Homework 02

yourname

Septemeber 16, 2018

```
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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.

```
heights <- na.omit(heights)
```

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.mean <- heights$earn - mean(heights$earn)
height.mean <- heights$height- mean(heights$height)
earn.height <- lm(earn~ height.mean, data = heights)
summary(earn.height)
```

```
##
## Call:
## lm(formula = earn ~ height.mean, data = heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30031 -12497  -3215   7474 174659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20014.9      507.7   39.42   <2e-16 ***
## height.mean   1563.1      133.4   11.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18850 on 1377 degrees of freedom
## Multiple R-squared:  0.09061,    Adjusted R-squared:  0.08995
## F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

```
new.heights <- data.frame(earn = earn.mean,height = height.mean)
coefficients(earn.height)
```

```
## (Intercept) height.mean
##   20014.859    1563.138
```

```
new.earn.height <- lm(earn ~ height, data = new.heights)
coefficients(new.earn.height)
```

```
##  (Intercept)       height
## 4.199057e-13 1.563138e+03
```

```
summary(new.earn.height)
```

```
##
## Call:
```

```
## lm(formula = earn ~ height, data = new.heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30031 -12497  -3215   7474 174659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.199e-13  5.077e+02    0.00        1
## height      1.563e+03  1.334e+02   11.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18850 on 1377 degrees of freedom
## Multiple R-squared:  0.09061,    Adjusted R-squared:  0.08995
## F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

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

4. Interpret all model coefficients.

```
# sex is either 1 or 2, sex -1 means 0 is male and 1 is female

earning.model1<- lm(earn ~ height + (sex-1),data = heights)
display(earning.model1)
```

```
## lm(formula = earn ~ height + (sex - 1), data = heights)
##        coef.est  coef.se
## height    571.62     22.80
## sex    -11123.73    897.83
## ---
## n = 1379, k = 2
## residual sd = 18453.75, R-Squared = 0.57
```

```
earning.model2 <- lm(log(earn+1)~ height + (sex-1), data = heights)
summary(earning.model2)
```

```
##
## Call:
## lm(formula = log(earn + 1) ~ height + (sex - 1), data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1639  -0.1417   0.9945   2.0326   4.6696
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## height  0.15621    0.00404  38.667  < 2e-16 ***
## sex    -1.23907    0.15909  -7.789 1.32e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.27 on 1377 degrees of freedom
## Multiple R-squared:  0.8703, Adjusted R-squared:  0.8701
## F-statistic:  4619 on 2 and 1377 DF,  p-value: < 2.2e-16
```

For earning model 2: The intercept is the predicted log earnings if height and male both equal zero. Because heights are never close to zero, the intercept has no direct interpretation.

. The coefficient for height is the predicted difference in log earnings corresponding to a 1-inch difference in height, if male equals zero. Thus, the estimated predictive difference per inch of height is 16%. The estimate is more than 2 standard errors from zero, indicating that the data are consistent with a positive predictive difference also.

. The coefficient for sex is the predicted difference in log earnings between women and men, if height equals 0. Heights are never close to zero, and so the coefficient for male has no direct interpretation in this model.

```
z.height <- (heights$height-mean(heights$height))/(2*sd(heights$height))
z.sex <- ((heights$sex-1)-mean(heights$sex-1))/(2*sd(heights$sex-1))
earning.model3 <- lm(earn ~ z.height + z.sex, data = heights)
summary(earning.model3)
```

```
##
## Call:
## lm(formula = earn ~ z.height + z.sex, data = heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30553 -12448  -3243   7451 171098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20014.9      497.1  40.262  < 2e-16 ***
## z.height      4190.7     1404.9   2.983  0.00291 **
## z.sex       -10913.2     1404.9  -7.768 1.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18460 on 1376 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1275
## F-statistic: 101.7 on 2 and 1376 DF,  p-value: < 2.2e-16
```

For earning model 3: tHE INTERCEPT means 20014.86 when z.height and z.sex are both 0. It means standarizing value can give more true value about prediction. One more inch means 4190 increase in earning female has 10913.16 more earning than male's.

```
earning.model4 <- lm(log(earn+1)~ z.height + sex,data = heights)
display(earning.model4)
```

```
## lm(formula = log(earn + 1) ~ z.height + sex, data = heights)
##             coef.est coef.se
## (Intercept) 11.04     0.43
## z.height     0.71     0.25
```

```
## sex         -1.63     0.26
## ---
## n = 1379, k = 3
## residual sd = 3.27, R-Squared = 0.10
```

For earning medel 4 The intercept is the predicted log earnings if z. height and male both equal zero. Thus, a 66.9-inch tall woman is predicted to have log earnings of 11.04, and thus earnings of exp(11.04) = 62317. . The coefficient for z.height is the predicted difference in log earnings corresponding to a 1 standard-deviation difference in height, if male equals zero. Thus, the estimated predictive difference for a 3.8-inch increase in height is 71% for women.

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

```
z.heights <- data.frame(earn = heights$earn, height = z.height,sex = z.sex)
confint(earning.model3, level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept)  19039.670 20990.047
## z.height      1434.668  6946.776
## z.sex       -13669.212 -8157.104
```

Analysis of mortality rates and various environmental factors

The folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

Variables, in order:

- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all facilities
- DENS Population per sq. mile in urbanized areas, 1960
- NONW % non-white population in urbanized areas, 1960
- WWDRK % employed in white collar occupations
- POOR % of families with income < $3000
- HC Relative hydrocarbon pollution potential
- NOX Same for nitric oxides
- SO@ Same for sulphur dioxide
- HUMID Annual average % relative humidity at 1pm
- MORT Total age-adjusted mortality rate per 100,000

For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations in regression.

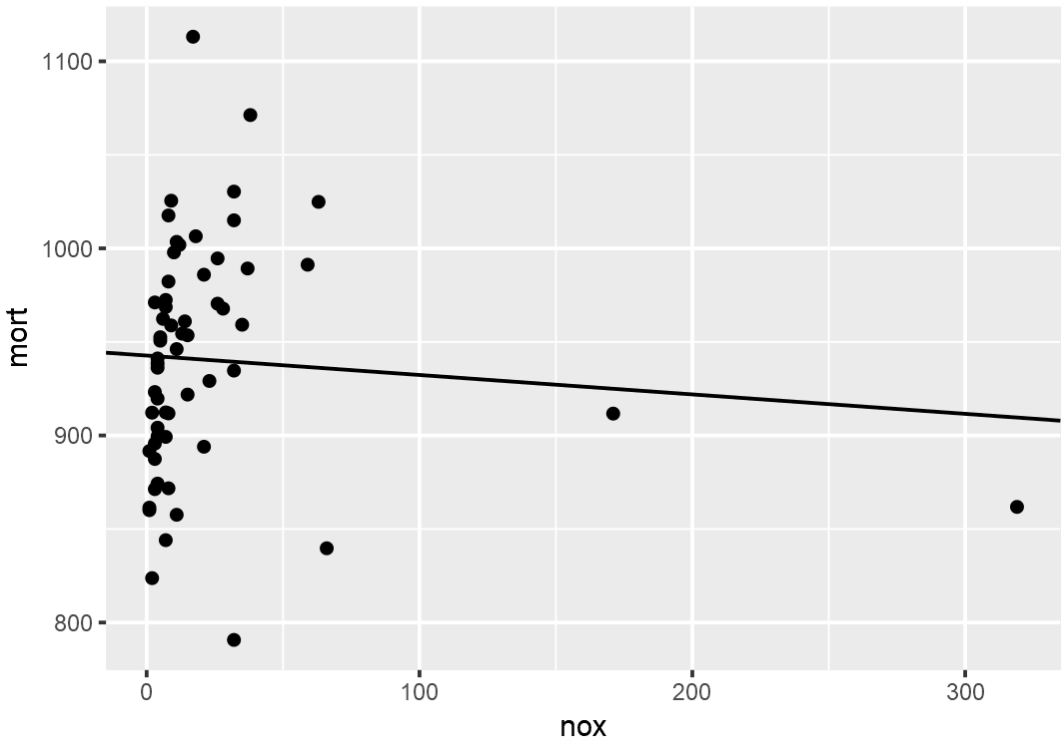
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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

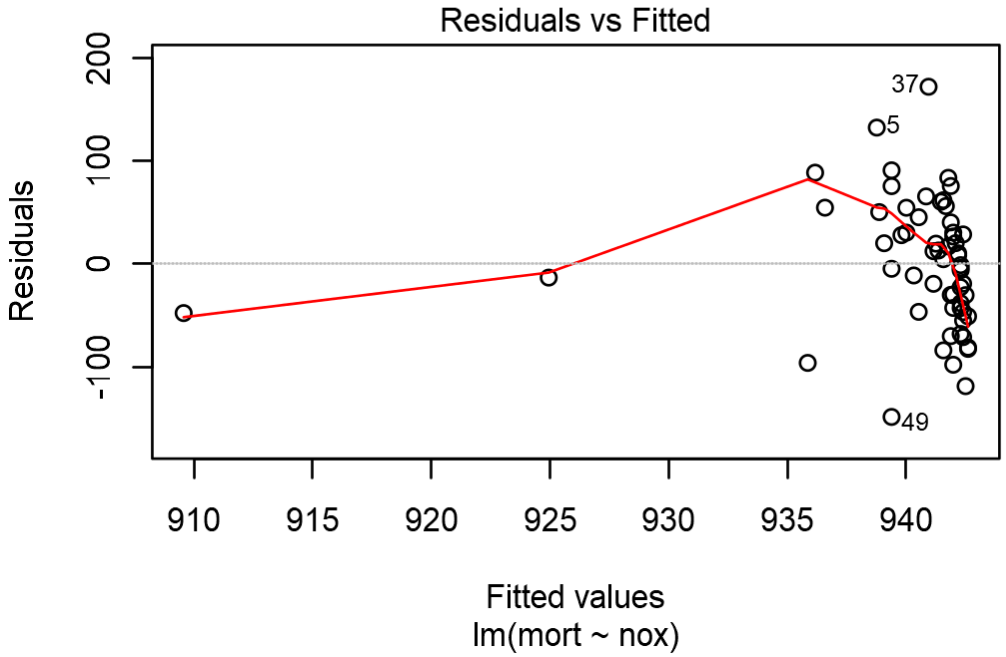
```
nox.mort <- lm(mort~nox,data = pollution)
summary(nox.mort)
```

```
##
## Call:
## lm(formula = mort ~ nox, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.654  -43.710    1.751   41.663  172.211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 942.7115     9.0034 104.706   <2e-16 ***
## nox          -0.1039     0.1758  -0.591    0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.55 on 58 degrees of freedom
## Multiple R-squared:  0.005987,   Adjusted R-squared:  -0.01115
## F-statistic: 0.3494 on 1 and 58 DF,  p-value: 0.5568
```

```
ggplot(data = pollution)+
  geom_point(mapping = aes(x = nox, y = mort))+
  geom_abline(mapping = aes(x = nox, y= mort),slope = -0.1039,intercept = 942.7115 )
```



```
plot(nox.mort,which = 1)
```



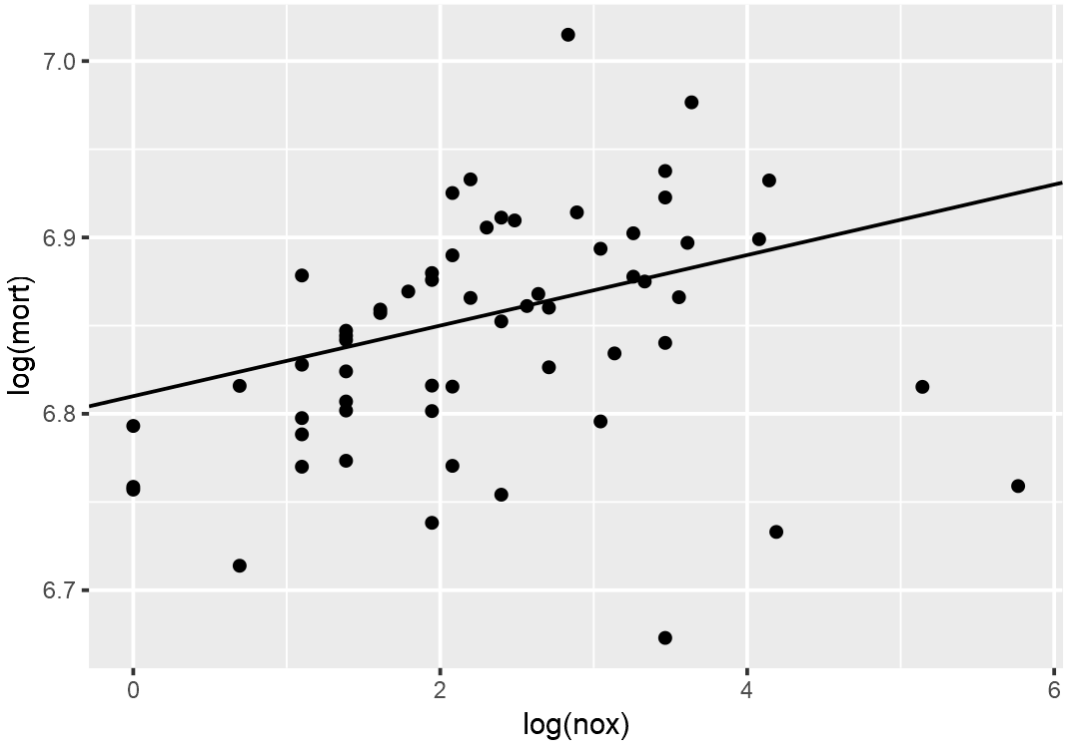
It does not fit that well.

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

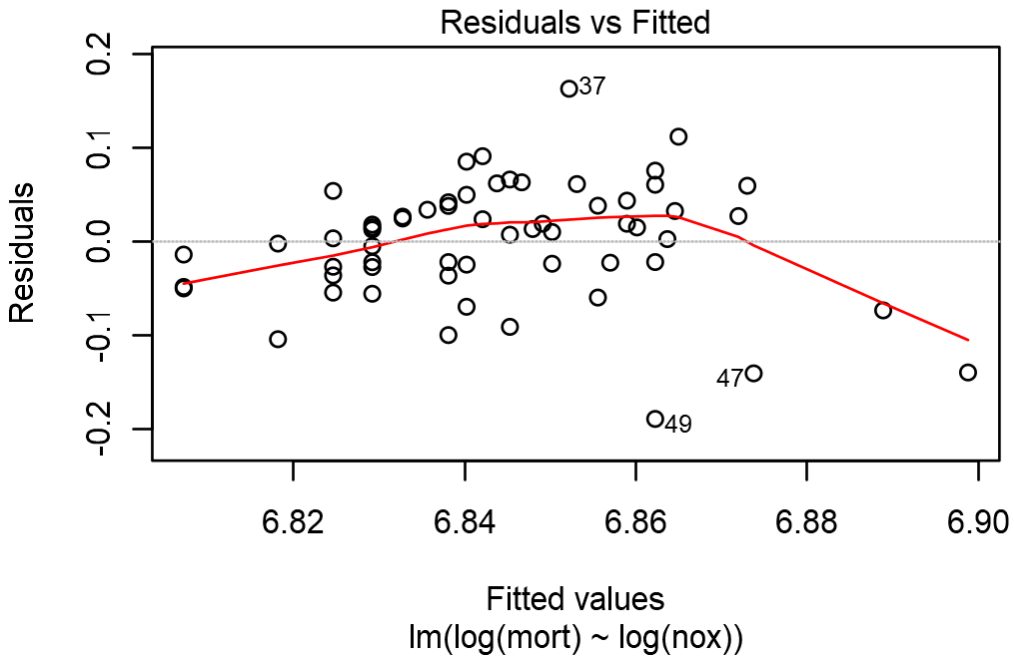
```
log.nox.mort <- lm(log(mort)~log(nox),data = pollution)
display(log.nox.mort)
```

```
## lm(formula = log(mort) ~ log(nox), data = pollution)
##             coef.est coef.se
## (Intercept) 6.81     0.02
## log(nox)    0.02     0.01
## ---
## n = 60, k = 2
## residual sd = 0.06, R-Squared = 0.08
```

```
ggplot(data = pollution)+
  geom_point(mapping = aes(x = log(nox), y = log(mort)))+
  geom_abline(mapping = aes(x = log(nox), y= log(mort)),slope = 0.02,intercept = 6.81)
```




```
plot(log.nox.mort,which = 1)
```



The log make linear regression fit data better 3. Interpret the slope coefficient from the model you chose in 2.

For each 1% difference in nox, the predicted difference in mort is 0.02%.

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

```
confint(object = log.nox.mort,level = 0.99 )
```

```
##                    0.5 %     99.5 %
## (Intercept)  6.758304991 6.85604444
## log(nox)    -0.002876882 0.03466334
```

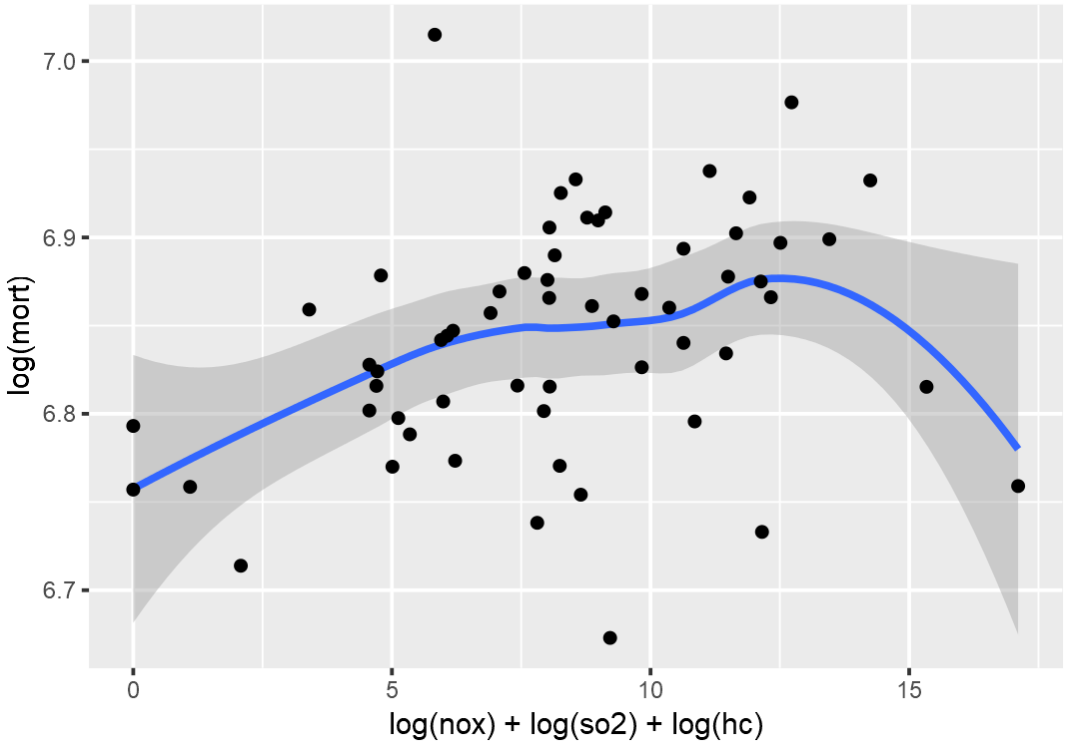
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
#z.so2 <- (pollution$so2 - mean(pollution$so2))/(2*sd(pollution$so2))
nsh.mort<- lm(log(mort)~ log(nox)+ log(so2) +log(hc),data = pollution)
summary(nsh.mort)
```

```
##
## Call:
## lm(formula = log(mort) ~ log(nox) + log(so2) + log(hc), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10874 -0.03574 -0.00218  0.03709  0.20085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.826749   0.022701 300.726  < 2e-16 ***
## log(nox)     0.059837   0.023021   2.599  0.01192 *
## log(so2)     0.014309   0.007584   1.887  0.06436 .
## log(hc)     -0.060812   0.020553  -2.959  0.00452 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05753 on 56 degrees of freedom
## Multiple R-squared:  0.2852, Adjusted R-squared:  0.2469
## F-statistic: 7.449 on 3 and 56 DF,  p-value: 0.0002777
```

```
ggplot(data= pollution)+
  geom_smooth(mapping = aes(x = log(nox)+ log(so2) +log(hc), y= log(mort)))+
  geom_point(mapping = aes(x = log(nox)+ log(so2) +log(hc), y= log(mort)))
```

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



For each 1% difference in nox, the predicted difference in mort is 0.059%. For each 1% difference in so2, the predicted difference in mort is 0.014%. For each 1% difference in hc, the predicted difference in mort is 0-0.061%.

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.)

```
nsh.mort2 <- lm(log(mort)~ log(nox)+ log(so2) +log(hc),data = pollution[1:30,])
display(nsh.mort2)
```

```
## lm(formula = log(mort) ~ log(nox) + log(so2) + log(hc), data = pollution[1:30,
##     ])
##             coef.est coef.se
## (Intercept)  6.80     0.03
## log(nox)     0.01     0.03
## log(so2)     0.02     0.01
## log(hc)     -0.02     0.03
## ---
## n = 30, k = 4
## residual sd = 0.06, R-Squared = 0.24
```

```
predict(nsh.mort2, newdata = pollution[31:60,])
```

```
##       31       32       33       34       35       36       37       38
## 6.865267 6.782537 6.883490 6.848612 6.879385 6.836847 6.778409 6.877503
##       39       40       41       42       43       44       45       46
## 6.895184 6.905029 6.848926 6.851197 6.887739 6.869825 6.848398 6.866064
##       47       48       49       50       51       52       53       54
## 6.825915 6.855637 6.780315 6.837748 6.856927 6.855909 6.853728 6.835775
##       55       56       57       58       59       60
```

```
## 6.868617 6.784731 6.866106 6.826688 6.876400 6.866190
```

Study of teenage gambling in Britain

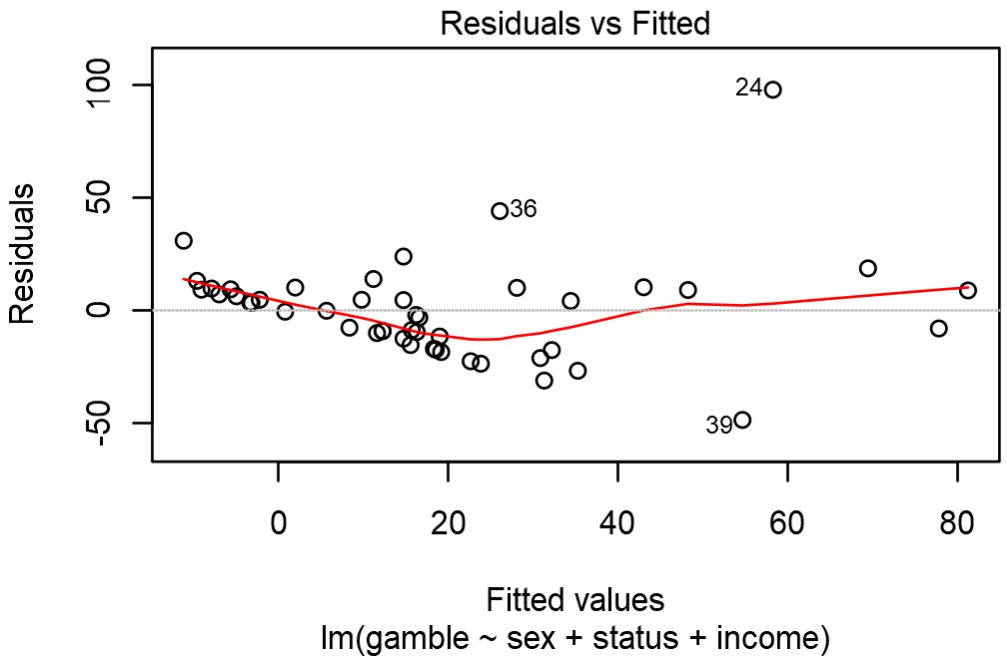
```
data(teengamb)
?teengamb
```

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.model <- lm(gamble~sex+status+income,data = teengamb)
display(teengamb.model)
```

```
## lm(formula = gamble ~ sex + status + income, data = teengamb)
##             coef.est coef.se
## (Intercept)  13.03    15.87
## sex         -24.34     8.13
## status       -0.15     0.24
## income        4.93     1.04
## ---
## n = 47, k = 4
## residual sd = 22.92, R-Squared = 0.51
```

```
plot(teengamb.model,which =1)
```




```
z.income = (teengamb$income - mean(teengamb$income ))/(2*sd(teengamb$income ))
z.status = (teengamb$status - mean(teengamb$status))/(2*sd(teengamb$status))
```

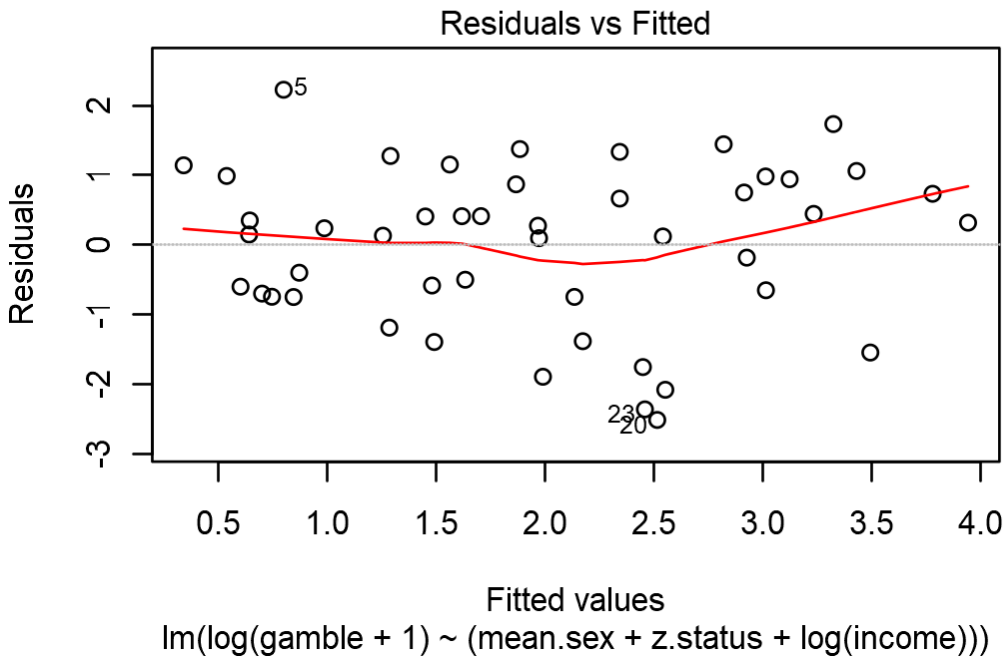
```
mean.sex = teengamb$sex-mean(teengamb$sex)
```

```
teengamb.mode2 <- lm(log(gamble+1)~(mean.sex + z.status+log(income)), data = teengamb)

display(teengamb.mode2)
```

```
## lm(formula = log(gamble + 1) ~ (mean.sex + z.status + log(income)),
##     data = teengamb)
##             coef.est coef.se
## (Intercept)  0.78     0.36
## mean.sex    -1.29     0.41
## z.status     0.24     0.42
## log(income)  0.93     0.25
## ---
## n = 47, k = 4
## residual sd = 1.18, R-Squared = 0.42
```

```
plot(teengamb.mode2,which =1)
```



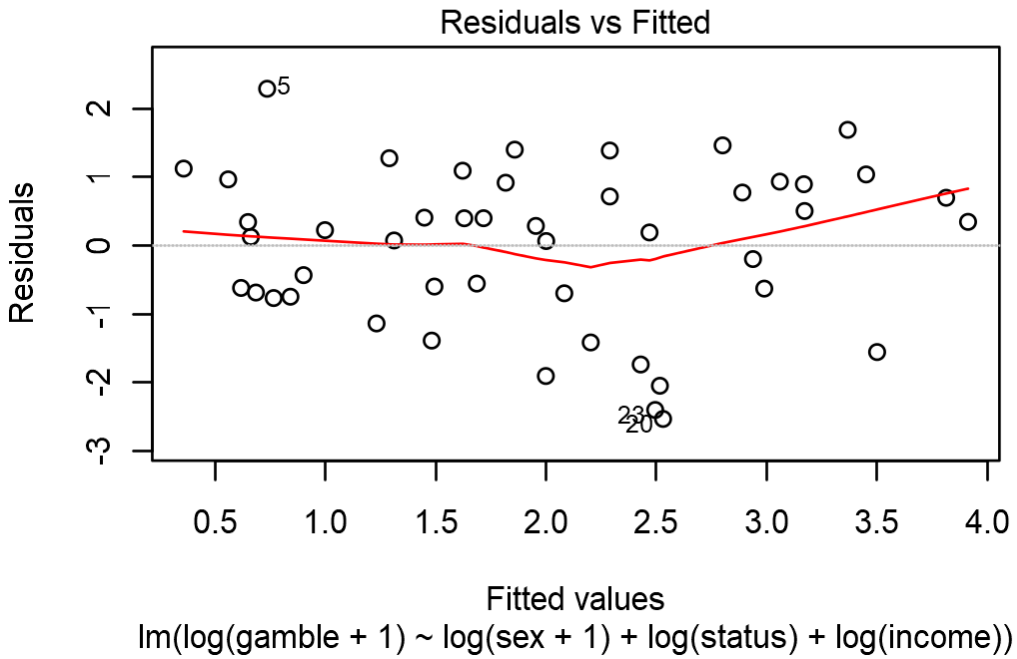
R-square value means this model does not work that good.

```
teengamb.mode3 <- lm(log(gamble+1)~log(sex+1)+log(status)+log(income),data = teengamb)
display(teengamb.mode3)
```

```
## lm(formula = log(gamble + 1) ~ log(sex + 1) + log(status) + log(income),
##     data = teengamb)
##              coef.est coef.se
## (Intercept)   0.55     2.10
## log(sex + 1) -1.92     0.58
```

```
## log(status)   0.21     0.50
## log(income)   0.92     0.25
## ---
## n = 47, k = 4
## residual sd = 1.18, R-Squared = 0.42
```

```
plot(teengamb.mode3,which =1 )
```



R-square value means this model does not work that good compare to orignial model.

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

```
round(confint(teengamb.mode2,level = 0.95),3)
```

```
##              2.5 % 97.5 %
## (Intercept)  0.056  1.512
## mean.sex    -2.110 -0.475
## z.status    -0.596  1.084
## log(income)  0.430  1.434
```

The range of [0.056, 1.512] is the 95% confidence interval within intercept. From this range, we know it does not cross zero. So this means it's significant.

The range of [-2.110, -0.475] is the 95% confidence interval within the vlue of centering sex. From this range, we know it does not cross zero. So this means the it's significant.

The range of [-0.586, 1.084] is the 95% confidence interval within the value of standarizing value. From this range, we know it crosses zero. So this means it's not significant.

The range of [0.430, 1.434] is the 95% confidence interval within log of income. From this range, we know it does not

cross zero. So this means it's 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?

```
average.male <- data.frame(sex=0, status = mean(teengamb$status),income = mean(teengamb$verbal
), verbal = mean(teengamb$verbal))

predict(teengamb.model, average.male, interval = "prediction")
```

```
##        fit       lwr     upr
## 1 39.08335 -8.100398 86.2671
```

```
max.male <- data.frame(sex=0, status = max(teengamb$status),income = max(teengamb$verbal), ver
bal = max(teengamb$verbal))
predict(teengamb.model, max.male, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 51.09234 0.7308213 101.4539
```

The max Male has wider CI than average male. In my opinion, this result is expected because max value is not a normal value. Use this calue to fit the model will cause the confidence interval become higher and wider. ### 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.

```
sat.model1 <- lm(total~expend+ratio+salary,data = sat)
display(sat.model1)
```

```
## lm(formula = total ~ expend + ratio + salary, data = sat)
##             coef.est coef.se
## (Intercept) 1069.23   110.92
## expend        16.47    22.05
## ratio          6.33     6.54
## salary        -8.82     4.70
## ---
## n = 50, k = 4
## residual sd = 68.65, R-Squared = 0.21
```

```
ratio.mean = 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)
sat.model2 <- lm(total~ z.expend+ratio.mean+ z.salary,data = sat)
display(sat.model2)
```

```
## lm(formula = total ~ z.expend + ratio.mean + z.salary, data = sat)
##             coef.est coef.se
## (Intercept) 965.92     9.71
## z.expend     24.17    32.36
## ratio.mean    6.33     6.54
## z.salary     -2.97     1.58
## ---
## n = 50, k = 4
## residual sd = 68.65, R-Squared = 0.21
```

```
sat.model3 <- lm(log(total)~ log(expend) + log(ratio)+log(salary),data= sat)
display(sat.model3)
```

```
## lm(formula = log(total) ~ log(expend) + log(ratio) + log(salary),
##     data = sat)
##             coef.est coef.se
## (Intercept)  7.48     0.31
## log(expend)  0.10     0.14
## log(ratio)   0.13     0.12
## log(salary) -0.32     0.17
## ---
## n = 50, k = 4
## residual sd = 0.07, R-Squared = 0.22
```

The last two models do not work that well. The coefficients are not significant except the intercept.

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

```
round(confint(sat.model3, level = 0.98),3)
```

```
##                1 %  99 %
## (Intercept)  6.734 8.232
## log(expend) -0.232 0.432
## log(ratio)  -0.175 0.428
## log(salary) -0.724 0.077
```

The range of [6.734, 8.232] is the 98% confidence interval within intercept. From this range, we know it does not cross zero. So this means it's significant.

The range of [-0.232, 0.432] is the 98% confidence interval within the log of expend. From this range, we know it does not cross zero. So this means the it's significant.

The range of [-0.175, 0.428] is the 98% confidence interval within the log of ratio. From this range, we know it crosses zero. So this means it's not significant.

The range of [-0.175, 0.428] is the 98% confidence interval within log of salary. From this range, we know it does cross zero. So this means it's not significant.

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

```
```

```
taker.model1<-lm(formula = total ~ expend + ratio + salary + takers, data = sat)
display(taker.model1)
```

```
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##             coef.est coef.se
## (Intercept) 1045.97    52.87
## expend         4.46    10.55
## ratio         -3.62     3.22
## salary         1.64     2.39
## takers        -2.90     0.23
## ---
## n = 50, k = 5
## residual sd = 32.70, R-Squared = 0.82
```

```
takers.mean <- sat$takers - mean(sat$takers)
takers.model2 <- lm(total~ z.expend+ratio.mean+ z.salary + takers.mean,data = sat)
display(takers.model2)
```

```
## lm(formula = total ~ z.expend + ratio.mean + z.salary + takers.mean,
##     data = sat)
##             coef.est coef.se
## (Intercept) 965.92     4.62
## z.expend      6.55    15.48
## ratio.mean   -3.62     3.22
## z.salary      0.55     0.80
## takers.mean  -2.90     0.23
## ---
## n = 50, k = 5
## residual sd = 32.70, R-Squared = 0.82
```

```
takers.model3 <- lm(log(total)~ log(expend) + log(ratio)+log(salary) + log(takers),data= sat)
display(takers.model3)
```

```
## lm(formula = log(total) ~ log(expend) + log(ratio) + log(salary) +
##     log(takers), data = sat)
##             coef.est coef.se
## (Intercept)  6.87     0.12
## log(expend)  0.07     0.05
## log(ratio)   0.01     0.05
## log(salary)  0.03     0.07
## log(takers) -0.08     0.00
## ---
## n = 50, k = 5
## residual sd = 0.03, R-Squared = 0.89
```

From the r-square value, the all three models are better than the original ones.

The log one is better than previous two.

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, $D_i - R_i$ This is the simplest method which compare two parties 's money directly. Advantage: it can see difference of two parties in different district directly. disadvantage: only number, no other way like proportin to search about these data
- The ratio, D_i/R_i Advantage: we know above 1, D get more money. We can understand this easier. Disadvantage:
- The difference on the logarithmic scale, $logD_i - logR_i$ advantage: the number after log become small and easier to calculate
- The relative proportion, $D_i/(D_i + R_i)$. know D's part is how many percent of whole part. Advantage: there is a comparsion to D's part to whole part. If this number is high means more money

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 $\mathrm{x}^\star = 10(\mathrm{x} - 1)$ and that y is regressed on $\mathrm{x}^\star$. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^\star)$ and $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star)$.
5. Now suppose that the response variable scores are transformed according to the formula $\mathrm{y}^{\star\star} = 5(\mathrm{y} + 2)$ and that $\mathrm{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^{\star\star})$ and $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star})$.
6. In general, how are the hypothesis tests and confidence intervals for β affected by linear transformations of y and x?

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your

opinions.