MA678 homework 01

Dave Anderson

Septemeber 10, 2018

Introduction

For homework 1 you will fit linear regression models and interpret them. You are welcome to transform the variables as needed. How to use `lm` should have been covered in your discussion session. Some of the code are written for you. Please remove `eval=FALSE` inside the knitr chunk options for the code to run.

This is not intended to be easy so please come see us to get help.

Data analysis

Pyth!

```
gelman_example_dir<-"http://www.stat.columbia.edu/~gelman/arm/examples/"
pyth <- read.table (paste0(gelman_example_dir,"pyth/exercise2.1.dat"),
                    header=T, sep=" ")
```

The folder pyth contains outcome `y` and inputs `x1`, `x2` for 40 data points, with a further 20 points with the inputs but no observed outcome. Save the file to your working directory and read it into R using the `read.table()` function.

1. Use R to fit a linear regression model predicting `y` from `x1`,`x2`, using the first 40 data points in the file. Summarize the inferences and check the fit of your model.

```
# First 40 data points
y <- pyth[1:40,1]
x1 <- pyth[1:40,2]
x2 <- pyth[1:40,3]


#Regression model
lmpyth <- lm(y~x1+x2)
lmpyth
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##      1.3151       0.5148       0.8069
```

```
#Check fit with Summary, R^2
summary(lmpyth)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.9585 -0.5865 -0.3356  0.3973  2.8548 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.31513    0.38769   3.392  0.00166 ** 
## x1           0.51481    0.04590  11.216 1.84e-13 ***
## x2           0.80692    0.02434  33.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9 on 37 degrees of freedom
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709 
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

Model fits well with intercept of 1.3151, slopes of 0.5148 and 0.8069 for x1 and x2 respectively, R^2 *of 0.97*

2. Display the estimated model graphically as in (GH) Figure 3.2.

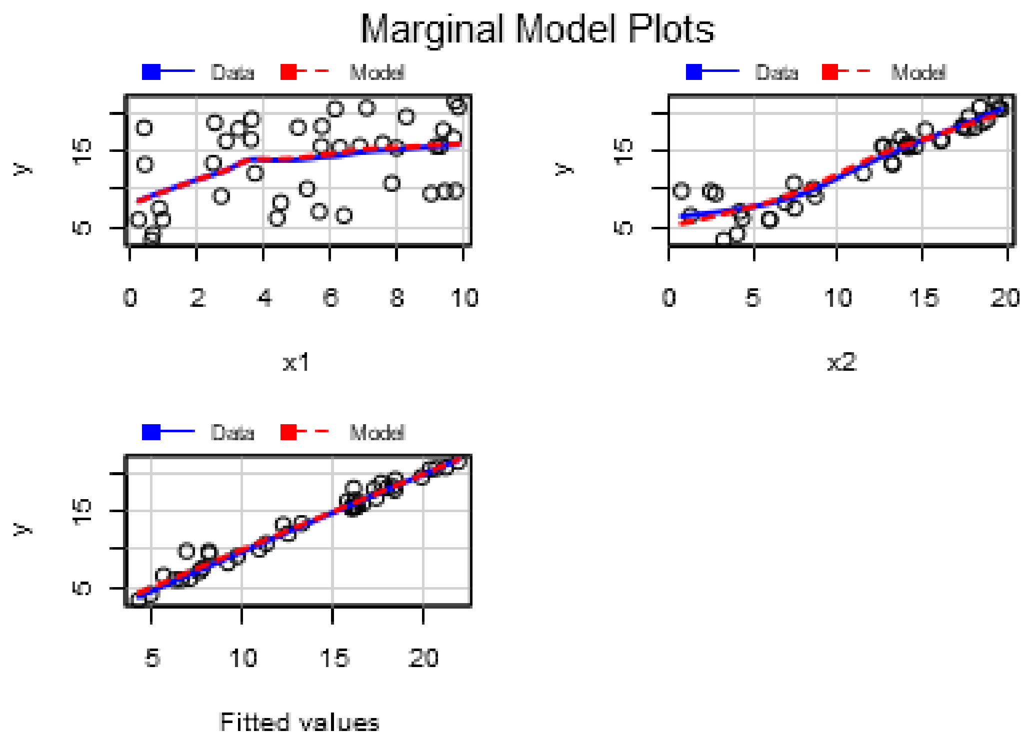
```r
#Drawing a regression plane made by x1 and x2
library(car)
```

```
## Loading required package: carData
```

```
## 
## Attaching package: 'car'
```

```
## The following object is masked from 'package:arm':
## 
##     logit
```

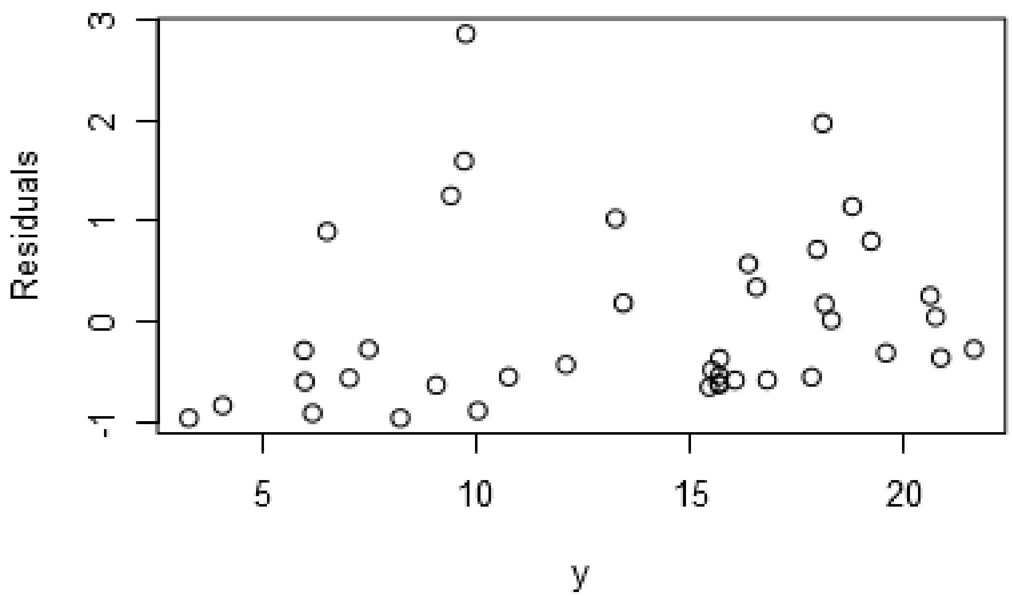
```r
par(mfrow=c(2,2))
car::marginalModelPlots(lmpyth)
```



3. Make a residual plot for this model. Do the assumptions appear to be met?

```
#Residual Plot
res <- resid(lmpyth)

plot(y, res, ylab="Residuals", xlab="y")
```




```r
#Residual plot is skewed
```

4. Make predictions for the remaining 20 data points in the file. How confident do you feel about these predictions?

```r
#Predictions
pythb <- pyth[41:60,]
pythpredict <- predict(lmpyth,pythb,level=0.95)

pythpredict
```

```
##        41        42        43        44        45        46        47 
## 14.812484 19.142865  5.916816 10.530475 19.012485 13.398863  4.829144 
##        48        49        50        51        52        53        54 
##  9.145767  5.892489 12.338639 18.908561 16.064649  8.963122 14.972786 
##        55        56        57        58        59        60 
##  5.859744  7.374900  4.535267 15.133280  9.100899 16.084900
```

After doing this exercise, take a look at Gelman and Nolan (2002, section 9.4) to see where these data came from. (or ask Masanao)

Earning and height

Suppose that, for a certain population, we can predict log earnings from log height as follows:

- A person who is 66 inches tall is predicted to have earnings of $30,000.
- Every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings.
- The earnings of approximately 95% of people fall within a factor of 1.1 of predicted values.

1. Give the equation the regression line and the residual standard deviation of the regression.

Answer 1 log(earning) = B0 + B1 log(height) B1 = .008/.01 = .8
B0 = log(30000) - 0.8 log(66) = 6.957229

$log(earning) = 6.957229 + .8 * log(height)$

Calculating residual standard deviation = Standard deviation in error of Beta

(1.1 - 1) * B1 = SD_Error * 1.96

0.1 * 0.8 / 1.96 = SD_Error = 0.0408

2. Suppose the standard deviation of log heights is 5% in this population. What, then, is the R^2 of the regression model described here?

```
Var_Error <- 0.0408^2
Var_Pop <- 0.05^2

R_squared <- 1- (Var_Error/Var_Pop)
```

$R^2 = 0.334144$

Beauty and student evaluation

The folder beauty contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, s
```

1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals versus fitted values.

```
#Both are continuos variables
x <- beauty.data$btystdave
y <- beauty.data$courseevaluation
lmbeaut <- lm(y~x)
lmbeaut
```

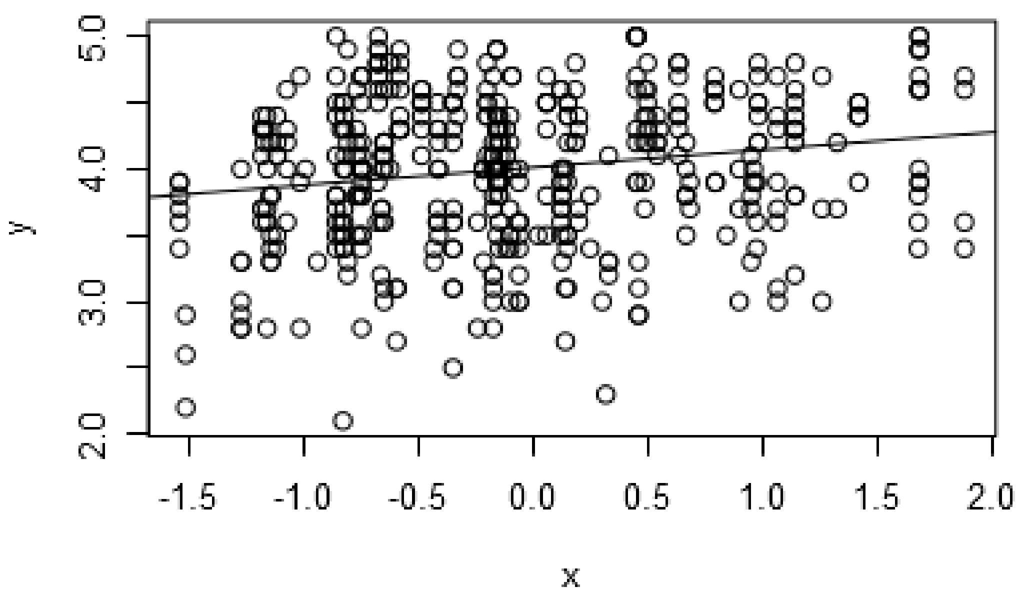
```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       4.010        0.133
```

```
summary(lmbeaut)
```

```
##
```

```
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## x            0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

```r
plot(x,y,abline(lmbeaut))
```

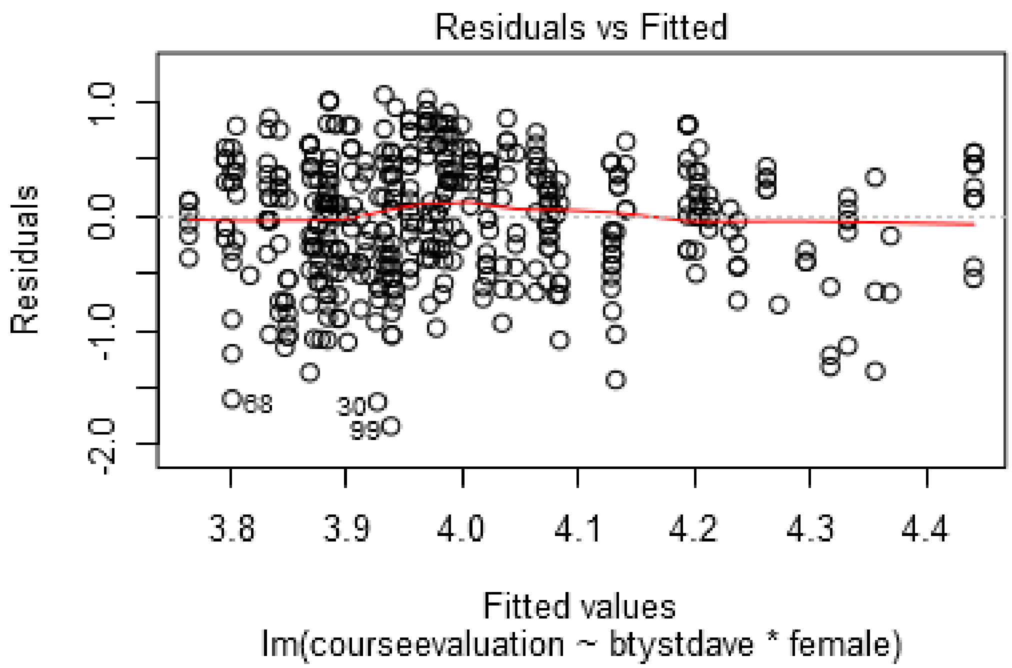


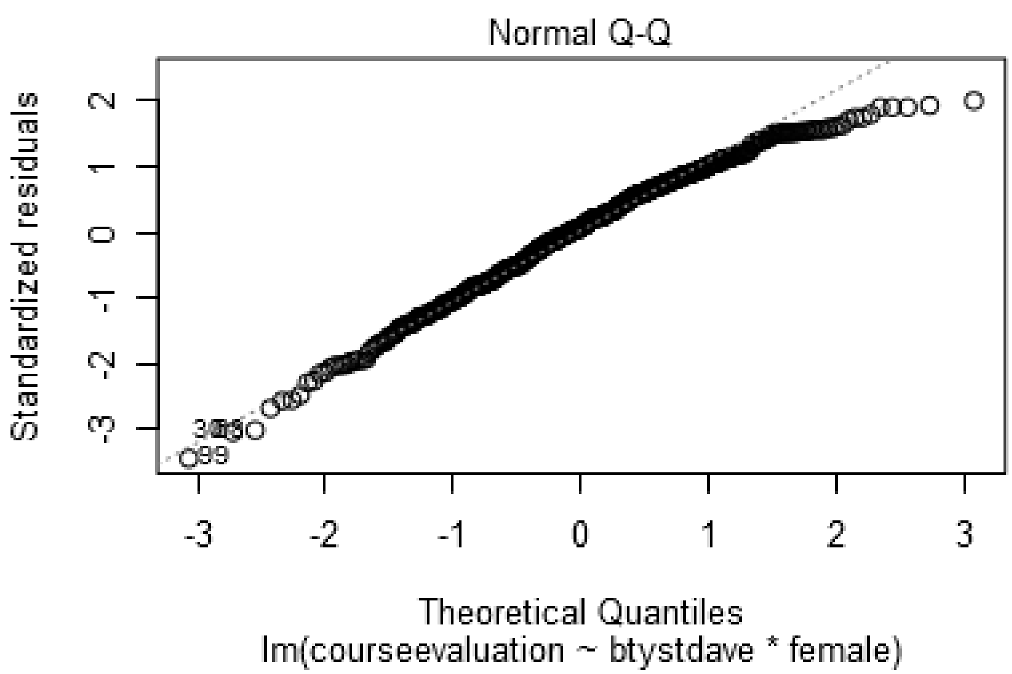
Intercept is 4 as in someone with a "beauty" of 0 (sad) will have an evaluation score of 4. Slope is 0.133 as in every added unit of "beauty" will produce a 0.133 higher evaluation score.

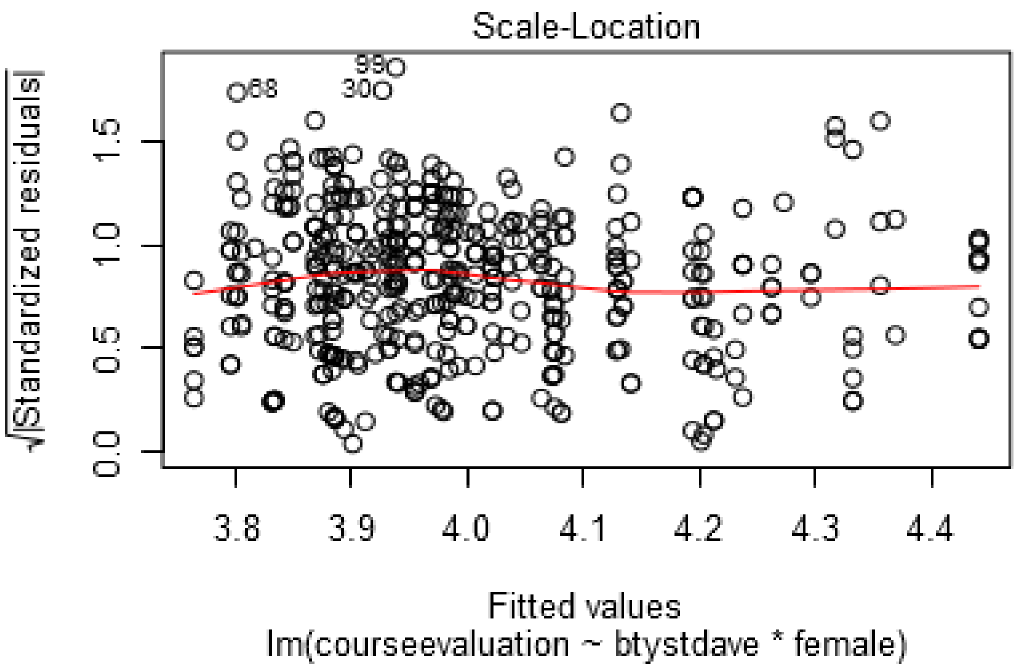
$R^2 = 0.03$ ##R^2 is very small. As we can see from the plot, the data is very wide-spread.

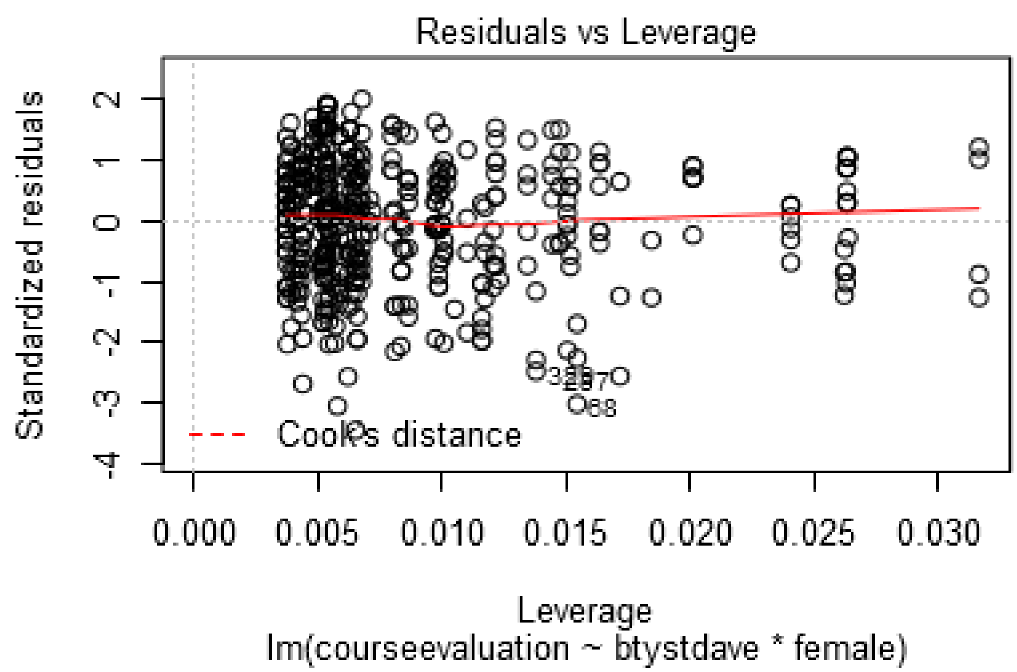
2. Fit some other models, including beauty and also other input variables. Consider at least one model with interactions. For each model, state what the predictors are, and what the inputs are, and explain the meaning of each of its coefficients.

```r
beauty_lm2 <- lm(courseevaluation ~ btystdave * female , data = beauty.data )
plot(beauty_lm2)
```









See also Felton, Mitchell, and Stinson (2003) for more on this topic link

Conceptula excercises

On statistical significance.

Note: This is more like a demo to show you that you can get statistically significant result just by random chance. We haven't talked about the significance of the coefficient so we will follow Gelman and use the approximate definition, which is if the estimate is more than 2 sd away from 0 or equivalently, if the z score is bigger than 2 as being "significant".

(From Gelman 3.3) In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.

1. First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing in R. Generate another variable in the same way (call it var2).

```
set.seed(2018)
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
```

Run a regression of one variable on the other. Is the slope coefficient statistically significant? [absolute value of the z-score(the estimated coefficient of var1 divided by its standard error) exceeds 2]

```
fit  <- lm (var2 ~ var1)
fit
z.scores <- coef(fit)[2]/se.coef(fit)[2]
z.scores
```

The absolute value of my z-score does not exceed 2, therefore not statistically significant

2. Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of var1 divided by its standard error). If the absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation:

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit  <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}

z.scores
sum( abs(z.scores) > 2 )
```

How many of these 100 z-scores are statistically significant?
3 values are statistically significant What can you say about statistical significance of regression coefficient?

If the estimate of the coefficient is more than 2 standard deviations away from the true coefficient value, it is statistically significant. In this situation with random data points, the coefficient is rarely significant.

Fit regression removing the effect of other variables

Consider the general multiple-regression equation

$$Y = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k + E$$

An alternative procedure for calculating the least-squares coefficient B_1 is as follows:

1. Regress Y on X_2 through X_k, obtaining residuals $E_{Y|2,\dots,k}$.
2. Regress X_1 on X_2 through X_k, obtaining residuals $E_{1|2,\dots,k}$.
3. Regress the residuals $E_{Y|2,\dots,k}$ on the residuals $E_{1|2,\dots,k}$. The slope for this simple regression is the multiple-regression slope for X_1 that is, B_1.

(a) Apply this procedure to the multiple regression of prestige on education, income, and percentage of women in the Canadian occupational prestige data (http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/Prestige.pdf), confirming that the coefficient for education is properly recovered.

```
fox_data_dir<-"http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/"
Prestige<-read.table(paste0(fox_data_dir,"Prestige.txt"))

#Regression of prestige on variables except education
lmprestige <- lm(prestige ~ income + women,Prestige)
lmprestige
```

```
##
## Call:
## lm(formula = prestige ~ income + women, data = Prestige)
##
## Coefficients:
## (Intercept)       income        women
##   20.326775     0.003334     0.132623
```

```
summary(lmprestige)
```

```
##
## Call:
## lm(formula = prestige ~ income + women, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.037  -7.109  -1.560   6.464  36.302
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.033e+01  2.996e+00   6.785 8.58e-10 ***
## income      3.334e-03  3.012e-04  11.067  < 2e-16 ***
## women       1.326e-01  4.032e-02   3.289  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.54 on 99 degrees of freedom
## Multiple R-squared:  0.5593, Adjusted R-squared:  0.5504
## F-statistic: 62.81 on 2 and 99 DF,  p-value: < 2.2e-16
```

```
residuals1 <- resid(lmprestige)

# Regression of education on variables except prestige
lmeducation <- lm(education ~ income + women,Prestige)
residuals2 <- resid(lmeducation)
#Regression of residuals
lmresid <- lm(residuals1~residuals2)
lmresid
```

```
##
```

```
## Call:
## lm(formula = residuals1 ~ residuals2)
##
## Coefficients:
## (Intercept)   residuals2
##  -2.992e-15    4.187e+00
```

```
#B1 = 4.187

#Regression using all the variables
lm_all <- lm(prestige ~ education + income + women, Prestige)
lm_all
```

```
##
## Call:
## lm(formula = prestige ~ education + income + women, data = Prestige)
##
## Coefficients:
## (Intercept)    education       income        women
##   -6.794334     4.186637     0.001314    -0.008905
```

```
#We get the same B1=4.187
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case?
The two sets of residuals of regressing both education and prestige on other variables should average each other out, as we are looking at the correlation of how education effects prestige, without actually fitting a model on the two.

(c) In light of this procedure, is it reasonable to describe B_1 as the "effect of X_1 on Y when the influence of $X_2, \cdots, X_k$ is removed from both X_1 and Y"?
Yes, I believe it is. We fit both variables against all other variables, took an average of their residuals, and still saw the same effect of B1 when all other variables were accounted for.

(d) The procedure in this problem reduces the multiple regression to a series of simple regressions (in Step 3). Can you see any practical application for this procedure?

Partial correlation

The partial correlation between X_1 and Y "controlling for" $X_2, \cdots, X_k$ is defined as the simple correlation between the residuals $E_{Y|2,\ldots,k}$ and $E_{1|2,\ldots,k}$, given in the previous exercise. The partial correlation is denoted $r_{y1|2,\ldots,k}$.

1. Using the Canadian occupational prestige data, calculate the partial correlation between prestige and education, controlling for income and percentage women.

```
#To find correlation between the residuals
cor(residuals1,residuals2)
```

```
## [1] 0.7362604
```

```
#The partial correlation between prestige and education is 0.736
```

2. In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{y1|2,\ldots,k} = 0$ if and only if B_1 is 0? *The correlation will be 0 if and only if the coefficient of x1 is 0 because both measures show no effect of x1 on y*

Mathematical exercises.

Prove that the least-squares fit in simple-regression analysis has the following properties:

1. $\sum \hat{y}_i \hat{e}_i = 0$

$\sum \hat{y}_i \hat{e}_i = \hat{Y}^T(Y - \hat{Y}) = [X(X^TX)^{-1}X^TY]^T(Y - X(X^TX)^{-1}X^TY)$

$= (HY)^T(Y - HY) = Y^THY - Y^TH^THY$

$= Y^THY - Y^THY = 0$

2. $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum \hat{e}_i(\hat{y}_i - \bar{y}) = 0$

We know that $\sum(y_i - \hat{y}_i)$ is our definition of the error estimates.

Suppose that the means and standard deviations of $\boldsymbol{y}$ and $\boldsymbol{x}$ are the same: $\bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}$ and $sd(\boldsymbol{y}) = sd(\boldsymbol{x})$.

1. Show that, under these circumstances

$$\beta_{y|x} = \beta_{x|y} = r_{xy}$$

   where $\beta_{y|x}$ is the least-squares slope for the simple regression of $\boldsymbol{y}$ on $\boldsymbol{x}$, $\beta_{x|y}$ is the least-squares slope for the simple regression of $\boldsymbol{x}$ on $\boldsymbol{y}$, and r_{xy} is the correlation between the two variables. Show that the intercepts are also the same, $\alpha_{y|x} = \alpha_{x|y}$.

$\bar{y} = \beta_0 + \beta_{x|y}\bar{x}$ $\beta_{x|y} = (\bar{y} - \beta_0)/\bar{x}$

similarly

$\beta_{y|x} = (\bar{x} - \beta_0)/\bar{y}$

Since $\bar{x} = \bar{y}$, *the two are equivalent*

2. Why, if $\alpha_{y|x} = \alpha_{x|y}$ and $\beta_{y|x} = \beta_{x|y}$, is the least squares line for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ different from the line for the regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ (when $r_{xy} < 1$)?

3. Imagine that educational researchers wish to assess the efficacy of a new program to improve the reading performance of children. To test the program, they recruit a group of children who are reading substantially vbelow grade level; after a year in the program, the researchers observe that the children, on average, have imporved their reading performance. Why is this a weak research design? How could it be improved?

This is a weak design because they chose only students who read below grade level. They have the most room to imrpove, so their scores should increase from a year. This design is fine if the program is made only for poor-achieving students, but if they want to apply the program to all students, they should include students of all levels.

Feedback comments etc.

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

I no longer have access to the bootcamp tutorials. Is it possible to open those back up for review?