Homework 02

Xinyi Wang

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.

```
a = select(heights,earn,sex,height,yearbn,ed)
#Exclude NA
new_heights = na.omit(a)
#Exclude zero earings
earn_zero = which(new_heights$earn==0)
new_heights = new_heights[-earn_zero,]
#Assume male = 1, female = 2, factorise 'sex' variable
new_heights$sex = factor(new_heights$sex, labels=c("male", "female"))
#new_heights
summary(new_heights)
```

```
##       earn            sex          height          yearbn
##  Min.   :   200   male  :505   Min.   :58.00   Min.   : 1.00
##  1st Qu.: 10000   female:687   1st Qu.:64.00   1st Qu.:39.75
##  Median : 20000                Median :66.00   Median :52.00
##  Mean   : 23155                Mean   :66.92   Mean   :48.87
##  3rd Qu.: 30000                3rd Qu.:70.00   3rd Qu.:61.00
##  Max.   :200000                Max.   :77.00   Max.   :99.00
##        ed
##  Min.   : 3.0
##  1st Qu.:12.0
##  Median :13.0
##  Mean   :13.5
##  3rd Qu.:16.0
##  Max.   :18.0
```

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?

```
h = new_heights$height
z.height = (h - mean(h)) / (2*sd(h))
height.lm = lm(earn ~ z.height,new_heights)
#display(height.lm)
summary(height.lm)
```

```
##
## Call:
## lm(formula = earn ~ z.height, data = new_heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30166 -11309  -3428   6527 172953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23154.8      546.4  42.376   <2e-16 ***
## z.height      9711.4     1093.3   8.883   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1190 degrees of freedom
## Multiple R-squared:  0.06218,    Adjusted R-squared:  0.06139
## F-statistic:  78.9 on 1 and 1190 DF,  p-value: < 2.2e-16
```

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.

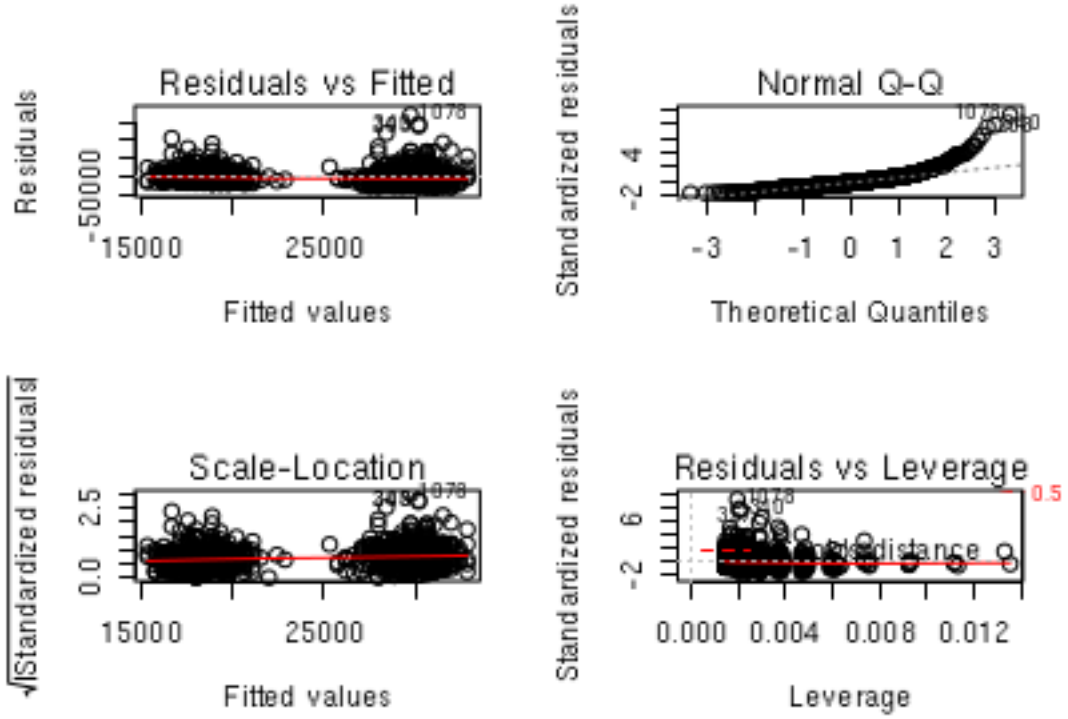
```
earn = new_heights$earn
height = new_heights$height
sex = new_heights$sex
ed = new_heights$ed
log.earn = log(new_heights$earn)

model1.lm = lm(earn ~ height + sex)
summary(model1.lm)
```

```
##
## Call:
## lm(formula = earn ~ height + sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30087 -11013  -3315   6128 170242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1246.8    13800.5  -0.090   0.9280
## height         442.9      196.6   2.253   0.0245 *
## sexfemale    -9087.9     1529.9  -5.940 3.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 18600 on 1189 degrees of freedom
## Multiple R-squared:  0.08921,    Adjusted R-squared:  0.08768
## F-statistic: 58.23 on 2 and 1189 DF,  p-value: < 2.2e-16
```

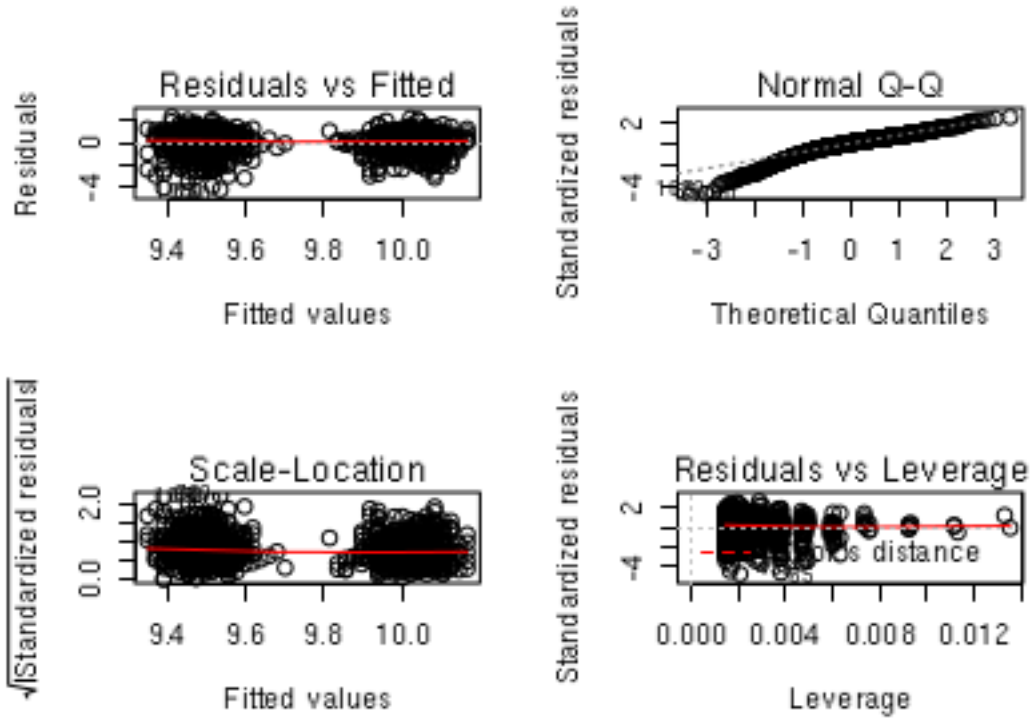
```
par(mfrow=c(2,2))
plot(model1.lm)
```




```
model2.lm = lm(log.earn ~ height + sex)
summary(model2.lm)
```

```
##
## Call:
## lm(formula = log.earn ~ height + sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2384 -0.3718  0.1410  0.5649  2.3071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.575907   0.653635  13.120  < 2e-16 ***
## height       0.020658   0.009312   2.218   0.0267 *
## sexfemale   -0.423217   0.072462  -5.841 6.71e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8809 on 1189 degrees of freedom
## Multiple R-squared:  0.08656,    Adjusted R-squared:  0.08502
## F-statistic: 56.34 on 2 and 1189 DF,  p-value: < 2.2e-16
```

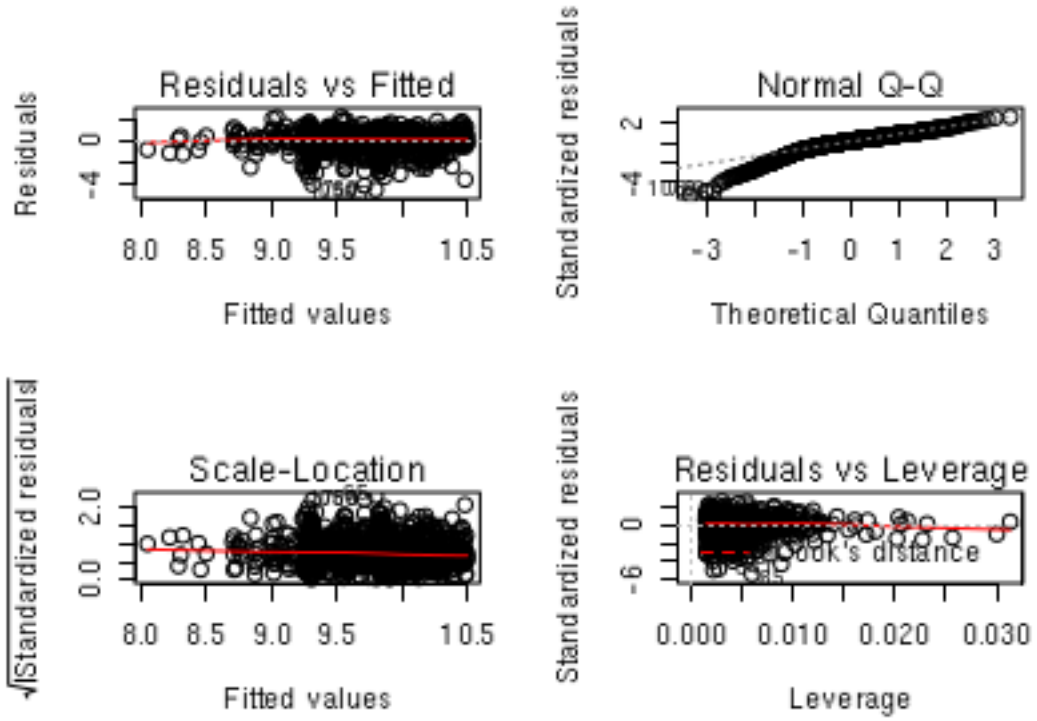
```
par(mfrow=c(2,2))
plot(model2.lm)
```




```
model3.lm = lm(log.earn ~ height + sex + ed + sex*ed)
summary(model3.lm)
```

```
##
## Call:
## lm(formula = log.earn ~ height + sex + ed + sex * ed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4962 -0.3293  0.1292  0.5179  2.1872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.106008   0.644869  12.570  < 2e-16 ***
## height        0.008505   0.008870   0.959 0.337810
## sexfemale    -0.983855   0.281612  -3.494 0.000494 ***
## ed            0.097185   0.014893   6.526    1e-10 ***
## sexfemale:ed  0.037940   0.020076   1.890 0.059031 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8334 on 1187 degrees of freedom
## Multiple R-squared:  0.1838, Adjusted R-squared:  0.1811
## F-statistic: 66.84 on 4 and 1187 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model3.lm)
```



Overall, model 3 is the preferred model since it has the largest r^2, residuals looks more likely equally spread around a horizontal line.

4. Interpret all model coefficients.

From model 3 we get,

log(earn) = 8.106 + 0.0085*height* - *0.984*sex + 0.097*ed* + *0.038*sex*ed

The intercept 8.106 is the average log earning for a male has 0 height and 0 rate of education.

The coefficient for height is the predicted difference in log earning corresponding to male with every 1 inch difference in height.

The coefficient for sex is the predicted difference in log earning bewteen male and female if height and education rate are both 0.

The coefficient for education is the predicted difference in log earning corresponding to male with every 1 unit change in education rate.

The coefficient for sex:ed = 0.038 means 1 unit of education rate corresponds to 3.8% more of an increase in earnings among female than male.

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

```
confint(model3.lm)
```

```
##                    2.5 %      97.5 %
## (Intercept)   6.840797466  9.37121926
## height       -0.008897412  0.02590838
## sexfemale    -1.536368540 -0.43134117
```

```
## ed            0.067966483  0.12640401
## sexfemale:ed -0.001449258  0.07732837
```

[6.84,9.37] is the range of values that you can be 95% certain contains the true value of intercept.

[-0.0089,0.0259] is the range of values that you can be 95% certain contains the true value coefficient of height.

[-1.536,-0.431] is the range of values that you can be 95% certain contains the true value coefficient of sex.

[0.0679,0.126] is the range of values that you can be 95% certain contains the true value coefficient of education.

[-0.00144,0.0773] is the range of values that you can be 95% certain contains the true value coefficient of sex:ed.

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:

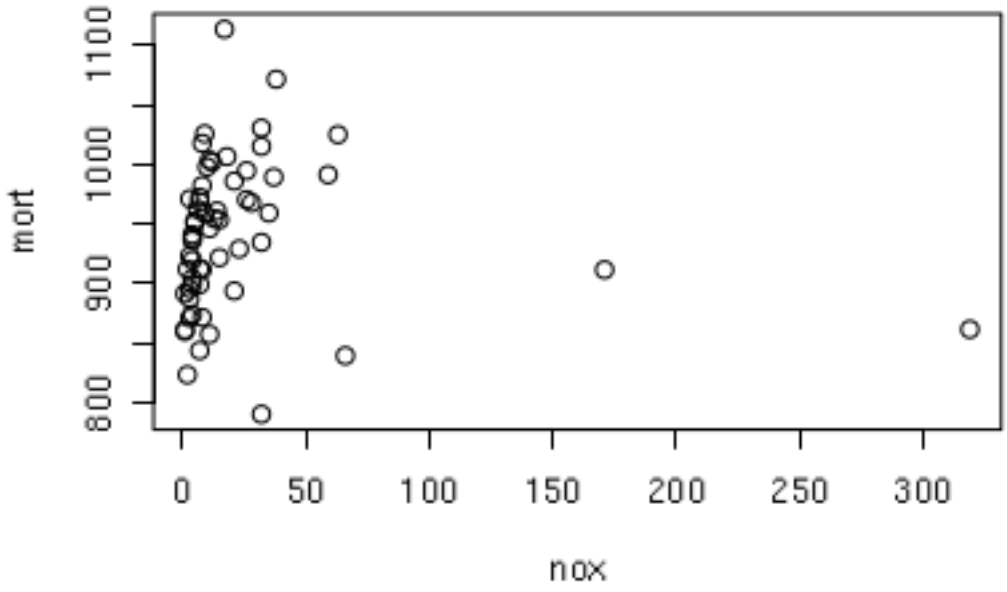
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all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

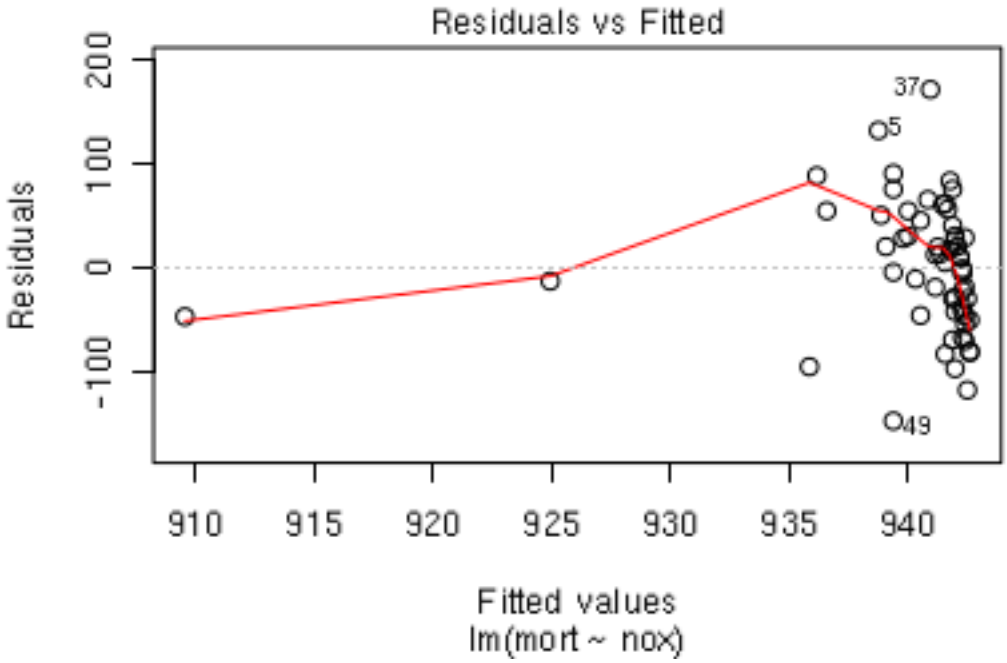
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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.

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




```
model1.lm = lm(mort ~ nox)
plot(model1.lm,which = 1)
```

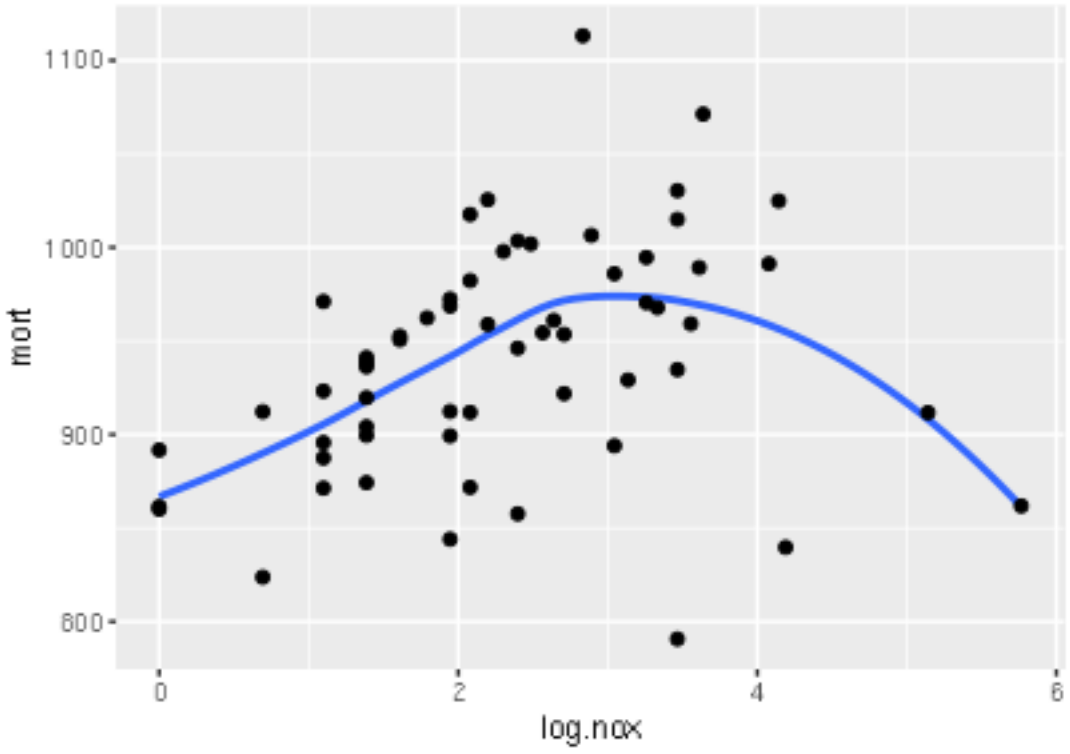


The model doesn't fit well. Residual plot also shows residuals are not symmetrically distributed around 0.

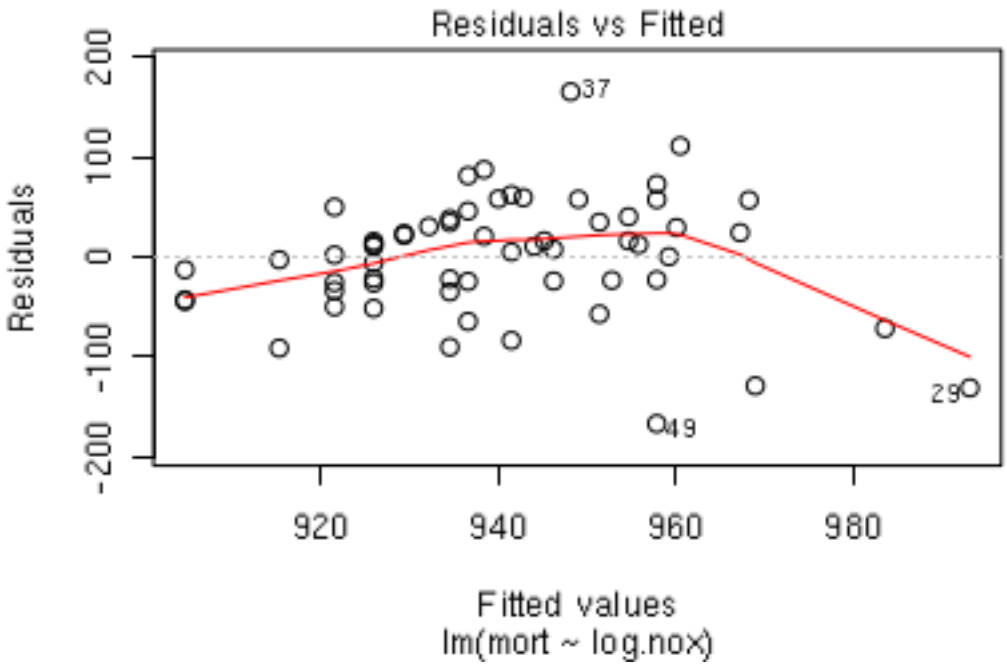
2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```
log.nox = log(nox)
model2.lm = lm(mort ~ log.nox)
ggplot(data = pollution, mapping = aes(x=log.nox,y=mort)) +
  geom_smooth(se = FALSE)+
  geom_point()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
plot(model2.lm,which=1)
```



In model 2, we use log transformation in nox and the residual plot looks evenly distributed around 0, which is better than model 1.

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

```
summary(model2.lm)
```

```
##
## Call:
## lm(formula = mort ~ log.nox)
##
## 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
```

For each increase of 10^x in NOX, there is respective change of 15.335x in mortality.

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

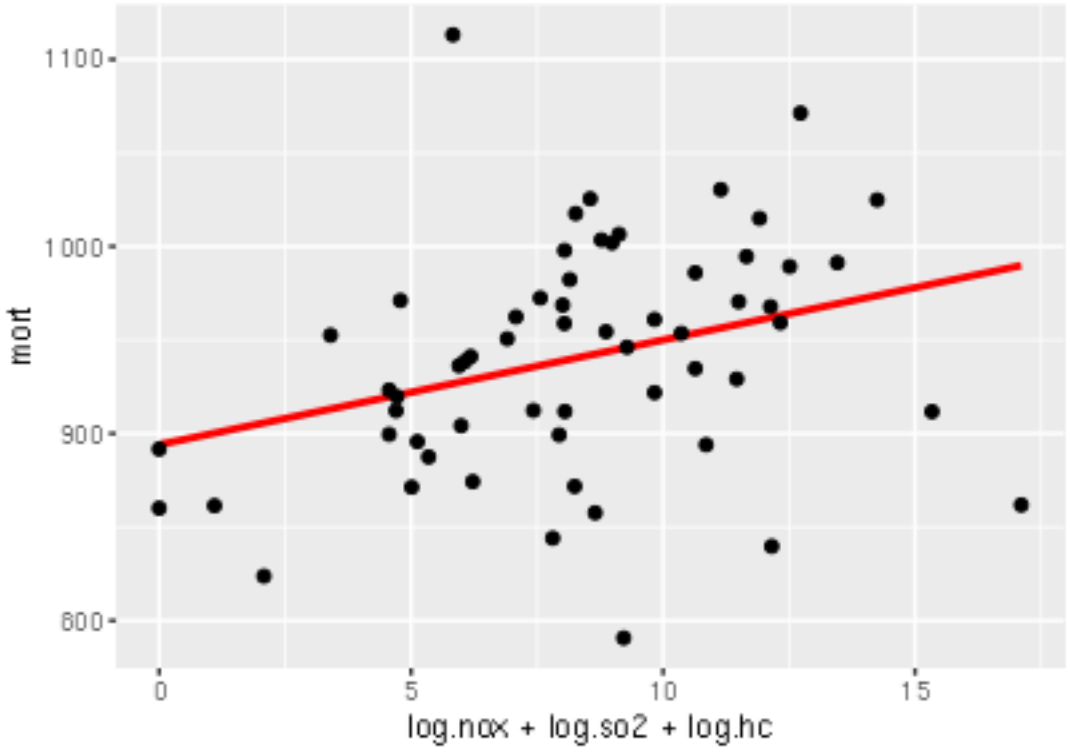
```
confint(model2.lm,level=0.99)
```

```
##                   0.5 %    99.5 %
## (Intercept) 858.988556 950.46037
## log.nox      -2.230963  32.90196
```

[-2.230,32.901] is the range of values that you can be 99% certain contains the true value coefficient of log(nox).

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.

```
so2 = pollution$so2
hc = pollution$hc
log.so2 = log(so2)
log.hc = log(hc)
model3.lm = lm(mort ~ log.nox + log.so2 + log.hc)
ggplot(data = pollution, mapping = aes(x=log.nox + log.so2 + log.hc,y=mort)) +
  geom_smooth(se = FALSE,method="lm",col="red")+
  geom_point()
```




```
summary(model3.lm)
```

```
##
## Call:
## lm(formula = mort ~ log.nox + log.so2 + log.hc)
##
## 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
```

The intercept means when nox, so2, hc level is 1, the average mortality rate is 924.965.

For each increase of 10^x in nox, so2, hc there is respective change of 58.336x, 11.762x, -57.3x in 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.)

```
# split dataset into training and test sets
train = pollution[1:(nrow(pollution)/2), ]
test = pollution[((nrow(pollution)/2)+1):nrow(pollution), ]

# fit linear model
log.nox = log(train$nox)
log.so2 = log(train$so2)
log.hc = log(train$hc)
train.lm = lm(mort ~ log.nox + log.so2 + log.hc, data=train)
display(train.lm)
```

```
## lm(formula = mort ~ log.nox + log.so2 + log.hc, data = train)
##             coef.est coef.se
## (Intercept) 899.97    25.71
## log.nox      10.57    29.59
## log.so2      21.87    12.32
## log.hc      -17.47    26.21
## ---
## n = 30, k = 4
## residual sd = 52.07, R-Squared = 0.25
```

```
# Predict test half
prediction = predict(train.lm, test)
prediction
```

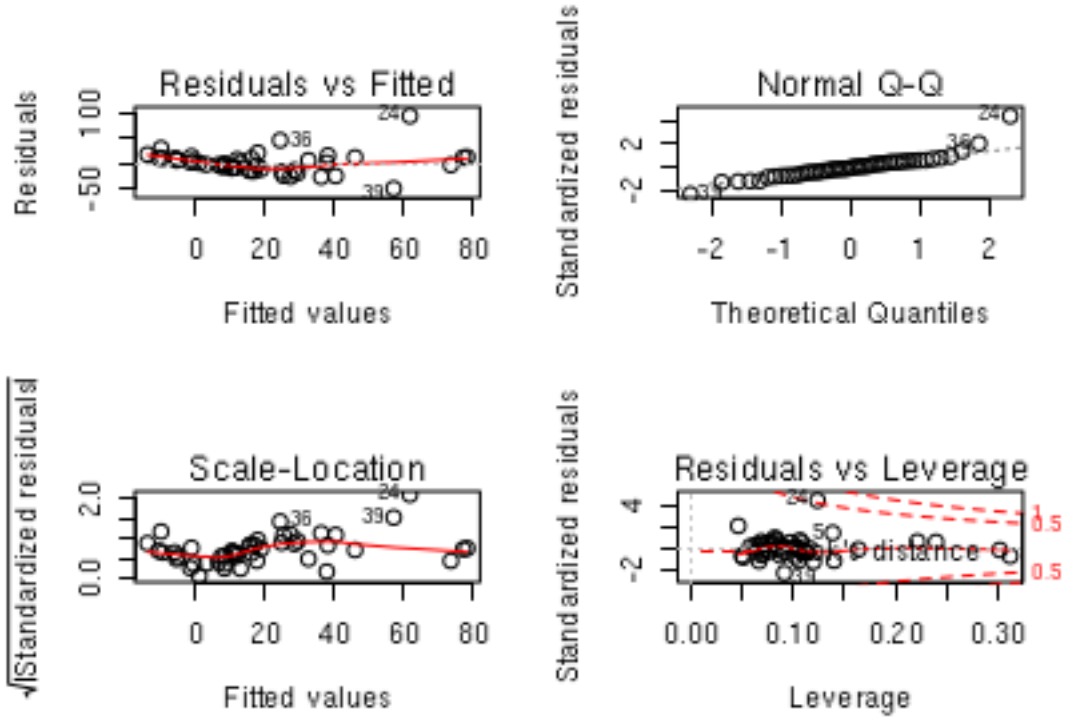
```
##       31       32       33       34       35       36       37       38
## 964.5427 968.0708 964.0479 940.9253 989.1818 970.6696 973.6326 913.6184
##       39       40       41       42       43       44       45       46
## 954.6750 942.6130 943.5009 988.5634 986.4356 963.0706 927.6083 899.9675
##       47       48       49       50       51       52       53       54
## 943.9335 945.2951 973.8920 925.1476 899.9675 933.8022 910.4314 927.9224
##       55       56       57       58       59       60
## 885.6610 952.1651 910.9247 954.4114 954.1901 989.6347
```

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.

```
status = teengamb$status
z.status = (status - mean(status)) / (2*sd(status))
model1.lm = lm(gamble ~ sex + z.status + income + verbal, data = teengamb)
par(mfrow=c(2,2))
plot(model1.lm)
```




```
summary(model1.lm)
```

```
##
## Call:
## lm(formula = gamble ~ sex + z.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)   24.918     15.515   1.606   0.1157
## sex          -22.118      8.211  -2.694   0.0101 *
## z.status       1.803      9.706   0.186   0.8535
```

```
## income         4.962      1.025   4.839 1.79e-05 ***
## verbal        -2.959      2.172  -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
```

The intercept means average spends on gambling for male with same status, income, and verbal is 24.918.

The coefficient of sex means the difference in predicted expenditure on gambling between male (sex=0) and female (sex=1) will be -22.118.

For someone has same income and verbal level, with every 1 unit increase in status, spends in gambling will increase 1.803.

For someone has same status and verbal level, with every 1 unit increase in income will spends 4.962 more in gambling.

For someone has same status and income level, with every 1 unit increase in verbal level will spends 2.595 less in gambling.

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

```
confint(model1.lm)
```

```
##                 2.5 %    97.5 %
## (Intercept)  -6.392223 56.229020
## sex         -38.689030 -5.547630
## z.status    -17.783326 21.390166
## income        2.892654  7.031305
## verbal       -7.343070  1.424083
```

[-6.392,56.229] is the range of values that you can be 95% certain contains the true value of intercept.

[-38.689,-5.547] is the range of values that you can be 95% certain contains the true value coefficient of sex.

[-17.78,21.39] is the range of values that you can be 95% certain contains the true value coefficient of mean of status.

[2.89,7.03] is the range of values that you can be 95% certain contains the true value coefficient of income.

[-7.34,1.42] is the range of values that you can be 95% certain contains the true value coefficient of verbal.

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?

```
income = teengamb$income
verbal = teengamb$verbal
avg.data = data.frame(
  sex=0,z.status=mean(z.status),income=mean(income),verbal=mean(verbal))

avg.prediction = predict(model1.lm,avg.data,interval="confidence", level=0.95)
avg.prediction
```

```
##        fit      lwr      upr
## 1 28.24252 18.78277 37.70227
```

```
max.data = data.frame(
  sex=0,z.status=max(z.status),income=max(income),verbal=max(verbal))
max.prediction = predict(model1.lm,max.data,interval="confidence", level=0.95)
max.prediction
```

```
##        fit      lwr      upr
## 1 71.30794 42.23237 100.3835
```

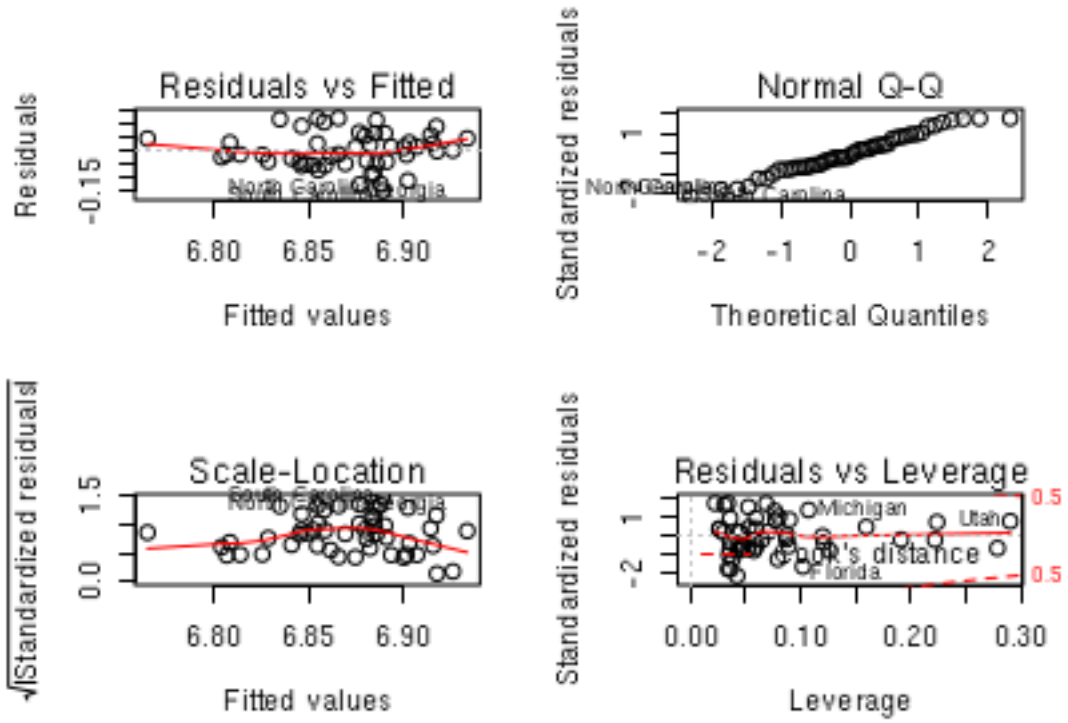
Because the standard deviation of max.data is bigger than avg.data.

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.

```
total = sat$total
log.total = log(sat$total)
expend = sat$expend
salary = sat$salary
ratio = sat$ratio
z.expend = (expend - mean(expend)) / (sd(expend))
z.salary = (salary - mean(salary)) / (sd(salary))
z.ratio = (ratio - mean(ratio)) / (sd(ratio))
model1.lm = lm(log.total ~ z.expend + z.ratio + z.salary,sat)
par(mfrow=c(2,2))
plot(model1.lm)
```




```
summary(model1.lm)
```

```
##
## Call:
## lm(formula = log.total ~ z.expend + z.ratio + z.salary, data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151140 -0.046616 -0.006997  0.046837  0.123402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.87016    0.01001 686.348   <2e-16 ***
## z.expend     0.02391    0.03098   0.772   0.4442
## z.ratio      0.01540    0.01529   1.008   0.3189
## z.salary    -0.05443    0.02877  -1.892   0.0648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07078 on 46 degrees of freedom
## Multiple R-squared:  0.2082, Adjusted R-squared:  0.1566
## F-statistic: 4.032 on 3 and 46 DF,  p-value: 0.01256
```

The intercept is the average total sat score when z.expend,z.ratio,and z.salary are 0 which is e^6.87 = 962.95.

The coefficient of z.expend means the difference in log sat score corresponding to a 1 standard-deviation difference in expend is 0.023.

The coefficient of z.ratio means the difference in log sat score corresponding to a 1 standard-deviation difference in ratio is 0.015.

The coefficient of z.salary means the difference in log sat score corresponding to a 1 standard-deviation difference in salary is -0.054.

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

```
confint(model1.lm,level=0.98)
```

```
##                     1 %       99 %
## (Intercept)  6.84603569 6.89428636
## z.expend    -0.05075940 0.09857730
## z.ratio     -0.02143981 0.05224324
## z.salary    -0.12377316 0.01490439
```

[6.84,6.89] is the range of values that you can be 98% certain contains the true value of intercept.

[-0.05,0.09] is the range of values that you can be 98% certain contains the true value coefficient of z.expend.

[-0.021,0.05] is the range of values that you can be 98% certain contains the true value coefficient of z.ratio.

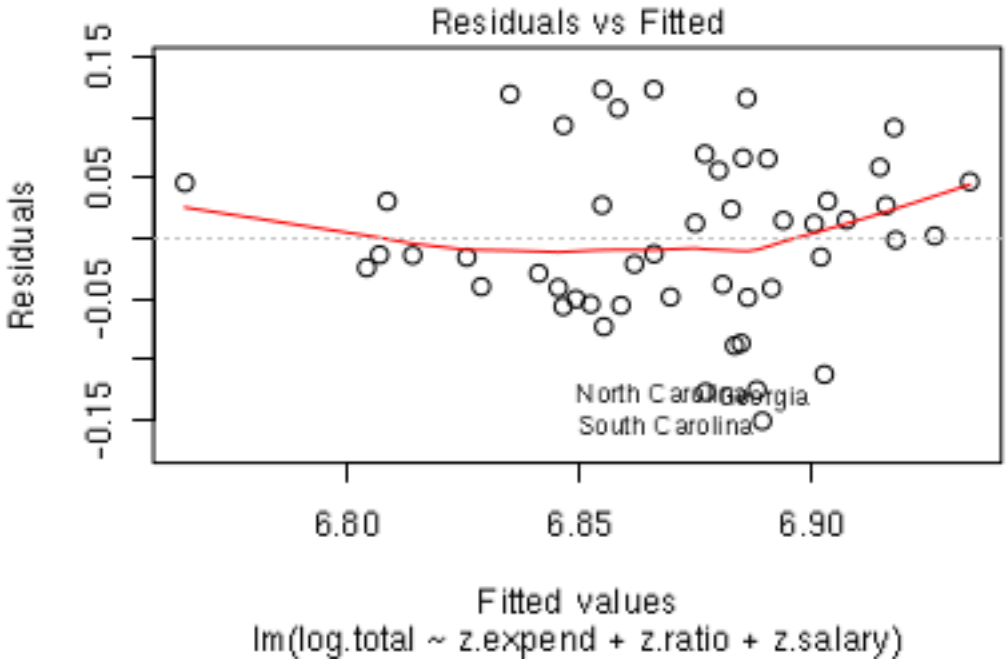
[-0.123,0.014] is the range of values that you can be 98% certain contains the true value coefficient of z.salary.

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?

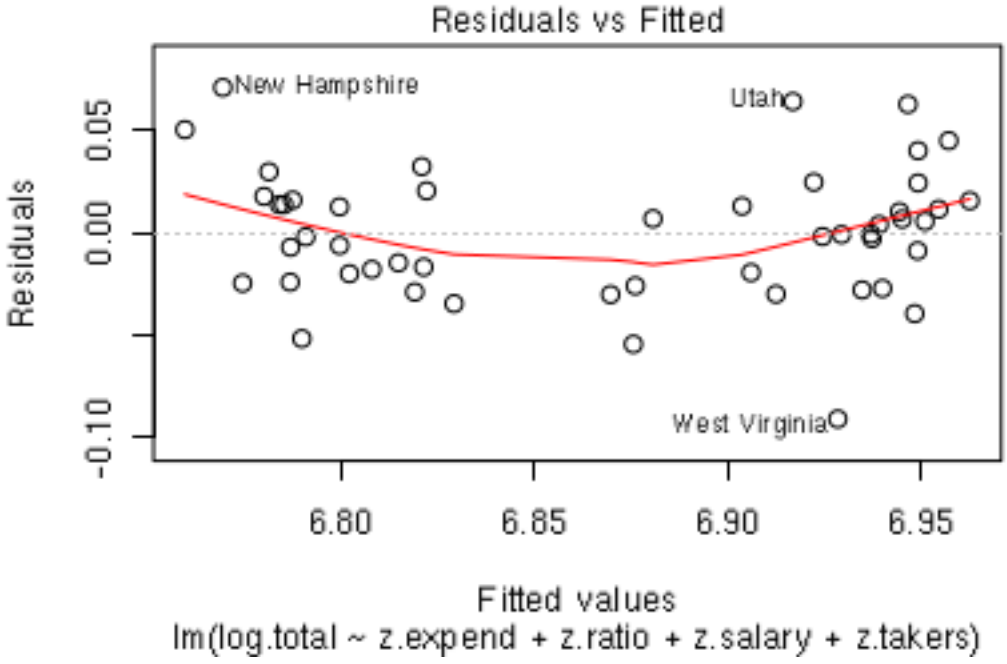
```
takers = sat$takers
z.takers = (takers - mean(takers)) / (sd(takers))
model2.lm = lm(log.total ~ z.expend + z.ratio + z.salary + z.takers,sat)
summary(model2.lm)
```

```
##
## Call:
## lm(formula = log.total ~ z.expend + z.ratio + z.salary + z.takers,
##     data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.091157 -0.023196 -0.000844  0.015822  0.070993
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  6.870161   0.004682 1467.352   <2e-16 ***
## z.expend     0.006954   0.014551    0.478    0.635
## z.ratio     -0.007976   0.007377   -1.081    0.285
## z.salary     0.009965   0.014359    0.694    0.491
## z.takers    -0.080545   0.006266  -12.855   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03311 on 45 degrees of freedom
## Multiple R-squared:  0.8305, Adjusted R-squared:  0.8155
## F-statistic: 55.13 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
plot(model1.lm,which=1)
```




```
plot(model2.lm,which=1)
```



The r^2 of model 2 is 0.83 which is much better than model 1. So the model with takers as predictor explain

the outcome better.

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, $D_i - R_i$

Advantage: Sems good transformation because is symmetric and centered at 0.

Disadvantage: Not proportional.This could limit the effectiveness of the predictor if districts differ widely in average money raised

- The ratio, D_i/R_i

Advantage: It is easy to interpret the effect of ratio.

Disadvantage: This transformation has the disadvantage of being centered at 1 and that is asymmetric. In particualar it tends to zero for case where the Republics have more money raised than Democrats, and tend to infinity on the opposite case.

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

Advantage:It is centered to zero and is symmetric; proportional to the magnitude of the difference.

Disadvantage: Similar to first transformation, not proportional.

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

Advantage: This transformation is centered at 0.5 and symmetric.

Disadvantage: Similar to transformation 2, only indicate the effect of the relative ratio.

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 $x^\star = 10(x - 1)$ and that y is regressed on $x^\star$. 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 $y^{\star\star} = 5(y + 2)$ and that $y^{\star\star}$ 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?

Transformation 1-6 please see attached pdf

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.