Homework 02

SHIYU ZHANG

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.

```
library(foreign)
library(arm)
library(ggplot2)

data.new <- read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/earnings/heights.dta")

# remove all the N/A
data.new <- data.new[complete.cases(data.new), ]

# label the sex variable (1 = male, 2 = female)
data.new$sex <- factor(data.new$sex, labels=c("male", "female"))

# remove observations where yearbn > 90
data.new <- data.new[data.new$yearbn <= 90,]

# change the scale of earnings to make the data more readable
data.new$earn <- data.new$earn / 1000

summary(data.new)
```

```
##       earn           height1        height2          sex
##  Min.   :  0.00   Min.   :4.000   Min.   : 0.000   male  :519
##  1st Qu.:  6.00   1st Qu.:5.000   1st Qu.: 3.000   female:857
##  Median : 16.02   Median :5.000   Median : 5.000
##  Mean   : 19.99   Mean   :5.129   Mean   : 5.048
##  3rd Qu.: 28.00   3rd Qu.:5.000   3rd Qu.: 8.000
##  Max.   :200.00   Max.   :6.000   Max.   :11.000
```

```
##       race            hisp             ed             yearbn
##  Min.   :1.000   Min.   :1.000   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:12.00   1st Qu.:39.00
##  Median :1.000   Median :2.000   Median :13.00   Median :52.00
##  Mean   :1.169   Mean   :1.942   Mean   :13.35   Mean   :48.78
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:15.00   3rd Qu.:61.00
##  Max.   :9.000   Max.   :2.000   Max.   :18.00   Max.   :72.00
##      height
##  Min.   :58.00
##  1st Qu.:64.00
##  Median :66.00
##  Mean   :66.59
##  3rd Qu.:69.00
##  Max.   :77.00
```

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

```r
# normalise `height` and `earn`
data.new$height <- (data.new$height - mean(data.new$height)) / (2 * sd(data.new$height))

model1<- lm(earn ~ height, data=data.new)

model1
```

```
##
## Call:
## lm(formula = earn ~ height, data = data.new)
##
## Coefficients:
## (Intercept)       height
##       19.99        11.95
```

```r
display(model1)
```

```
## lm(formula = earn ~ height, data = data.new)
##             coef.est coef.se
## (Intercept) 19.99    0.51
## height      11.95    1.02
## ---
## n = 1376, k = 2
## residual sd = 18.85, R-Squared = 0.09
```

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
model2 <- lm(earn ~ sex * ed + height + yearbn, data=data.new)

display(model2)
```

```
## lm(formula = earn ~ sex * ed + height + yearbn, data = data.new)
##              coef.est coef.se
## (Intercept)  -9.22     4.39
## sexfemale     1.89     5.39
## ed            3.36     0.30
## height        3.36     1.31
```

```
## yearbn       -0.18     0.03
## sexfemale:ed -0.99     0.39
## ---
## n = 1376, k = 6
## residual sd = 17.07, R-Squared = 0.26
```

4. Interpret all model coefficients.

```
"Intercept: the intercept represent the average salary for a male of average age
and height which has no education

Sex: female who didn't earn any degree and have average age and height, earn $1,890
(becasue i used the scale of 1000 in previous question) more than males
with similar characteristic.

Education: better education rates corresponds to higher earnings.

Sex : Education: women's average salary is $9,900 less than what a male
individual would have"
```

```
## [1] "Intercept: the intercept represent the average salary for a male of average age \nand height whi
```

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

```
confint(model2,level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept)  -17.8219180 -0.6119841
## sexfemale     -8.6853701 12.4745684
## ed             2.7657697  3.9465424
## height         0.7830980  5.9360500
## yearbn        -0.2353121 -0.1175249
## sexfemale:ed  -1.7547644 -0.2328923
```

```
"it means that we are 95% confident that the intercept lies between 6.577 and 6.964
we are 95% confident that the coefficient of ratio lies between -0.005 and 0.0013
we are 95% confident that the coefficient of log(salary) lies between 0.0478 and 0.1682
we are 95% confident that the coefficient of sat taker lies between -0.0936 and -0.0735"
```

```
## [1] "it means that we are 95% confident that the intercept lies between 6.577 and 6.964\nwe are 95% c
```

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

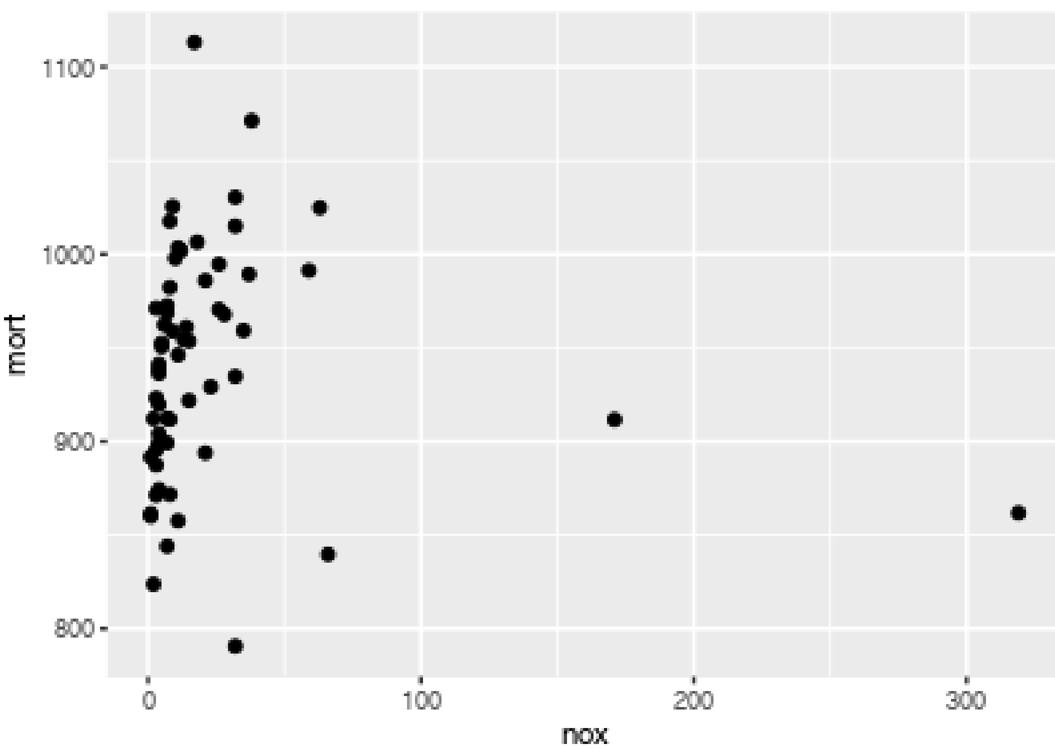
- WWDRK % employed in white collar occupations
- POOR % of families with income < $3000
- HC Relative hydrocarbon pollution potential
- NOX Same for nitric oxides
- SO@ Same for sulphur dioxide
- HUMID Annual average % relative humidity at 1pm
- MORT Total age-adjusted mortality rate per 100,000

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(aes(x=nox,y=mort))
```

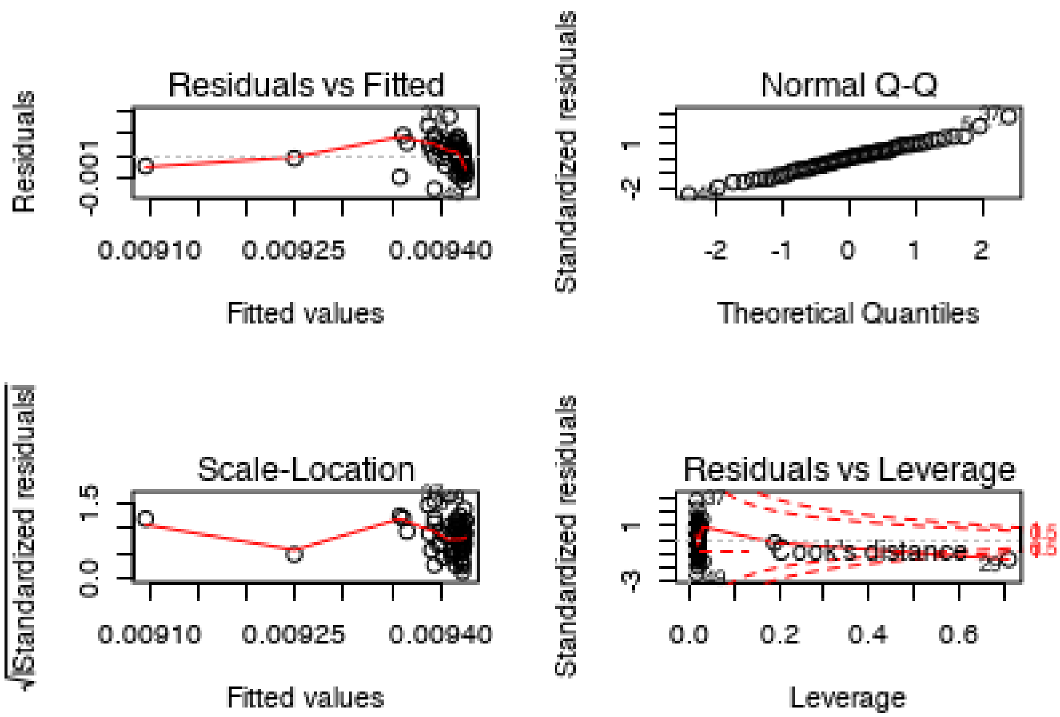



```
# we can see outliners from the graph

pollution$mort <- pollution$mort / 100000

a1<-lm(mort~nox,data=pollution)

a1
```

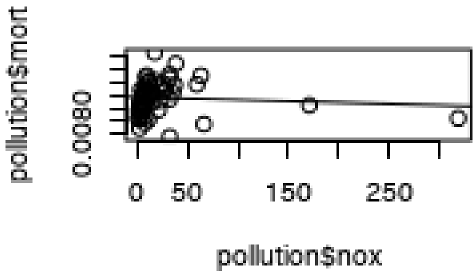
```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
```

```
## Coefficients:
## (Intercept)          nox
##   9.427e-03   -1.039e-06
```

```r
par(mfrow=c(2,2))
plot(a1)
```




```r
plot(y=pollution$mort,x=pollution$nox)
abline(a1)
```

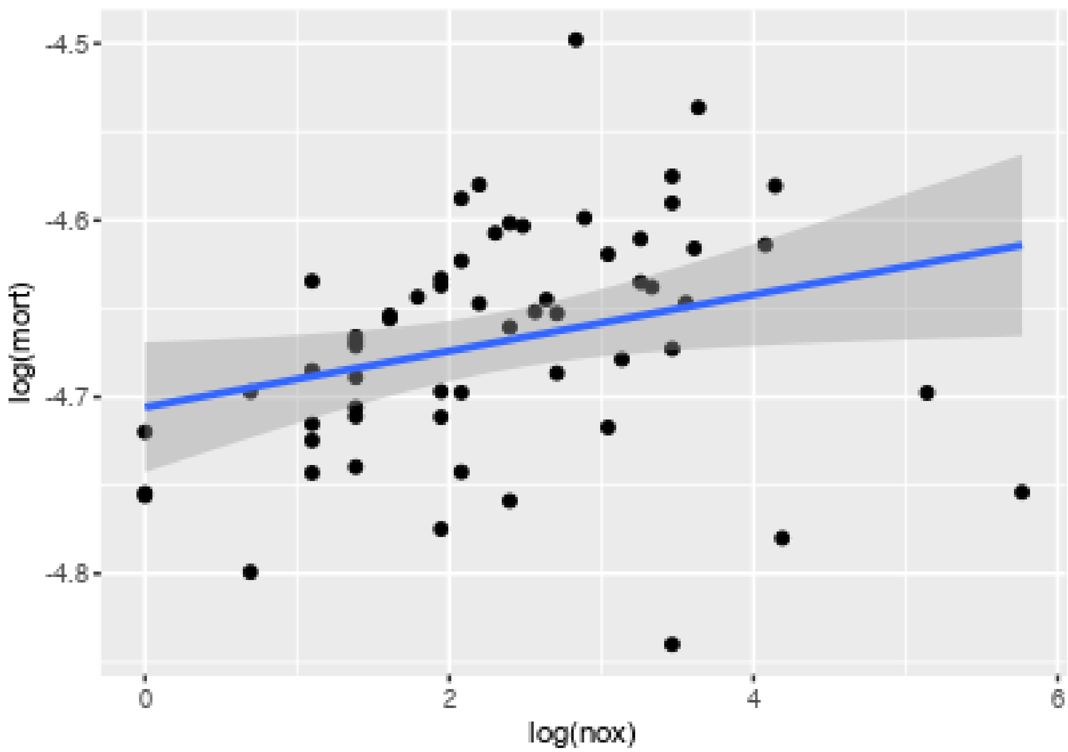


2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

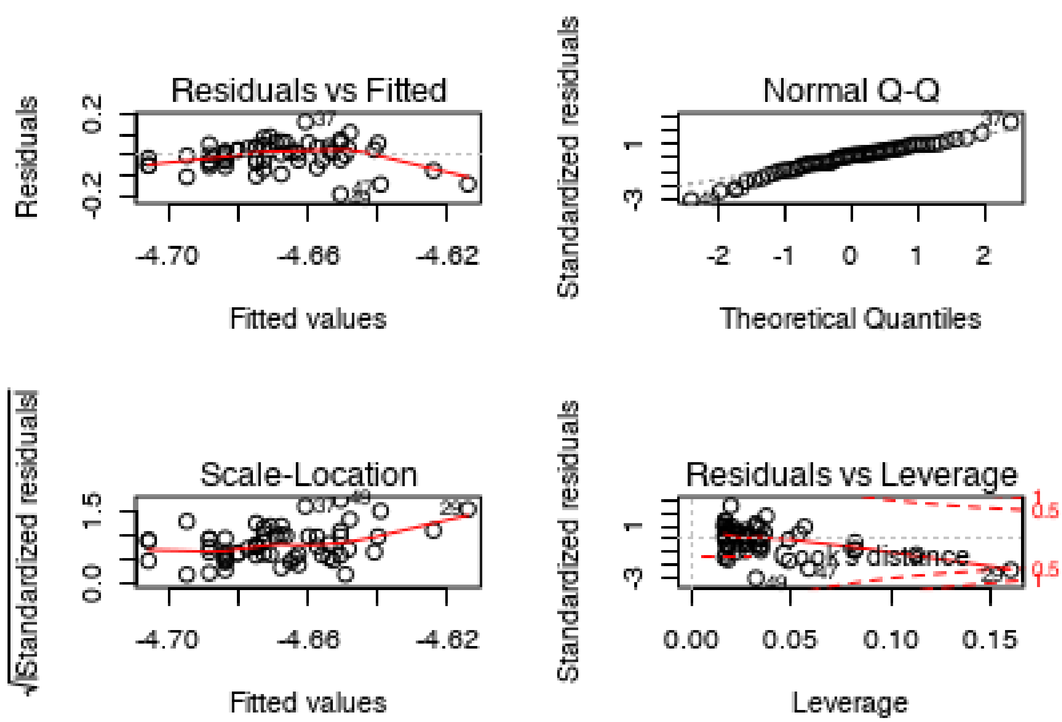
```
# use log to improve the model
a2 <- lm(log(mort) ~ log(nox), data=pollution)

display(a2)
```

```
## lm(formula = log(mort) ~ log(nox), data = pollution)
##             coef.est coef.se
## (Intercept) -4.71     0.02
## log(nox)     0.02     0.01
## ---
## n = 60, k = 2
## residual sd = 0.06, R-Squared = 0.08
```

```
ggplot(data=pollution, aes(x=log(nox), y=log(mort))) + geom_point() +
  stat_smooth(method="lm", formula=y ~ x, se=TRUE)
```




```
# from the new plot ouput, the residuals are evenly distributed around the line.

par(mfrow=c(2,2))
plot(a2)
```



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

```
"Intercept: when is nitric oxides doesnt exist, the overall mortality rate is 6.81%.
log(nox): For each 1% difference in nitric oxides, the predicted difference
in mortality rate is 0.02%."
```

```
## [1] "Intercept: when is nitric oxides doesnt exist, the overall mortality rate is 6.81%.\nlog(nox): F
```

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

```
confint(a2,level=0.99)
```

```
##                    0.5 %      99.5 %
## (Intercept) -4.754620474 -4.65688103
## log(nox)    -0.002876882  0.03466334
```

```
"we are 99% confident that the intercept lies between -4.742 and -4.669
we are 99% confident that the coefficient of log(nox) lies between 0.0017 and 0.0300
"
```

```
## [1] "we are 99% confident that the intercept lies between -4.742 and -4.669\nwe are 99% confident tha
```

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.

```
# check IQR
apply(pollution[, c("hc", "nox", "so2")], FUN=IQR, MARGIN = 2)
```

```
##    hc   nox   so2
## 23.25 19.75 58.00
```

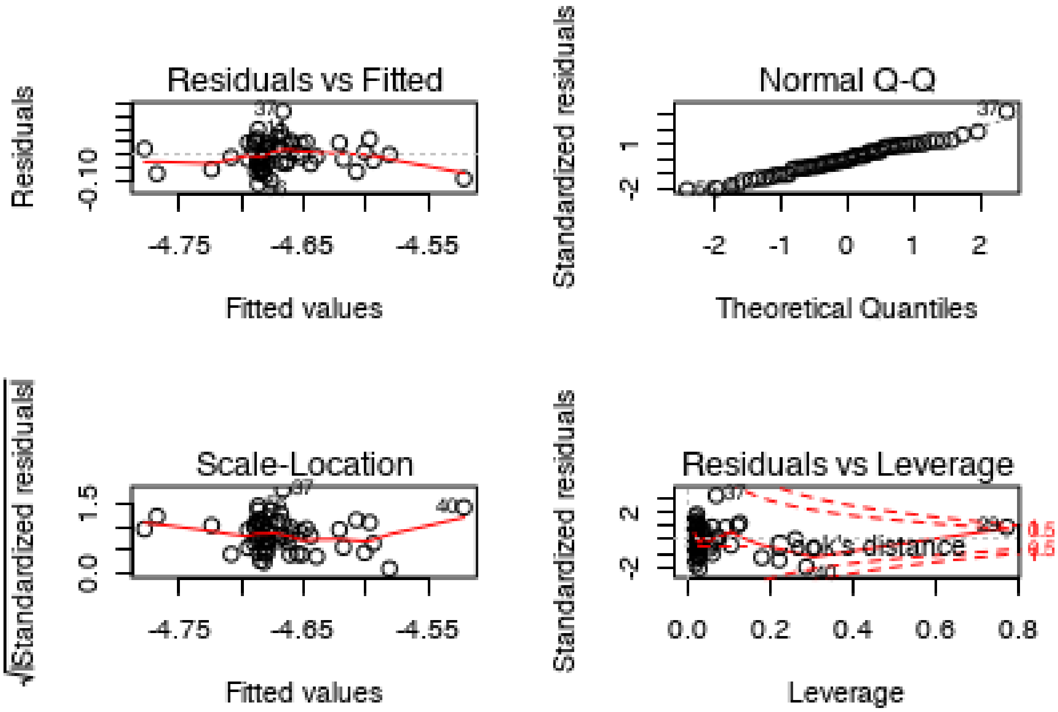
```
# scale predictors
s2 <- function(X) (X - mean(X)) / (2*sd(X))
pollution[, c("hc_new", "nox_new", "so2_new")] <- apply(pollution[, c("hc", "nox", "so2")], FUN=s2, MARG

apply(pollution[, c("hc_new", "nox_new", "so2_new")], FUN=IQR, MARGIN = 2)
```

```
##    hc_new   nox_new   so2_new
## 0.1263894 0.2131297 0.4574820
```

```
a3 <- lm(log(mort) ~ hc_new + nox_new + so2_new, data=pollution)

a3
```

```
##
## Call:
## lm(formula = log(mort) ~ hc_new + nox_new + so2_new, data = pollution)
##
## Coefficients:
## (Intercept)       hc_new      nox_new      so2_new
##    -4.66882     -0.32315      0.29622      0.02643
```

```
par(mfrow=c(2,2))

plot(a3)
```




```
"from the residual plot output, we can say that the model  fits well.

Intercept: The mortality rate for an individual exposed to average levels of nitric oxides,
sulfur dioxide, and hydrocarbons is exp(-39.2076)

hc_new : when one unit increase in hydrocarbons, the mortality rate would decrease by 27%
(becasue it's $exp(-0.32) = 0.726$ times lower)

nox_new: when one unit nitric oxides increases, the mortality rate would be $exp(0.30) = 1.35$
times higher, which is 35% more.

so2_new: one unit difference for sulfur dioxide corresponds to 0.03% increase in mortality rate."
```

```
## [1] "from the residual plot output, we can say that the model  fits well. \n\nIntercept: The mortali
```

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.)

```
first_half <- pollution[1:30, ]
sec_half<- pollution[31:60, ]

a4 <- lm(log(mort) ~ hc_new + nox_new + so2_new, data=pollution)

display(a4)
```

```
## lm(formula = log(mort) ~ hc_new + nox_new + so2_new, data = pollution)
##             coef.est coef.se
## (Intercept) -4.67     0.01
## hc_new      -0.32     0.12
## nox_new      0.30     0.12
```

```
## so2_new      0.03     0.02
## ---
## n = 60, k = 4
## residual sd = 0.05, R-Squared = 0.35
```

```r
predictions <- predict(a4, sec_half)

cbind(predictions=exp(predictions), observed=sec_half$mort)
```

```
##    predictions   observed
## 31 0.009583954 0.01006490
## 32 0.009206867 0.00861439
## 33 0.009615940 0.00929150
## 34 0.009274448 0.00857622
## 35 0.009515885 0.00961009
## 36 0.009263160 0.00923234
## 37 0.009404904 0.01113156
## 38 0.009539248 0.00994648
## 39 0.010041460 0.01015023
## 40 0.010871946 0.00991290
## 41 0.009022669 0.00893991
## 42 0.009279549 0.00938500
## 43 0.009546799 0.00946185
## 44 0.009388724 0.01025502
## 45 0.009263262 0.00874281
## 46 0.009295040 0.00953560
## 47 0.008881718 0.00839709
## 48 0.009393588 0.00911701
## 49 0.008501802 0.00790733
## 50 0.009145171 0.00899264
## 51 0.009299741 0.00904155
## 52 0.009290156 0.00950672
## 53 0.009300596 0.00972464
## 54 0.009223156 0.00912202
## 55 0.009192579 0.00967803
## 56 0.009220133 0.00823764
## 57 0.009404031 0.01003502
## 58 0.009214471 0.00895696
## 59 0.009426718 0.00911817
## 60 0.009458429 0.00954442
```

Study of teenage gambling in Britain

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

```
##    sex status income verbal gamble
## 1    1     51   2.00      8   0.00
## 2    1     28   2.50      8   0.00
## 3    1     37   2.00      6   0.00
## 4    1     28   7.00      4   7.30
## 5    1     65   2.00      8  19.60
## 6    1     61   3.47      6   0.10
```

```
## 7    1      28   5.50      7   1.45
## 8    1      27   6.42      5   6.60
## 9    1      43   2.00      6   1.70
## 10   1      18   6.00      7   0.10
## 11   1      18   3.00      6   0.10
## 12   1      43   4.75      6   5.40
## 13   1      30   2.20      4   1.20
## 14   1      28   2.00      6   3.60
## 15   1      38   3.00      6   2.40
## 16   1      38   1.50      8   3.40
## 17   1      28   9.50      8   0.10
## 18   1      18  10.00      5   8.40
## 19   1      43   4.00      8  12.00
## 20   0      51   3.50      9   0.00
## 21   0      62   3.00      8   1.00
## 22   0      47   2.50      9   1.20
## 23   0      43   3.50      5   0.10
## 24   0      27  10.00      4 156.00
## 25   0      71   6.50      7  38.50
## 26   0      38   1.50      7   2.10
## 27   0      51   5.44      4  14.50
## 28   0      38   1.00      6   3.00
## 29   0      51   0.60      7   0.60
## 30   0      62   5.50      8   9.60
## 31   0      18  12.00      2  88.00
## 32   0      30   7.00      7  53.20
## 33   0      38  15.00      7  90.00
## 34   0      71   2.00     10   3.00
## 35   0      28   1.50      1  14.10
## 36   0      61   4.50      8  70.00
## 37   0      71   2.50      7  38.50
## 38   0      28   8.00      6  57.20
## 39   0      51  10.00      6   6.00
## 40   0      65   1.60      6  25.00
## 41   0      48   2.00      9   6.90
## 42   0      61  15.00      9  69.70
## 43   0      75   3.00      8  13.30
## 44   0      66   3.25      9   0.60
## 45   0      62   4.94      6  38.00
## 46   0      71   1.50      7  14.40
## 47   0      71   2.50      9  19.20
```

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

```r
gamble_log<-log(teengamb$gamble+1)
sex<-teengamb$sex

# center the status data
status_new<-(teengamb$status-mean(teengamb$status))/sd(teengamb$status)
income<-teengamb$income
verbal<-teengamb$verbal
m1<-lm(gamble_log~sex+status_new+income+verbal)
```

```r
display(m1)
```

```
## lm(formula = gamble_log ~ sex + status_new + income + verbal)
##             coef.est coef.se
## (Intercept)  3.07     0.74
## sex         -0.87     0.39
## status_new   0.51     0.23
## income       0.22     0.05
## verbal      -0.26     0.10
## ---
## n = 47, k = 5
## residual sd = 1.09, R-Squared = 0.52
```

```r
"from the model output, the r-squared is 0.52 which means that the model is okay in general.
intercept: a male teenager with  Socioeconomic status score, no income and 0 verbal score spend exp(3.0
pounds per year for gambling.

sex: when Socioeconomic status score is 0, female teenager with no income and 0 verbal score spend exp(
pounds on gambling less than male on the same characteristic.

status_new : one unit increase in Socioeconomic status score, the overall spend on gambling for male
teen increase exp(0.51) pounds per year.

income: when one unit increase in income, male teenager with 0  Socioeconomic status score, no income
and 0 verbal score tends to spend exp(0.22) pounds more on gambling per year.

verbal: one unit increase in verbal score increase correspond to the expenditure on gambling decrease
exp(0.26) pounds per year for a male teen with no  Socioeconomic status score and no income "
```

```
## [1] "from the model output, the r-squared is 0.52 which means that the model is okay in general. \ni
```

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

```r
confint(m1,level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept)  1.56816814  4.56290788
## sex         -1.66365707 -0.07873377
## status_new   0.04660771  0.98330592
## income       0.11668468  0.31460764
## verbal      -0.47128110 -0.05200895
```

```r
"we are 95% confident that the intercept lies between 1.568 and 4.5629
we are 95% confident that the coefficint of sex lies between -1.66 to -0.07
we are 95% confident that the coefficient of Socioeconomic status score lies between 0.04 to 0.983
we are 95% confident that the coefficient of income lies between 0.1166 to 0.3146
we are 95% confident that the coefficient of verbal score lies between -0.47 to -0.052 "
```

```
## [1] "we are 95% confident that the intercept lies between 1.568 and 4.5629\nwe are 95% confident tha
```

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

```r
# p1 stands for the prediction of average score
p1<-predict(m1,newdata = data.frame(sex=0,status_new=0,income=mean(teengamb$income),verbal=mean(teengam
p1
```

```
##        fit       lwr     upr
## 1 2.324105 0.0879996 4.56021
```

```
length1<-p1[3]-p1[2]
length1
```

```
## [1] 4.47221
```

```
#p2 stands for the prediction of max scores
p2<-predict(m1,newdata = data.frame(sex=0,status_new=max(teengamb$status)-mean(teengamb$status),income=
p2
```

```
##        fit      lwr     upr
## 1 19.01196 5.043522 32.9804
```

```
length2<-p2[3]-p2[2]
length2
```

```
## [1] 27.93688
```

```
# the width for a male with max scores is larger than the width of a male with average scores, which
#stands for the standard error of male with max scores is greater
```

School expenditure and test scores from USA in 1994-95

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

```
##                expend ratio salary takers verbal math total
## Alabama         4.405  17.2 31.144      8    491  538  1029
## Alaska          8.963  17.6 47.951     47    445  489   934
## Arizona         4.778  19.3 32.175     27    448  496   944
## Arkansas        4.459  17.1 28.934      6    482  523  1005
## California      4.992  24.0 41.078     45    417  485   902
## Colorado        5.443  18.4 34.571     29    462  518   980
## Connecticut     8.817  14.4 50.045     81    431  477   908
## Delaware        7.030  16.6 39.076     68    429  468   897
## Florida         5.718  19.1 32.588     48    420  469   889
## Georgia         5.193  16.3 32.291     65    406  448   854
## Hawaii          6.078  17.9 38.518     57    407  482   889
## Idaho           4.210  19.1 29.783     15    468  511   979
## Illinois        6.136  17.3 39.431     13    488  560  1048
## Indiana         5.826  17.5 36.785     58    415  467   882
## Iowa            5.483  15.8 31.511      5    516  583  1099
## Kansas          5.817  15.1 34.652      9    503  557  1060
## Kentucky        5.217  17.0 32.257     11    477  522   999
## Louisiana       4.761  16.8 26.461      9    486  535  1021
## Maine           6.428  13.8 31.972     68    427  469   896
## Maryland        7.245  17.0 40.661     64    430  479   909
## Massachusetts   7.287  14.8 40.795     80    430  477   907
## Michigan        6.994  20.1 41.895     11    484  549  1033
## Minnesota       6.000  17.5 35.948      9    506  579  1085
## Mississippi     4.080  17.5 26.818      4    496  540  1036
```

```
## Missouri        5.383  15.5 31.189      9     495  550  1045
## Montana         5.692  16.3 28.785     21     473  536  1009
## Nebraska        5.935  14.5 30.922      9     494  556  1050
## Nevada          5.160  18.7 34.836     30     434  483   917
## New Hampshire   5.859  15.6 34.720     70     444  491   935
## New Jersey      9.774  13.8 46.087     70     420  478   898
## New Mexico      4.586  17.2 28.493     11     485  530  1015
## New York        9.623  15.2 47.612     74     419  473   892
## North Carolina  5.077  16.2 30.793     60     411  454   865
## North Dakota    4.775  15.3 26.327      5     515  592  1107
## Ohio            6.162  16.6 36.802     23     460  515   975
## Oklahoma        4.845  15.5 28.172      9     491  536  1027
## Oregon          6.436  19.9 38.555     51     448  499   947
## Pennsylvania    7.109  17.1 44.510     70     419  461   880
## Rhode Island    7.469  14.7 40.729     70     425  463   888
## South Carolina  4.797  16.4 30.279     58     401  443   844
## South Dakota    4.775  14.4 25.994      5     505  563  1068
## Tennessee       4.388  18.6 32.477     12     497  543  1040
## Texas           5.222  15.7 31.223     47     419  474   893
## Utah            3.656  24.3 29.082      4     513  563  1076
## Vermont         6.750  13.8 35.406     68     429  472   901
## Virginia        5.327  14.6 33.987     65     428  468   896
## Washington      5.906  20.2 36.151     48     443  494   937
## West Virginia   6.107  14.8 31.944     17     448  484   932
## Wisconsin       6.930  15.9 37.746      9     501  572  1073
## Wyoming         6.160  14.9 31.285     10     476  525  1001
```

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.

```
# center ratio data

ratio_center<-sat$ratio-mean(sat$ratio)/sd(sat$ratio)
salary_new<-log(sat$salary)

model1<-lm(log(total)~ratio_center+log(salary),data=sat)
model1
```

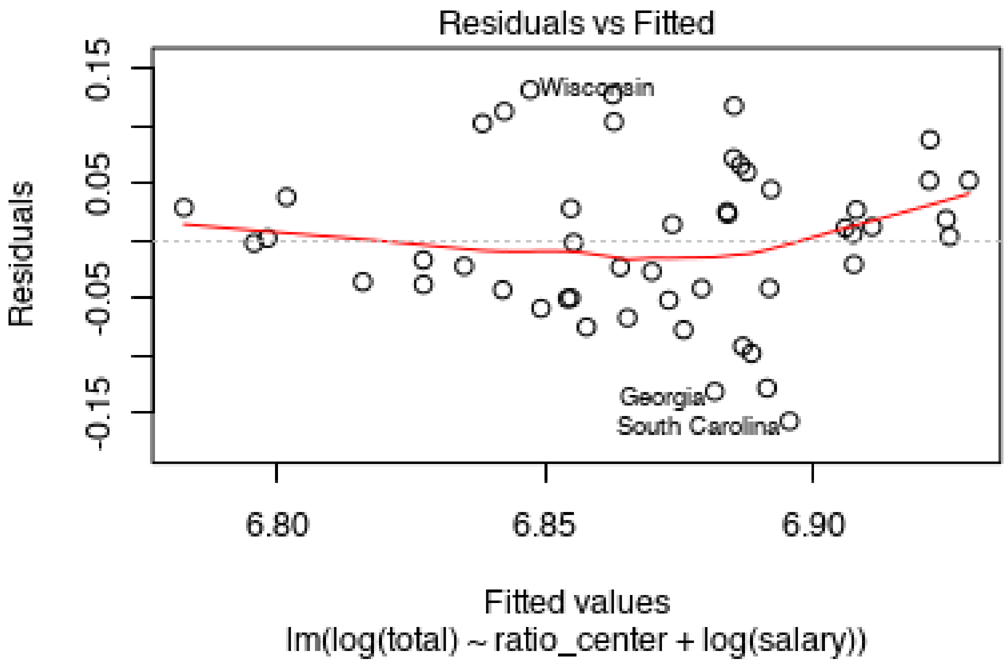
```
##
## Call:
## lm(formula = log(total) ~ ratio_center + log(salary), data = sat)
##
## Coefficients:
##  (Intercept)  ratio_center   log(salary)
##     7.589832      0.003145     -0.211852
```

```
display(model1)
```

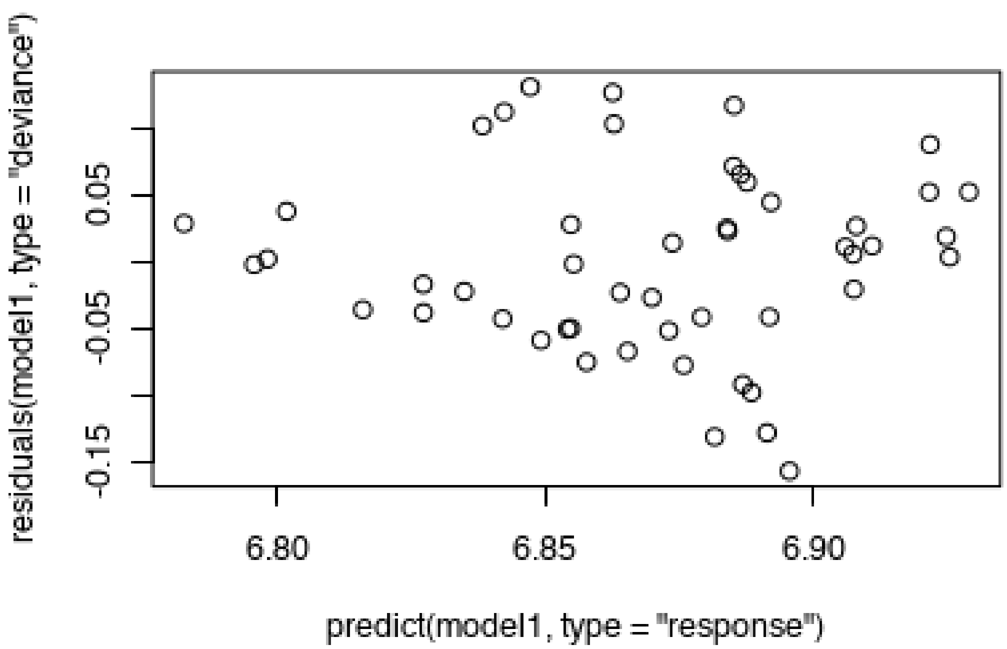
```
## lm(formula = log(total) ~ ratio_center + log(salary), data = sat)
##               coef.est coef.se
## (Intercept)    7.59     0.22
## ratio_center   0.00     0.00
## log(salary)   -0.21     0.06
## ---
## n = 50, k = 3
```

```
## residual sd = 0.07, R-Squared = 0.21
```

```r
plot(model1,which=1)
```




```r
plot(predict(model1,type = "response"),residuals(model1,type = "deviance"))
```



```
"When ratio increase 1 unit, the total score will be exp(0.003) times of the original overall score

When log(salary) increase 1 unit, the total score will be exp(-0.212) times of the original overall sco
```

```
## [1] "When ratio increase 1 unit, the total score will be exp(0.003) times of the original overall sc
```

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

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

```
##                        1 %        99 %
## (Intercept)   7.065419225  8.11424391
## ratio_center -0.007458157  0.01374726
## log(salary)  -0.357684974 -0.06601843
```

```
"we are 98% confident that the intercept lies between 7.0654 and 8.114
we are 98% confident that the coefficient of ratio(after center) lies between -0.007 and 0.013
we are 98% confident that the coefficient of log(salary) lies between -0.357 and -0.066
"
```

```
## [1] "we are 98% confident that the intercept lies between 7.0654 and 8.114\nwe are 98% confident tha
```

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?

```
model2<-lm(log(total)~ratio_center+log(salary)+log(sat$taker),data=sat)
display(model2)
```

```
## lm(formula = log(total) ~ ratio_center + log(salary) + log(sat$taker),
##     data = sat)
##                coef.est coef.se
## (Intercept)     6.77     0.10
## ratio_center    0.00     0.00
## log(salary)     0.11     0.03
## log(sat$taker) -0.08     0.00
## ---
## n = 50, k = 4
## residual sd = 0.03, R-Squared = 0.89
```

```
 "the r-squared is 0.89 which means that 89% of the realtionsip between the dependent variables
can be explained by the model, the r-squared of the new model is greater than the previous model (0.21)
means that this model is much better than the previous one."
```

```
## [1] "the r-squared is 0.89 which means that 89% of the realtionsip between the dependent variables\n
```

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 : it shows the difference between the two parties found and know how much democrats less/ more than the republicans. disadvantage : it only shows the difference, the base of the two parties is unknown (for example, the difference between 2millions and 1million is the same as the difference between 4 millions and 3 millions)
- The ratio, D_i/R_i

advantages : it shows how much multiple the two parties differ. disadvantages : the same as the previous question, it only shows the ratio, but not the nunber base.

- The difference on the logarithmic scale, $logD_i - logR_i$
- The relative proportion, $D_i/(D_i + R_i)$. it shows the relative proportion of the two parties when compare together. however, it fails to show individual's advantages and disadvantages compare to the other competitor.

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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$?

$x^\star = x - 10$, $\hat{\alpha}^\star = \hat{\alpha} + 10\hat{\beta} = 10$, $\hat{\beta}^\star = \hat{\beta} = 0.9$, $r^\star = r = 0.3$,$\hat{\sigma}^\star = \hat{\sigma} = 2$.

$x^\star = 10x$, $\hat{\alpha}^\star = \hat{\alpha} = 1$, $\hat{\beta}^\star = \frac{\hat{\beta}}{10} = 0.09$, $r^\star = r = 0.3$,$\hat{\sigma}^\star = \hat{\sigma} = 2$.

$x^\star = 10(x - 1)$, $\hat{\alpha}^\star = \hat{\alpha} + \hat{\beta} = 1.9$, $\hat{\beta}^\star = \frac{\hat{\beta}}{10} = 0.09$, $r^\star = r = 0.3$,$\hat{\sigma}^\star = \hat{\sigma} = 2$.

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

(1)$y^{\star\star} = y + 10$, $\hat{\alpha}^{\star\star} = \hat{\alpha} + 10 = 11$,

$\hat{\beta}^{\star\star} = \hat{\beta} = 0.9$, $r^\star = r = 0.3$

$\hat{\sigma}^{\star\star} = \hat{\sigma} = 2$.

(2)$y^{\star\star} = 5y$, $\hat{\alpha}^{\star\star} = 5\hat{\alpha} = 5$,

$\hat{\beta}^{\star\star} = 5\hat{\beta} = 4.5$, $r^\star = r = 0.3$, $\hat{\sigma}^{\star\star} = 5\hat{\sigma} = 10$.

(3)$y^{\star\star} = 5(y + 2)$, $\hat{\alpha}^{\star\star} = 5(\hat{\alpha} + 2) = 15$,

$\hat{\beta}^{\star\star} = 5\hat{\beta} = 4.5$, $r^\star = r = 0.3$, $\hat{\sigma}^{\star\star} = 5\hat{\sigma} = 10$.

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

when x plus or minus a constant number, it only changes the intercept. when changes the scale of x, it only change the slope coefficients.

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

$SE(\hat{\beta}^\star) = SE(\hat{\beta}) = 0.03$ $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star) = 0.09/0.03 = 3$.

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

$SE(\hat{\beta}^{\star\star}) = 5 * SE(\hat{\beta}) = 0.15$ and $t_0^{\star\star} = \hat{\beta}^{\star\star}/SE(\hat{\beta}^{\star\star}) = 30$

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

(a) $\frac{\bar{\beta}-\mu_0}{SE(\beta)}\sim t(n-1)$

Confidence Interval is $[\bar{\beta} - t_{\alpha/2} * SE(\beta), \bar{\beta} + t_{\alpha/2} * SE(\beta)]$

if $x^{=}cx$, then $\bar{\beta}^{*} = \bar{\beta}/c$, CI is $[\bar{\beta}/c - t_{\alpha/2} * SE(\beta)/c, \bar{\beta}/c + t_{\alpha/2} * SE(\beta)/c]$

If $y^{=}dy$, then $\bar{\beta}^{*} = \bar{\beta} * d$, CI is $[\bar{\beta} * d - t_{\alpha/2} * SE(\beta) * d, \bar{\beta} * d + t_{\alpha/2} * SE(\beta) * d]$

(b) In hypothesis test, H_0:$\mu = 0$, H_1:$\mu \neq 0$

$T = \frac{\bar{\beta}}{SE(\beta)}\sim t(n-1)$

And if $x^{=}cx$, then $\bar{\beta}^{*} = \bar{\beta}/c$,$T$.

If $y^{=}dy$, then $\bar{\beta}^{*} = \bar{\beta} * d$,$T$

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.