Homework 02

Longhao Chen

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.

```
# First thing I do is to add another column of age by using 90 minus year born
# since this data is collected on 1990.
heights$age <- exp <- 90 - heights$yearbn
# Then I select people whose age is between 18 and 65
age18to65 <- heights[heights$age >= 18 & heights$age <= 65, ]
# Next, I filter out all the rows that have na
x <- na.omit(age18to65)
# The last step is to filter out people whose earning is 0
# because they are not our target population to study.
y <- x[x$earn != "0", ]
```

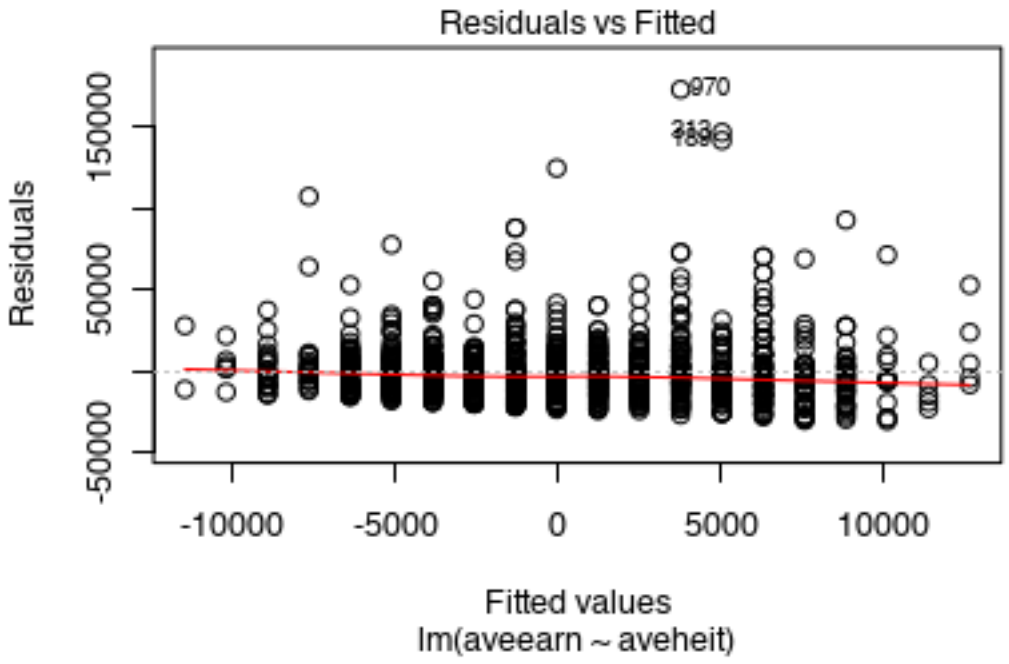
2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

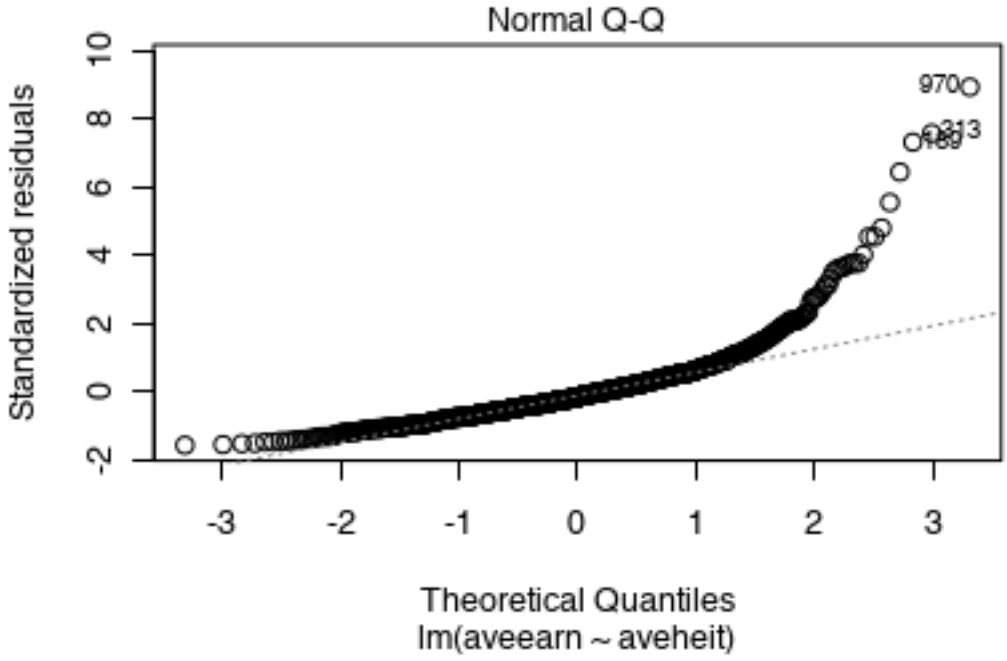
```
aveearn <- y$earn - mean(y$earn)
aveheit <- y$height - mean(y$height)
lmfit <- lm(aveearn ~ aveheit)
summary(lmfit)
```

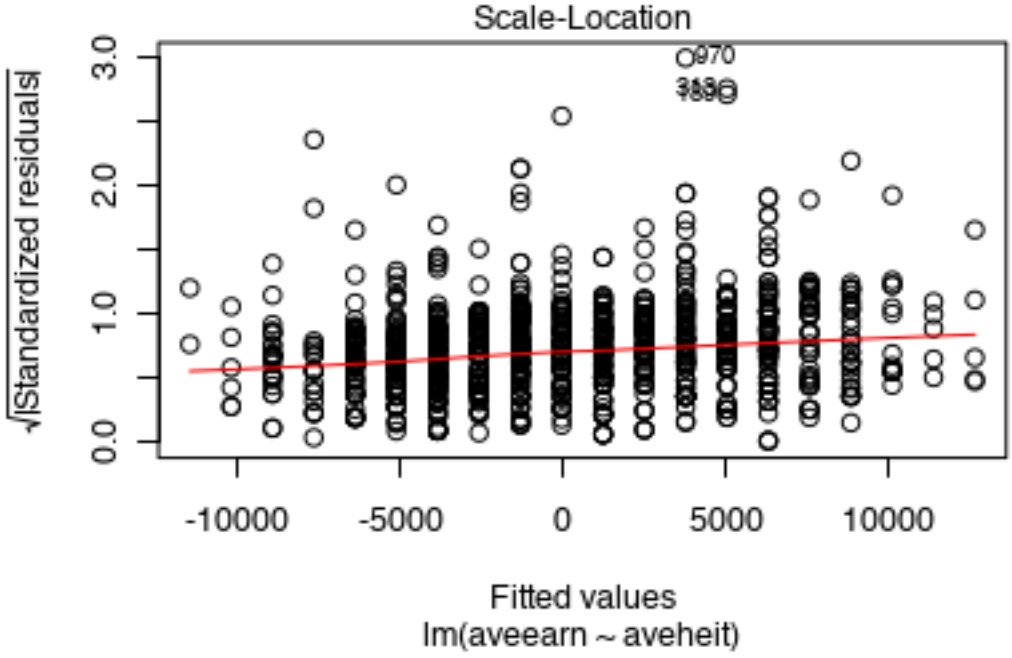
```
##
## Call:
## lm(formula = aveearn ~ aveheit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30623 -11380  -3568   6432 172538
##
```

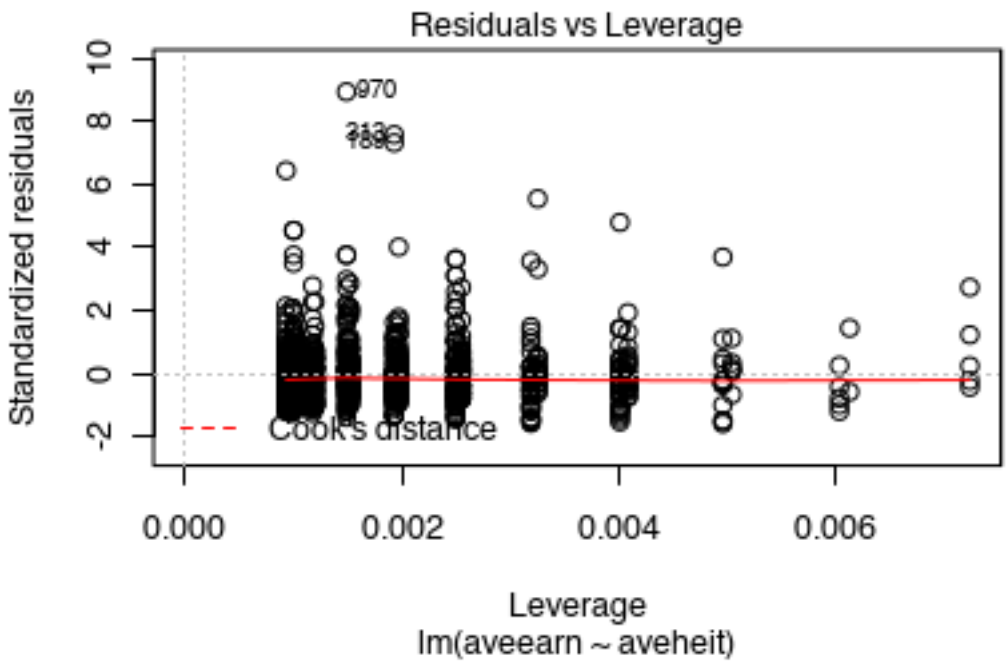
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.720e-13  5.904e+02   0.000        1
## aveheit      1.271e+03  1.541e+02   8.247 4.73e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19330 on 1070 degrees of freedom
## Multiple R-squared:  0.05976,    Adjusted R-squared:  0.05888
## F-statistic: 68.01 on 1 and 1070 DF,  p-value: 4.734e-16
```

```
plot(lmfit)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(aveearn ~ aveheit)
970


Scale-Location
√|Standardized residuals|
Fitted values
lm(aveearn ~ aveheit)
970



I use a linear transformation for both the earnings and height by deduct the average of earning and average of height from each data points. If permitted, I would rather regress the mean of height on the earn but not the center of earn.

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.

```
# First fit is the regress education and sex to log(earning)
newgender <- abs(y$sex - 2)
# I reordered the sex to female = 0 and male =1
# so that it is easier to interpreate in the linear model.
aveed <- y$ed - mean(y$ed) # This is to center the education years to its average.
earnpredit <- lm(log(y$earn) ~ newgender + aveed)
summary(earnpredit)
```

```
##
## Call:
## lm(formula = log(y$earn) ~ newgender + aveed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4922 -0.3216  0.1531  0.5196  2.1680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.50683    0.03431  277.10   <2e-16 ***
## newgender    0.53237    0.05243   10.15   <2e-16 ***
## aveed        0.11931    0.01100   10.85   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8494 on 1069 degrees of freedom
## Multiple R-squared:  0.1733, Adjusted R-squared:  0.1717
## F-statistic:   112 on 2 and 1069 DF,  p-value: < 2.2e-16
```

```
# Second fit incorporates the height into predictor variable
earnpredit2 <- lm(log(aveearn) ~ newgender + aveed + aveheit)
```

```
## Warning in log(aveearn): NaNs produced
```

```
summary(earnpredit2)
```

```
##
## Call:
## lm(formula = log(aveearn) ~ newgender + aveed + aveheit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9356 -0.8374  0.1312  0.8957  3.2150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.40222    0.11482  73.179  < 2e-16 ***
## newgender    0.66475    0.17947   3.704 0.000239 ***
## aveed        0.12292    0.02679   4.589 5.81e-06 ***
## aveheit      0.02611    0.02306   1.133 0.257983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.284 on 442 degrees of freedom
##   (626 observations deleted due to missingness)
## Multiple R-squared:  0.1163, Adjusted R-squared:  0.1103
## F-statistic:  19.4 on 3 and 442 DF,  p-value: 7.853e-12
```

I prefer the first model, which is to predict earning through years of education and sex. The reason is that the standard error is relatively small and P value also close to 0 for the intercept and coefficients. The confidence interval is possitive for both the intercepts and coefficients, which means that they have statistical significance. I choose not to put age down as a variable because it is not a linear relationship between age and earning. Old people (>60 ages) generally don't make as much as mid age people (40~60 ages).

4. Interpret all model coefficients.
5. The intercept of the model falls at 9.49, which indicates that at average education year, the female's earning is at the e^9.49=13226
6. The coefficient of new gender is 0.517, which indicates that if the person is a man, he is predicted to have an increase of log(earn)=0.517. This is to say that the earning of a man is e^(9.49+0.517)=22181 compared to a woman e^(9.49)=13227.
7. The coefficient of average education is 0.119. This means that the earning of a person is predicted to have an increase of log(earn)=0.119 for every extra year of education the person has.
8. Construct 95% confidence interval for all model coefficients and discuss what they mean.

```
confint(object = earnpredit, parm = "newgender", level = 0.95)
```

```
##               2.5 %    97.5 %
## newgender 0.4294806 0.6352501
```

```
confint(object = earnpredit, parm = "aveed", level = 0.95)
```

```
##              2.5 %    97.5 %
## aveed 0.09772828 0.1408953
```

The confidence interval for newgender means that if we run the model many times, the probability that the true value of coefficient of newgender falls between 0.43 and 0.64 is 95%. The other one means that 95% chance the coefficient of aveed will falls between 0.097 and 0.14.

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:

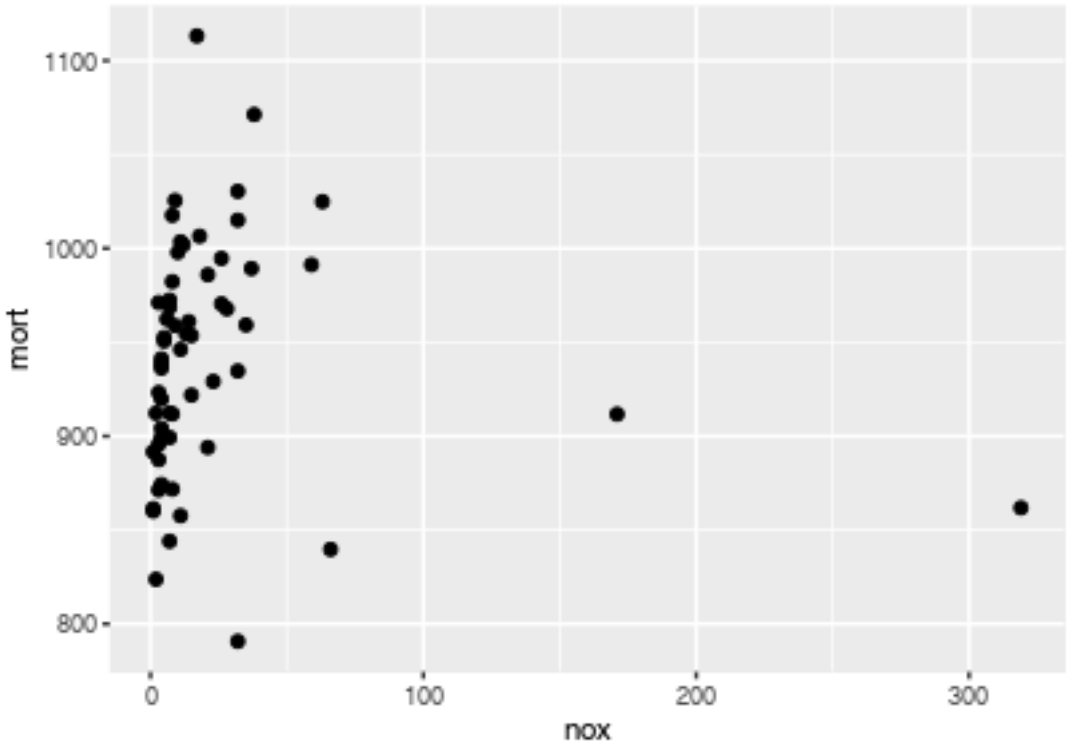
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all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

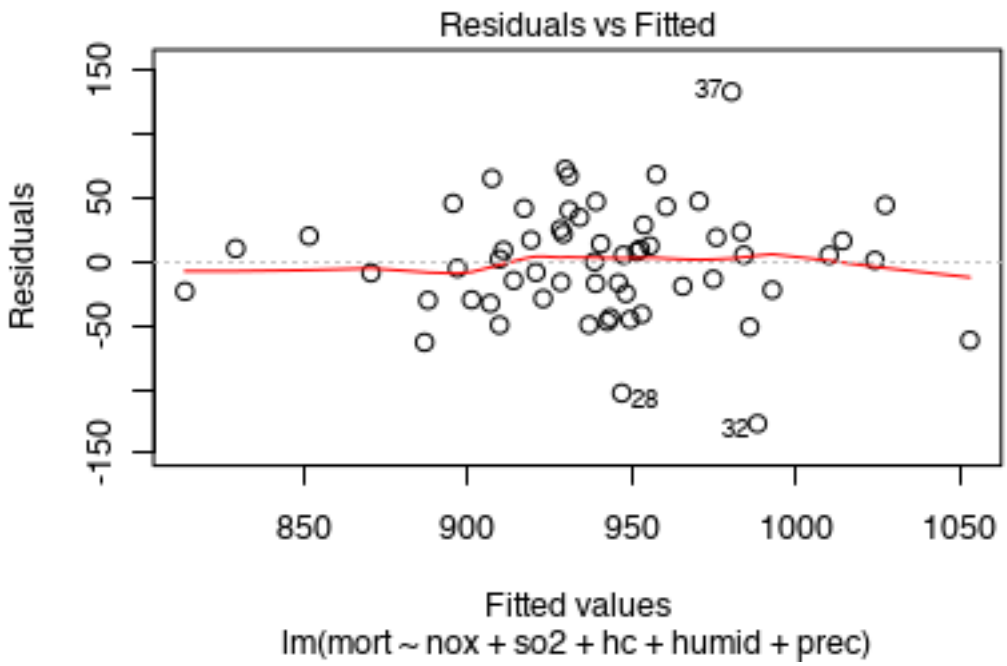
1. Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

```
ggplot(data = pollution) +
  geom_point(mapping = aes(x = nox, y = mort))
```

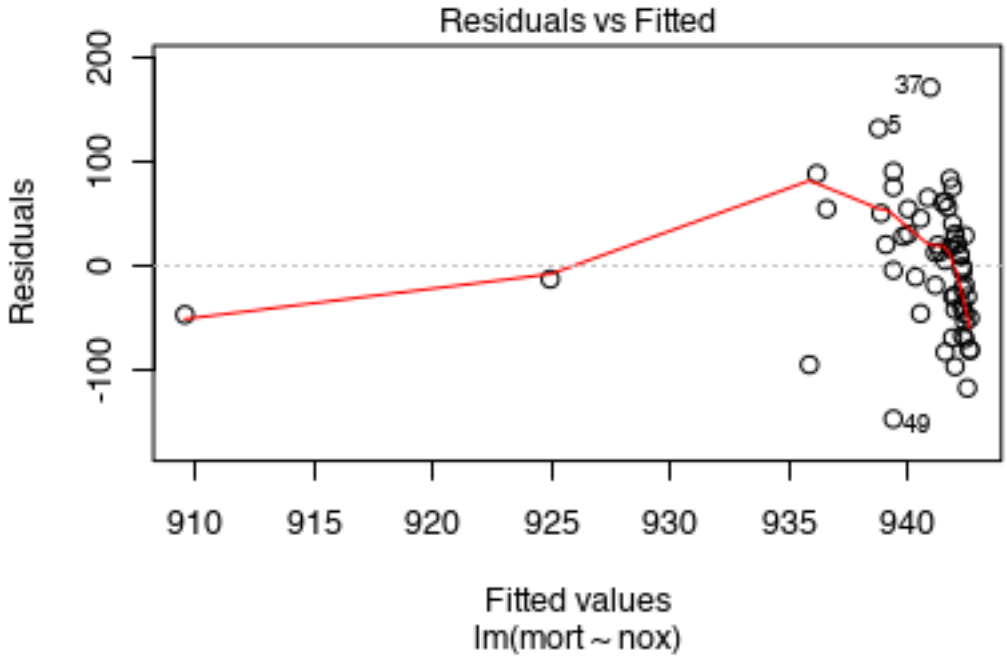


I don't think a linear regression will fit nox and mortality data well because of the outlier. The trend of the scatter plot also looks like a curve line rather than a straight line. However, if we introduce the variable of precipitation and humidity, it will improve the linear model. The first residual plot is spreaded out evenly on both sides of 0 line but cluster around the 950 value.The second residual plot is clustered to the right.

```
lfit <- lm(mort ~ nox + so2 + hc + humid + prec, data = pollution)
lfit1 <- lm(mort ~ nox , data = pollution)
plot(lfit, which =1 )
```




```
plot(lfit1, which =1 )
```




```
coefficients(lfit)
```

```
## (Intercept)         nox         so2          hc       humid        prec
```

```
## 789.8251747   1.4248061   0.3365357  -0.7089897   0.1558816   3.1582449
```

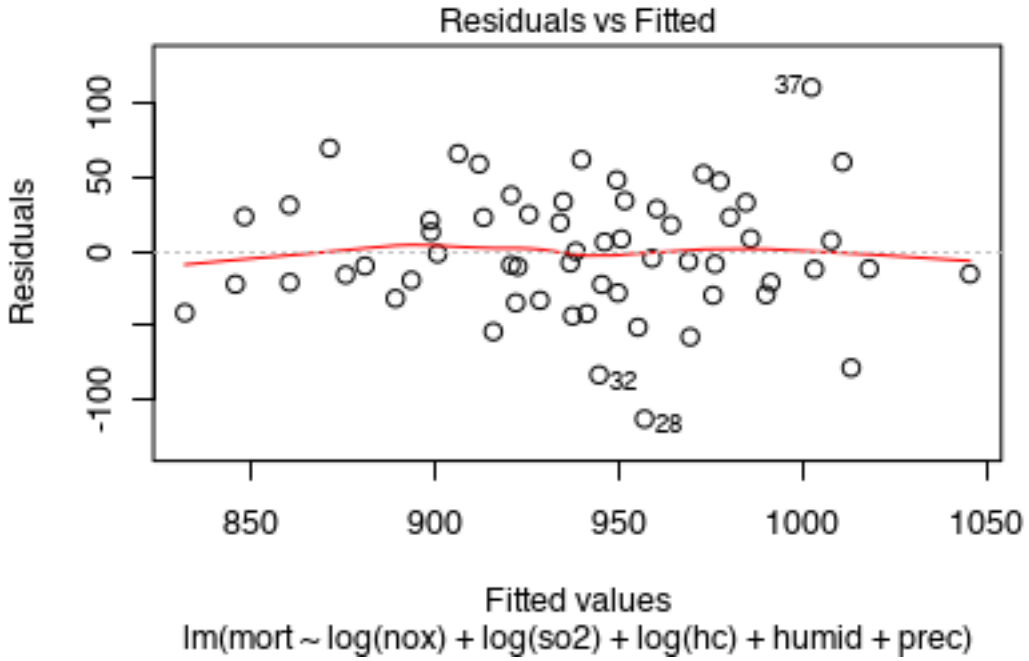
```
coefficients(lfit1)
```

```
## (Intercept)         nox
## 942.7114753  -0.1038871
```

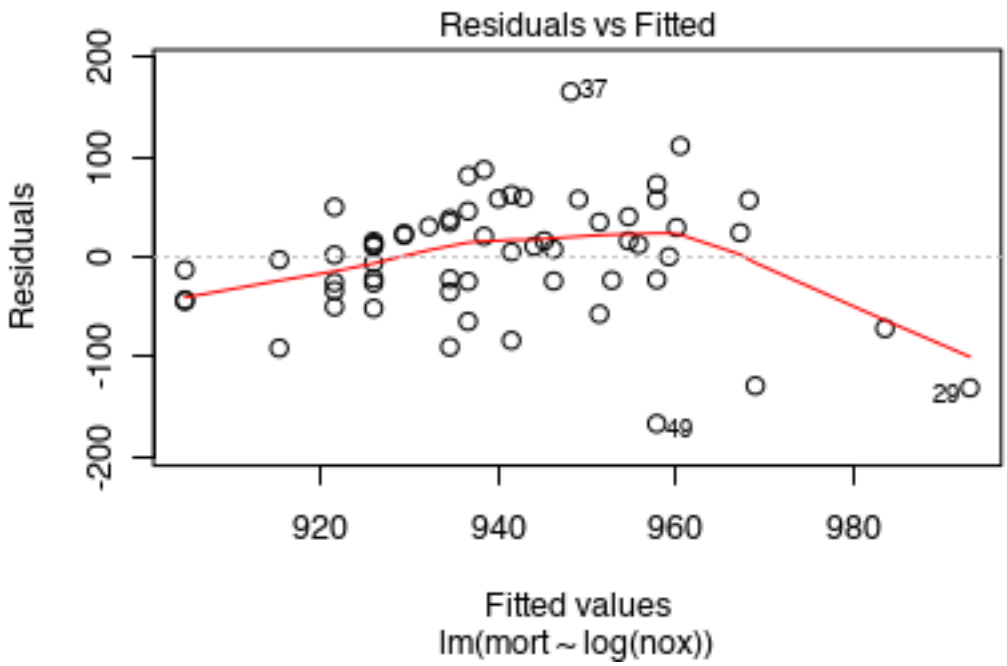
2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

```
# I use log transformation on nox so2 and hc variables because their value are skewed to the right.

newlogfit <- lm(mort ~ log(nox) + log(so2) + log(hc) + humid + prec, data = pollution)
newlogfit1 <- lm(mort ~ log(nox) , data = pollution)
plot(newlogfit, which = 1)
```




```
plot(newlogfit1, which = 1)
```




```
#This residual plot is much better than previous one because the residual points are not clustering aro
```

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

```
coefficients(newlogfit)
```

```
## (Intercept)    log(nox)    log(so2)     log(hc)       humid        prec
## 760.6761887  45.3889072   5.8235169 -24.4728976  -0.4147969   3.9312384
```

```
coefficients(newlogfit1)
```

```
## (Intercept)    log(nox)
##    904.7245     15.3355
```

The coefficient of pollution#nox indicates that every increase of 1 log(nox) level is associated with an increase of 45 mortality unit. The coefficient of pollution#so2 indicates that every increase of 1 log(nox) level is associated with an increase of 5.824 mortality unit. The coefficient of pollution#hc indicates that every increase of 1 log(hc) level is associated with a decrease of 24 mortality unit. The coefficient of humid indicates that every increase of 1% relative humidity level is associated with a decrease of 0.415 mortality unit. The coefficient of prec indicates that every increase of 1 inch precipitation is associated with an increase of 3.931 mortality unit.

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

```
confint(object = newlogfit, level = 0.99)
```

```
##                  0.5 %     99.5 %
## (Intercept) 578.048113 943.304265
## log(nox)     -1.361427  92.139242
## log(so2)     -9.758121  21.405155
## log(hc)     -69.600402  20.654606
## humid        -3.416591   2.586997
```

```
## prec          2.172800   5.689677
```

```
confint(object = newlogfit, parm = "log(nox)", level = 0.9)
```

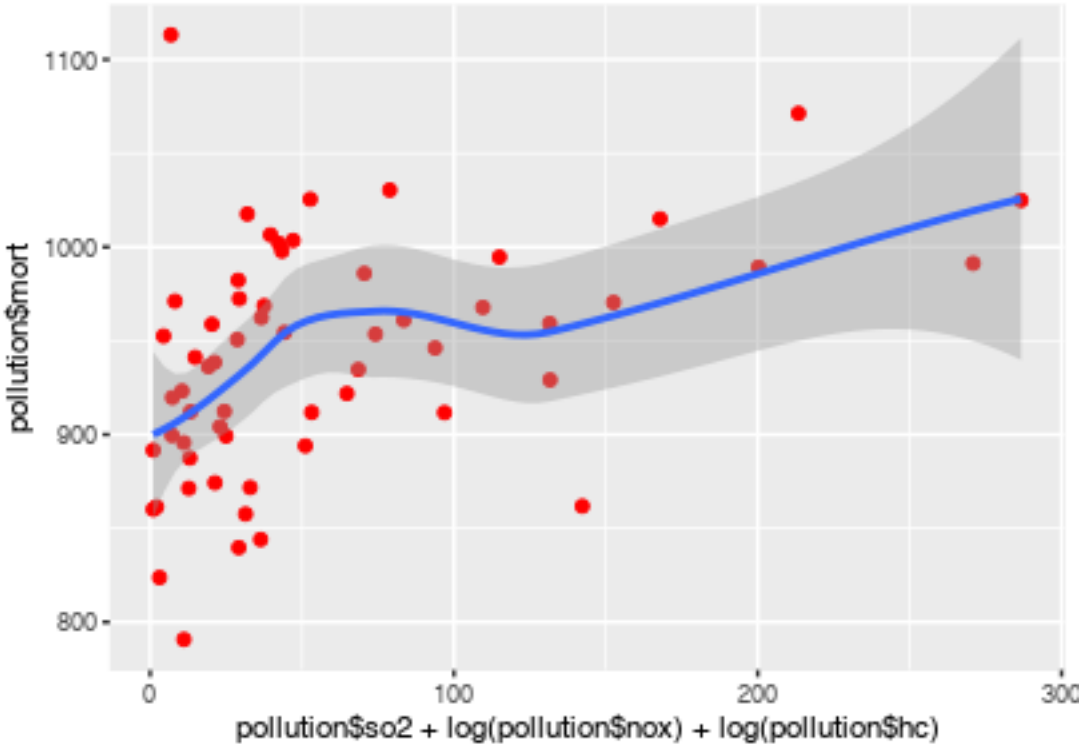
```
##                5 %     95 %
## log(nox) 16.08547 74.69234
```

Let's take log(nox) as an example: The confidence interval gives us two boundaries from -1.3614 to 92.139. What it means is that if we are going to conduct the modelling process many times, 99% chance that the slope coefficient is going to fall between the value of -1.3614 to 92.139. Notice that the confidence interval crosses 0, this implies that it may not have statistics significance. However, if I change the confidence level from 99% to 90%, the ranges changes to 16 to 75. So it looks like it indeed has an effect on the mortality. Likewise, we can summerize the meaning of confidence interval for other coefficients. Notice that the precipitaion level definately has an important effect on the mortality as it ranges from 2.17 to 5.69 positive value.

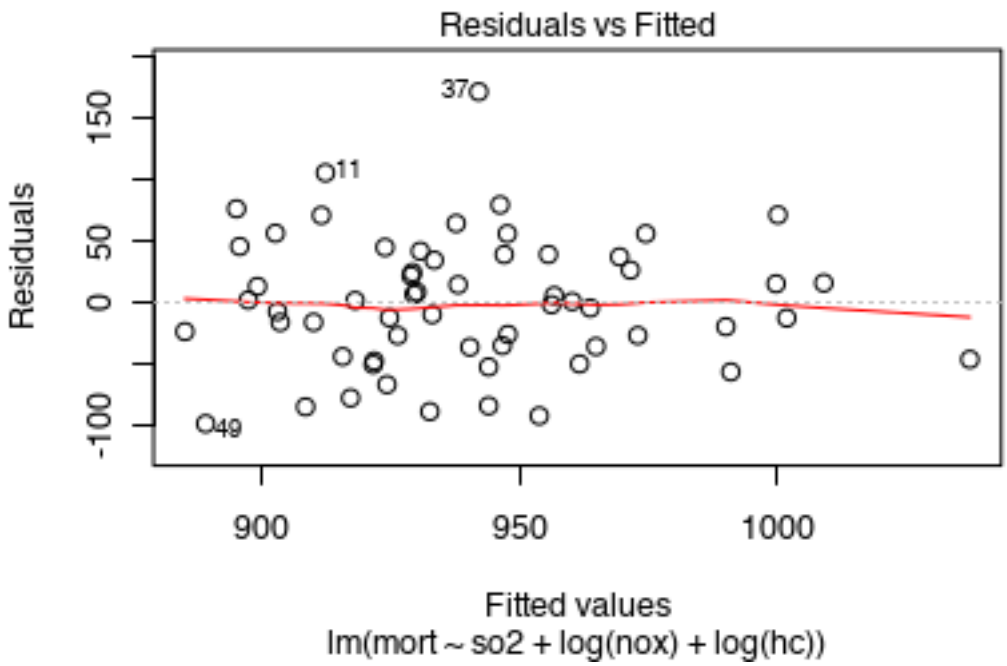
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
# I take a log of nox level because of some extremlly large value.
threepredictors <- lm(mort ~ so2  + log(nox) + log(hc), data = pollution)
# The ggplot uses loess or GAM to capture the nonlinear trend.
ggplot(data = threepredictors) + aes(y = pollution$mort, x = pollution$so2 + log(pollution$nox) + log(p
```

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




```
# I plot the
plot(threepredictors, which =1)
```



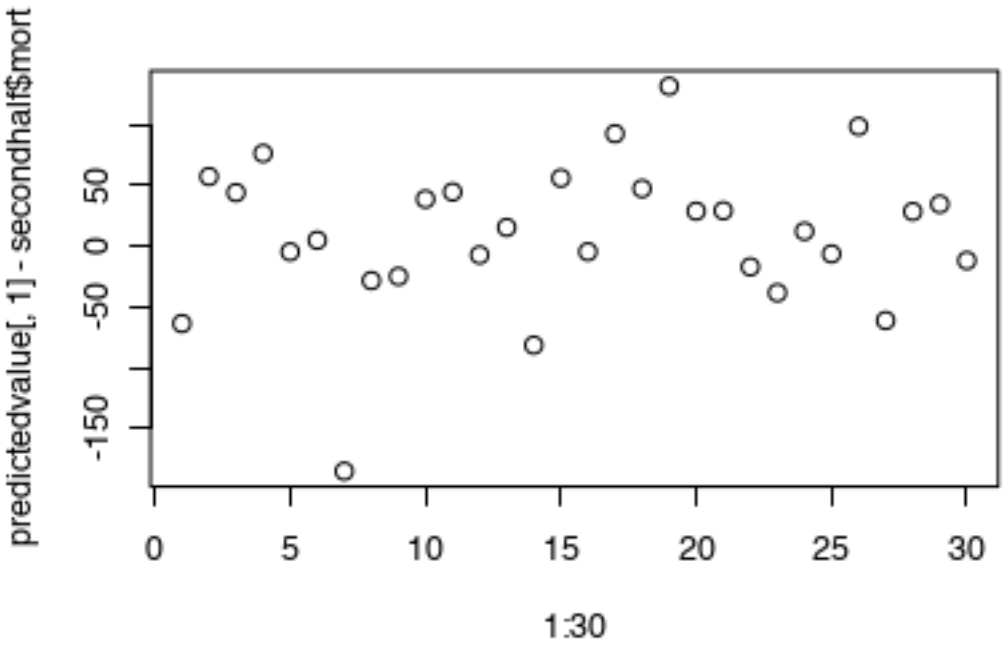

```
coefficients(threepredictors)
```

```
## (Intercept)         so2    log(nox)     log(hc)
## 943.6587585   0.2637597  56.0778719 -53.6761000
```

The first coefficient so2 has a value of 0.264 which indicates that an increase of 1 unit in so2 level is associated with .264 mortality unit. The log(nox) coefficient means that an crease of 1 log(nox) unit is associated with increase of 56 mortality unit. The log(hc) coefficient means that an crease of 1 log(hc) unit is associated with decrease of 53 mortality unit.

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.)

```
firsthalf<-pollution[1:30, ]
secondhalf<-pollution[31:60, ]
likeabove <- lm(mort ~ so2  + log(nox) + log(hc), data = firsthalf)
predictedvalue <- predict (likeabove, newdata=secondhalf, interval="confidence", level=0.95)
# This is the plot for the difference between predicted value and real data
plot(y = predictedvalue[ ,1]-secondhalf$mort,x = 1:30)
```



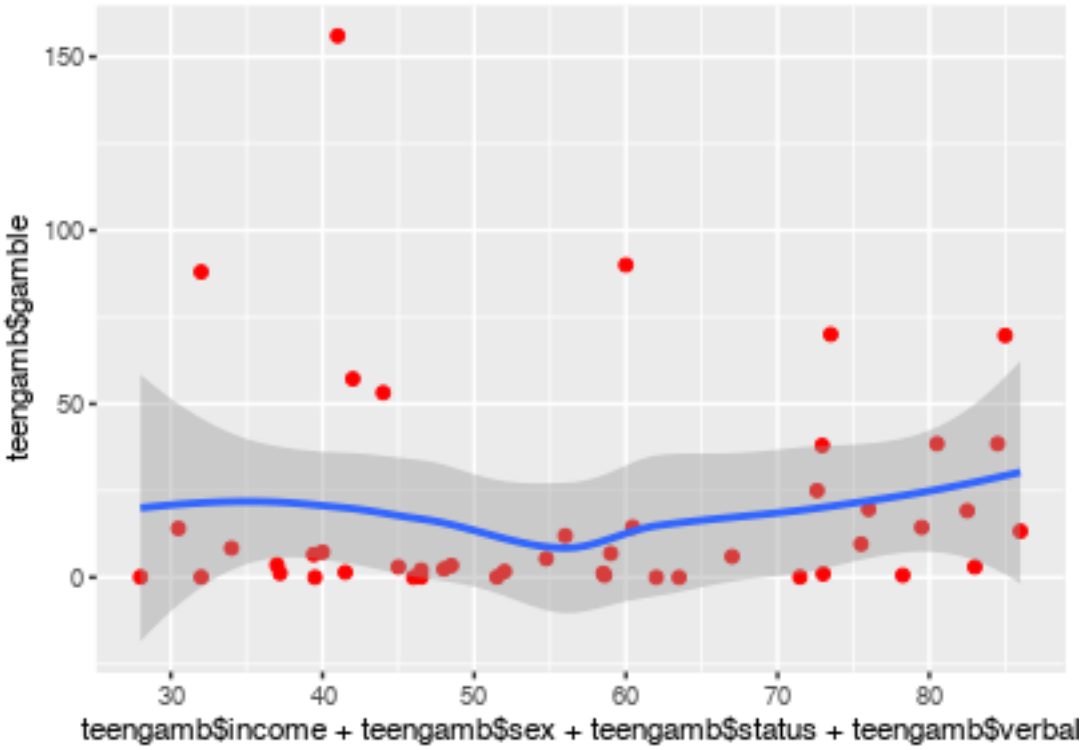
Study of teenage gambling in Britain

```
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.

```
#
aveverbal<-teengamb$verbal-mean(teengamb$verbal)
avestatus<-teengamb$status-mean(teengamb$status)
aveincome<-teengamb$income-mean(teengamb$income)
gam<-lm(gamble~aveincome + sex  + avestatus + aveverbal,data=teengamb)
ngam<-lm(gamble~income+sex+status+verbal,data = teengamb)
ggplot(data = gam) + aes(y = teengamb$gamble, x = teengamb$income+teengamb$sex  +teengamb$status +teenga
```

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




```
coefficients(gam)
```

```
##  (Intercept)    aveincome          sex    avestatus    aveverbal
##  28.24251642   4.96197922 -22.11833009   0.05223384  -2.95949350
```

The meaning of each coefficient is aveincome:every extra pound per week income is assosiated with 4.96 pounds extra expenditure on gambling each year. teengamb$sex: If the person is a girl, he is predicted to spend 22.118 pounds less each year on gambling. avestatus:Every extra 10 socioeconomic status scores the kid has, he/she is predicted to spend 0.0522 pounds more each year. aveverbal: Every extra score the kid has on verbal, he/she is predicted to spend 2.96 pounds less each year.

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

```
confint(object = gam, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept)  18.7827668 37.7022661
## aveincome     2.8926538  7.0313047
## sex         -38.6890301 -5.5476301
## avestatus    -0.5150722  0.6195399
## aveverbal    -7.3430703  1.4240833
```

For the first two coefficients, they have statistical significance since the interval does not cross 0. The other coefficients, they cross the 0, especially avestatus. So they probably don't have statistical significance.

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?

```
mean(teengamb$income)
```

```
## [1] 4.641915
```

```r
mean(teengamb$status)
```

```
## [1] 45.23404
```

```r
mean(teengamb$verbal)
```

```
## [1] 6.659574
```

```r
ave<-data.frame(income = mean(teengamb$income), status=mean(teengamb$status), verbal = mean(teengamb$ver
max<-data.frame(income = 15, status=75, verbal = 10, sex = 0)
predict(ngam, newdata = (ave), interval = 'prediction')
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

```r
predict(ngam, newdata = (max), interval = 'prediction')
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

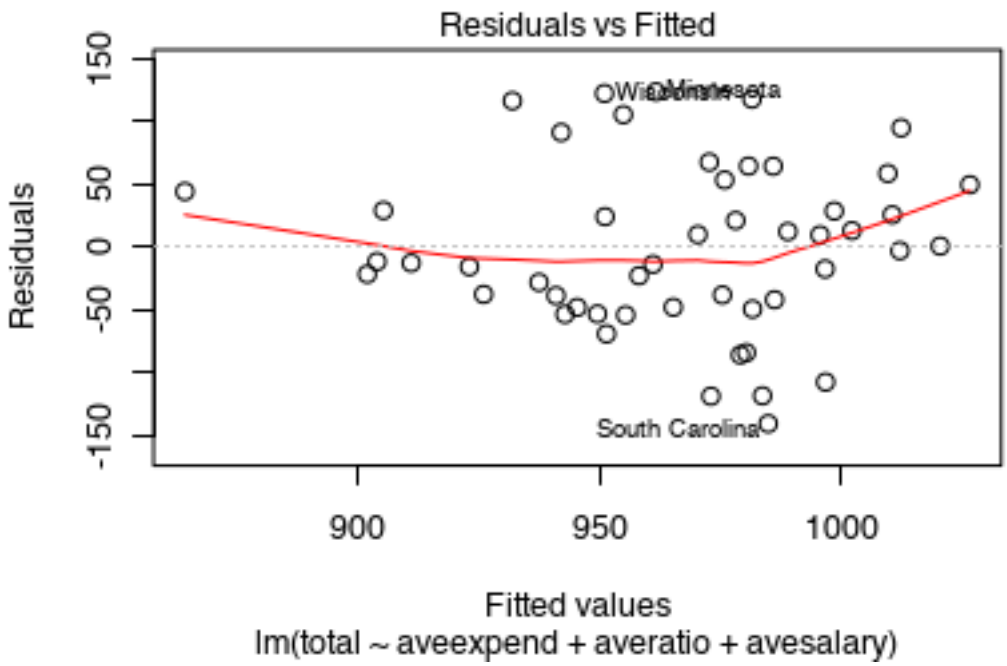
The one with maximum value is wider because the standard error is not the same, actually larger. length=2*s/sqrt(n)*t

School expenditure and test scores from USA in 1994-95

```r
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
aveexpend<-sat$expend-mean(sat$expend)
averatio<-sat$ratio-mean(sat$ratio)
avesalary<-sat$salary-mean(sat$salary)
sat1<-lm(total~aveexpend+averatio+avesalary, data = sat)
plot(sat1, which = 1)
```



Coeffi of aveexpend means that every extra thousand of dollars is associated with increase of 16 points in total SAT test score. Coeffi of averatio means that every extra 1 unit of ratio pupil/teacher is associated with an increase of 6.33 points in SAT test. Coeffi of avesalary means that every extra thousand of dollars the teacher's salary is associated with an decrease of 8.82 points in SAT test.

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

```
confint(object = sat1, level = 0.98)
```

```
##                    1 %       99 %
## (Intercept) 942.519313 989.320687
## aveexpend   -36.675540  69.613271
## averatio     -9.437308  22.097842
## avesalary   -20.142788   2.497524
```

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?

```
sat2<-lm(total~takers+aveexpend+averatio+avesalary, data = sat)
summary(sat2)
```

```
##
## Call:
## lm(formula = total ~ takers + aveexpend + averatio + avesalary,
##     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) 1068.2739     9.3704 114.005  < 2e-16 ***
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## aveexpend      4.4626    10.5465   0.423    0.674
## averatio      -3.6242     3.2154  -1.127    0.266
## avesalary      1.6379     2.3872   0.686    0.496
## ---
## 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
```

```
summary(sat1)
```

```
##
## Call:
## lm(formula = total ~ aveexpend + averatio + avesalary, 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 ***
## aveexpend     16.469     22.050   0.747   0.4589
## averatio       6.330      6.542   0.968   0.3383
## avesalary     -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
```

I think the one with taker works better because the p-value is smaller. The adjusted R- squared is 0.809 considering SAT1 compared to R- squared in SAT2 0.158

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, $D_i - R_i$

Advantages: We know the concrete number of differences and we can add each district up and find out the total differences. Disadvantages: We loose track of the ratio of these two values.

- The ratio, D_i/R_i

Advantages: We know the ratio of two candidates on each district. Disadvantages: We can not calculate the total ratio difference.

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

Advantages: The value of this variable is not too big and easy to compare from district to another Disadvantages: It might be hard to interprete.

- The relative proportion, $D_i/(D_i + R_i)$. Advantages: We can easily see the democrate portion from the value. Disadvantages: Again, we can not calculate the total ratio. Also, it is hard to intepret the republic 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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?

For x1=x-10, the corresponding values of $\hat{\alpha}^\star$, $\hat{\beta}^\star$, $\hat{\sigma}^\star$, and $r^\star$ are

10, 0.9, 2, 0.3

for x2=10x, the corresponding values are

1, 0.09, 0.2, 0.3

for x3=10(x-1), the corresponding values are

1.9, 0.09, 0.2, 0.3

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 $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$?

for y1=y+10 the corresponding values are

11, 0.9, 2, 0.3

for y2=5y, the corresponding values are

5, 4.5, 10, 0.3

for y3=5(y+2), the corresponding values are 15, 4.5, 10, 0.3

3. In general, how are the results of a simple regression analysis affected by linear transformations of y and x?

Linear transformations will not change the value of r. Adding or subtracting value from x or y will not affect the slope and standard deviation. Multiple and deviding value of x or y will not affect the intercept.

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 $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star)$.

The standard error of b is 0.003 The t value of b is 0.09/0.003=30

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 $t_0^{\star\star} = \hat{\beta}^{\star\star}/SE(\hat{\beta}^{\star\star})$.

The standard error of b is 5*0.03=0.15 The t value of b is 0.9*5/0.15=30

6. In general, how are the hypothesis tests and confidence intervals for β affected by linear transformations of y and x?

If we multiple or devide the value of x or y, the confidence interval will change becasue the standard error changes. However, adding or subtracting values from x or y will NOT change the confidence interval.

From both examples, we can see that the T value of b does not change. Therefore, the hypothesis test does not change under linear transformation.

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.