Homework 02

Chaoqun Yin

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.

```
head(complete.cases(heights)) #check the missing data
```

```
## [1] FALSE FALSE  TRUE  TRUE  TRUE FALSE
```

```
heights.complete <- na.omit(heights) #omit missing data

head(heights.complete == 0) #check the 0 data
```

```
##     earn height1 height2   sex  race  hisp    ed yearbn height
## 3  FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
## 4  FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
## 5  FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
## 7  FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
## 9  FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
## 10 FALSE   FALSE   FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE
```

```
zerodata <- which(heights.complete == 0)#omit 0
heights.complete <- heights.complete[-zerodata,]
```

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?

```
earn <- heights.complete$earn
sumheight <- heights.complete$height1 + heights.complete$height2
weight <- heights.complete$height
```

In order to interpret the intercept as average earning for people with average height, we should make a center transformation to the linear regression using earnings subtracting the mean of earnings.

```r
center.height <- sumheight - mean(sumheight)
regout.1 <- lm(earn ~ center.height)
summary(regout.1)
```

```
##
## Call:
## lm(formula = earn ~ center.height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25398 -12175  -4156   6299 174002
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23154.8      561.7  41.226  < 2e-16 ***
## center.height    607.6      183.4   3.313 0.000951 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19390 on 1190 degrees of freedom
## Multiple R-squared:  0.009138,   Adjusted R-squared:  0.008305
## F-statistic: 10.97 on 1 and 1190 DF,  p-value: 0.0009515
```

The formula is

$$earn = 607.6 height_{centered} + 23154.8$$

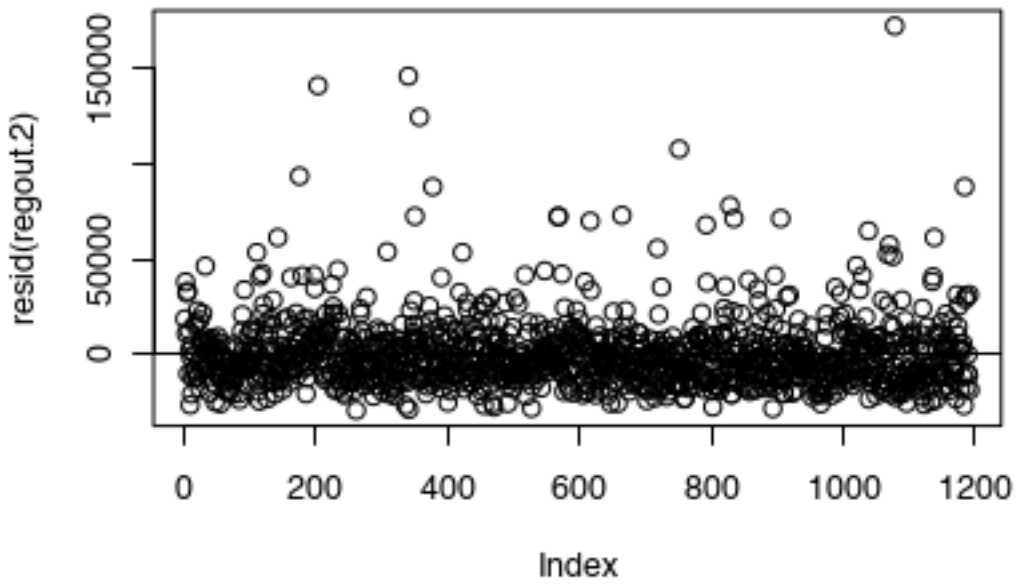
a person with average height is supposed to have the earning of 23154.8.

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
#Regress earnings on centered height and centered weight
center.weight <- weight - mean(weight)
regout.2 <- lm(earn ~ center.height + center.weight, heights.complete)
summary(regout.2)
```

```
##
## Call:
## lm(formula = earn ~ center.height + center.weight, data = heights.complete)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29704 -11385  -3494   6392 172397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23154.8      546.5  42.368  < 2e-16 ***
## center.height    141.4      187.2   0.755     0.45
## center.weight   1228.3      149.1   8.237  4.6e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1189 degrees of freedom
## Multiple R-squared:  0.06263,    Adjusted R-squared:  0.06106
## F-statistic: 39.72 on 2 and 1189 DF,  p-value: < 2.2e-16
```

```r
#Plot the residuals
plot(resid(regout.2))
abline(h = 0)
```

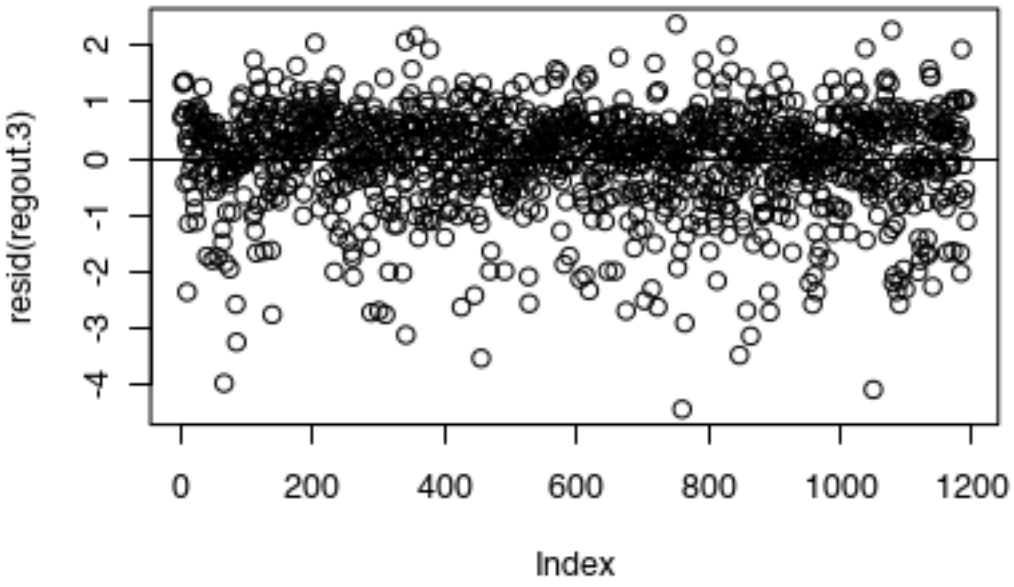


The coefficients are large so we make a log transformation to the linear regression.

```r
#Make a log transformation on the earn
regout.3 <- lm(log(earn) ~ center.height + center.weight, heights.complete)
sumary(regout.3)
```

```
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)   9.7143494  0.0258667 375.5546 < 2.2e-16
## center.height 0.0091583  0.0088617   1.0335    0.3016
## center.weight 0.0566120  0.0070575   8.0215 2.484e-15
##
## n = 1192, p = 3, Residual SE = 0.89306, R-Squared = 0.06
```

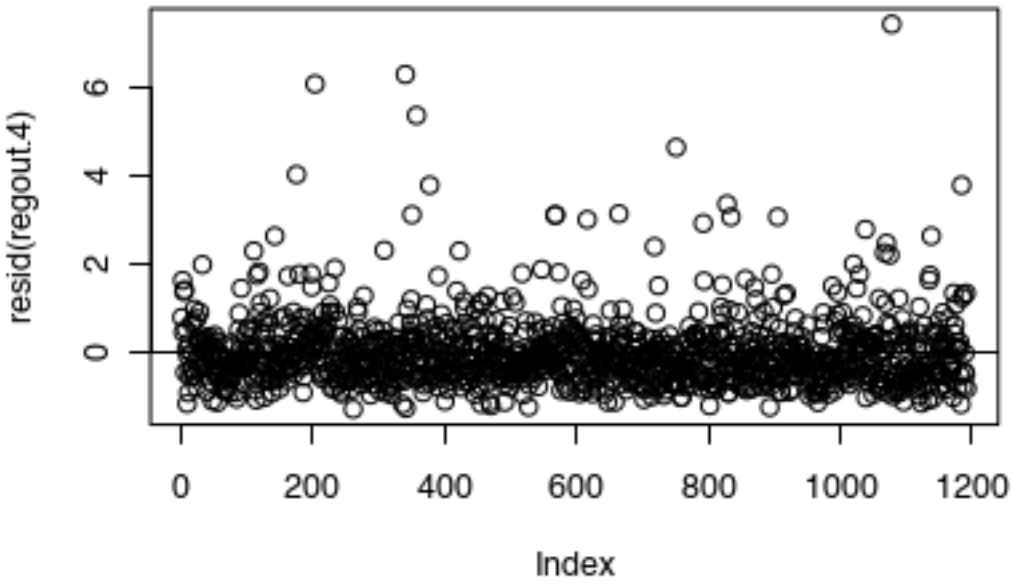
```r
#Plot the residuals
plot(resid(regout.3))
abline(h = 0)
```




```
#Transform earnings by dividing it by its mean
regout.4 <- lm((earn/mean(earn)) ~ center.height + center.weight, heights.complete)
summary(regout.4)
```

```
##
## Call:
## lm(formula = (earn/mean(earn)) ~ center.height + center.weight,
##     data = heights.complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2829 -0.4917 -0.1509  0.2761  7.4454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.000000   0.023603  42.368  < 2e-16 ***
## center.height 0.006106   0.008086   0.755     0.45
## center.weight 0.053047   0.006440   8.237  4.6e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8149 on 1189 degrees of freedom
## Multiple R-squared:  0.06263,    Adjusted R-squared:  0.06106
## F-statistic: 39.72 on 2 and 1189 DF,  p-value: < 2.2e-16
```

```
plot(resid(regout.4))
abline(h = 0)
```



Upon the transformation above, I prefer the regout.3. It makes a log transformation for earn to make a linear regression on centered height and centered weight. The results have the proper coefficients and the residuals are small.

4. Interpret all model coefficients. As regression 1 for example, for regressing earnings on centered height in regout.1: When the height is the average value, the earning is supposed to be 23154.8. After the height increases 1 unit, the earning increases 607.6.

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

```
confint(regout.1, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept)   22052.8332 24256.7138
## center.height   247.7608   967.4621
```

- As regressing earnings on centered height for example: We have the confidence of 95% that the range [22052.83,24256.71] includes the true value of intercept of the regression. It is not acrossing the 0, so it is statistically significant. Similarly, We have the confidence of 95% that the range [247.76,967,46] includes the true value of centered height's true value. It is not acrossing the 0, so it is statistically significant.

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

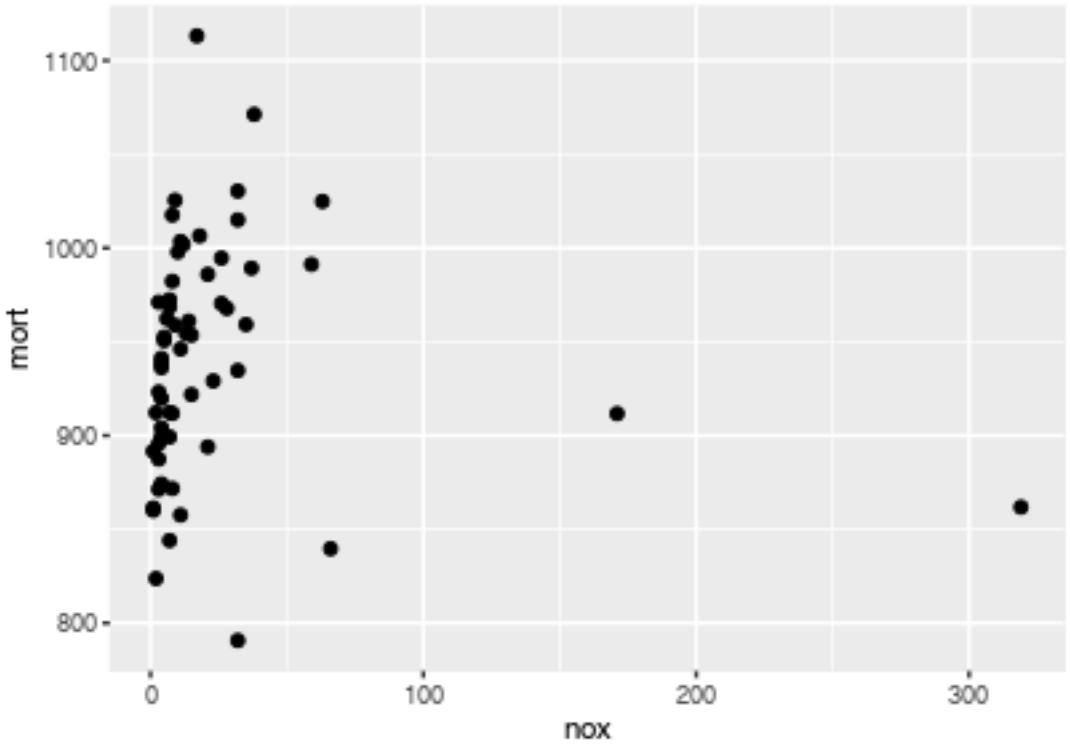
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all facilities
- DENS Population per sq. mile in urbanized areas, 1960
- NONW % non-white population 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
- SO2 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.

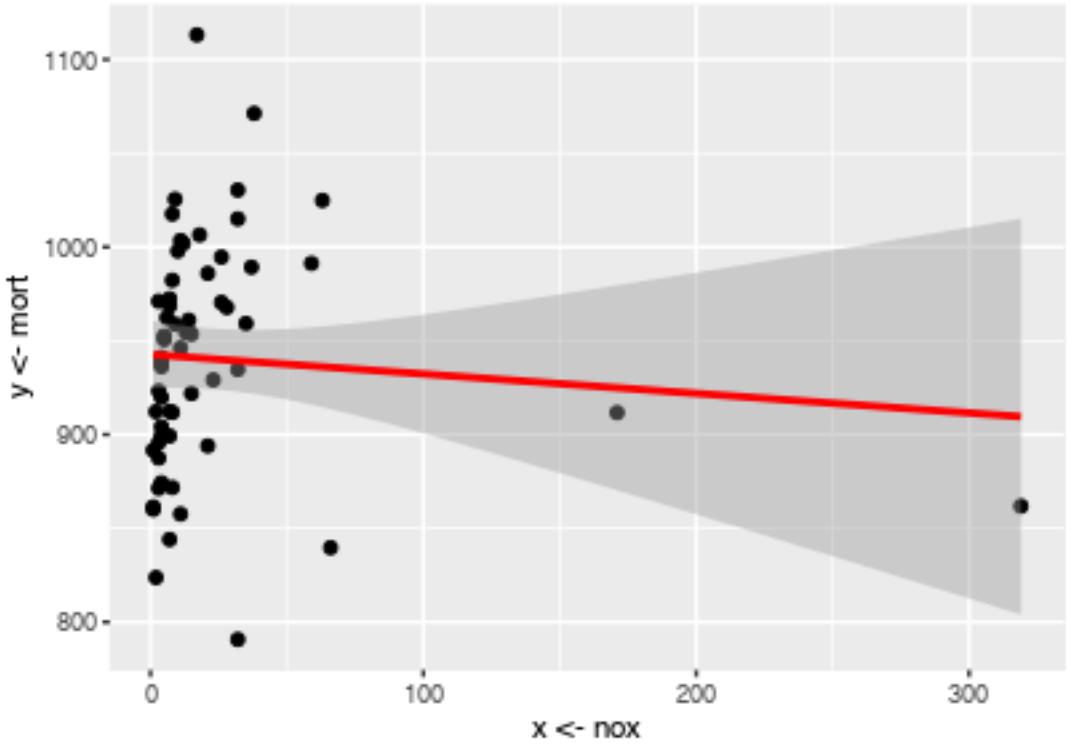
```
#Make the scatterplot
library("ggplot2")
ggplot(data = pollution, aes(x = nox, y = mort)) +
    geom_point()
```



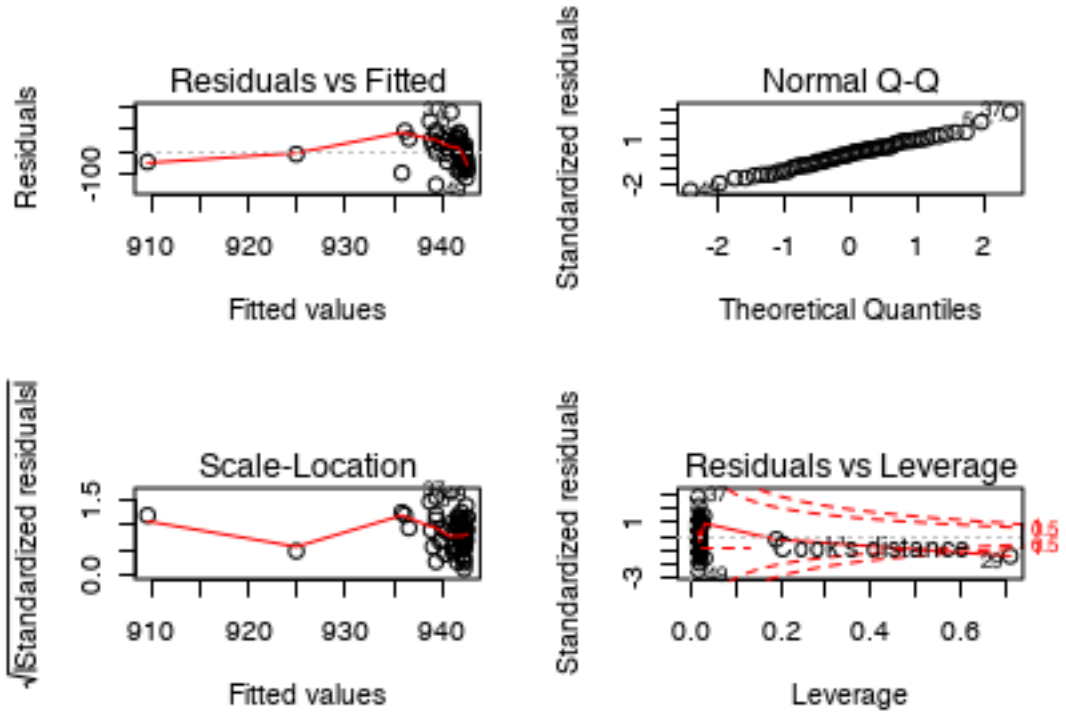
```r
(regout.5 <- lm(mort ~ nox, pollution))
```

```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## Coefficients:
## (Intercept)          nox
##    942.7115      -0.1039
```

```r
#Make the plot
ggplot(pollution, aes(x <- nox, y <- mort)) +
  geom_point() +
  stat_smooth(method = lm, col = "red")
```




```r
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
plot(regout.5) # Plot the residuals
```



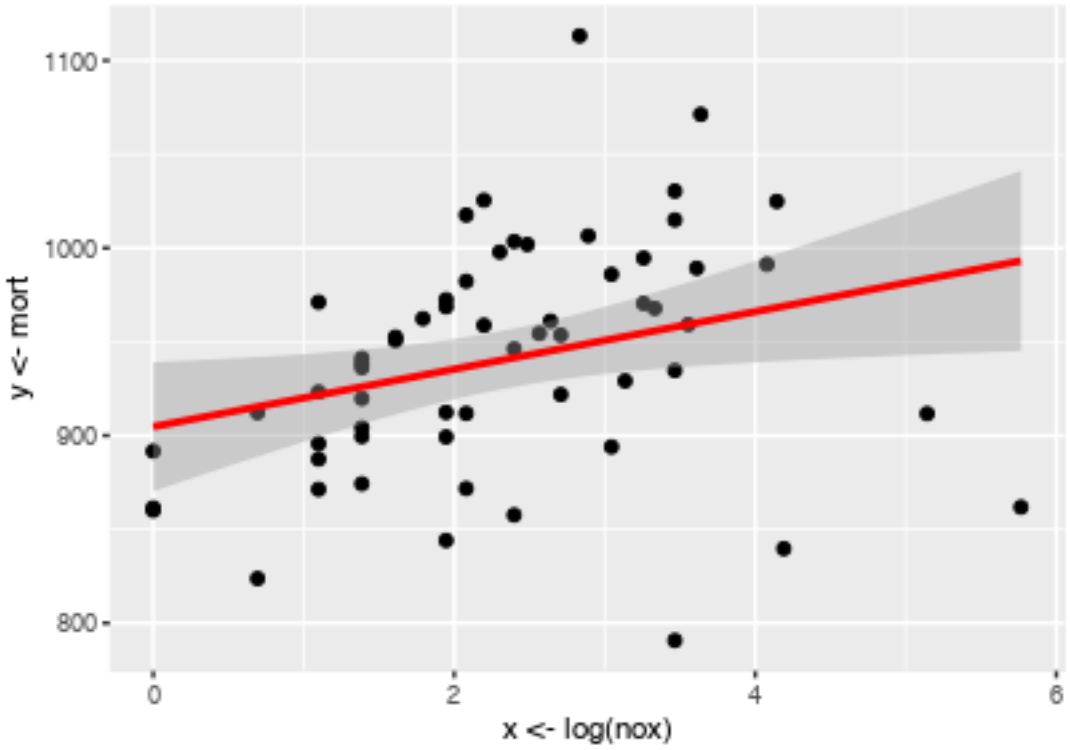
It seems that it does not fit the linear regression model well. The residuals seems not good for the model. Maybe we should make some transformation for the variables of the model.

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. **We can find that the points are gathered in the left of the scatter plot, so we can use log(nox) to fit the model.**

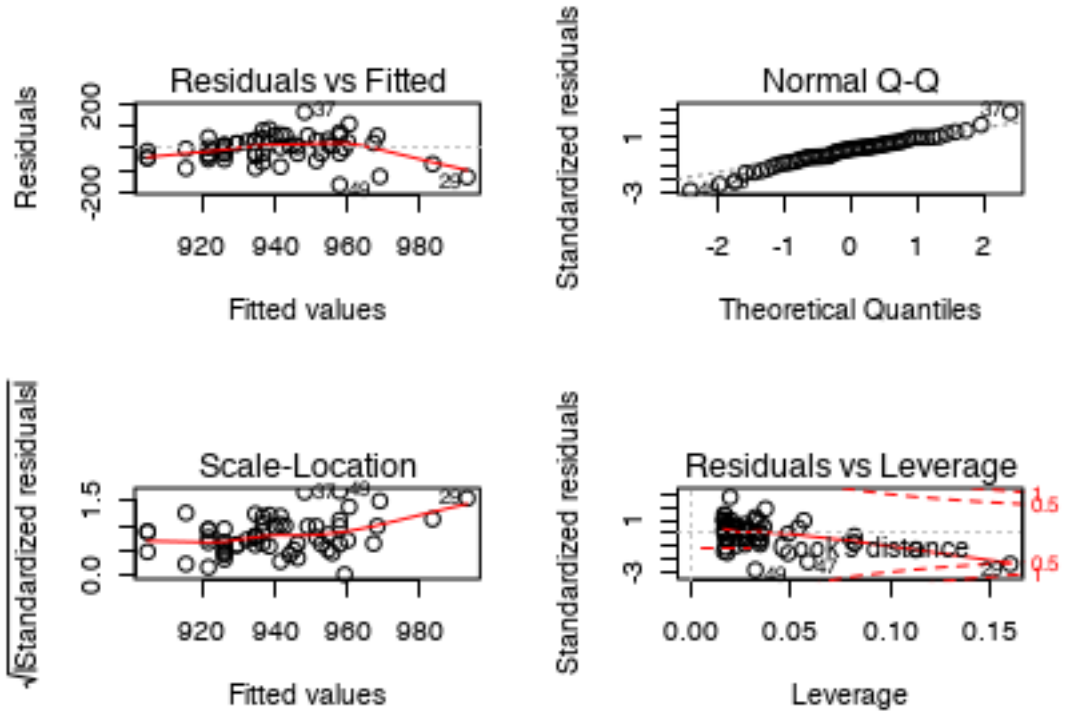
```
#Regress mort on log(nox)
regout.6 <- lm(mort ~ log(nox), pollution)
summary(regout.6)
```

```
##
## 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
#Make the new plot
ggplot(pollution, aes(x <- log(nox), y <- mort)) +
  geom_point() +
  stat_smooth(method = lm, col = "red")
```




```r
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
plot(regout.6) # Plot
```



It seems that the regression has better fit for the variables and the residuals versus fitted value looks better than the formal regression.

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

$$mort = 15log(nox) + 905$$

. Controlling other variables, 1 unit change in

$$log(nox)$$

will lead to 15 unit change in mortality.

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

```
confint(regout.6, level = 0.99)
```

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

We have the confidence of 99% that the range [-2, 33] includes the true value of the slope coeffiecient of the regression. It is acrossing the 0, so we cannot say it is statistically significant.

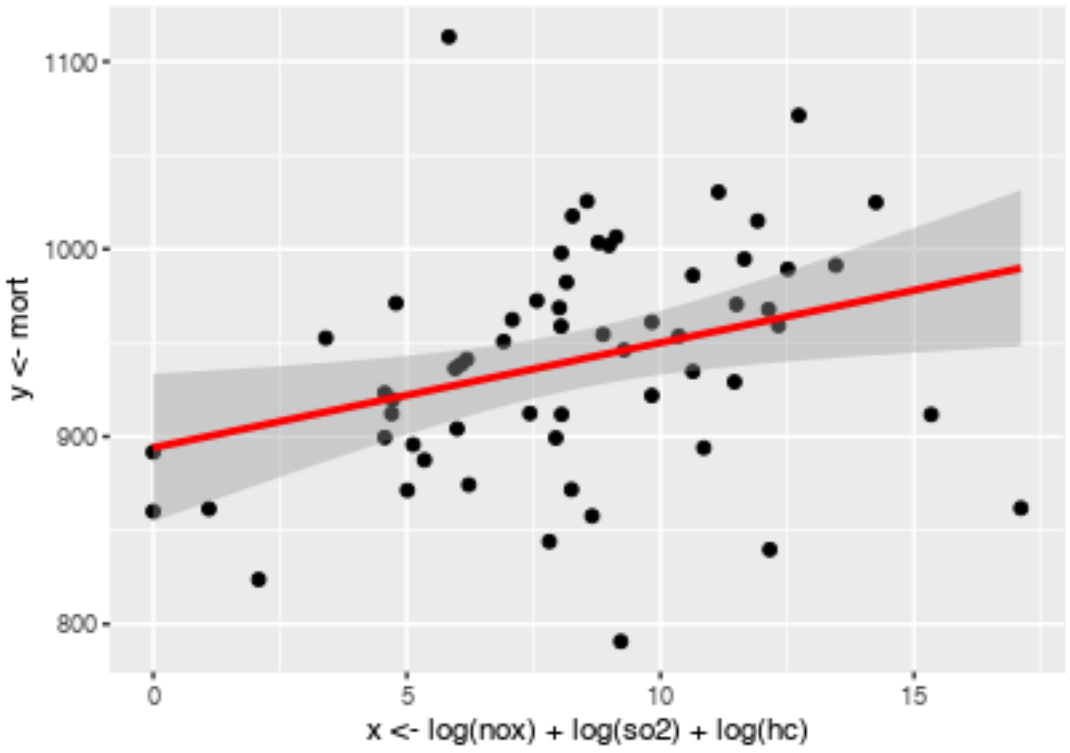
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

Similarly to above analysis, we should make a log transformation to the right of the equation. As follows:

```
#Make a linear regression
regout.7 = lm(mort ~ log(nox) + log(so2) + log(hc), data = pollution)
summary(regout.7)
```

```
##
## 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
```

```
#Plot it
ggplot(pollution, aes(x <- log(nox) + log(so2) + log(hc), y <- mort)) +
  geom_point() +
  stat_smooth(method = lm, color = "red")
```



When the other variables are all 0, the total age-adjusted mortality rate per 100,000 is 925. When $log(nox)$ increases 1 unit, the mortality rate increases 58; when $log(so2)$ increases 1 unit, the mortality rate increases 12; when $log(hc)$ increases 1 unit, the mortality rate decreases 57.

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.)

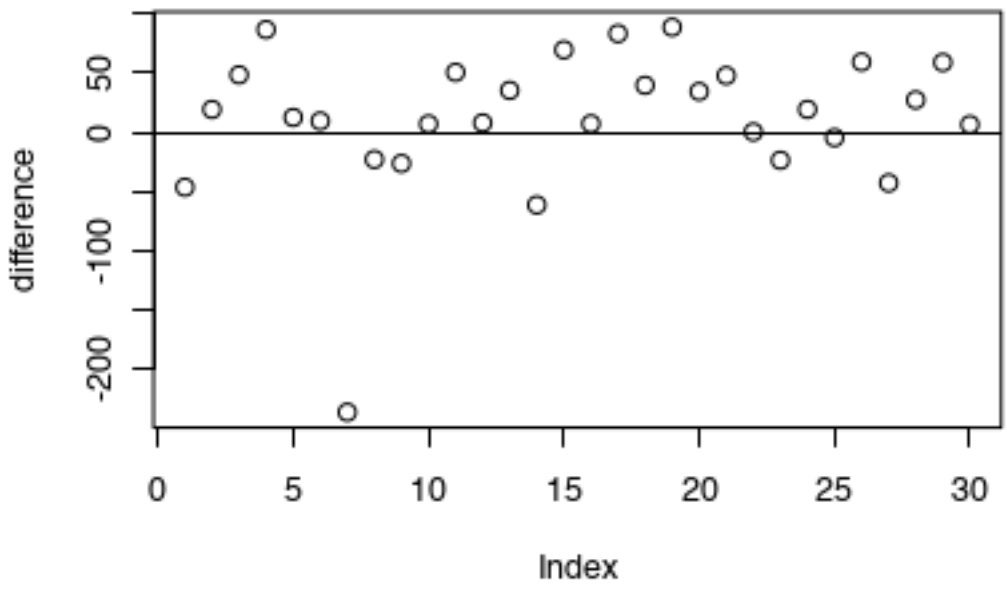
```
#Fit the model
dim(pollution)
```

```
## [1] 60 16
```

```
pollution.fit <- pollution[1:30,]
pollution.pred <- pollution[31:60,]
regout.8 <- lm(mort ~ log(nox) + log(so2) + log(hc), data = pollution.fit)
summary(regout.8)
```

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

```
#Make prediction
pred <- predict(object = regout.8, newdata = data.frame(mort = pollution.pred$mort,
                                                        nox = pollution.pred$nox,
                                                        so2 = pollution.pred$so2,
                                                        hc = pollution.pred$hc), interval = "prediction")
#the difference between the actual values and the predicted values
difference <- pred[,1] - pollution[31:60,]$mort
plot(difference)
abline(h = 0)
```



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.

```
teengamb <- teengamb
status.center <- teengamb$status - mean(teengamb$status)
verbal.center <- teengamb$verbal - mean(teengamb$verbal)
regout.8 <- lm(log(gamble+1) ~ sex + status.center + log(income) + verbal.center, data = teengamb)
summary(regout.8)
```

```
##
## Call:
## lm(formula = log(gamble + 1) ~ sex + status.center + log(income) +
##     verbal.center, data = teengamb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67231 -0.56225  0.08561  0.80866  2.11944
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.22334    0.39911   3.065 0.003793 **
## sex           -1.10598    0.39296  -2.814 0.007404 **
## status.center  0.02430    0.01354   1.795 0.079887 .
```

```
## log(income)    0.93798    0.23640   3.968 0.000278 ***
## verbal.center -0.25520    0.10704  -2.384 0.021713 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.118 on 42 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.4423
## F-statistic: 10.12 on 4 and 42 DF,  p-value: 7.906e-06
```

The fomula is as follows:

$$log(gamble + 1) = 1.22 - 1.10sex + 0.02status.center + 0.93log(income) - 0.25verbal.center$$

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

```
confint(regout.8, level = 0.95)
```

```
##                      2.5 %      97.5 %
## (Intercept)    0.417908244  2.02876371
## sex           -1.898998675 -0.31295814
## status.center -0.003023852  0.05163342
## log(income)    0.460910955  1.41504488
## verbal.center -0.471208407 -0.03918313
```

We have the confidence of 95% that the range [0.41,2.02] will include the intercept's true value; the range doesn't cross 0 so the intercept is significant. Similarly, the range [-1.89,-0.31] will include the sex coefficient's true value; the range [-0.003,0.05] will include the centered status coefficient's true value; the range [0.46,1.41] will include the log income coefficient' true value; the range [-0.47,-0.039] will include the centered verbal coefficient's true value; the range doesn't cross zero so the coefficient is significant.

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?

```
prediction.average <- predict(object = regout.8, newdata = data.frame(sex=0, status.center = mean(teenga
prediction.average
```

```
##        fit        lwr      upr
## 1 1.625366 -0.7226398 3.973371
```

```
prediction.max <- predict(object = regout.8,newdata = data.frame(sex = 0, status.center = max(teengamb$s
prediction.max
```

```
##        fit        lwr      upr
## 1 2.028769 -0.3495968 4.407135
```

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.

```r
#Regress total sat score on expend, rario and aslary
regout.9 <- lm(total ~ expend + ratio + salary, data = sat)
summary(regout.9)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```r
#Make z-score transformation
ratio.center = sat$ratio - mean(sat$ratio)
z.expend = (sat$expend - mean(sat$expend)) / 2*sd(sat$expend)
z.salary = (sat$salary -mean(sat$salary))/2*sd(sat$salary)
regout.10 <- lm(total ~ z.expend + ratio.center + z.salary, data = sat)
summary(regout.10)
```

```
##
## Call:
## lm(formula = total ~ z.expend + ratio.center + z.salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   965.920      9.709  99.486   <2e-16 ***
## z.expend       24.169     32.360   0.747   0.4589
## ratio.center    6.330      6.542   0.968   0.3383
## z.salary       -2.970      1.581  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```r
#Make log transformation on expend, ratio and salary
regout.11 <- lm(total ~ log(expend) + log(ratio) + log(salary), data = sat)
```

```r
summary(regout.11)
```

```
##
## Call:
## lm(formula = total ~ log(expend) + log(ratio) + log(salary),
##     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(expend)     92.9      133.7   0.695   0.4905
## log(ratio)     117.3      121.2   0.968   0.3381
## log(salary)   -311.1      161.2  -1.930   0.0598 .
## ---
## 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
```

```r
#Make log tranformation on total sat score
regout.12 <- lm(log(total) ~ expend + ratio + salary, data = sat)
summary(regout.12)
```

```
##
## Call:
## lm(formula = log(total) ~ expend + ratio + 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)  6.971101   0.114360  60.958   <2e-16 ***
## expend       0.017544   0.022733   0.772   0.4442
## ratio        0.006796   0.006745   1.008   0.3189
## salary      -0.009162   0.004842  -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 z-score transformation seems the best among the four types of transformation.

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

```r
confint(regout.10, level = 0.98)
```

```
##                     1 %        99 %
```

```
## (Intercept)  942.519313 989.3206872
## z.expend     -53.823545 102.1616330
## ratio.center  -9.437308  22.0978419
## z.salary      -6.780639   0.8407381
```

Only the intercept does not across 0 and is statistically significant. Others all across 0 and are not statistically significiant.

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?

```
#Regress total sat score on expend, rario, aslary and takers
regout.13 <- lm(total ~ expend + ratio + salary + takers, data = sat)
summary(regout.13)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
#Make z-score transformation
takers.center <- sat$takers - mean(sat$takers)
regout.14 <- lm(total ~ z.expend + ratio.center + z.salary + takers.center, data = sat)
summary(regout.14)
```

```
##
## Call:
## lm(formula = total ~ z.expend + ratio.center + z.salary + takers.center,
##     data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   965.9200     4.6248 208.858  < 2e-16 ***
## z.expend        6.5491    15.4777   0.423    0.674
## ratio.center   -3.6242     3.2154  -1.127    0.266
## z.salary        0.5514     0.8036   0.686    0.496
```

```
## takers.center  -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
#Make log transformation on expend, ratio, salary and takers
regout.15 <- lm(total ~ log(expend) + log(ratio) + log(salary) + log(takers), data = sat)
summary(regout.15)
```

```
##
## Call:
## lm(formula = total ~ log(expend) + log(ratio) + log(salary) +
##     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(expend)   61.583     49.995   1.232    0.224
## log(ratio)     5.454     45.799   0.119    0.906
## log(salary)   33.024     63.616   0.519    0.606
## 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
```

```
#Make log tranformation on total sat score
regout.16 <- lm(log(total) ~ expend + ratio + salary + takers, data = sat)
summary(regout.16)
```

```
##
## Call:
## lm(formula = log(total) ~ expend + ratio + salary + 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)  6.9469966  0.0535243 129.791   <2e-16 ***
## expend       0.0051029  0.0106771   0.478    0.635
## ratio       -0.0035191  0.0032552  -1.081    0.285
## salary       0.0016772  0.0024168   0.694    0.491
## takers      -0.0030096  0.0002341 -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
```

Compared with the formal regressions, the regressions in 3 are better for the better r square.

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, $D_i - R_i$

It is the numeric difference beteen two variables. It can be input to compare the vote share based on the numeric difference of the amount of money raised. But it ignores the proportion.

- The ratio, D_i/R_i

Using this variable, we can compare when one party has the 1 percent more money raised how much the party has vote share more than the other one. But it ignores the actual number of each party.

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

$logD_i - logR_i = log(D_i/R_i)$, so it can reflect the proportion of the relation. In some case, it can simplify the problem. But it ignores the actual number too.

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

It uses the whole money raised for the two parties, takes the whole volumn into account to compare the difference of the money raised related the vote share. However, it can be hard to interpret.

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?

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.