Homework 02

Guangyan Yu

Septemeber 21, 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.

```
#Remove NA
heights_new<-na.omit(heights)
rownames(heights_new)<-seq(1,nrow(heights_new))
#Remove earning=0
zero<-which(heights_new$earn==0)
heights_new<-heights_new[-zero,]
#Remove outliers by clustering(kmeans)
set.seed(12)
km = kmeans(heights_new,center=3)
r = nrow(heights_new)
c = ncol(heights_new)
#distance
x1=matrix(km$centers[1,],nrow=r,ncol=c,byrow=T)
juli1=sqrt(rowSums((heights_new-x1)^2))
x2=matrix(km$centers[2,],nrow=r,ncol=c,byrow=T)
juli2=sqrt(rowSums((heights_new-x2)^2))
x3=matrix(km$centers[3,],nrow=r,ncol=c,byrow=T)
juli3=sqrt(rowSums((heights_new-x3)^2))
dist=data.frame(juli1,juli2,juli3)
#minimum of distance
y=apply(dist,1,min)
y_new<-sort(y)
q<-y_new[ceiling(r*0.8)]

sub<-which(y>q)
heights_nn <- heights_new[-sub,]
```

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?

```
y <- heights_nn$earn
x <- heights_nn$height
model<-lm(y~x)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21696  -8189  -1812   7515  64534
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25589.13    6402.61  -3.997 6.92e-05 ***
## x              673.03      95.77   7.028 4.00e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11310 on 952 degrees of freedom
## Multiple R-squared:  0.04932,    Adjusted R-squared:  0.04832
## F-statistic: 49.39 on 1 and 952 DF,  p-value: 3.997e-12
```

We should centering the earnings and heights

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.

```
sex<-heights_nn$sex
earning<-heights_nn$earn
height<-heights_nn$height
#log
model1<-lm(log(earning+1)~sex+height)
summary(model1)
```

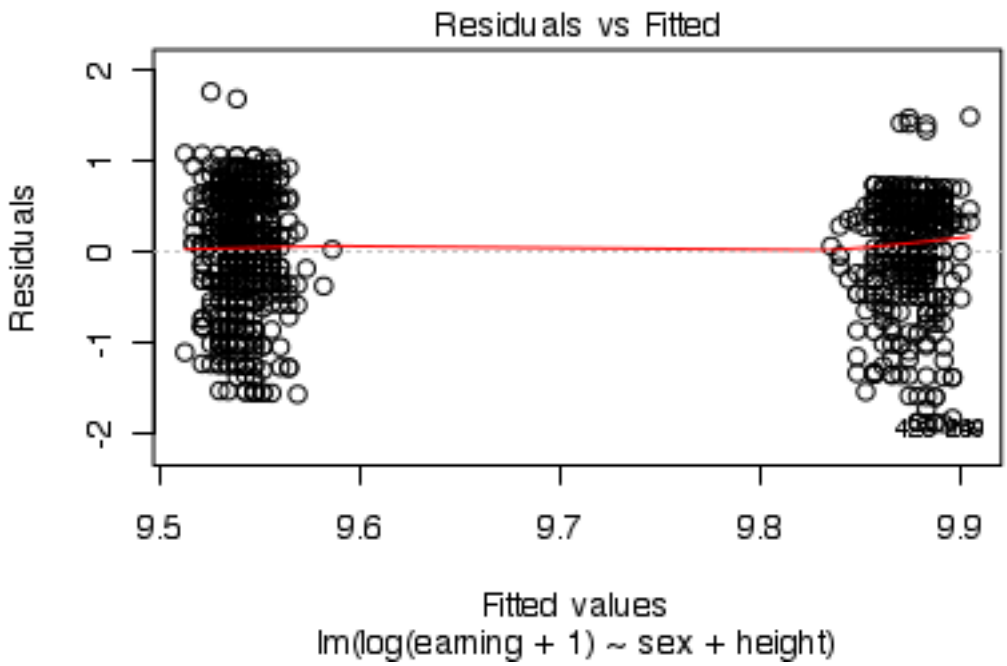
```
##
## Call:
## lm(formula = log(earning + 1) ~ sex + height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88503 -0.33639  0.07772  0.52005  1.76465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.880915   0.584099  16.917  < 2e-16 ***
## sex         -0.310195   0.059935  -5.176 2.77e-07 ***
## height       0.004338   0.007679   0.565    0.572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.647 on 951 degrees of freedom
## Multiple R-squared:  0.06057,    Adjusted R-squared:  0.0586
## F-statistic: 30.66 on 2 and 951 DF,  p-value: 1.248e-13
```

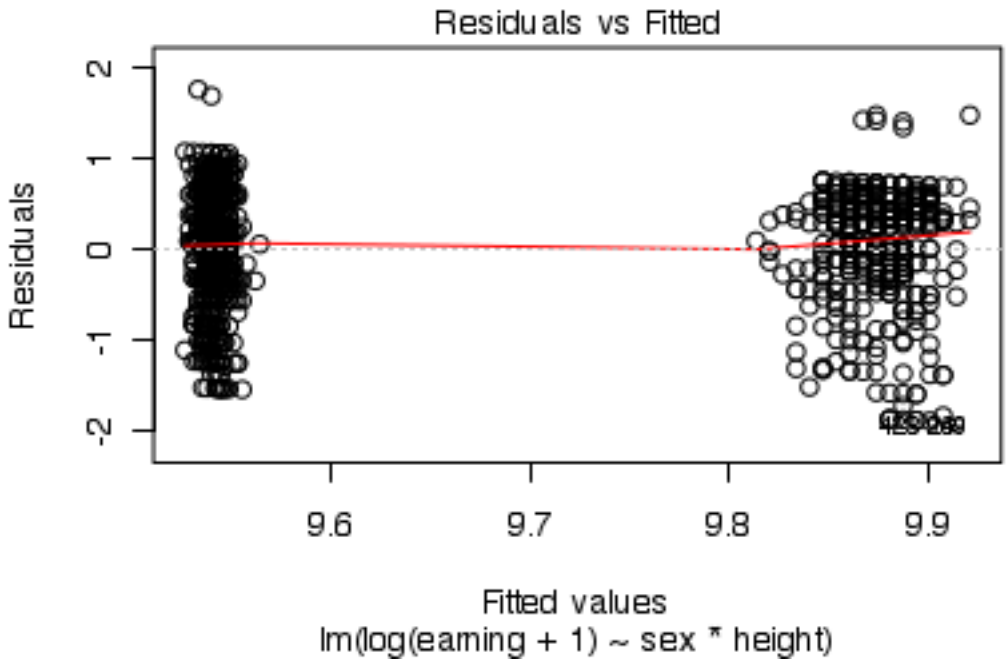
```
model2<-lm(log(earning+1)~sex*height)
summary(model2)
```

```
##
## Call:
## lm(formula = log(earning + 1) ~ sex * height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89454 -0.33334  0.07654  0.51882  1.75713
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.409305   1.713019   5.493 5.08e-08 ***
## sex         -0.006072   1.040125  -0.006    0.995
## height       0.011238   0.024780   0.454    0.650
## sex:height  -0.004508   0.015393  -0.293    0.770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6474 on 950 degrees of freedom
## Multiple R-squared:  0.06066,    Adjusted R-squared:  0.05769
## F-statistic: 20.45 on 3 and 950 DF,  p-value: 7.598e-13
```

```
#pyth_gp<-ggplot(model1)
#pyth_gp + aes(x=sex+height+race,earning) + geom_point() + stat_smooth(method = "lm",col = "red")
plot(model1,which=1)
```




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




```
#zscore
earning_z<-(earning-mean(earning))/sd(earning)
height_z<-(height-mean(height))/sd(height)
```

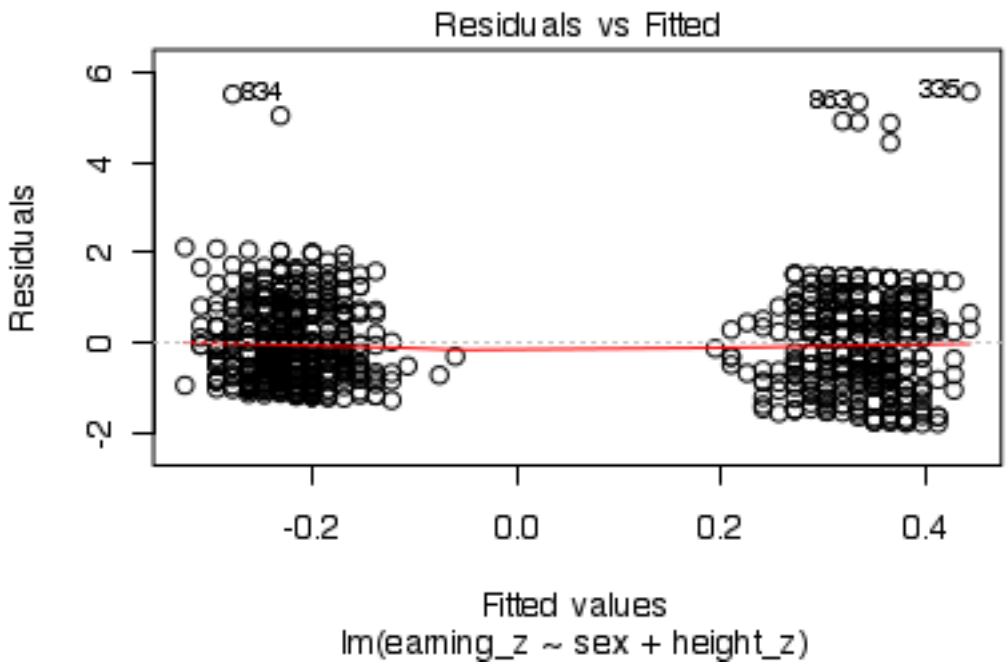
```r
model3<-lm(earning_z~sex+height_z)
model4<-lm(earning_z~sex*height_z)
summary(model3)
```

```
##
## Call:
## lm(formula = earning_z ~ sex + height_z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8050 -0.6969 -0.1415  0.6311  5.5642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75631    0.14589   5.184 2.65e-07 ***
## sex         -0.47282    0.08910  -5.307 1.39e-07 ***
## height_z     0.05964    0.04368   1.365    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9619 on 951 degrees of freedom
## Multiple R-squared:  0.07666,    Adjusted R-squared:  0.07472
## F-statistic: 39.48 on 2 and 951 DF,  p-value: < 2.2e-16
```

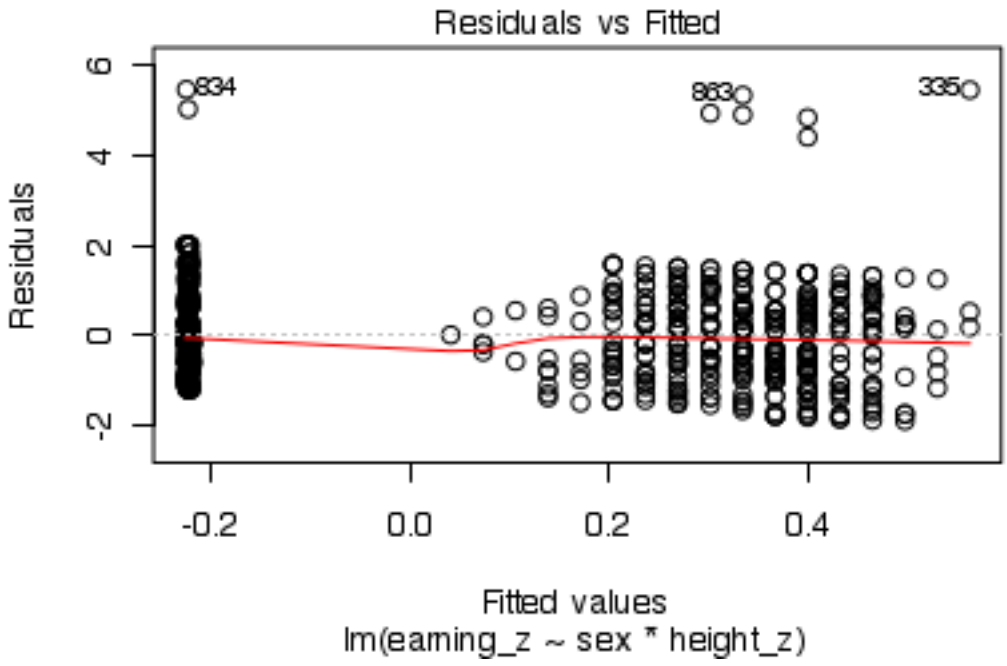
```r
summary(model4)
```

```
##
## Call:
## lm(formula = earning_z ~ sex * height_z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8887 -0.6681 -0.1499  0.6266  5.4567
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.67764    0.15621   4.338 1.59e-05 ***
## sex          -0.44998    0.09053  -4.971 7.92e-07 ***
## height_z      0.24756    0.14083   1.758   0.0791 .
## sex:height_z -0.12278    0.08748  -1.404   0.1608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9614 on 950 degrees of freedom
## Multiple R-squared:  0.07857,    Adjusted R-squared:  0.07566
## F-statistic:    27 on 3 and 950 DF,  p-value: < 2.2e-16
```

```r
plot(model3,which=1)
```




```
plot(model4,which=1)
```



I think model3 is the best, which is $earning\_z = 0.756 - 0.473 * sex + 0.059 * height\_z$. *Because* R^2 *of this model is the biggest one, and the residual plot indicate that the residuals spread evenly on both sides of 0*

4. Interpret all model coefficients.

In model3, the regression function is $earning\_z = 0.756 - 0.473 * sex + 0.059 * height\_z$

When a woman has the average height and race=0, her earning is 0.756 more than the average earning.

At same condition, a man's earning is $0.473 * sd(earning)$ *less than a woman*

With height increases 1 sd(height), earning increases $0.059 * sd(earning)$

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

```
confint(model3,level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept)  0.47001321  1.0426081
## sex         -0.64767508 -0.2979611
## height_z    -0.02608321  0.1453607
```

[0.47,1.04] contains the true intercept with 95% probability

[-0.648,-0.298] contains the true coefficient of sex with 95% probability

[-0.026,0.145] contains the true coefficient of height_z with 95% probability

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:

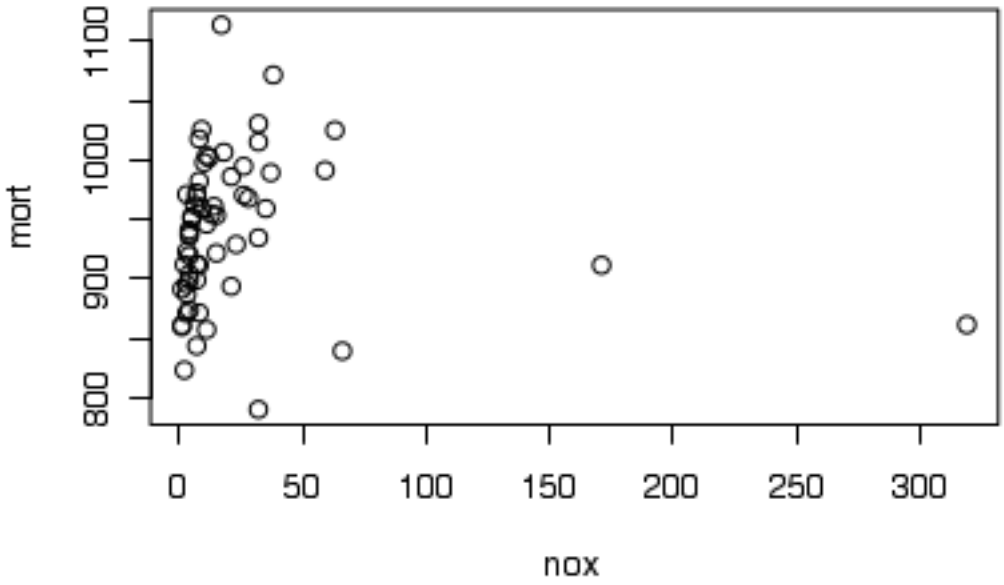
- PREC Average annual precipitation in inches
- JANT Average January temperature in degrees F
- JULT Same for July
- OVR65 % of 1960 SMSA population aged 65 or older
- POPN Average household size
- EDUC Median school years completed by those over 22
- HOUS % of housing units which are sound & with all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

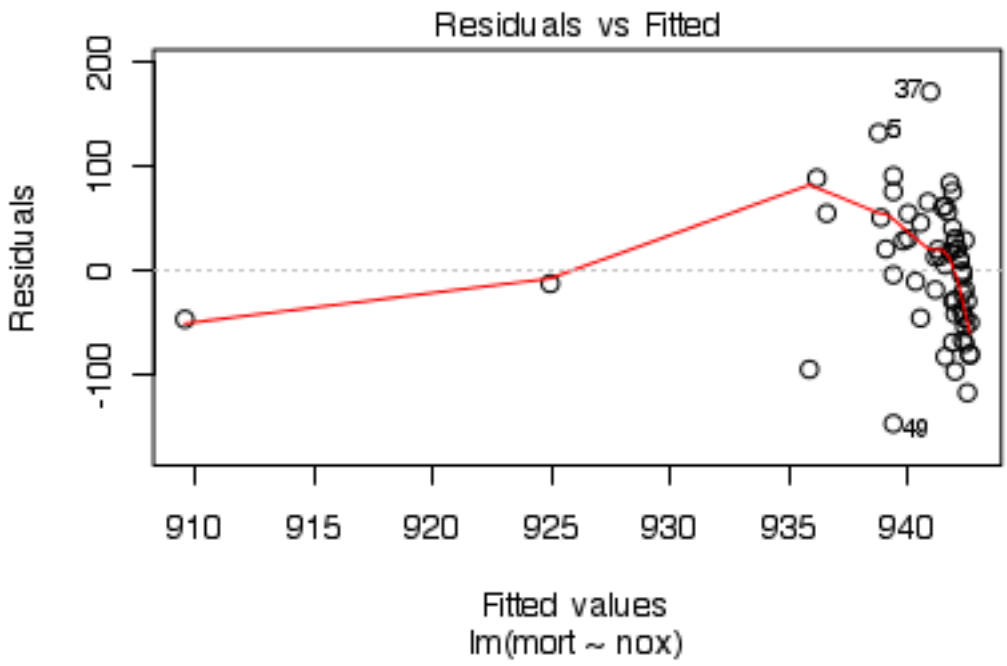
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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.

```r
mort<-pollution$mort
nox<-pollution$nox
plot(x=nox,y=mort)
```



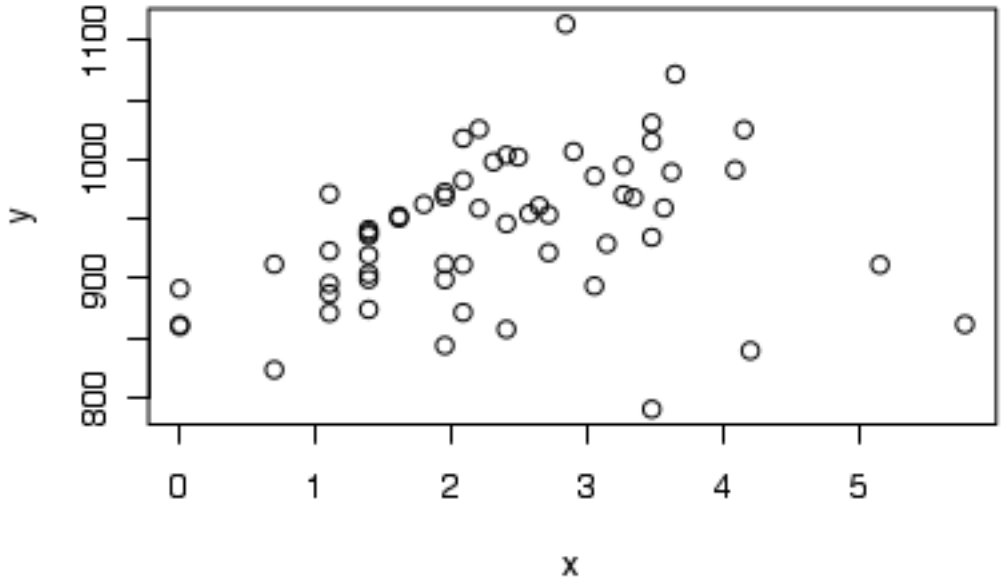

```r
model<-lm(mort~nox)
plot(model,which=1)
```



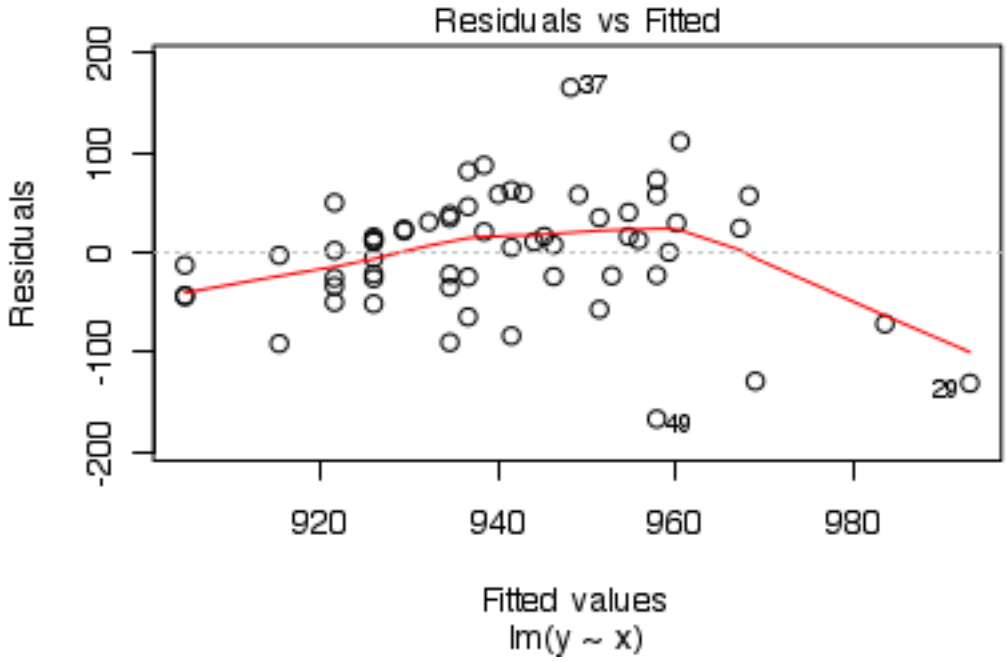
It will not fit well, because the data has right-skewness The residuals are not evenly distrbuted in the plot, so that the regression does not fit well.

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.

```
y<-mort
x<-log(nox)
plot(x,y)
```




```
model<-lm(y~x)
plot(model,which=1)
```



In this case, the residuals evenly distributed on both sides of 0 in the plot, so that the regression is more appropriate 3. Interpret the slope coefficient from the model you chose in 2.

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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 ***
## x             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
```

The regression function is mort $= 904.724 * 15.335 * log(nox)$ *If x times 2, exp(y) will times* $2^{15.335}$ 4. Construct 99% confidence interval for slope coefficient from the model you chose in 2 and interpret them.

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

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

[-2.231,32.902] contains the true slop of log(nox) with 99% probability

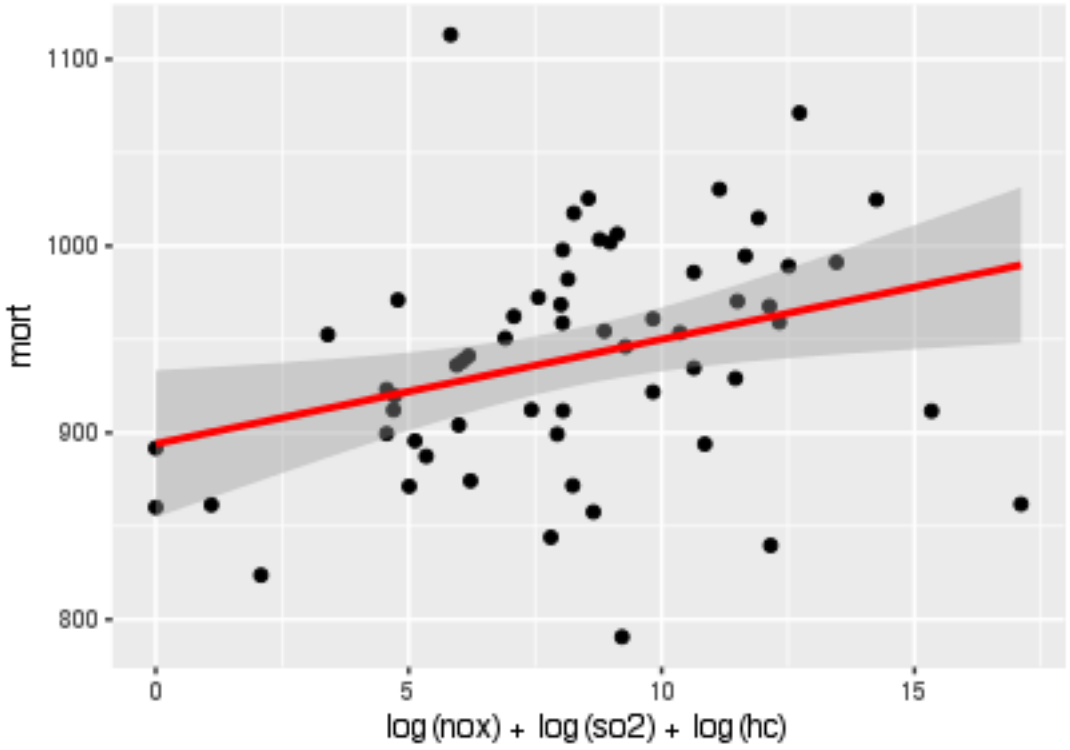
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
so2<-pollution$so2
hc<-pollution$hc
model<-lm(mort~log(nox)+log(so2)+log(hc))
summary(model)
```

```
##
## Call:
## lm(formula = mort ~ log(nox) + log(so2) + log(hc))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.793 -34.728  -3.118  34.148 194.567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  924.965     21.449  43.125  < 2e-16 ***
## log(nox)      58.336     21.751   2.682  0.00960 **
## log(so2)      11.762      7.165   1.642  0.10629
## log(hc)      -57.300     19.419  -2.951  0.00462 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 54.36 on 56 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2363
## F-statistic: 7.086 on 3 and 56 DF,  p-value: 0.0004044
```

```
library(ggplot2)
pyth_gp<-ggplot(model)
pyth_gp + aes(x=log(nox)+log(so2)+log(hc),mort) + geom_point() + stat_smooth(method = "lm",col = "red")
```



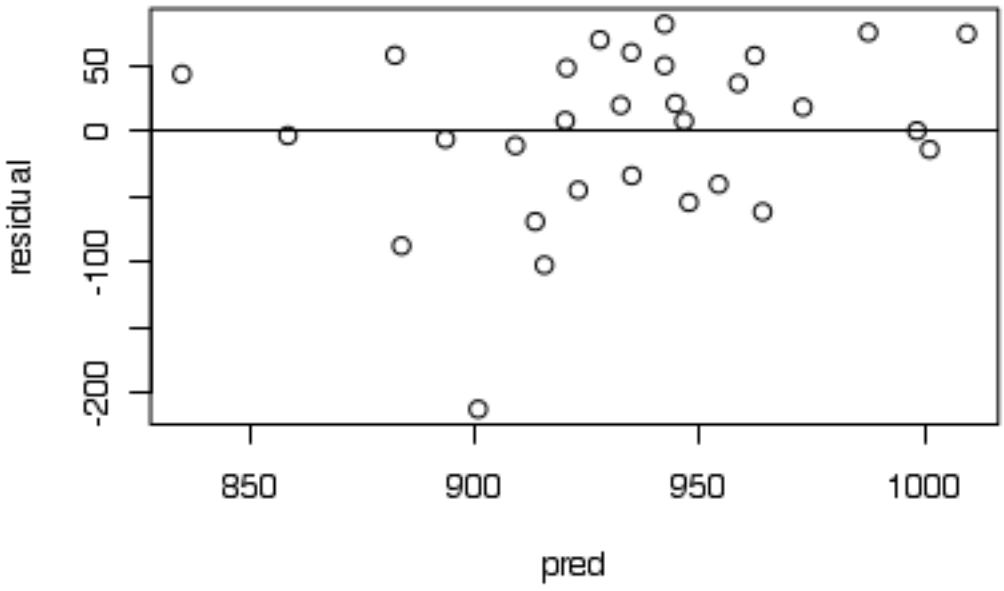
The regression function ismort $= 924.965 + 58.336 * log(nox) + 11.762 * log(so2) - 57.3 * log(hc)$

When nox=so2=hc=1,mort=924.965

When log(nox) or log(so2) or log(hc) increases 1 unit, mort will increase 1 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.)

```
set.seed(10)
index<-sample(nrow(pollution), 0.5*nrow(pollution),replace=F)
data1<-pollution[index,]
data2<-pollution[-index,]
model<-lm(mort~log(nox)+log(so2)+log(hc),data=data1)
coef<-model$coefficients
pred<-coef[1]+coef[2]*log(data2$nox)+coef[3]*log(data2$so2)+coef[4]*log(data2$hc)
residual<-pred-data2$mort
plot(x=pred,y=residual)
abline(a=0,b=0)
```



The residual plot show that except one point, other residuals are evenly distributed on both sides of 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.

```
gamble_log<-log(teengamb$gamble+1)
sex<-teengamb$sex
status_center<-(teengamb$status-mean(teengamb$status))/sd(teengamb$status)
income<-teengamb$income
verbal<-teengamb$verbal
model<-lm(gamble_log~sex+status_center+income+verbal)
summary(model)
```

```
##
## Call:
## lm(formula = gamble_log ~ sex + status_center + income + verbal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35012 -0.56865  0.00413  0.71512  1.90319
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    3.06554    0.74198   4.132 0.000168 ***
## sex           -0.87120    0.39268  -2.219 0.031975 *
## status_center  0.51496    0.23208   2.219 0.031951 *
## income         0.21565    0.04904   4.398 7.33e-05 ***
## verbal        -0.26165    0.10388  -2.519 0.015673 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 42 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.475
## F-statistic:  11.4 on 4 and 42 DF,  p-value: 2.347e-06
```

The regression function is $gamble\_log = 3.06554 - 0.87120sex + 0.51496status\_center + 0.215654income - 0.26165verbal$ *At the same condition, a male spends exp(-0.871) times that of a female on gambling in pounds per year*

A female with average status and 0 income per week and 0 verbal score spends exp(3.066) in pounds on gambling per year.

With socioeconomic status score increasing one unit, the expenditure on gambling in pounds per year times exp(0.515)

With income per week increasing 1 pound, the expenditure on gambling in pounds per year times exp(0.216)

With verbal score increasing 1 unit, the expenditure on gambling in pounds per year times exp(-0.262) 2. Create a 95% confidence interval for each of the estimated coefficients and discuss how you would interpret this uncertainty.

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

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

[1.57,4.56] contains the true intercept with 95% probability

[-1.66,-0.08] contains the true coefficient of sex with 95% probability

[0.05,0.98] contains the true coefficient of status_center with 95% probability

[0.11,0.31] contains the true coefficient of income with 95% probability

[-0.47,-0.05] contains the true coefficient of verbal with 95% probability

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 = model,newdata = data.frame(sex=0,status_center=0,income=mean(teenga
l1<-prediction_average[3]-prediction_average[2]
prediction_max<-predict(object = model,newdata = data.frame(sex=0,status_center=max(teengamb$status)-mea
l2<-prediction_max[3]-prediction_max[2]
l1
```

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

```
l2
```

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

Because the length of CI is $2 * \frac{s}{\sqrt{n}} * t_{\frac{\alpha}{2}}$ *and for a male with maximal values of status, income and verbal score, the standard error* s *is bigger than the* s *of a male with average status, income and verbal score.*

School expenditure and test scores from USA in 1994-95

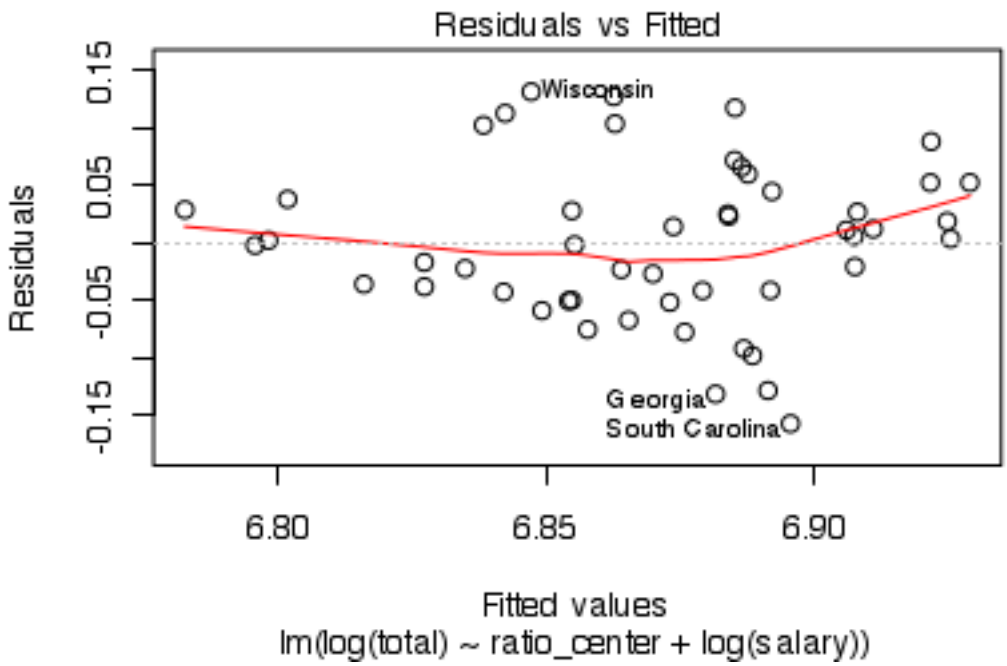
```
data(sat)
?sat
```

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.

```
ratio_center = sat$ratio-mean(sat$ratio)
salary_log = log(sat$salary)
model<-lm(log(total)~ratio_center+log(salary),data=sat)
summary(model)
```

```
##
## Call:
## lm(formula = log(total) ~ ratio_center + log(salary), data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.157349 -0.042730  0.002532  0.042619  0.130982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.619452   0.214397  35.539  < 2e-16 ***
## ratio_center  0.003145   0.004402   0.714  0.47859
## log(salary)  -0.211852   0.060553  -3.499  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06984 on 47 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1789
## F-statistic: 6.338 on 2 and 47 DF,  p-value: 0.003655
```

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



The regression model is $log(total) = 7.619 + 0.003 * ratio\_center - 0.212 * log(salary)$ *With the average teacher ratio and 1 salary, the log(total) is 7.619*

With ratio increasing 1 unit, the total score will be exp(0.003) times of the original total score

With log(salary) increasing 1 unit, the total score will be exp(-0.212) times of the original total score

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

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

```
##                       1 %        99 %
## (Intercept)   7.103111221  8.13579293
## ratio_center -0.007458157  0.01374726
## log(salary)  -0.357684974 -0.06601843
```

[7.10,8.14] contains the true intercept with 98% probability

[-0.0075,0.0137] contains the true coefficient of ratio_cneter with 98% probability

[-0.36,-0.066] contains the true coefficient of log(salary) with 98% probability

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?

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

```
##
## Call:
## lm(formula = log(total) ~ ratio_center + log(salary) + log(sat$taker),
##     data = sat)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.056268 -0.016657  0.000805  0.012664  0.057878
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.751891   0.096560  69.924  < 2e-16 ***
## ratio_center   -0.002054   0.001701  -1.208 0.233377
## log(salary)     0.108052   0.029916   3.612 0.000749 ***
## log(sat$taker) -0.083582   0.004998 -16.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02653 on 46 degrees of freedom
## Multiple R-squared:  0.8888, Adjusted R-squared:  0.8815
## F-statistic: 122.5 on 3 and 46 DF,  p-value: < 2.2e-16
```

The R^2 of this regression is 0.88, which is bigger than the earlier regression, illustrating that the fittness of the regression adding log(takers) is better than the original model.

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 is easy to interpret how the difference between money raised by two candidates can effect the result.*

Disadvantage: It can only indicate the effect of difference. For example, when $D_i = 200$,$R_i = 100$, *the result is same with the situation in which* $ D_i=400,R_i=200$

- The ratio, D_i/R_i *Advantage: It is easy to interpret how the ratio of money raised by two candidates can effect the result.*

Disadvantage: It can only indicate the effect of ratio. For example, when $D_i = 200$,$R_i = 100$, *the result is same with the situation in which* $ D_i=400,R_i=200$

- The difference on the logarithmic scale, $logD_i - logR_i$ skewness *Advantage: It can decrease the right-skewness of data*

Disadvantage: It can only indicate the effect of ratio. For example, when $D_i = 200$,$R_i = 100$, *the result is same with the situation in which* $ D_i=400,R_i=200$*

- The relative proportion, $D_i/(D_i + R_i)$. *Advantage: It can indicate the effect of the relation between a candidate and the both tow candidates*

Disadvantage: It can only indicate the effect of relationship between a individual and the ensable. For example, when $D_i = 200$,$R_i = 100$, *the result is same with the situation in which* $ D_i=400,R_i=200$*

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

(a) $x^* = x - 10$

$y = \hat{\alpha} + \hat{\beta}(x^* + 10) + \hat{\epsilon}$

so $\hat{\alpha^*} = \hat{\alpha} + 10 * \hat{\beta} = 10$

$\hat{\beta^*} = \hat{\beta} = 0.9$

$\hat{\epsilon^*} = \hat{\epsilon} \text{—>} \hat{\sigma^*} = \hat{\sigma} = 2$

becuase $R^2 = 1 - \frac{\sum(y_i - \hat{y_i})}{\sum(y_i - \bar{y})}$

and $\hat{y_i}$ *wil not change*

so$R^{2^*} = R^2$, $r^* = 0.3$

(b) $x^* = 10x$ $y = \hat{\alpha} + \hat{\beta}x^*/10 + \hat{\epsilon}$

so $\hat{\alpha^*} = \hat{\alpha} = 1$

$\hat{\beta^*} = \hat{\beta}/10 = 0.09$

$\hat{\epsilon^*} = \hat{\epsilon} \text{—>} \hat{\sigma^*} = \hat{\sigma} = 2$

and $\hat{y_i}$ *wil not change*

so $R^{2^*} = R^2$,$r^* = 0.3$

(c) $x^* = 10(x - 1)$

$y = \hat{\alpha} + \hat{\beta} + \hat{\beta}x^*/10 + \hat{\epsilon}$

so $\hat{\alpha^*} = \hat{\alpha} + \hat{\beta} = 1.9$

$\hat{\beta^*} = \hat{\beta}/10 = 0.09$

$\hat{\epsilon^*} = \hat{\epsilon} \text{—>} \hat{\sigma^*} = \hat{\sigma} = 2$

and $\hat{y_i}$ *wil not change*

so $R^{2^*} = R^2$,$r^* = 0.3$

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

(a)$y^{**} = y + 10$

$y^{**} = \hat{\alpha} + 10 + \hat{\beta}x + \hat{\epsilon}$

so $\hat{\alpha^{**}} = \hat{\alpha} + 10 = 11$

$\hat{\beta^{**}} = \hat{\beta} = 0.9$

$\hat{\epsilon^{**}} = \hat{\epsilon} \text{—>} \hat{\sigma^{**}} = \hat{\sigma} = 2$

because $\hat{y_i^{**}} = \hat{y}_i + 10$ *and* $\bar{y_i^{**}} = \bar{y}_i + 10$ *and* $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)}{\sum(y_i - \bar{y})}$

we can tell that $\hat{R^{2**}} = \hat{R^2}, r^{**} = 0.3$

(b)$y^{**} = 5y$

$y^{**} = 5\hat{\alpha} + 5\hat{\beta} + 5\hat{\epsilon}$

$\hat{\alpha^{**}} = 5\hat{\alpha} = 5$

$\hat{\beta^{**}} = 5\hat{\beta} = 4.5$

$\hat{\epsilon^{**}} = 5\hat{\epsilon} \text{—>} \hat{\sigma^{**}} = 5\hat{\sigma} = 10$

because $\hat{y_i^{**}} = 5\hat{y}_i$ *and* $\bar{y_i^{**}} = 5\bar{y}_i$ *and* $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)}{\sum(y_i - \bar{y})}$

we can tell that $\hat{R^{2**}} = \hat{R^2}, r^{**} = 0.3$

(c)$y^{**} = 5(y + 2)$

$y^{**} = 5\hat{\alpha} + 10 + 5\hat{\beta} + 5\hat{\epsilon}$

$\hat{\alpha^{**}} = 5\hat{\alpha} + 10 = 15$

$\hat{\beta^{**}} = 5\hat{\beta} = 4.5$

$\hat{\epsilon^{**}} = 5\hat{\epsilon} \text{—>} \hat{\sigma^{**}} = 5\hat{\sigma} = 10$

because $\hat{y_i^{**}} = 5\hat{y}_i + 10$ *and* $\bar{y_i^{**}} = 5\bar{y}_i + 10$ *and* $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)}{\sum(y_i - \bar{y})}$

we can tell that $\hat{R^{2**}} = \hat{R^2}, r^{**} = 0.3$

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

linear transformations of x *do not affect* ϵ *and* R^2

$\text{x} + c$ *will result the intercept change to* $\hat{\alpha} - c\hat{\beta}$, *and* $\hat{\beta}$ *do not change*

$\text{x} * d$ *will result the* $\hat{\beta}$ *change to* $\hat{\beta}/d$, *and* $\hat{\alpha}$ *do not change*

linear transformations of y *do not affect* R^2

$\text{y} + c$ *will result the intercept change to* $\hat{\alpha} + c$, *and* $\hat{\beta}$ *do not change*

$\text{x} * d$ *will result the* $\hat{\alpha}$ *change to* $\hat{\alpha} * d$, *and* $\hat{\beta}$ *will change to* $\hat{\beta} * d$,*and* $\hat{\sigma}$ *will change to* $\hat{\sigma} * 5$

4. Suppose that the explanatory variable values in a regression are transformed according to the $\text{x}^\star = 10(\text{x} - 1)$ and that y is regressed on $\text{x}^\star$. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^\star)$ and $t_0^\star = \hat{\beta}^\star / SE(\hat{\beta}^\star)$.

$\hat{\beta^*} = \hat{\beta}/10$

$SE(\hat{\beta^*}) = SE(\hat{\beta})/10 = 0.003$

$t_0^* = t_0 = 30$

5. Now suppose that the response variable scores are transformed according to the formula $\text{y}^{\star\star} = 5(\text{y} + 2)$ and that $\text{y}^{\star\star}$ is regressed on x. Without redoing the regression calculation in detail, find $SE(\hat{\beta}^{\star\star})$ and $t_0^{\star\star} = \hat{\beta}^{\star\star} / SE(\hat{\beta}^{\star\star})$.

$y^{**} = 5(y + 2)$

$\hat{\beta^{**}} = 5\hat{\beta}$

$SE(\hat{\beta^*}) = 5 * SE(\hat{\beta}) = 0.15$

$t_0^{**} = t_0 = 30$

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

We can tell that both addition or substraction on x or y will not change the CI

And if $x^* = c * x$, *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 $y^* = d * y$, *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:$\mu \neq 0$

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

We can tell that both addition or substraction on x or y will not change T so that will not change the result of test.

And if $x^* = c * x$, *then* $\bar{\beta^*} = \bar{\beta}/c$,*T does not change.*

If $y^* = d * y$, *then* $\bar{\beta^*} = \bar{\beta} * d$,*T does not change*

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.