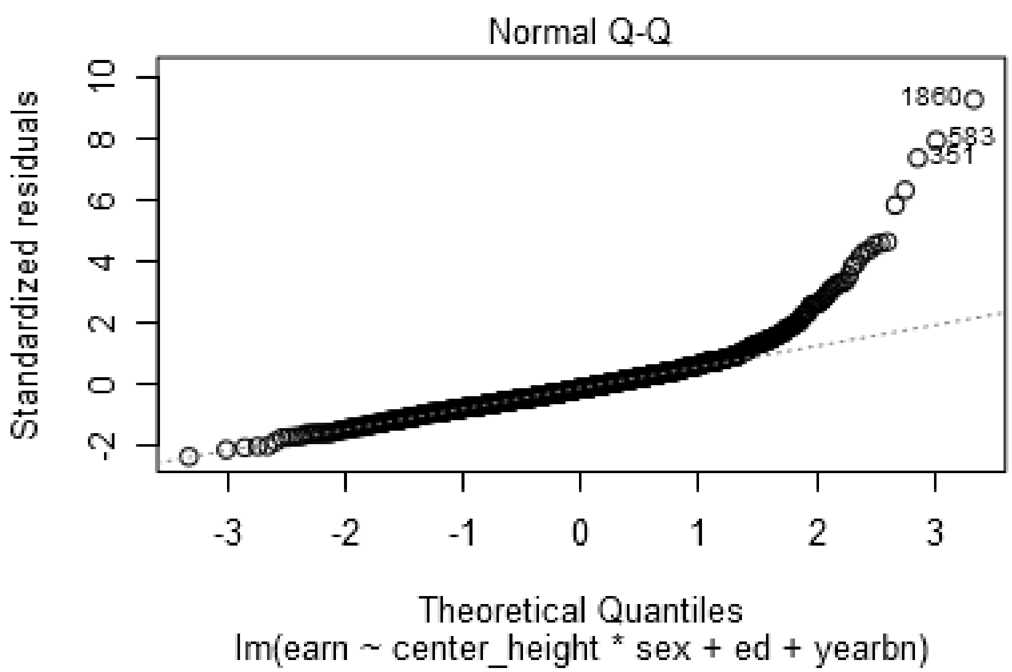
```
##
## Call:
## lm(formula = earn ~ center_height + sex + ed + yearbn, data = heights_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39127  -9667  -2341   5954 159146
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      300.7     3172.7   0.095   0.9245
## center_height    308.7      185.1   1.668   0.0956 .
## sex            -9802.4     1424.2  -6.883 9.48e-12 ***
## ed              2771.6      209.9  13.207  < 2e-16 ***
## yearbn          -183.7       32.5  -5.653 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17270 on 1184 degrees of freedom
## Multiple R-squared:  0.2168, Adjusted R-squared:  0.2141
## F-statistic: 81.91 on 4 and 1184 DF,  p-value: < 2.2e-16
```

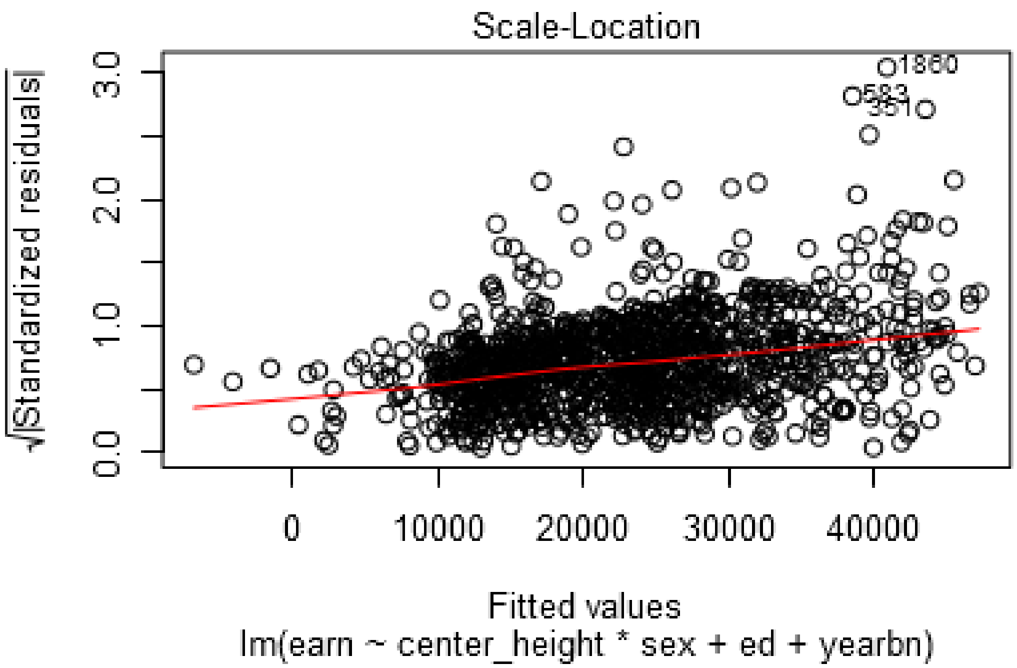
```r
# Test 4. Consider there are nteraction between
reg_t4 <- lm(earn~center_height*sex+ed+yearbn, data = heights_clean)
summary(reg_t4)
```

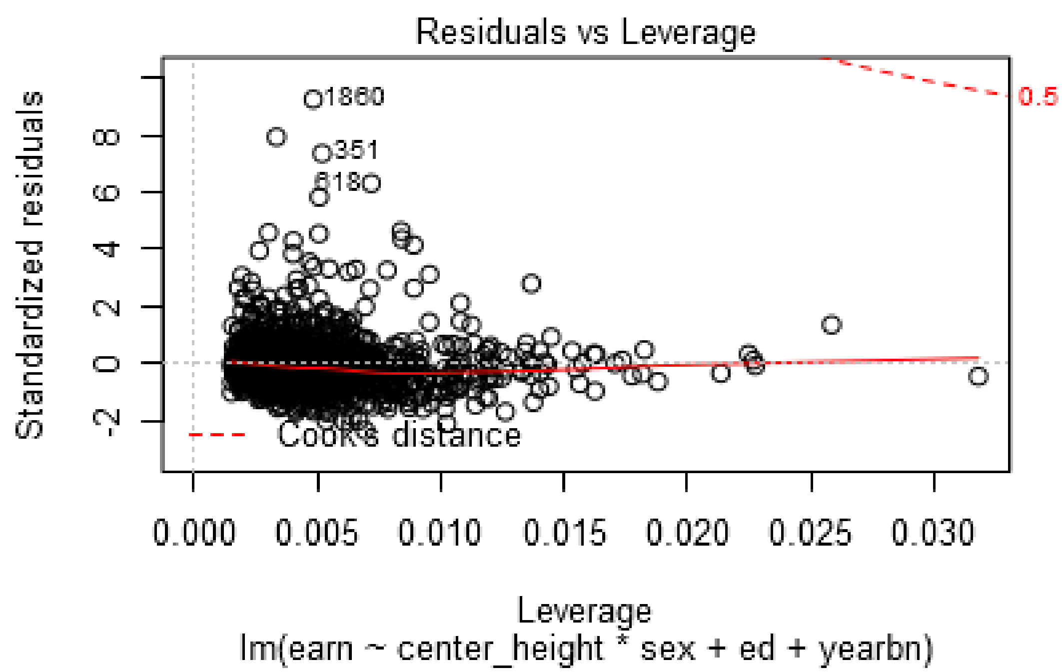
```
##
## Call:
## lm(formula = earn ~ center_height * sex + ed + yearbn, data = heights_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40236  -9578  -2224   6140 159088
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1090.17    3256.83  -0.335   0.7379
## center_height       651.80     261.46   2.493   0.0128 *
## sex               -9493.08    1432.41  -6.627 5.18e-11 ***
## ed                 2786.02     209.79  13.280  < 2e-16 ***
## yearbn             -181.35      32.49  -5.582 2.95e-08 ***
## center_height:sex  -678.65     365.67  -1.856   0.0637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17250 on 1183 degrees of freedom
## Multiple R-squared:  0.219,  Adjusted R-squared:  0.2157
## F-statistic: 66.35 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```r
plot(reg_t4)
```


Residuals vs Fitted
Residuals
150000
50000
-50000
0
10000
20000
30000
40000
1860
583
351
Fitted values
lm(earn ~ center_height * sex + ed + yearbn)

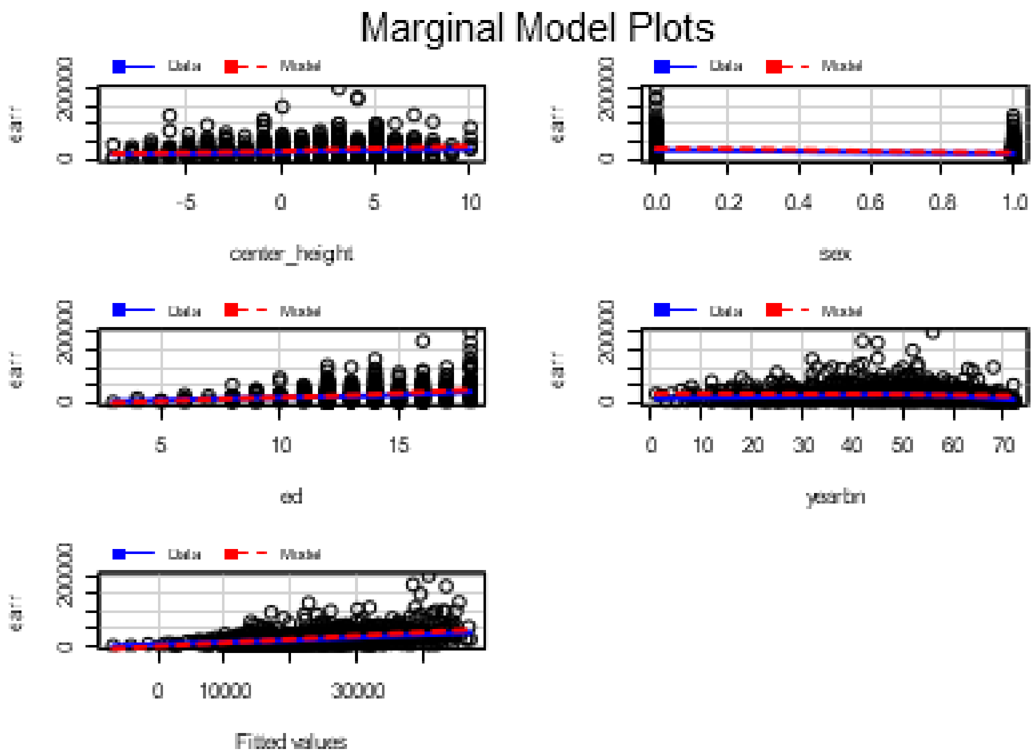

Normal Q-Q
Standardized residuals
10
8
6
4
2
0
-2
-3
-2
-1
0
1
2
3
1860
583
351
Theoretical Quantiles
lm(earn ~ center_height * sex + ed + yearbn)






```
marginalModelPlots(reg_t4)
```

```
## Warning in mmps(...): Interactions and/or factors skipped
```




```
# Test 5. Use log transformation for earn
reg_t5 <- lm(log(earn)~center_height*sex+ed+yearbn, data = heights_clean)
summary(reg_t5)
```

```
##
## Call:
## lm(formula = log(earn) ~ center_height * sex + ed + yearbn, data = heights_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5698 -0.3503  0.1395  0.5292  2.0824
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.740975   0.154944  56.414  < 2e-16 ***
## center_height      0.021332   0.012439   1.715   0.0866 .
## sex               -0.447716   0.068147  -6.570 7.53e-11 ***
## ed                 0.125660   0.009981  12.590  < 2e-16 ***
## yearbn            -0.009895   0.001546  -6.402 2.21e-10 ***
## center_height:sex -0.011262   0.017397  -0.647   0.5175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8208 on 1183 degrees of freedom
## Multiple R-squared:  0.2098, Adjusted R-squared:  0.2064
## F-statistic:  62.8 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```
# According the p-value and plots, I prefer the model 5, which include the centered height, sex, ed and
# First of all, I think it's very close to the real world cases. People with better education and older
```

4. Interpret all model coefficients.

```r
summary(reg_t5)
```

```
##
## Call:
## lm(formula = log(earn) ~ center_height * sex + ed + yearbn, data = heights_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5698 -0.3503  0.1395  0.5292  2.0824
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.740975   0.154944  56.414  < 2e-16 ***
## center_height      0.021332   0.012439   1.715   0.0866 .
## sex               -0.447716   0.068147  -6.570 7.53e-11 ***
## ed                 0.125660   0.009981  12.590  < 2e-16 ***
## yearbn            -0.009895   0.001546  -6.402 2.21e-10 ***
## center_height:sex -0.011262   0.017397  -0.647   0.5175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8208 on 1183 degrees of freedom
## Multiple R-squared:  0.2098, Adjusted R-squared:  0.2064
## F-statistic:  62.8 on 5 and 1183 DF,  p-value: < 2.2e-16
```

```r
# Intercept: Intercept represents the average income for a male person with average heights, no educati
# the year of 1900.

# sex: Females earn less than males by 44%.

# education: People with higher education will earn more than people that are less educated. The differ
# of each level of education is 12%

# yearbn: older people tend to make more money than younger people.

# center height: height has positive correlation with earn, every unit taller in height will result in
# in income.
```

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

```r
confint(reg_t5, level = 0.95)
```

```
##                          2.5 %       97.5 %
## (Intercept)        8.436978733  9.044970965
## center_height     -0.003073164  0.045736375
## sex               -0.581418873 -0.314013606
## ed                 0.106078453  0.145241903
## yearbn            -0.012927102 -0.006862073
## center_height:sex -0.045393148  0.022870020
```

```r
# The confidence intervals for "intercept", "sex", "ed", and "yearbn" are not across 0, therefore, I wo
# these predictors are more statistically significant than others. Although the CI of height across the
# I would still consider its influence since it could be a important variable.
```

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:

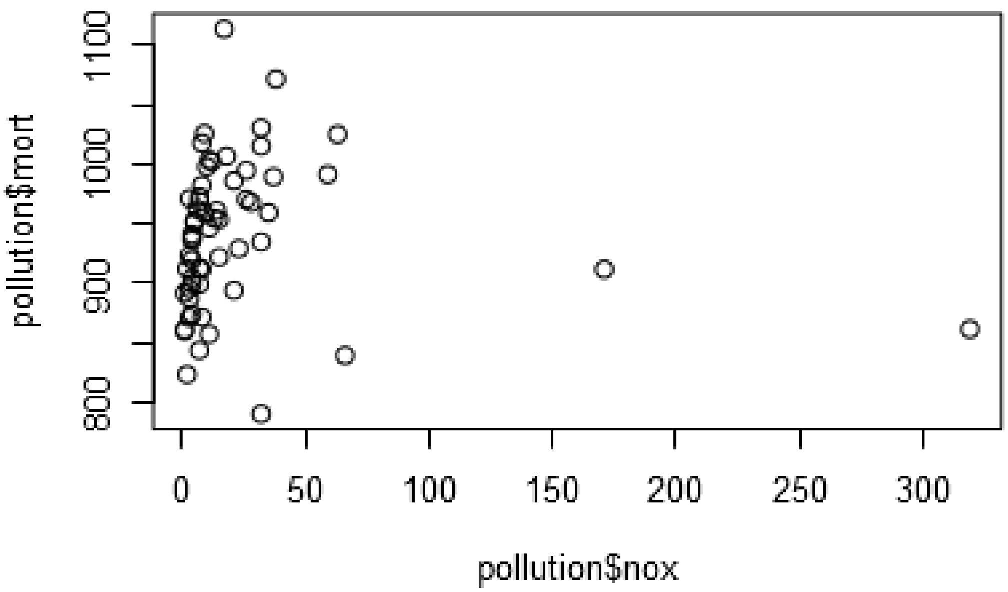
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all 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.

```
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.

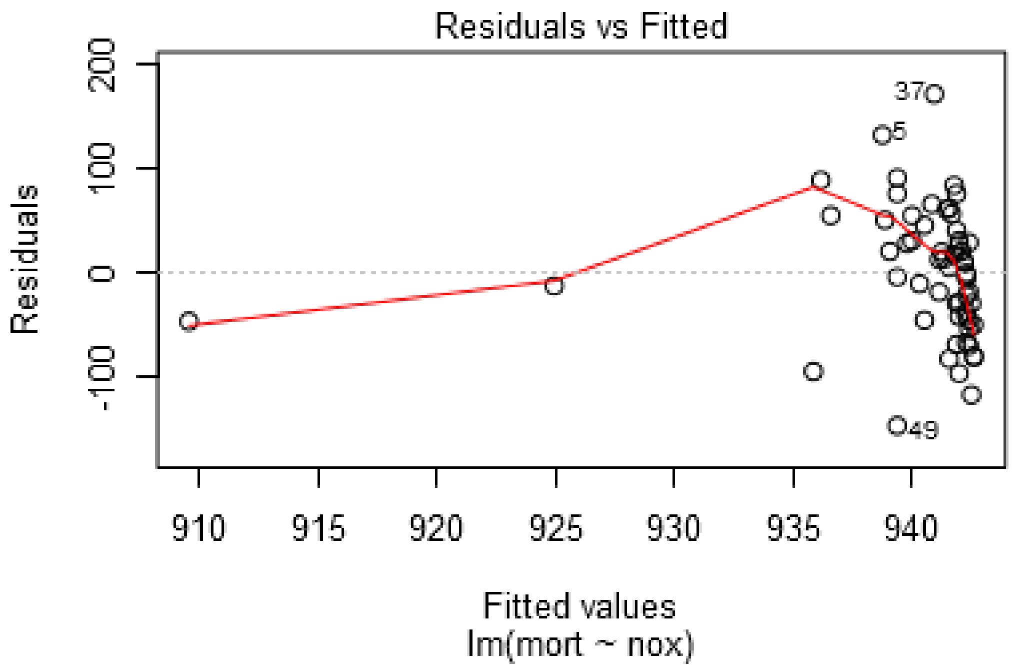
```
plot(x=pollution$nox,y=pollution$mort)
```

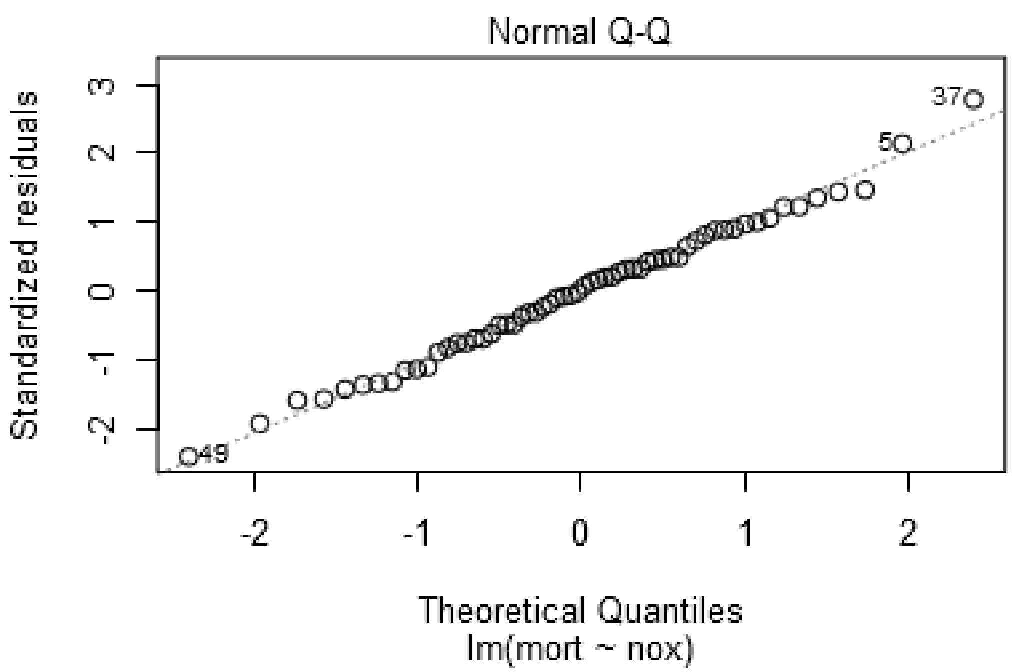


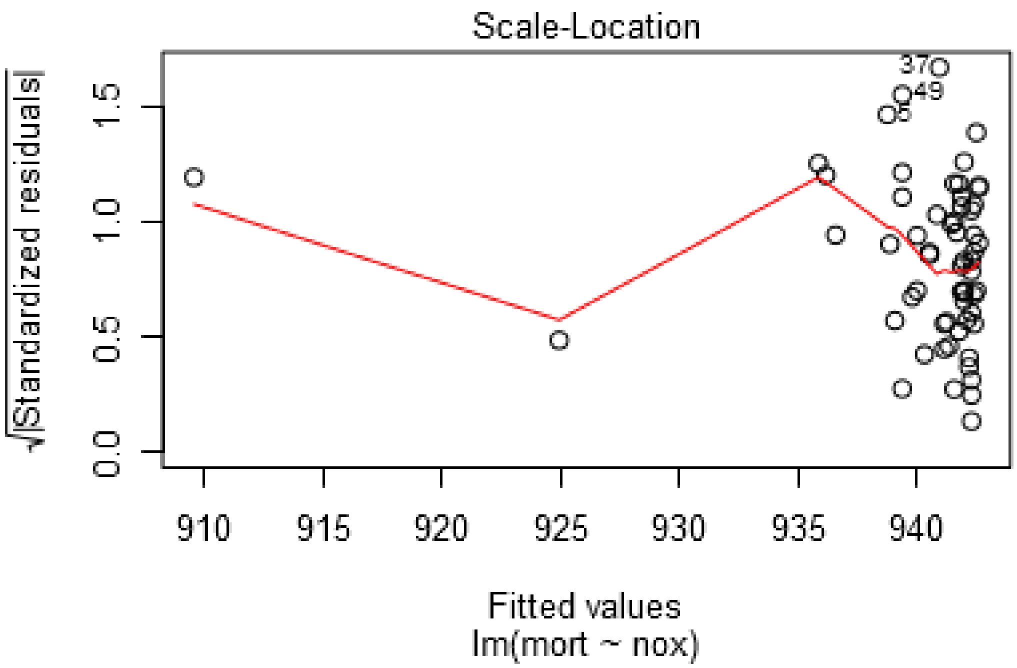

```r
# Based on the plot, I would say the regression could fit these data, but let's try
pol_t1 <- lm(mort~nox, data = pollution)
summary(pol_t1)
```

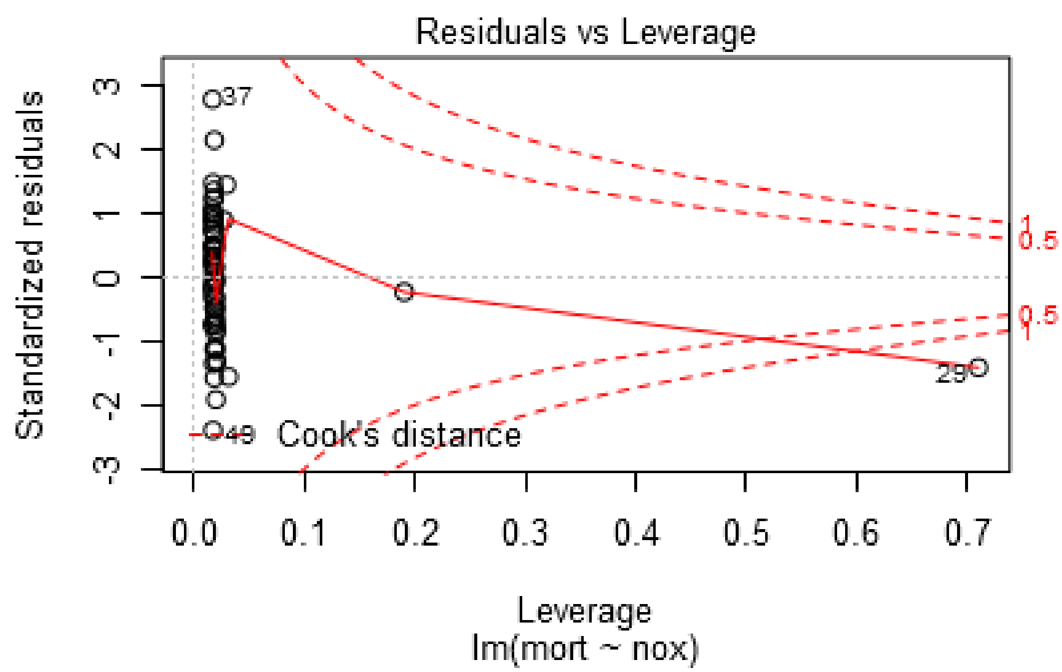
```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.654  -43.710    1.751   41.663  172.211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 942.7115     9.0034 104.706   <2e-16 ***
## 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
plot(pol_t1)
```


Residuals vs Fitted
Residuals
200
100
0
-100
37
5
49
910
915
920
925
930
935
940
Fitted values
lm(mort ~ nox)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
37
5
49
-2
-1
0
1
2
Theoretical Quantiles
lm(mort ~ nox)

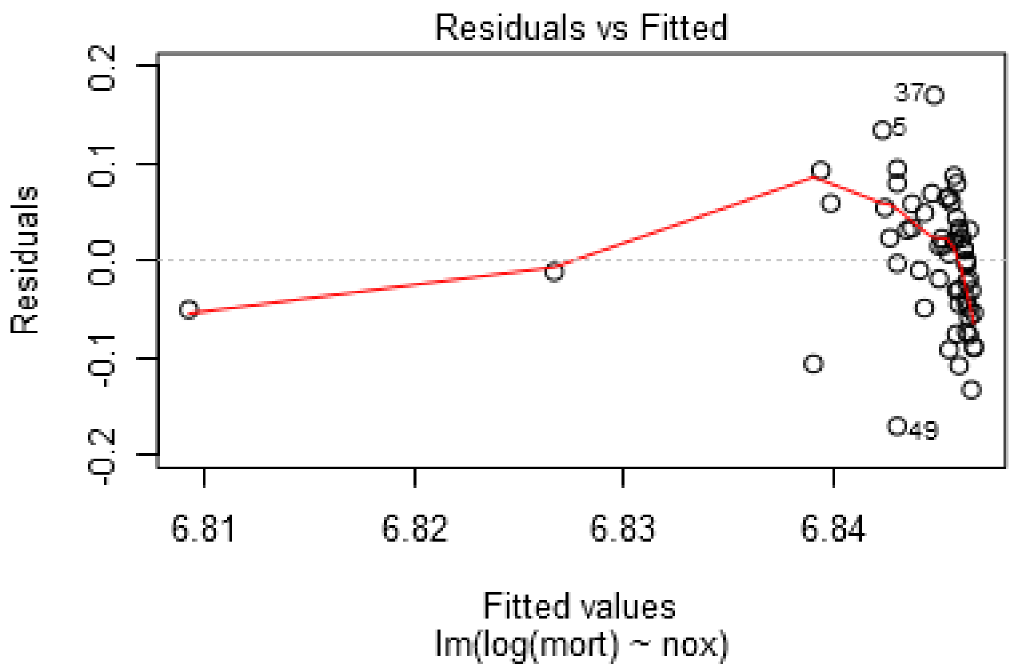


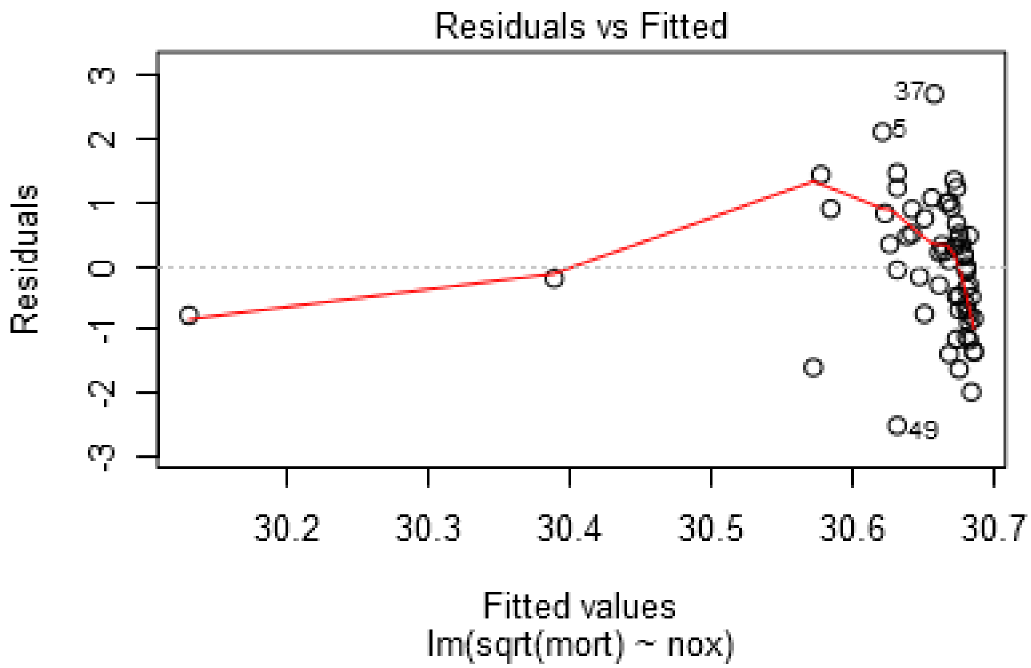


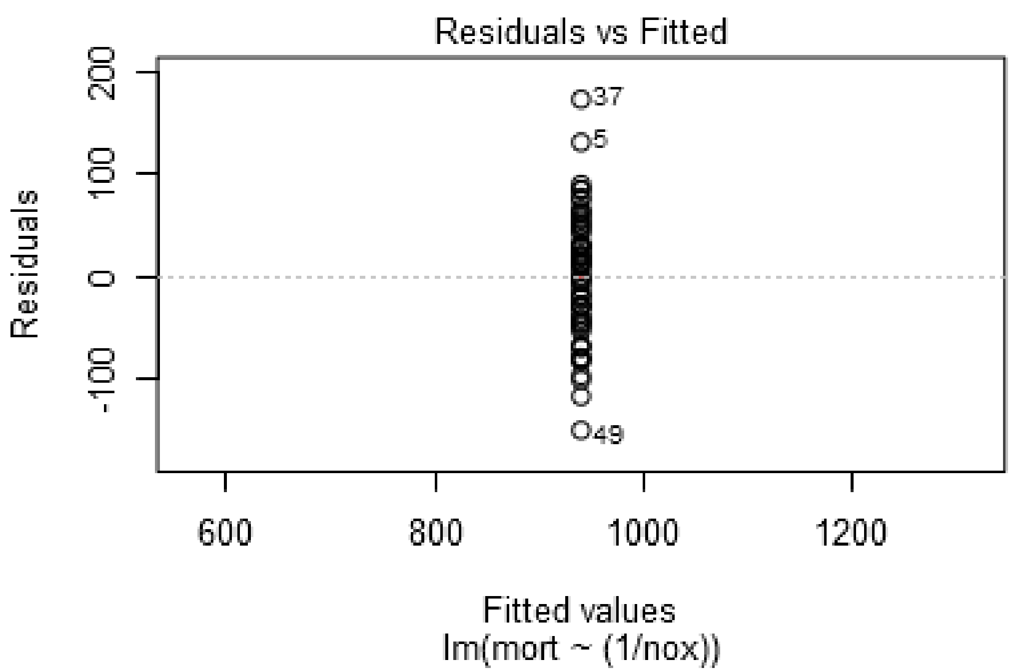

```
# The residual plot looks horrible
```

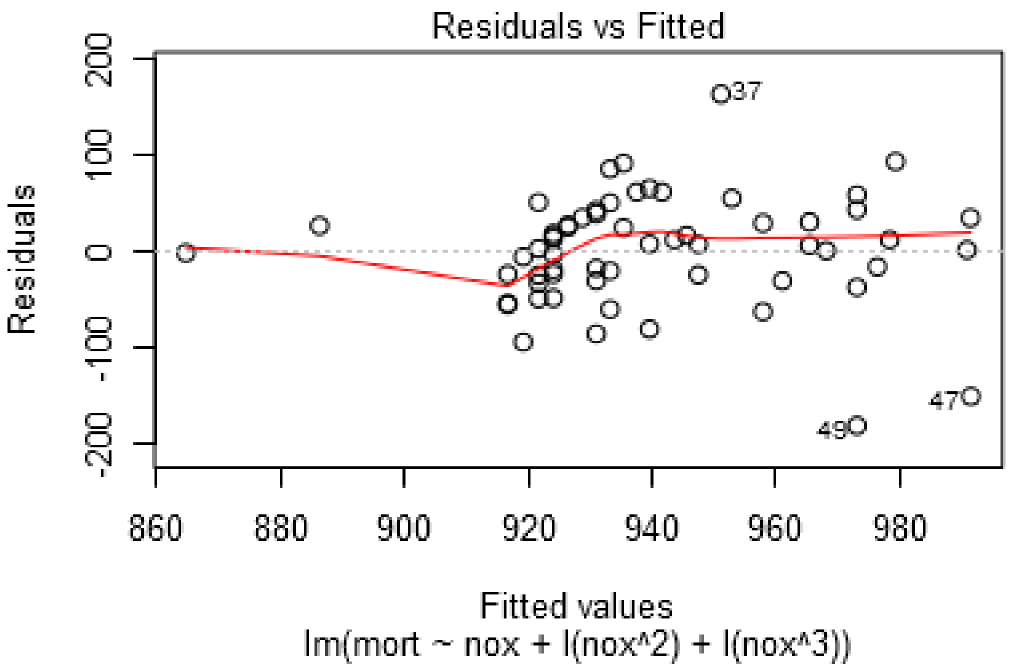
2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```r
# We could try the log transformation, square root transormations and reciprocal transformation
pol_t2 <- lm(log(mort)~nox, data = pollution)
pol_t3 <- lm(sqrt(mort)~nox, data = pollution)
pol_t4 <- lm(mort~(1/nox), data=pollution)
pol_t5 <- lm(mort~nox+I(nox^2)+I(nox^3), data=pollution)
plot(pol_t2, which = 1);plot(pol_t3, which = 1);plot(pol_t4, which = 1);plot(pol_t5,which = 1)
```




Residuals vs Fitted
Residuals
3
2
1
0
-1
-2
-3
37
5
49
30.2
30.3
30.4
30.5
30.6
30.7
Fitted values
lm(sqrt(mort) ~ nox)


Residuals vs Fitted
Residuals
200
100
0
-100
37
5
49
600
800
1000
1200
Fitted values
lm(mort ~ (1/nox))




```
# In the residual plot of "pol_t5", although the red line is still not close enough to line 0, but the
# is much better the others.
```

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

```
summary(pol_t5)
```

```
##
## Call:
## lm(formula = mort ~ nox + I(nox^2) + I(nox^3), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -182.430  -27.441    5.511   33.449  161.985
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.142e+02  1.235e+01  74.016  < 2e-16 ***
## nox           2.582e+00  8.902e-01   2.900  0.00532 **
## I(nox^2)     -2.468e-02  9.784e-03  -2.523  0.01451 *
## I(nox^3)      5.048e-05  2.352e-05   2.146  0.03622 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.62 on 56 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1121
## F-statistic: 3.483 on 3 and 56 DF,  p-value: 0.02165
```

```
# The slope coefficient tells that nox has a positive and significant relationship with mortality rate.
# For every unit increase in nox, the mortality rate will increase 2.582e+00
```

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

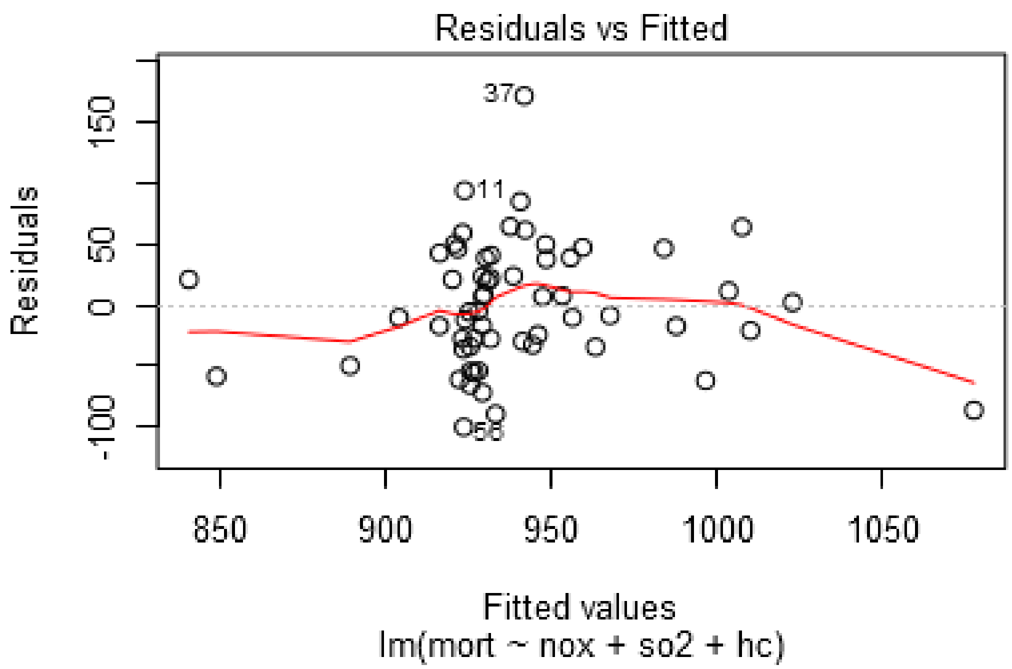
```
confint(pol_t5, level = 0.99)
```

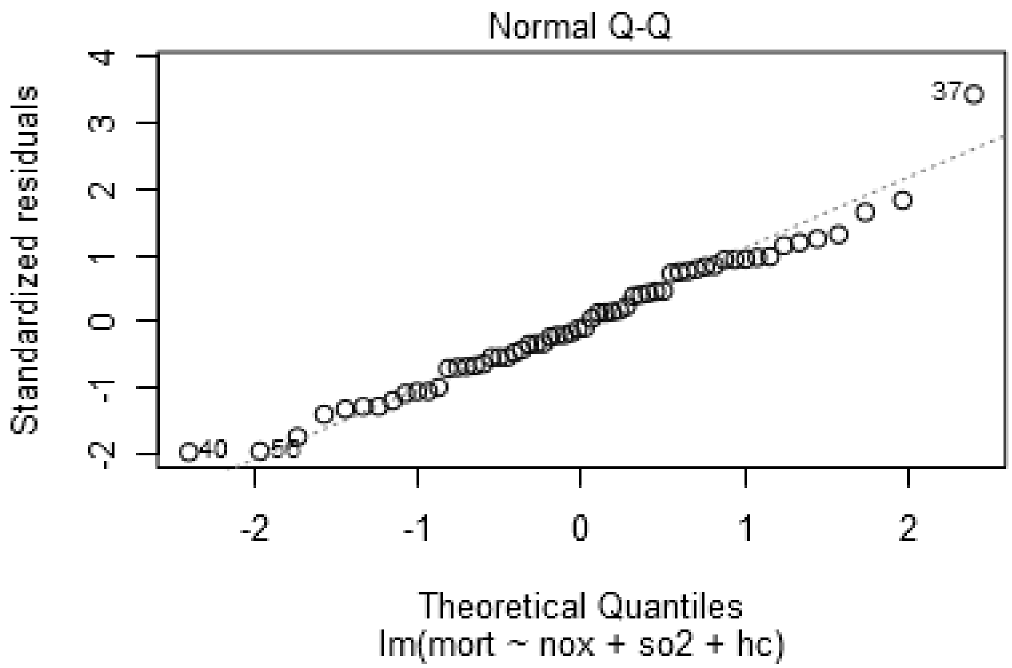
```
##                      0.5 %       99.5 %
## (Intercept)  8.812313e+02 9.470992e+02
## nox          2.081142e-01 4.955574e+00
## I(nox^2)    -5.077140e-02 1.406317e-03
## I(nox^3)    -1.224329e-05 1.132031e-04
```

```
# As we can see in ths table, the CI of nox doesn't across 0, therefore, it's statistically significant
```

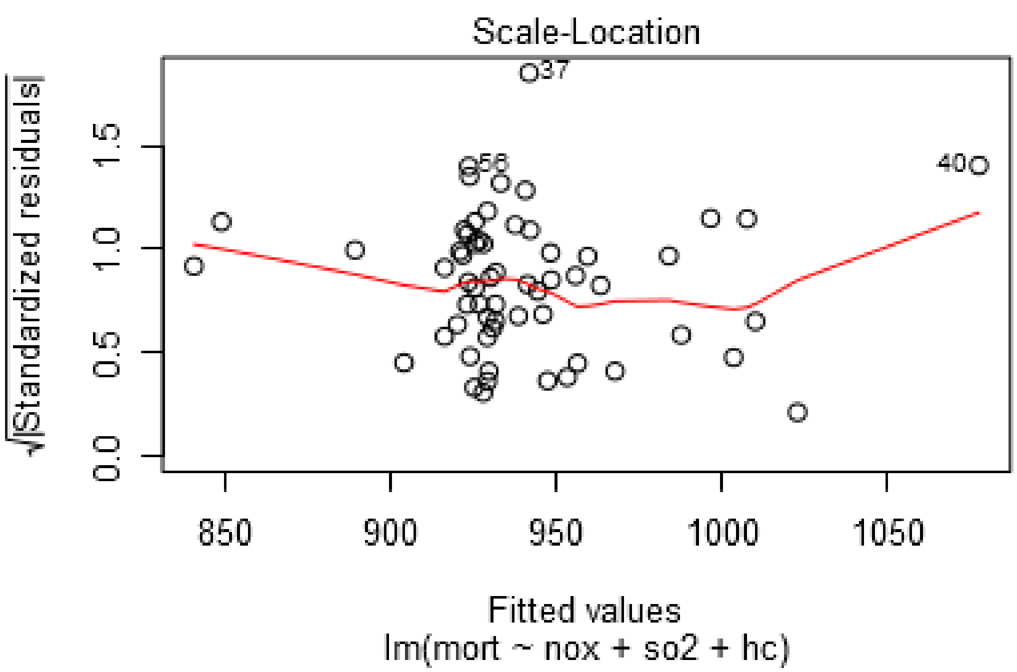
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
pol_t6 <- lm(mort~nox+so2+hc, data = pollution)
plot(pol_t6)
```

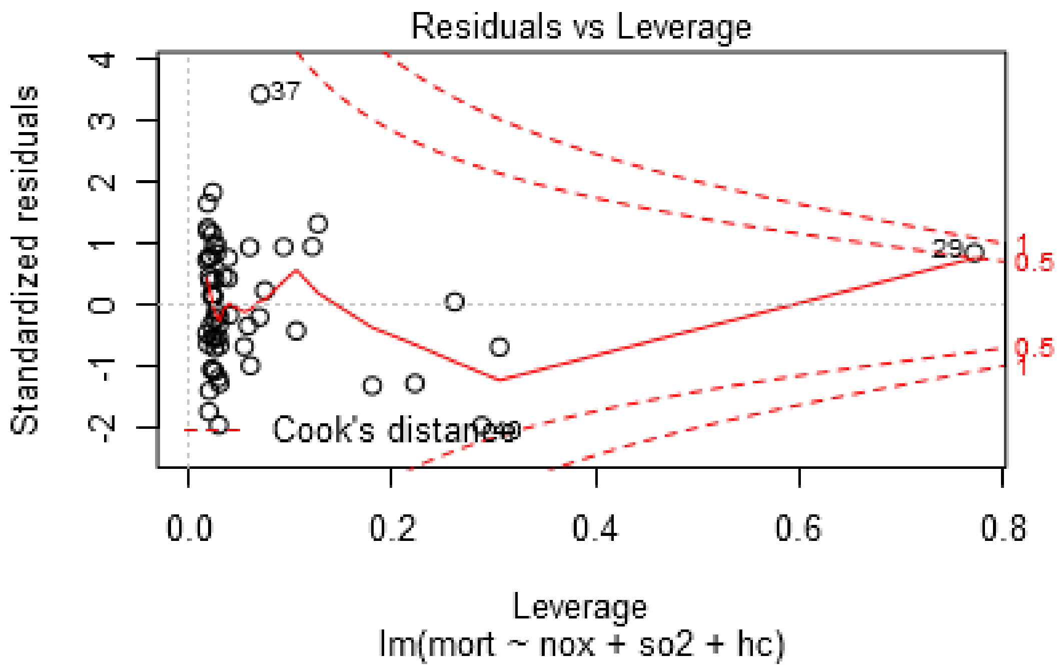




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mort ~ nox + so2 + hc)
37
40
56
-2
-1
0
1
2
3
4



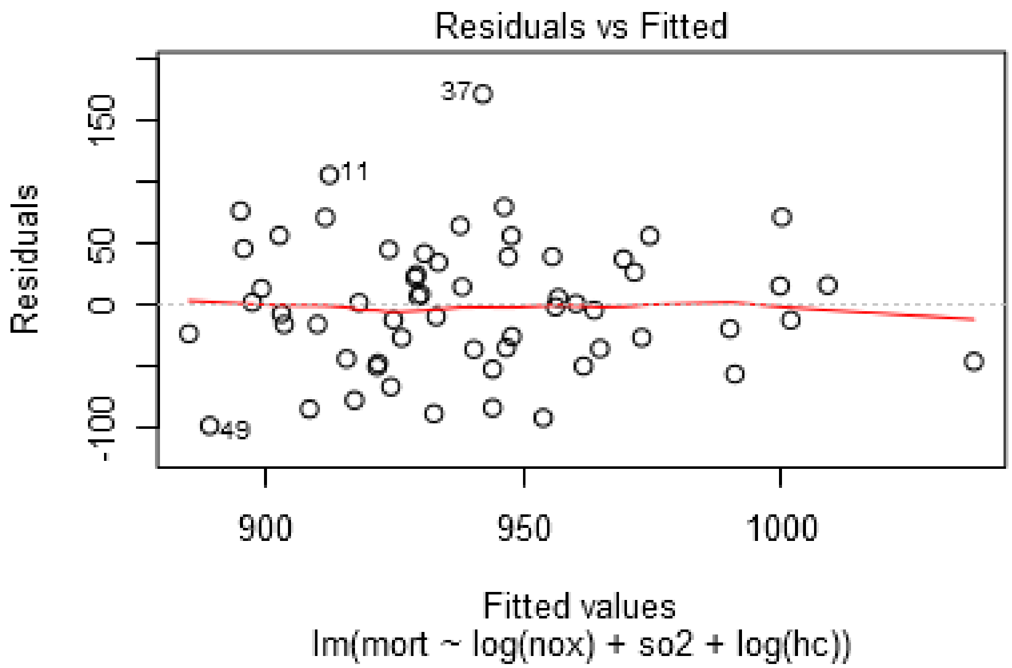
Scale-Location
√|Standardized residuals|
Fitted values
lm(mort ~ nox + so2 + hc)
37
58
40
0.0
0.5
1.0
1.5
850
900
950
1000
1050

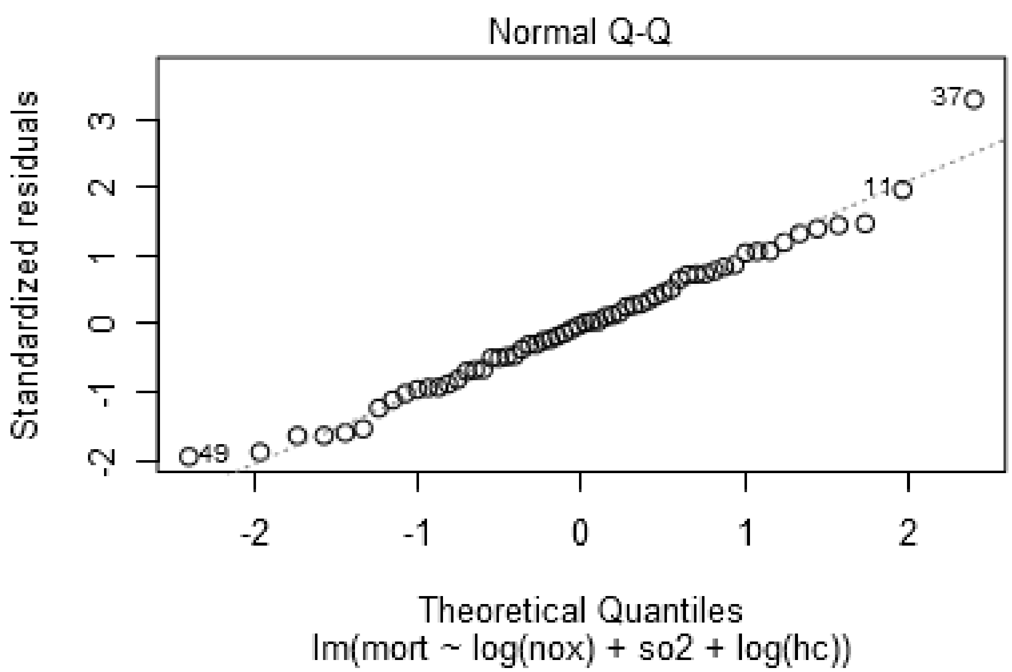


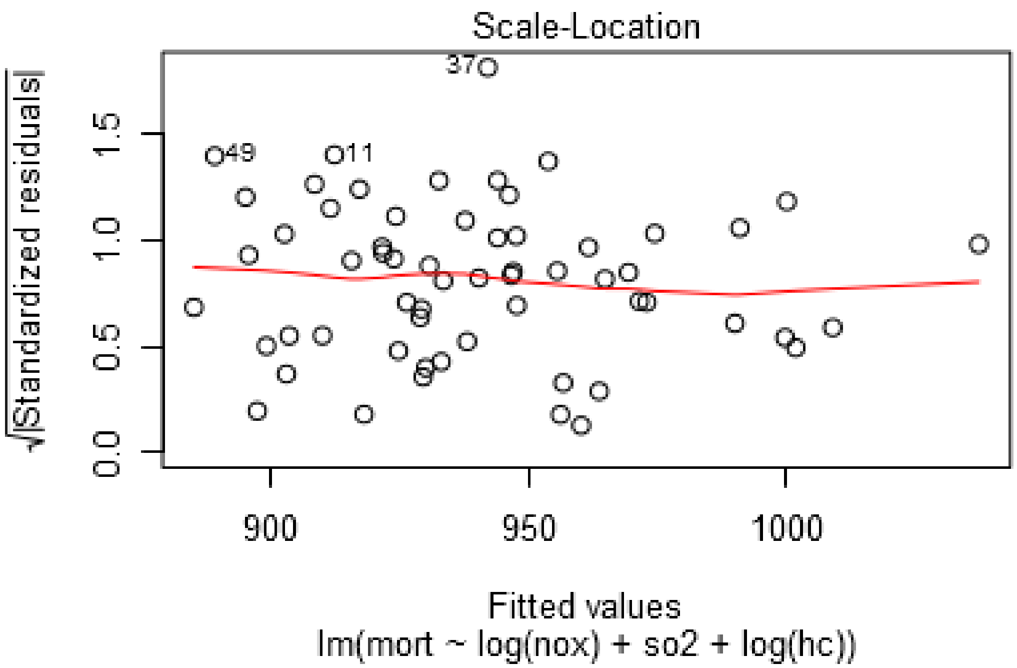

```r
# By observing the dataset, I found some extreme large value in "nox" and "hc", therefore I decided to
# on those variables.
pol_t7 <- lm(mort~log(nox)+so2+log(hc), data = pollution)
summary(pol_t7)
```

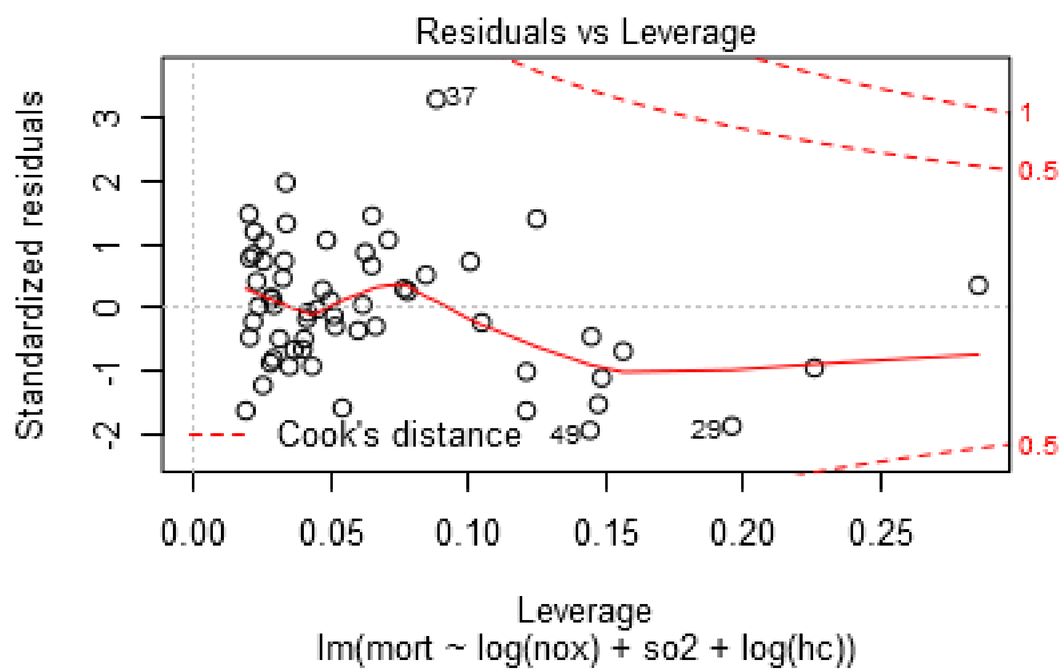
```
##
## Call:
## lm(formula = mort ~ log(nox) + so2 + log(hc), data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.262 -35.757  -0.413  37.602 171.152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 943.6588    18.9365  49.833   <2e-16 ***
## log(nox)     56.0779    22.7132   2.469   0.0166 *
## so2           0.2638     0.1654   1.594   0.1165
## log(hc)     -53.6761    20.1715  -2.661   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.43 on 56 degrees of freedom
## Multiple R-squared:  0.2733, Adjusted R-squared:  0.2344
## F-statistic:  7.02 on 3 and 56 DF,  p-value: 0.0004336
```

```r
plot(pol_t7)
```

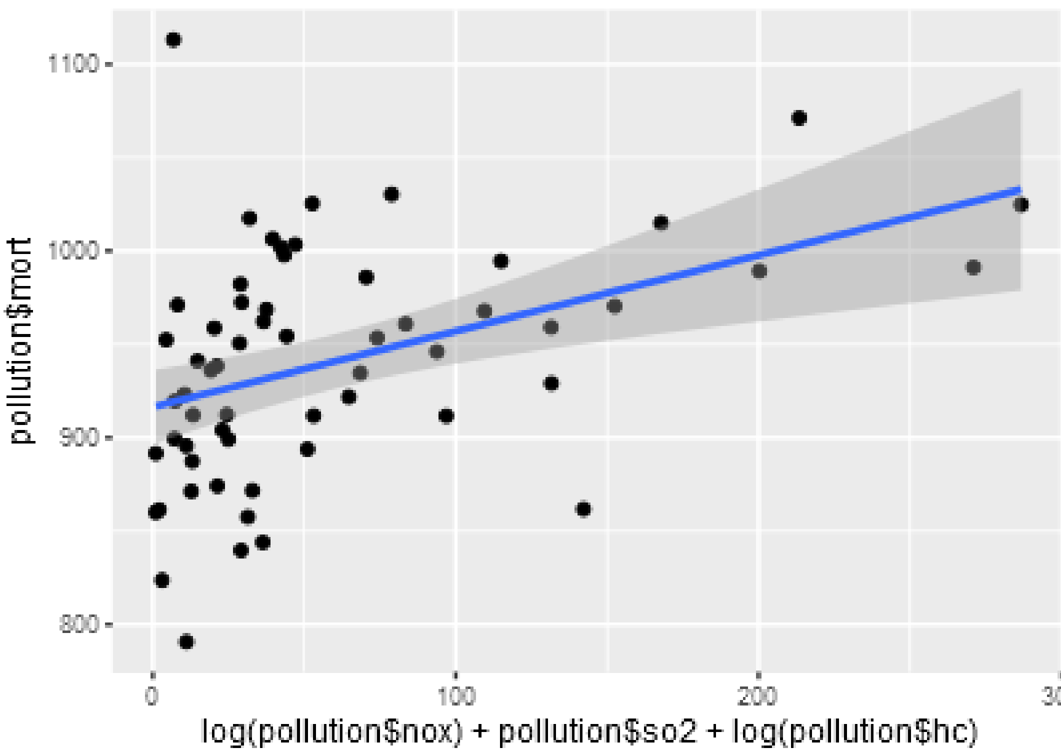

Residuals vs Fitted
Residuals
Fitted values
lm(mort ~ log(nox) + so2 + log(hc))
37
11
49
-100
0
50
150
900
950
1000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mort ~ log(nox) + so2 + log(hc))
37
11
49
-2
-1
0
1
2
3






```
ggplot(pol_t7)+aes(y=pollution$mort, x=log(pollution$nox)+pollution$so2+log(pollution$hc))+geom_point()
```




```
# log(nox) and so2 have positive correlation with mortality while log(hc) ahs negative correlation.
# Area that has higher nox and so2 tend to has higher mortality rate.
```

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.)

```
poll_30 <- pollution[1:30,]
poll_60 <- pollution[31:60,]
cv_1 <- lm(mort~log(nox)+so2+log(hc), data = poll_30)
pred <- predict(object = cv_1, poll_60, interval="prediction")
pred[,1]-poll_60$mort
```

```
##          31          32          33          34          35          36
##  -63.455818   57.431564   44.209766   76.457708   -4.477980    5.034742
##          37          38          39          40          41          42
## -184.898365  -28.164929  -24.692688   38.860770   44.741529   -7.049147
##          43          44          45          46          47          48
##   15.571054  -81.243127   56.085779   -4.426199   92.626336   47.446375
##          49          50          51          52          53          54
##  131.682429   28.954350   29.323184  -16.846178  -38.178466   12.362454
##          55          56          57          58          59          60
##   -6.201560   98.644006  -60.952806   28.637210   34.721020  -11.670375
```

Study of teenage gambling in Britain

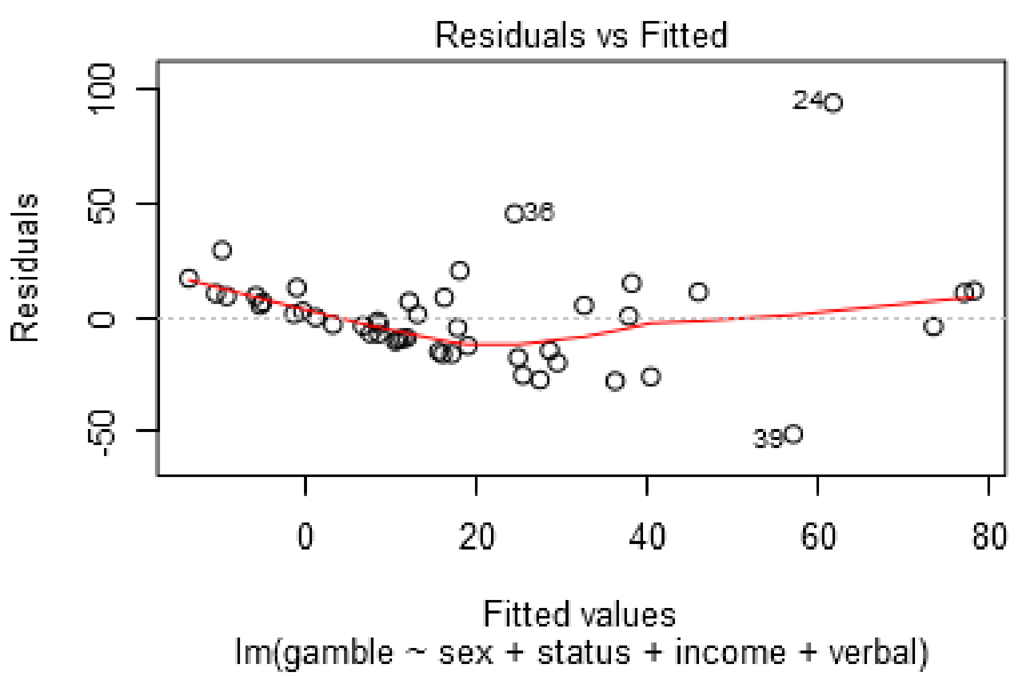
```
data(teengamb)
?teengamb
```

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.

```
gamb_1 <- lm(gamble~sex+status+income+verbal, data = teengamb)
summary(gamb_1)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

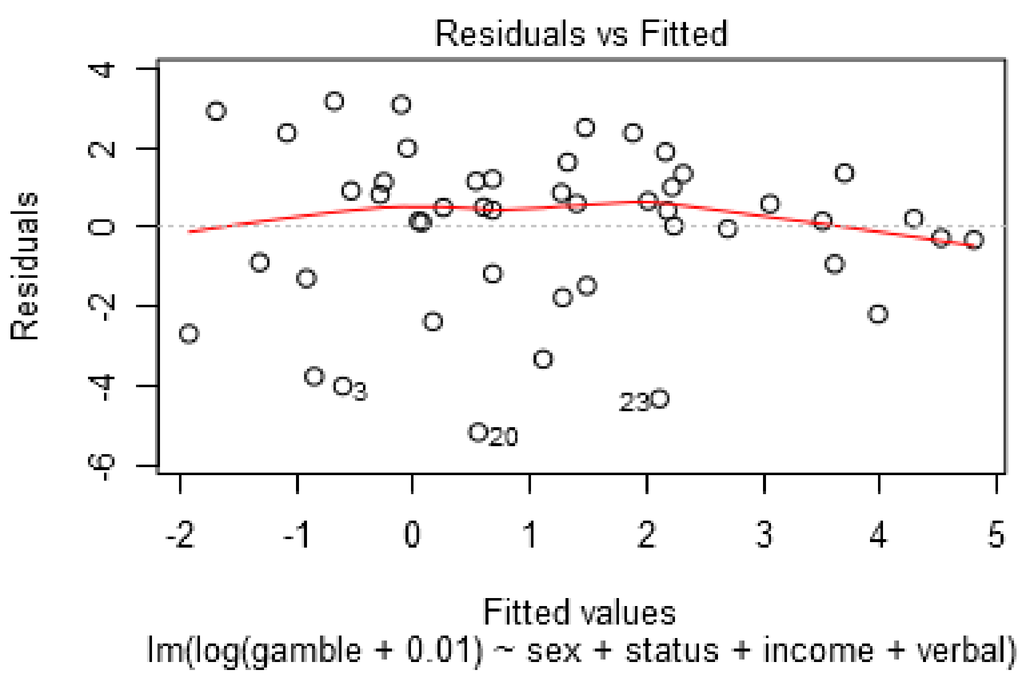
```
plot(gamb_1, which = 1)
```



```r
gamb_2 <- lm(log(gamble+0.01)~sex+status+income+verbal, data = teengamb)
summary(gamb_2)
```

```
##
## Call:
## lm(formula = log(gamble + 0.01) ~ sex + status + income + verbal,
##     data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1612 -1.0537  0.4244  1.1809  3.1625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.24952    1.58150   0.790  0.43392
## sex         -1.45110    0.75513  -1.922  0.06145 .
## status       0.05320    0.02585   2.058  0.04583 *
## income       0.29859    0.09430   3.166  0.00287 **
## verbal      -0.49467    0.19976  -2.476  0.01739 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.087 on 42 degrees of freedom
## Multiple R-squared:  0.4226, Adjusted R-squared:  0.3676
## F-statistic: 7.685 on 4 and 42 DF,  p-value: 9.675e-05
```

```r
plot(gamb_2, which = 1)
```




```r
# I take log on the respondent variable "gamble". For a male with zero income, zero verbal score and ze
# score, the average expenditure on gambling is 1.2495
```

```
# Females tend to spend less on gambling than males. For every dollar more in income, the expenditure w
# increase by 29.8%
```

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

```
confint(gamb_2, level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) -1.942074412  4.44110876
## sex         -2.975016117  0.07282055
## status       0.001032372  0.10537656
## income       0.108286334  0.48889584
## verbal      -0.897803034 -0.09153497
```

```
# "status" "verbal" and "income" are significant while "sex" might not seem to be as significant as the
```

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?

```
# Model for an "average guy"
c_status <- mean(teengamb$status)
c_income <- mean(teengamb$income)
c_verbal <- mean(teengamb$verbal)
agdata <- data.frame(status=c_status,income=c_income,verbal=c_verbal,sex=0)
ag <- predict(gamb_2, newdata = (agdata),level=0.95, interval="confidence")
summary(ag)
```

```
##       fit              lwr              upr
##  Min.   :1.748   Min.   :0.878   Min.   :2.618
##  1st Qu.:1.748   1st Qu.:0.878   1st Qu.:2.618
##  Median :1.748   Median :0.878   Median :2.618
##  Mean   :1.748   Mean   :0.878   Mean   :2.618
##  3rd Qu.:1.748   3rd Qu.:0.878   3rd Qu.:2.618
##  Max.   :1.748   Max.   :0.878   Max.   :2.618
```

```
# The average guy tends to spend 1.748 on gambling per week.

# Model for a "rich guy"
rgdata <- data.frame(status=max(teengamb$status),income=max(teengamb$income),verbal=max(teengamb$verbal)
rg <- predict(gamb_2, newdata = (rgdata),level=0.95, interval="confidence")
summary(rg)
```

```
##       fit              lwr              upr
##  Min.   :4.772   Min.   :2.098   Min.   :7.446
##  1st Qu.:4.772   1st Qu.:2.098   1st Qu.:7.446
##  Median :4.772   Median :2.098   Median :7.446
##  Mean   :4.772   Mean   :2.098   Mean   :7.446
##  3rd Qu.:4.772   3rd Qu.:2.098   3rd Qu.:7.446
##  Max.   :4.772   Max.   :2.098   Max.   :7.446
```

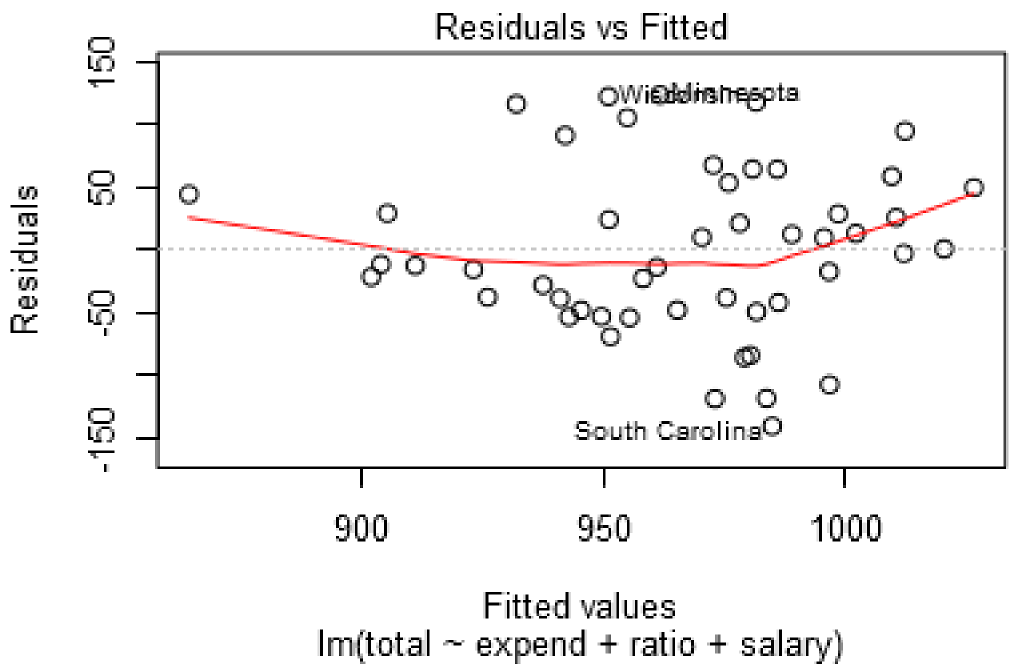
```
# A guy with maximal status, income and verbal score tends to spend 4.77 dollars on gambling.
```

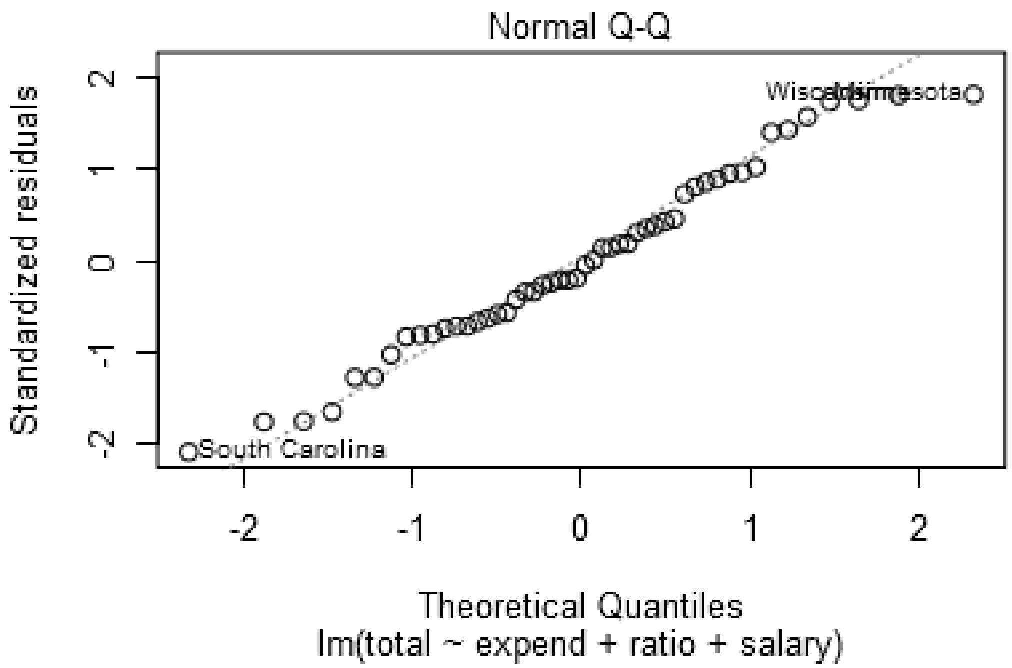
School expenditure and test scores from USA in 1994-95

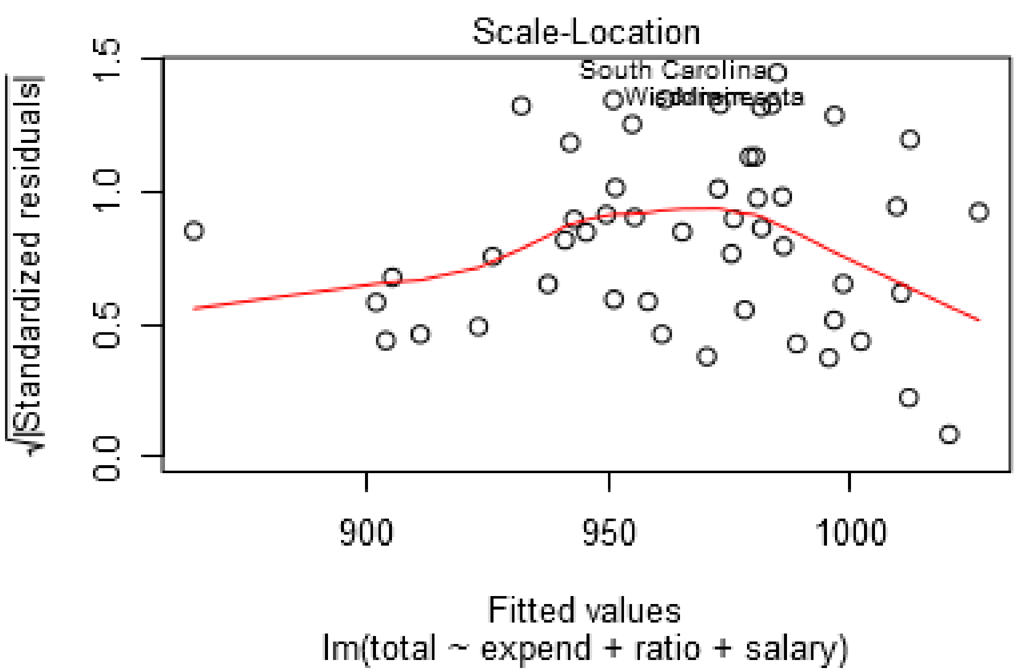
```
data(sat)
?sat
```

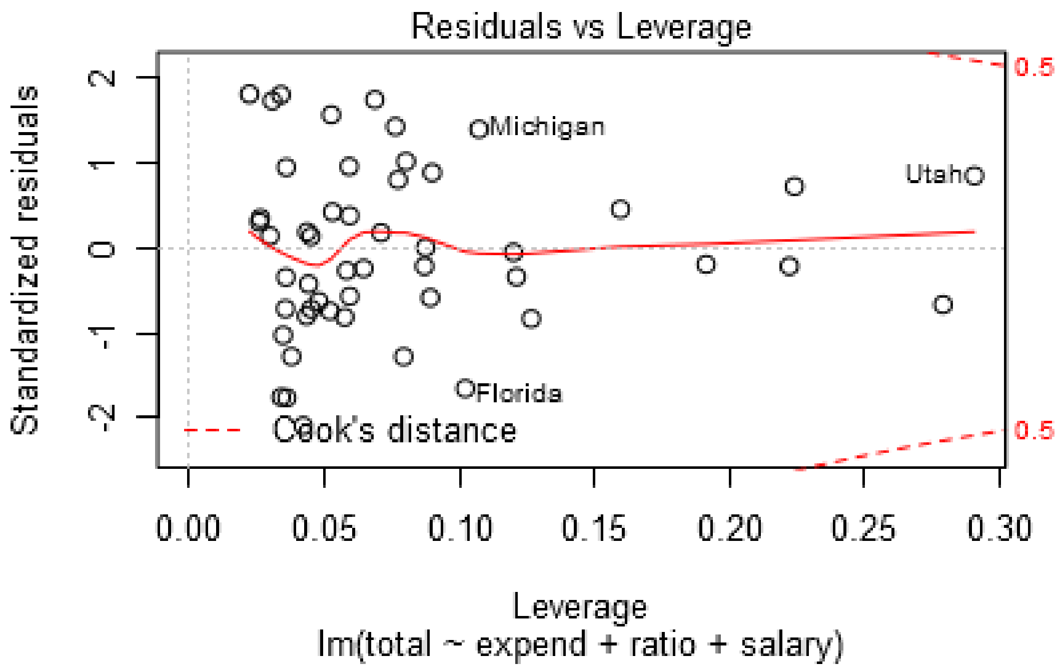
1. Fit a model with total sat score as the outcome and expend, ratio and salary as predictors. Make necessary transformation in order to improve the interpretability of the model. Interpret each of the coefficient.

```
regsat <- lm(total~expend+ratio+salary, data = sat)
plot(regsat)
```

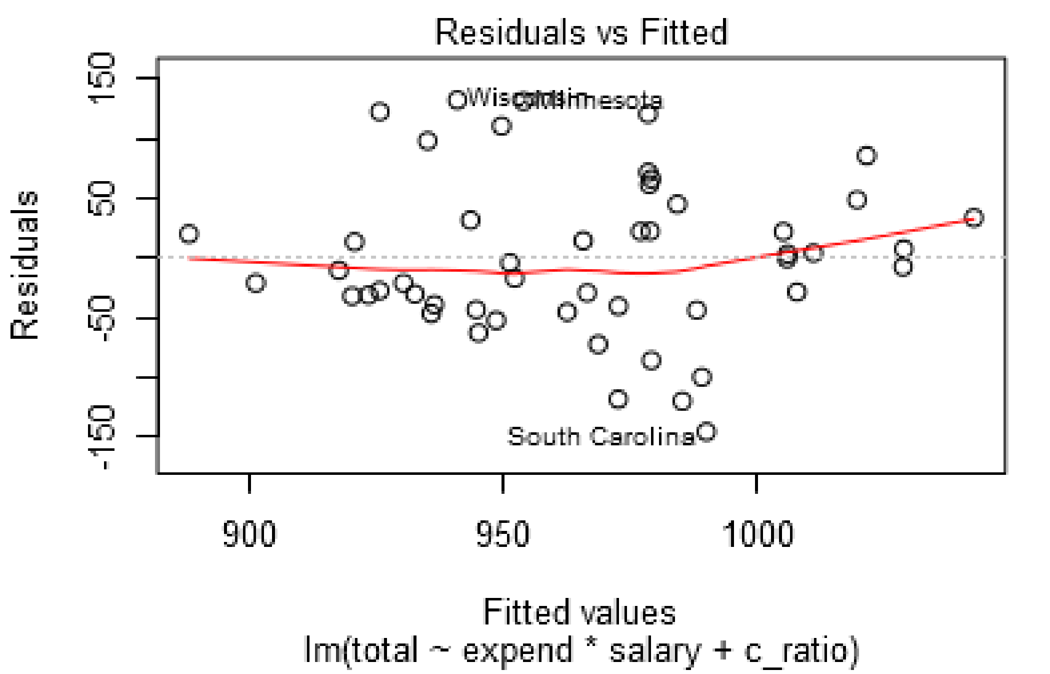



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(total ~ expend + ratio + salary)
South Carolina
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
lm(total ~ expend + ratio + salary)
South Carolina
0.0
0.5
1.0
1.5
900
950
1000




```
# I assume there are interactions between expend and salary
c_ratio <- sat$ratio - mean(sat$ratio)
regsat_2 <- lm(total~expend*salary+c_ratio, data = sat)
plot(regsat_2, which = 1)
```



```r
summary(regsat_2)
```

```
##
## Call:
## lm(formula = total ~ expend * salary + c_ratio, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.97  -40.36   -5.99   32.91  132.04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1411.455    255.833   5.517 1.62e-06 ***
## expend         -27.042     50.864  -0.532   0.5976
## salary         -14.485      7.594  -1.907   0.0629 .
## c_ratio          5.630      6.590   0.854   0.3975
## expend:salary    1.029      1.083   0.950   0.3474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.73 on 45 degrees of freedom
## Multiple R-squared:  0.2251, Adjusted R-squared:  0.1563
## F-statistic: 3.269 on 4 and 45 DF,  p-value: 0.01952
```

```r
# Intercept: a student from a zero income family, goes to average ratio school and doesn't spend money
# likely to have SAT score of 1411. With more expenditure at school will decrease the student's SAT sco
# student's family make more maney, his or her SAT score will also be decrease. However, if the student
# school that has higher student/teacher ratio, the student tend to have higher SAT score.
```

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

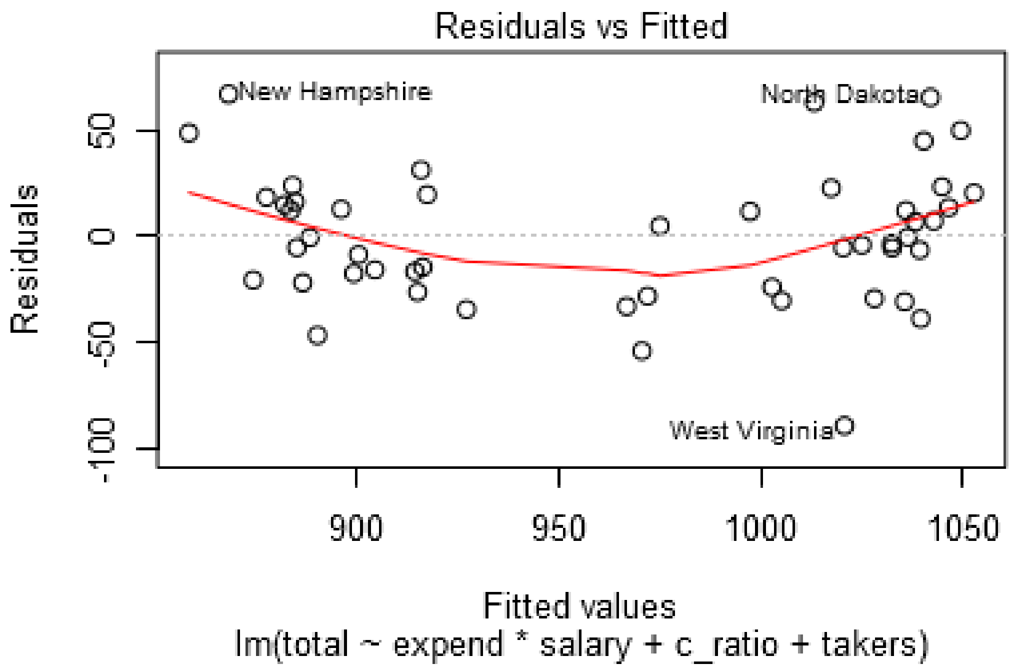
```r
confint(regsat_2, level = 0.98)
```

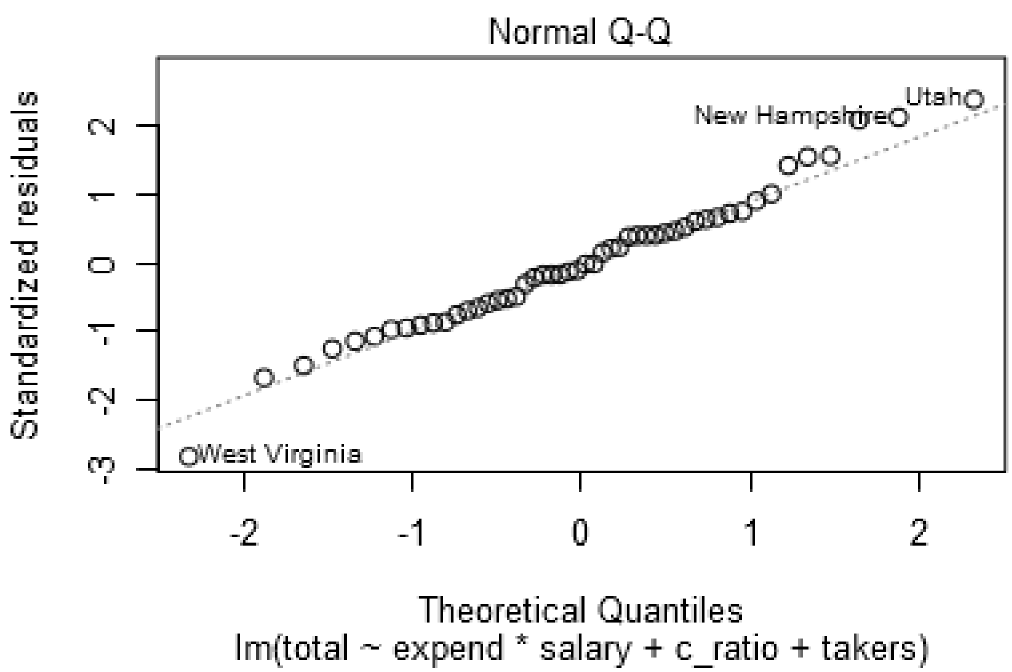
```
##                        1 %        99 %
## (Intercept)    794.355747 2028.553440
## expend        -149.731507   95.646677
## salary         -32.801273    3.832194
## c_ratio        -10.266755   21.527043
## expend:salary   -1.584281    3.641325
```

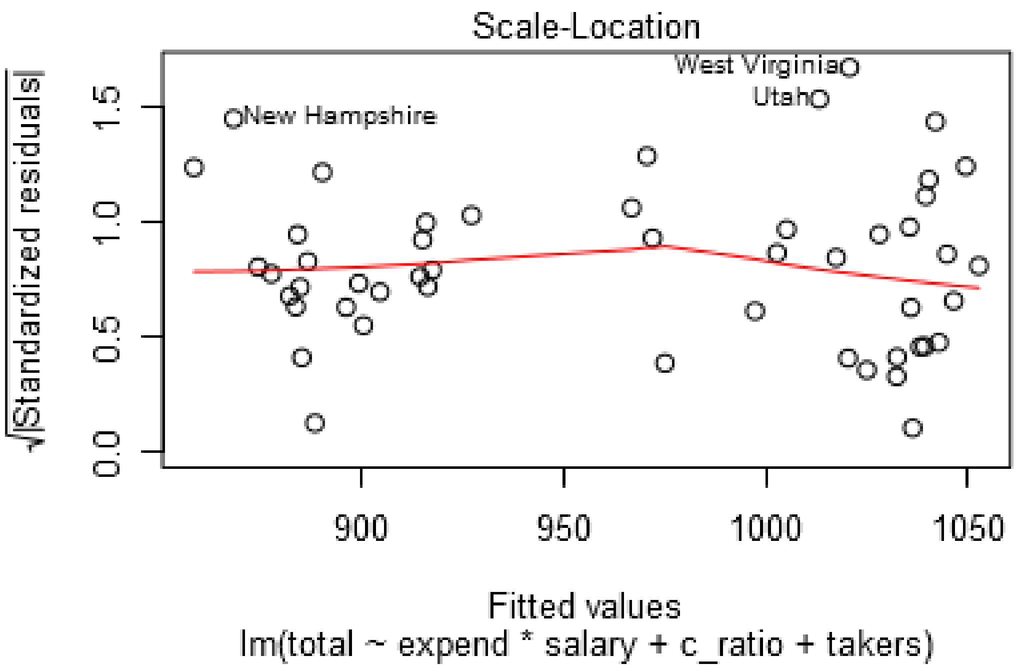
```r
# All of the variables are not statistically significant
```

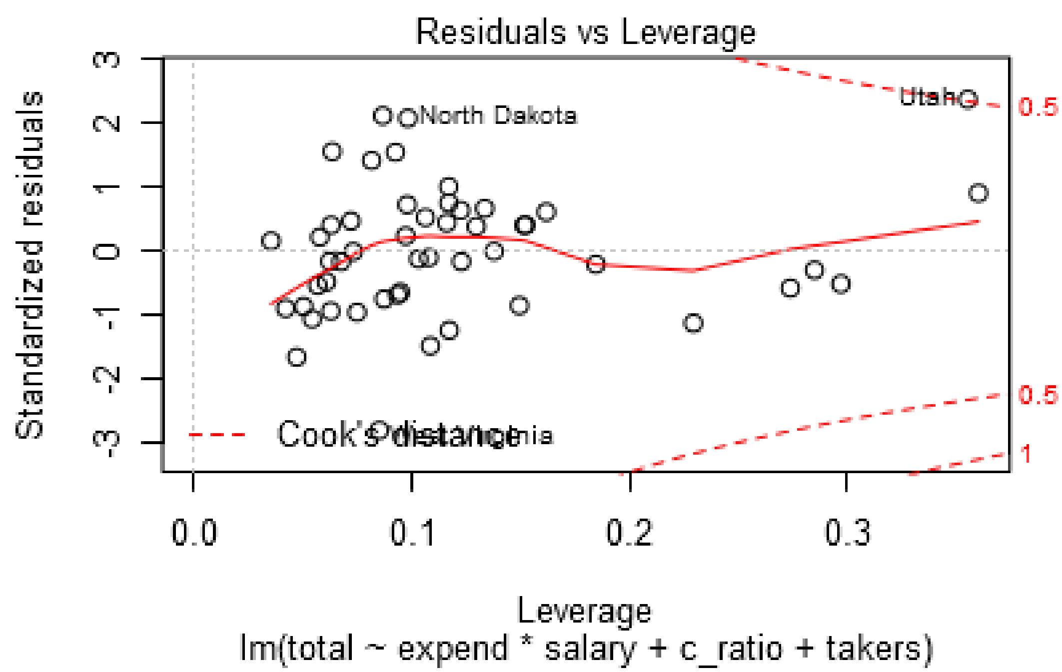
3. Now add takers to the model. Compare the fitted model to the previous model and discuss which of the model seem to explain the outcome better?

```r
regsat_3 <- lm(total~expend*salary+c_ratio+takers, data = sat)
plot(regsat_3)
```


Residuals vs Fitted
New Hampshire
North Dakota
West Virginia
Residuals
50
0
-50
-100
900
950
1000
1050
Fitted values
lm(total ~ expend * salary + c_ratio + takers)


Normal Q-Q
Utah
New Hampshire
West Virginia
Standardized residuals
2
1
0
-1
-2
-3
-2
-1
0
1
2
Theoretical Quantiles
lm(total ~ expend * salary + c_ratio + takers)






```
summary(regsat_3)
```

```
##
## Call:
## lm(formula = total ~ expend * salary + c_ratio + takers, data = sat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.714 -21.418  -1.957  17.739  66.616
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1029.0964   126.8034   8.116 2.75e-10 ***
## expend          -3.5273    24.5142  -0.144    0.886
## salary           0.5524     3.8485   0.144    0.887
## c_ratio         -3.7155     3.2567  -1.141    0.260
## takers          -2.8934     0.2355 -12.285 8.13e-16 ***
## expend:salary    0.1899     0.5249   0.362    0.719
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.02 on 44 degrees of freedom
## Multiple R-squared:  0.8251, Adjusted R-squared:  0.8052
## F-statistic: 41.51 on 5 and 44 DF,  p-value: 1.409e-15
```

```
# I personally prefer this model since it shows "takers" has significant incluence on the outcome, alth
# the residual plot is still bad.
```

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, $D_i - R_i$

The difference tells the the difference in amount of money raised by two indivisual candidates. By using this formula, we could easily tell who raise more money and how much in difference.

- The ratio, D_i/R_i

The ratio tells the proportion of amount of money raised by two indivisual candidates. By using this formula, we could easily find out the comparison of "efficiency". In other words, we could know that for every one dollar candidate D raised, how much candidate R could raise.

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

We could transfor $logD_i - logR_i$ to \$ log(D_i/R_i)\$. The formula tells us the percentage change in one candidate's fund raise will influence how much on the other candidate's fund raise.

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

The formula tells us the weight of amount money of D raised in the total money raised by both person. By using this method, we could track the fund raising dynamically.

Transformation

See attched photos

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.