Homework 02

yourname

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.

```
#delete the row with NA
na.in.row <- apply(heights, 1, function(x){any(is.na(x))})
sum(na.in.row)
```

```
## [1] 650
```

```
finalheights <- heights[!na.in.row,]

#delete rows contains 0 in earning
zero.in.earning <- which(finalheights$earn == 0)
sum(zero.in.earning)
```

```
## [1] 131722
```

```
finalheights2 <- finalheights[-zero.in.earning,]
```

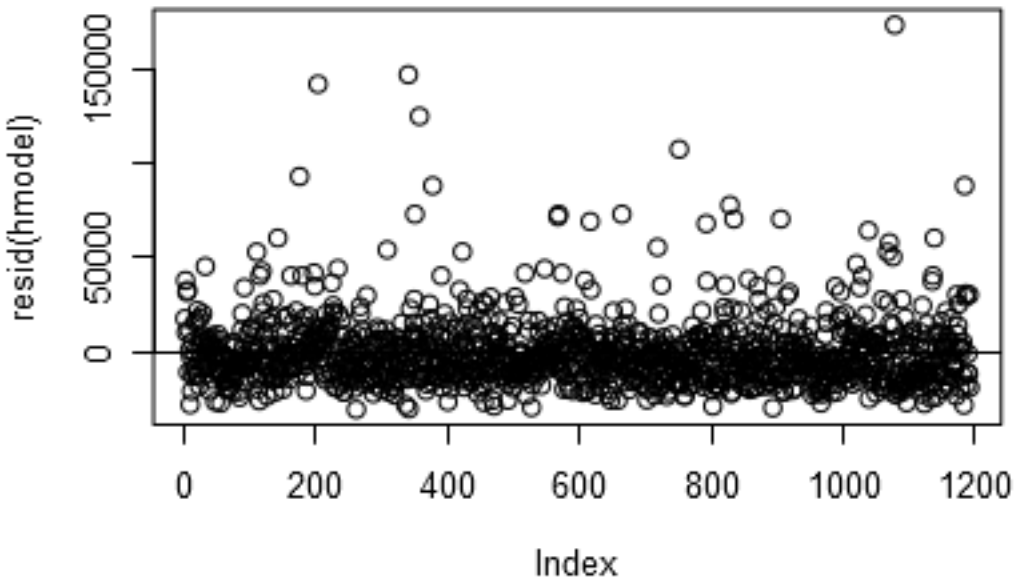
2. Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

centering transformation should be used in order to interpret the intercept from this model as average earnings for people with average height.

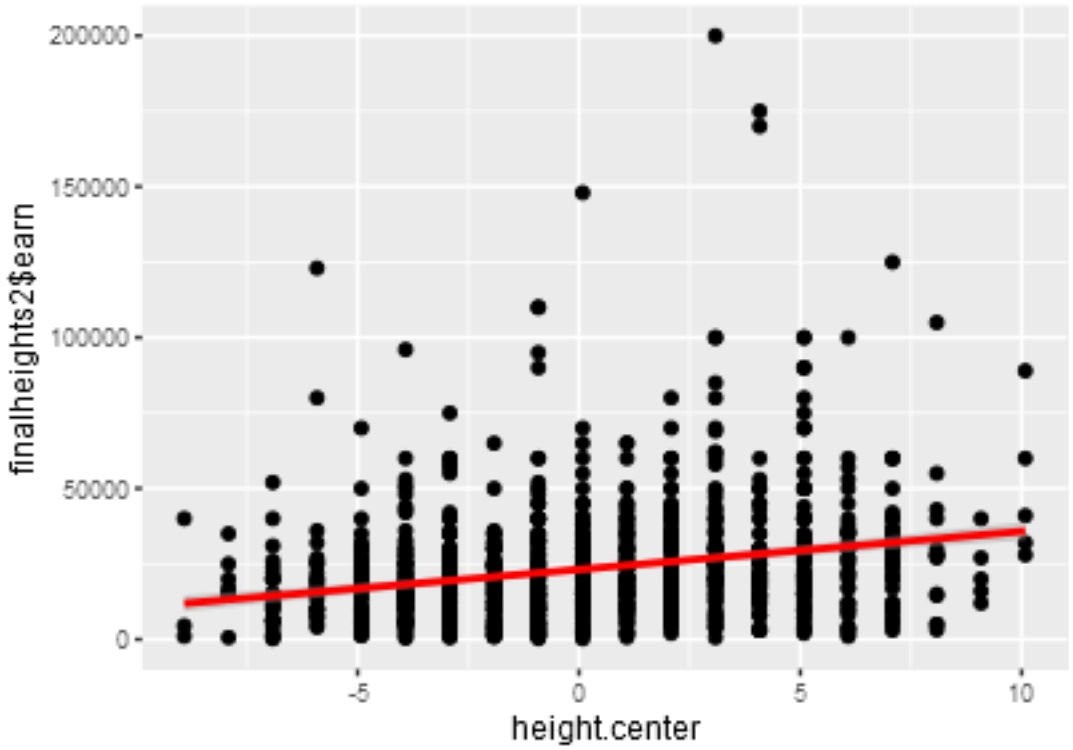
```
#calculating the center and using centering transformation
height.center <- finalheights2$height - mean(finalheights2$height)

hmodel <- lm(earn ~ height.center, data = finalheights2)
summary(hmodel)
```

```
##
## Call:
## lm(formula = earn ~ height.center, data = finalheights2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30166 -11309  -3428   6527 172953
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23154.8      546.4  42.376   <2e-16 ***
## height.center   1262.3      142.1   8.883   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18870 on 1190 degrees of freedom
## Multiple R-squared:  0.06218,    Adjusted R-squared:  0.06139
## F-statistic:  78.9 on 1 and 1190 DF,  p-value: < 2.2e-16
```

```
plot(resid(hmodel))
abline(h = 0)
```




```
ggplot(hmodel) + geom_point() + aes(x= height.center, y= finalheights2$earn) + geom_smooth(method = 'lm
```

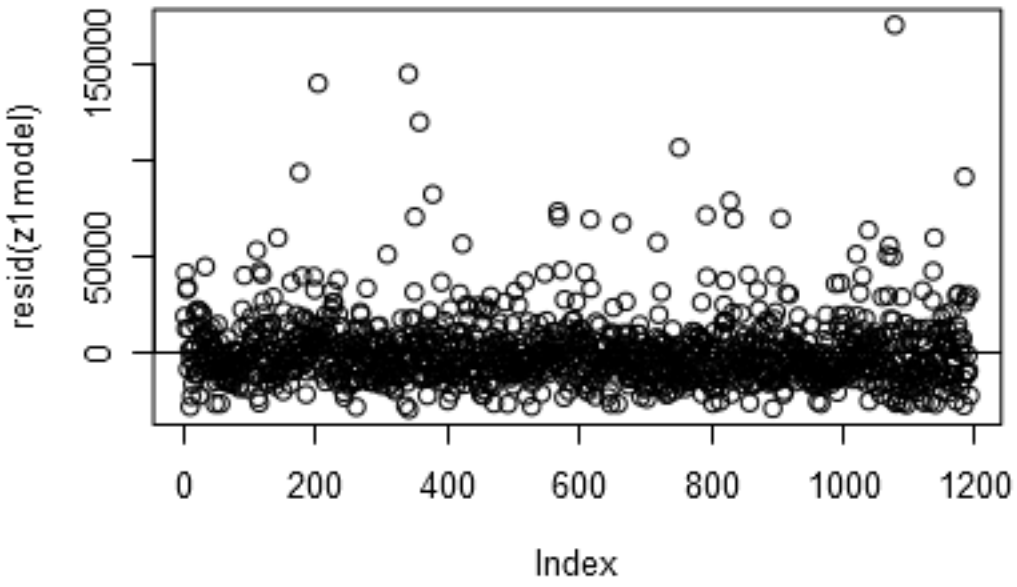


3. Fit some regression models with the goal of predicting earnings from some combination of sex, height, and weight. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

```
# zscore transformation
zheight <- height.center / sd(finalheights$height)

z1model <- lm(finalheights2$earn ~ zheight + finalheights2$sex)

summary(z1model)
```

```
##
## Call:
## lm(formula = finalheights2$earn ~ zheight + finalheights2$sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30087 -11013  -3315   6128 170242
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        37480.4     2471.1  15.167  < 2e-16 ***
## zheight             1685.8      748.3   2.253   0.0245 *
## finalheights2$sex  -9087.9     1529.9  -5.940 3.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18600 on 1189 degrees of freedom
## Multiple R-squared:  0.08921,    Adjusted R-squared:  0.08768
## F-statistic: 58.23 on 2 and 1189 DF,  p-value: < 2.2e-16
```

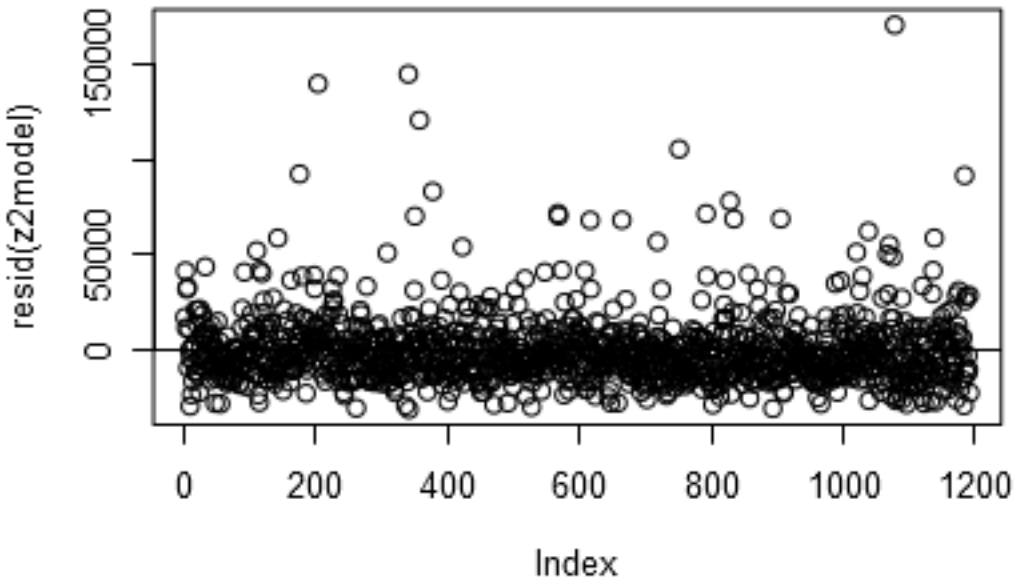
```r
plot(resid(z1model))
abline(h=0)
```




```r
# zscore transformation2

z2model <- lm(finalheights2$earn ~ zheight * finalheights2$sex)

summary(z2model)
```

```
##
## Call:
## lm(formula = finalheights2$earn ~ zheight * finalheights2$sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30943 -11164  -3182   6365 170260
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   36295       2601  13.953  < 2e-16 ***
## zheight                        4970       2380   2.089    0.037 *
## finalheights2$sex             -8819       1540  -5.725 1.31e-08 ***
## zheight:finalheights2$sex     -2175       1496  -1.454    0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18590 on 1188 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.08853
## F-statistic: 39.56 on 3 and 1188 DF,  p-value: < 2.2e-16
```

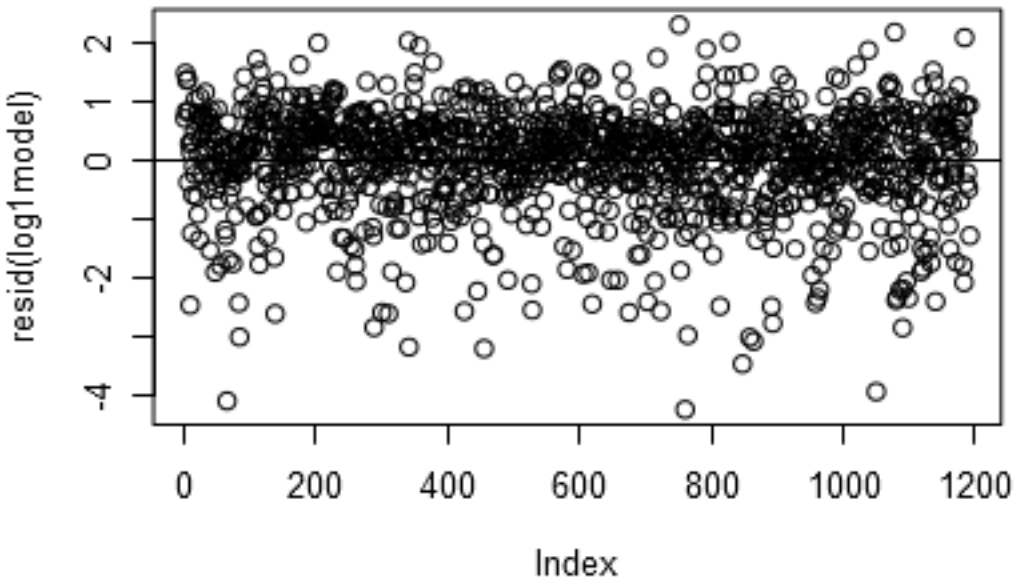
```r
plot(resid(z2model))
abline(h=0)
```



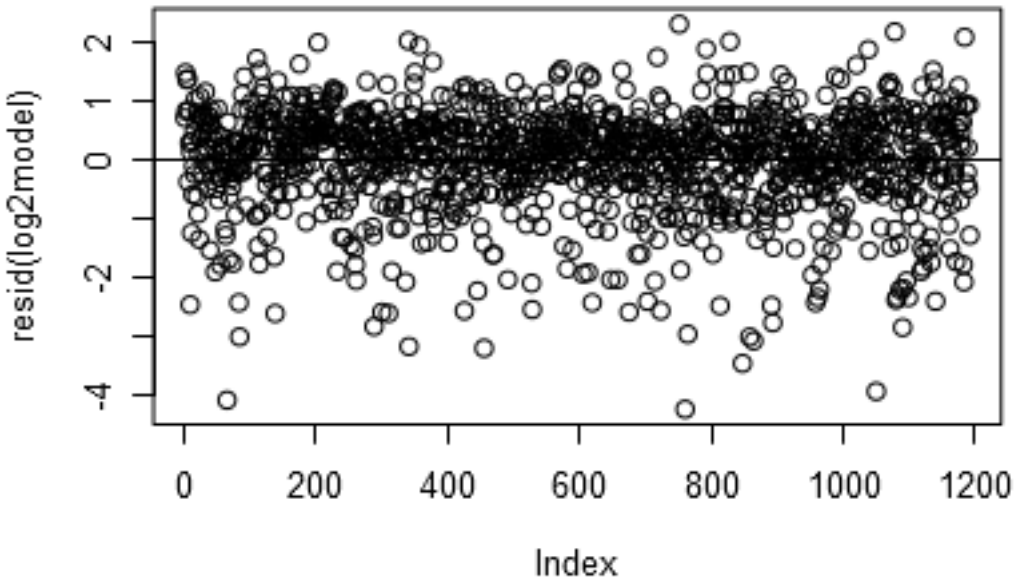

```r
# log transformation for log(earn)

log1model <- lm(log(finalheights2$earn) ~ finalheights2$height + finalheights2$sex)

summary(log1model)
```

```
##
## Call:
## lm(formula = log(finalheights2$earn) ~ finalheights2$height +
##     finalheights2$sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2384 -0.3718  0.1410  0.5649  2.3071
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.999125   0.708452  12.703  < 2e-16 ***
## finalheights2$height  0.020658   0.009312   2.218   0.0267 *
## finalheights2$sex    -0.423217   0.072462  -5.841 6.71e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8809 on 1189 degrees of freedom
## Multiple R-squared:  0.08656,    Adjusted R-squared:  0.08502
## F-statistic: 56.34 on 2 and 1189 DF,  p-value: < 2.2e-16
```

```r
plot(resid(log1model))
abline(h=0)
```




```r
# log transformation for log(earn) and log(height)

log2model <- lm(log(finalheights2$earn) ~ log(finalheights2$height) + finalheights2$sex)



plot(resid(log2model))
abline(h=0)
```



Based on the summary, log1model model would be better to select as a prederred model.Because it has relatively larger R square and larger F-statistics.

4. Interpret all model coefficients.

For log1model, the regression fn is $log(earning) = 8.999 - 0.42*sex + 0.02*height $ 1. When a female is in average height, she earns 8.999 more than average earning.

2. In the same height, a male's earning is 0.42 less that a female's earning.

3. Earning increases 0.02 when height is increased by 1 unit.

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

```
confint(log1model, level = 0.95)
```

```
##                              2.5 %      97.5 %
## (Intercept)           7.609170021 10.38907910
## finalheights2$height  0.002387267  0.03892842
## finalheights2$sex    -0.565385712 -0.28104860
```

We are 95% confident that [7.609, 10.389] contains the intercept coefficient. We are 95% confident that [0.002,0.038] contains the height coefficient. We are 95% confident that [-0.565, -0.281] contains the sex coefficient.

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.

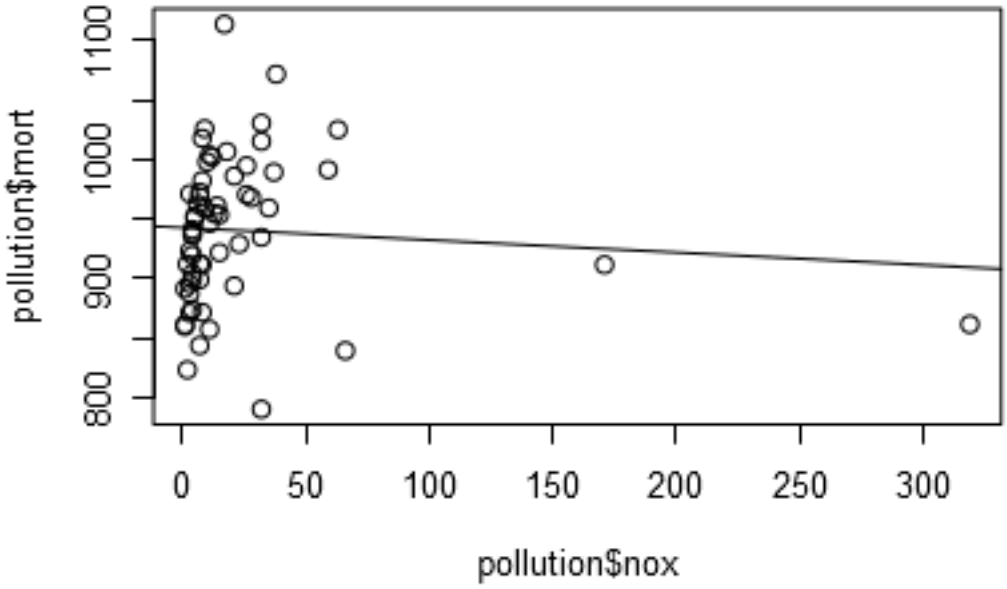
```
gelman_dir   <- "http://www.stat.columbia.edu/~gelman/arm/examples/"
pollution    <- read.dta (paste0(gelman_dir,"pollution/pollution.dta"))
```

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.

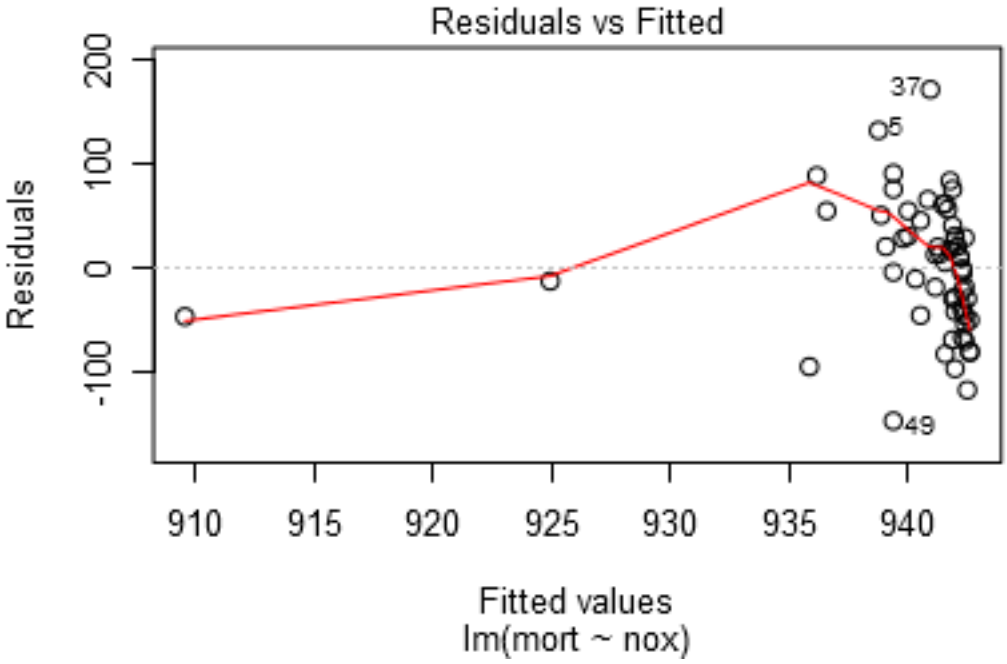
```
pmodel <- lm(mort ~ nox, data= pollution)
summary(pmodel)
```

```
##
## 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
```

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




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

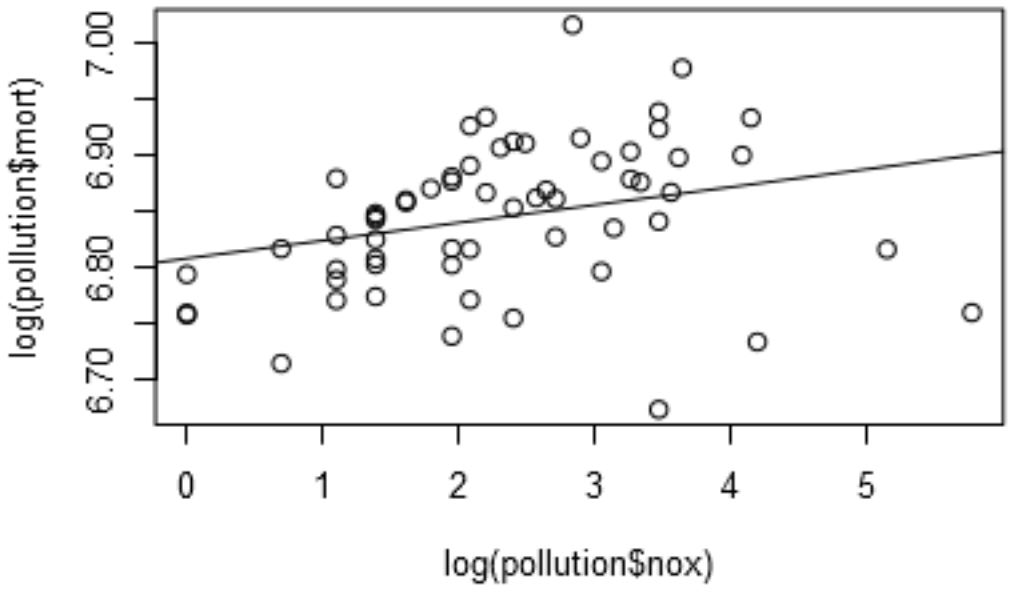


2. Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

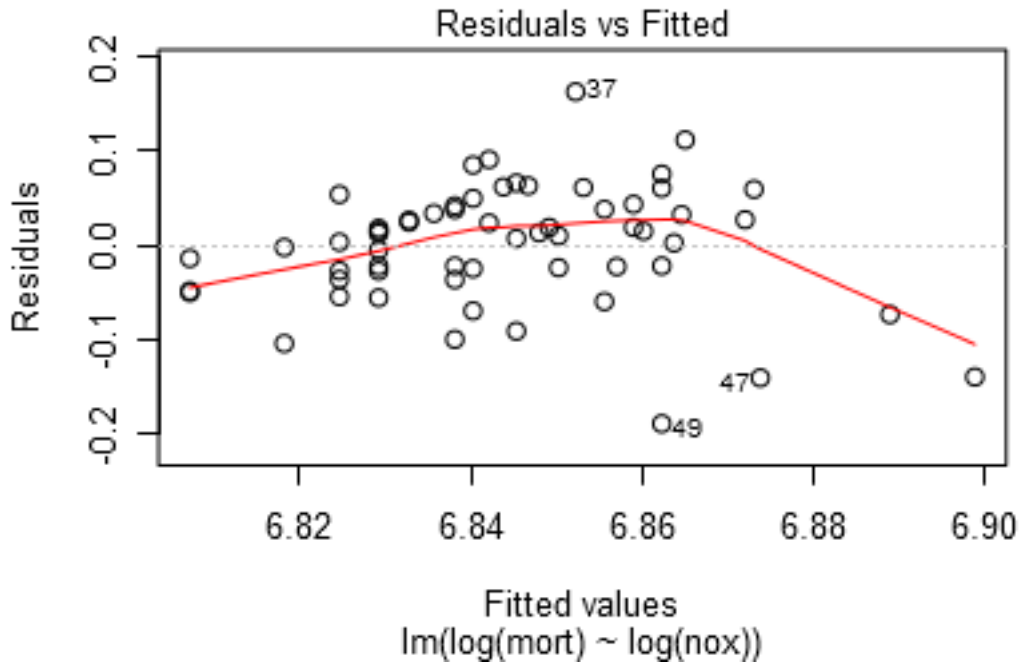
```r
newpmodel <- lm(log(mort)~ log(nox), data = pollution)
summary(newpmodel)
```

```
##
## Call:
## lm(formula = log(mort) ~ log(nox), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18930 -0.02957  0.01132  0.03897  0.16275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.807175   0.018349 370.975   <2e-16 ***
## log(nox)    0.015893   0.007048   2.255   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06412 on 58 degrees of freedom
## Multiple R-squared:  0.08061,    Adjusted R-squared:  0.06476
## F-statistic: 5.085 on 1 and 58 DF,  p-value: 0.02792
```

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




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



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

New model: $log(y) = 6.8 + 0.015 * log(x)$ for $y = mort, x = nox$ That is, $y = e^(6.8 + 0.015 * log(x))$

we can see that, when $nox = 1$, the log(y) has an unchange value 6.8. As log(x) increases by 1 unit, log(y) increases by 0.015 unit.

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

```
confint(newpmodel, parm = c("log(nox)"), level = 0.99)
```

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

We are 99% confident that log(nox) is in the range [-0.0028, 0.0346]. The range across zero from negative to positive, we cannot say it is a significant coefficient.

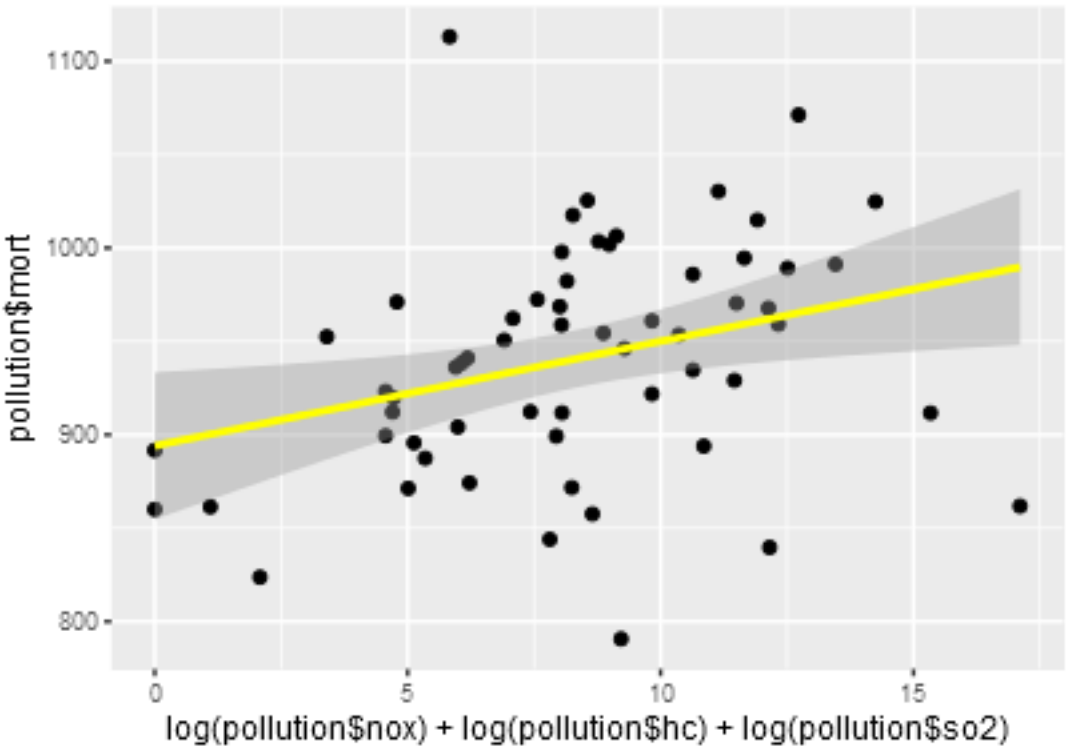
5. Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

```
pmodel.nsh <- lm(mort ~ log(nox) + log(hc) + log(so2), data = pollution)
summary(pmodel.nsh)
```

```
##
## Call:
## lm(formula = mort ~ log(nox) + log(hc) + log(so2), data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.793 -34.728  -3.118  34.148 194.567
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  924.965     21.449  43.125  < 2e-16 ***
## log(nox)      58.336     21.751   2.682  0.00960 **
## log(hc)      -57.300     19.419  -2.951  0.00462 **
## log(so2)      11.762      7.165   1.642  0.10629
## ---
## 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
```

```
ggplot(pmodel.nsh) + aes(y = pollution$mort, x = log(pollution$nox)+ log(pollution$hc) + log(pollution$s
```



We use log transformation for our model. $Y_mort = 924.965 + \beta_1 * log(nox) + \beta_2 * log(hc) + \beta_3 * log(so2)$

1 unit increase in log(nox) leads to 58.336 unit increases in mort. 1 unit increase in log(hc) leads to 57.300 unit decreases in mort. 1 unit increase in log(so2) leads to 11.762 unit increase in mort.

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.)

```
firstdata <- pollution[1:30,]
seconddata <- pollution[31:60,]

pmodel.nsh.half <- lm(mort ~ log(nox) + log(hc) + log(so2), data = firstdata)

predict(pmodel.nsh.half, newdata = seconddata, interval = "prediction" )
```

```
##         fit      lwr      upr
```

```
## 31 960.2959 846.8019 1073.790
## 32 880.7693 760.4481 1001.090
## 33 977.5782 866.3685 1088.788
## 34 944.1991 834.7796 1053.619
## 35 973.6112 862.4354 1084.787
## 36 932.8220 821.9672 1043.677
## 37 877.5513 731.7304 1023.372
## 38 971.8836 861.1764 1082.591
## 39 988.8598 874.9130 1102.807
## 40 998.3774 881.0720 1115.683
## 41 944.5429 829.8558 1059.230
## 42 946.5092 835.4347 1057.584
## 43 981.5553 866.8926 1096.218
## 44 964.4128 853.6797 1075.146
## 45 943.8154 832.5331 1055.098
## 46 960.8414 849.1453 1072.537
## 47 922.8955 793.7769 1052.014
## 48 951.3951 823.1784 1079.612
## 49 879.2809 732.8795 1025.682
## 50 933.6864 819.3189 1048.054
## 51 951.9992 839.9335 1064.065
## 52 951.0232 839.9610 1062.085
## 53 949.0132 839.5827 1058.444
## 54 931.6025 814.8003 1048.405
## 55 963.3640 849.6654 1077.063
## 56 883.0655 764.8971 1001.234
## 57 960.9191 851.2785 1070.560
## 58 923.0427 810.4831 1035.602
## 59 970.7047 857.3333 1084.076
## 60 961.0636 850.8359 1071.291
```

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.

```
centerstatus <- teengamb$status - mean(teengamb$status)
centerverbal <- teengamb$verbal - mean(teengamb$verbal)
lincome <- log(teengamb$income)

gammodel <- lm(log(teengamb$gamble + 1)~ teengamb$sex+ centerstatus + lincome + centerverbal, data =tee

summary(gammodel)
```

```
##
## Call:
## lm(formula = log(teengamb$gamble + 1) ~ teengamb$sex + centerstatus +
##     lincome + centerverbal, data = teengamb)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.67231 -0.56225  0.08561  0.80866  2.11944
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.22334    0.39911   3.065 0.003793 **
## teengamb$sex -1.10598    0.39296  -2.814 0.007404 **
## centerstatus  0.02430    0.01354   1.795 0.079887 .
## lincome       0.93798    0.23640   3.968 0.000278 ***
## centerverbal -0.25520    0.10704  -2.384 0.021713 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.118 on 42 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.4423
## F-statistic: 10.12 on 4 and 42 DF,  p-value: 7.906e-06
```

The model equation is $Y = 1.22334 + \beta_{sex} * -1.1059 + \beta_{status} * 0.0243 + \beta_{income} * 0.9379 + \beta_{verbal} * -0.2552$

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

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

```
##                     2.5 %      97.5 %
## (Intercept)   0.417908244  2.02876371
## teengamb$sex -1.898998675 -0.31295814
## centerstatus -0.003023852  0.05163342
## lincome       0.460910955  1.41504488
## centerverbal -0.471208407 -0.03918313
```

We are 95% confident that intercept is within the range [0.4179, 2.0287] We are 95% confident that the coefficient of sex is within the range [-1.8989, -0.3129] We are 95% confident that the coefficient of status is within the range [-0.0030, 0.0516] We are 95% confident that the coefficient of income is within the range [0.4609, 1.4150] We are 95% confident that the coefficient of verbal is within the range [-0.4712, -0.0391]

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?

```
averagestatus <- mean(teengamb$status)
averageincome <- mean(teengamb$income)
averageverbal <- mean(teengamb$verbal)

averagedata <- data.frame(averagestatus, averageincome,  averageverbal, sex =0)

averagepredit <- predict(object = gammodel, newdata = averagedata, level = 0.95, interval = "confidence
```

```
## Warning: 'newdata' had 1 row but variables found have 47 rows
```

```
maxstatus <- max(teengamb$status)
maxincome <- max(teengamb$income)
maxverbal <- max(teengamb$verbal)

maxdata <- data.frame(maxstatus, maxincome, maxverbal, sex =0)

maxpredit <- predict(gammodel, newdata = maxdata, level= 0.95, interval = "confidence")
```

```
## Warning: 'newdata' had 1 row but variables found have 47 rows
```

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.

```
centerratio <- sat$ratio - mean(sat$ratio)
logexpand <- log(sat$expend)
logsalary <- log(sat$salary)

satmodel <- lm(log(total) ~ logexpand + centerratio + logsalary, data = sat)
summary(satmodel)
```

```
##
## Call:
## lm(formula = log(total) ~ logexpand + centerratio + logsalary,
##     data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151468 -0.045848 -0.006242  0.046301  0.126984
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.840096   0.375170  20.897   <2e-16 ***
## logexpand    0.098837   0.137563   0.718   0.4761
## centerratio  0.007020   0.006977   1.006   0.3196
## logsalary   -0.323193   0.166493  -1.941   0.0584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0702 on 46 degrees of freedom
## Multiple R-squared:  0.2212, Adjusted R-squared:  0.1704
## F-statistic: 4.354 on 3 and 46 DF,  p-value: 0.008808
```

The regression model satmodel: $Y = 7.840096 + \beta_{logexpand} * 0.098837 + \beta_{centerratio} * 0.007020 + \beta_{logsalary} * -0.323193$

The log(total) remains 7.84 when the ratio in average ratio, expand =1 in log(expand) and salary = 1 in log(salary).

1. centerratio increases 1 unit leads to log(total) increases by 0.007020
2. logexpand increases 1 unit leads to log(total) increases by 0.098837
3. logsalary increases 1 unit leads to log(total) decreases by 0.323193
4. Construct 98% CI for each coefficient and discuss what you see.

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

```
##                      1 %       99 %
## (Intercept)  6.935865071 8.74432684
```

```
## logexpand   -0.232716751 0.43039028
## centerratio -0.009796151 0.02383659
## logsalary   -0.724471688 0.07808545
```

We are 98% confident that intercept of coefficient is within the range of [6.9358, 8.7443], so that coefficient is significant. We are 98% confident that coefficient of logexpand is within the range of [-0.2327, 0.4303] We are 98% confident that coefficient of centerratio is within the range of [-0.0097, 0.0238] We are 98% confident that coefficient of logsalary is within the range of [-0.7244, 0.0780]

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?

```
logtakers <- log(sat$takers)

addmodel <- lm(log(total) ~ logtakers + logexpand + centerratio + logsalary, data = sat)
summary(addmodel)
```

```
##
## Call:
## lm(formula = log(total) ~ logtakers + logexpand + centerratio +
##     logsalary, data = sat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060164 -0.016702  0.001594  0.013257  0.060659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.9035744  0.1514040  45.597   <2e-16 ***
## logtakers   -0.0833452  0.0049647 -16.787   <2e-16 ***
## logexpand    0.0668436  0.0516444   1.294    0.202
## centerratio  0.0005817  0.0026456   0.220    0.827
## logsalary    0.0318434  0.0659458   0.483    0.632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02634 on 45 degrees of freedom
## Multiple R-squared:  0.8928, Adjusted R-squared:  0.8832
## F-statistic: 93.66 on 4 and 45 DF,  p-value: < 2.2e-16
```

The new fitted model which adds the variable takers seems much better. We can see that the R-squared is much higher which is 0.8928. Also F-statistic is much larger compare to the previous model. And p-value is approximately equal to zero.

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, $D_i - R_i$

- The ratio, D_i/R_i
- The difference on the logarithmic scale, $logD_i - logR_i$
- The relative proportion, $D_i/(D_i + R_i)$.

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.