MA678 homework 01

Chaoqun Yin

Septemeber 16, 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.

```
pyth_40 = pyth[1:40,]
y = pyth_40$y
x1 = pyth_40$x1
x2 = pyth_40$x2
regout1 = lm(y~ x1 + x2)
summary(regout1)
```

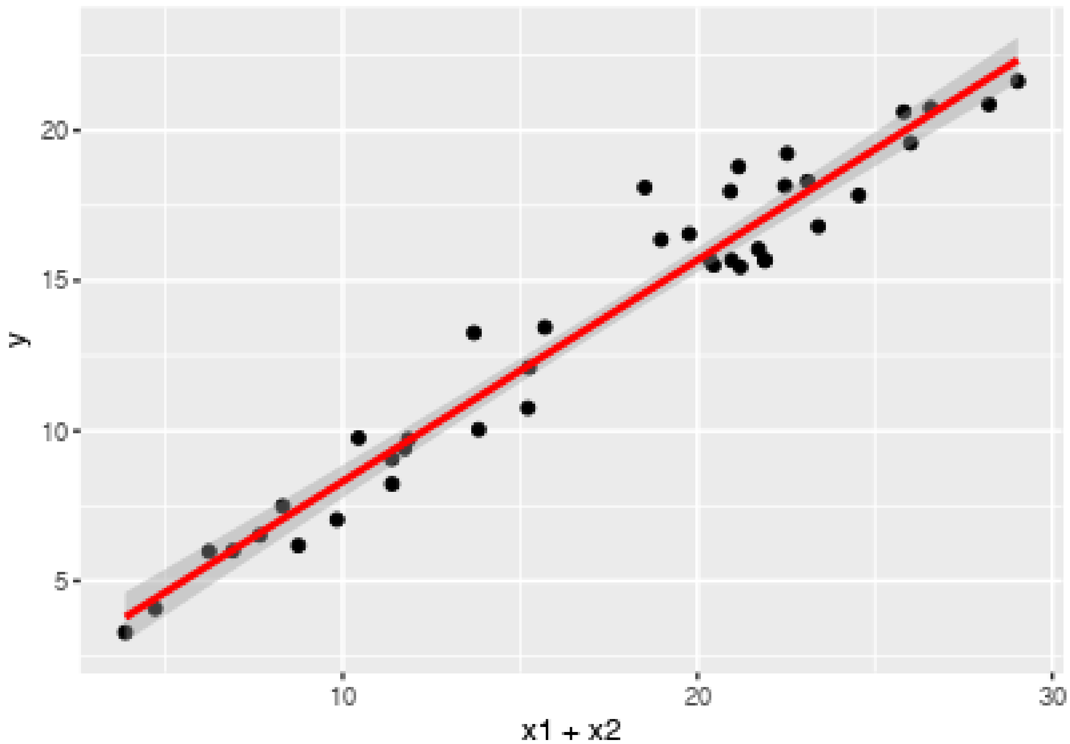
```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           0.51481    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
```

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

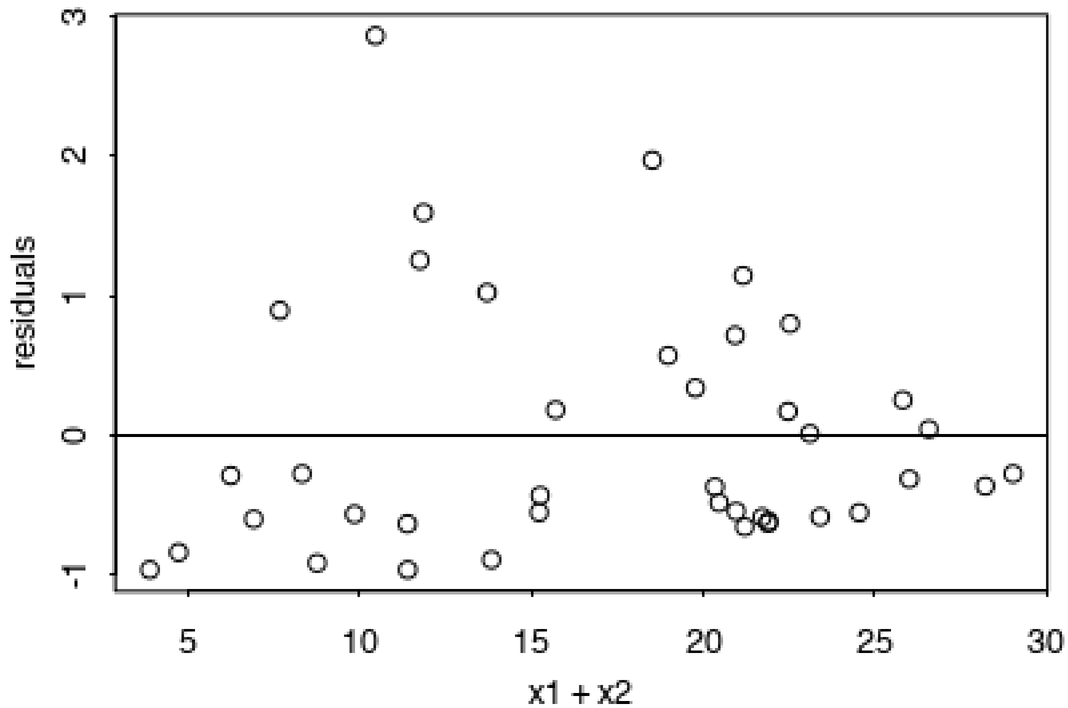
```r
library(ggplot2)
pyth_gp = ggplot(pyth_40)

pyth_gp + aes(x=x1 + x2, y)+ geom_point()+
  stat_smooth(method = "lm", col = "red")
```



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

```r
par (mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
plot(pyth_40$x1 + x2,regout1$residuals, xlab = "x1 + x2", ylab="residuals")
abline(h=0)
```



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

```
new <- pyth[41:60,]
pred = predict(lm(y ~ x1 + x2), new, se.fit = TRUE)
pred
```

```
## $fit
##        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
##
## $se.fit
##        41        42        43        44        45        46        47
## 0.2551744 0.2655255 0.3520552 0.2530248 0.2522090 0.1446663 0.2816065
##        48        49        50        51        52        53        54
## 0.2923014 0.2858877 0.2840251 0.2388084 0.1605835 0.2235294 0.2226090
##        55        56        57        58        59        60
## 0.2632822 0.2523755 0.2936852 0.1559494 0.2203609 0.2410461
##
## $df
## [1] 37
##
## $residual.scale
## [1] 0.9000333
```

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 of the regression line and the residual standard deviation of the regression.

The equation of the regression line can be expressed as (1) log(y) = B1log(x) + B0. Then from the information in the questions, we can easily get that: (2) log(30000) = B1log(66) + B0 (3) Log(1.008y) = B1log(1.01x) + B0

Then calculate the equations of (1)(2)(3), we can get that B1 = log(1.008)/log(1.01) = 0.80. Substitute B1 into equation (2), we can get that B0 = 6.96.Therefore the equation of the regression line is

$$\lg y = 0.8 \cdot lgx + 6.96$$

.From the question we can know that

$$2\sigma = lg1.1$$

, so

$$\sigma = 0.0477$$

.

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?

Beauty and student evaluation

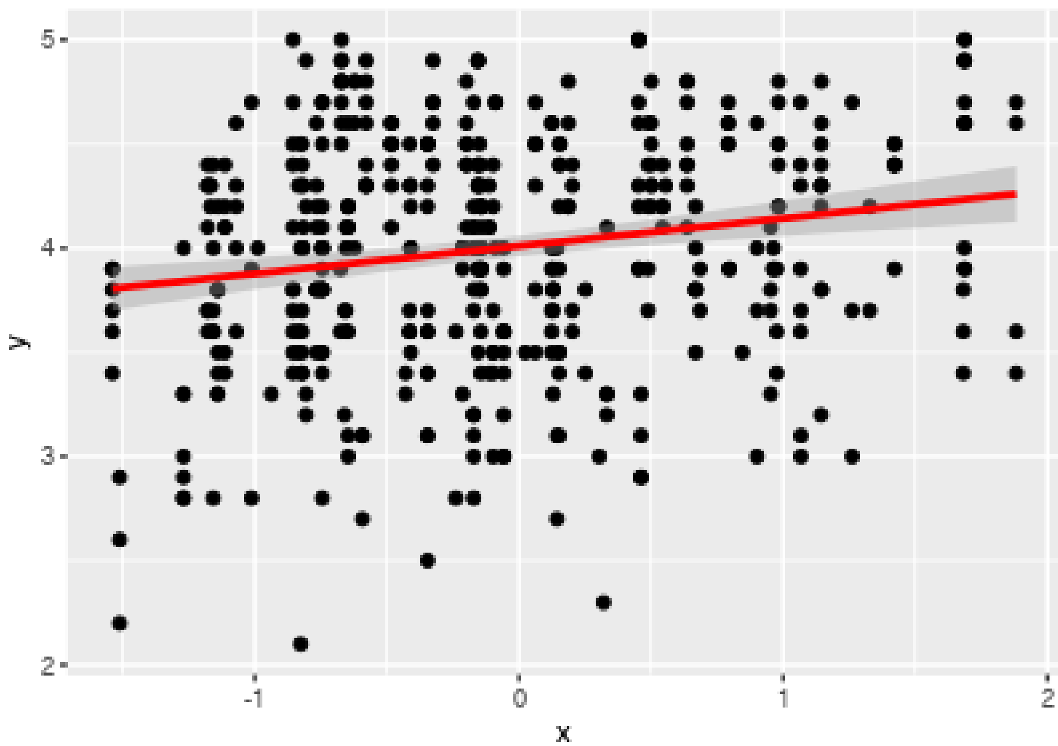
The folder beauty contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students 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, sep=",")
```

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.

```
#Use linear regression model among the beauty and the course evaluation
y = beauty.data$courseevaluation
x = beauty.data$btystdave
regout3 = lm(y ~ x, beauty.data)

#Make the plot of resuduals versus fitted values
beauty.data_gp = ggplot(beauty.data)
beauty.data_gp + aes(x, y)+ geom_point()+
  stat_smooth(method = "lm", col = "red")
```



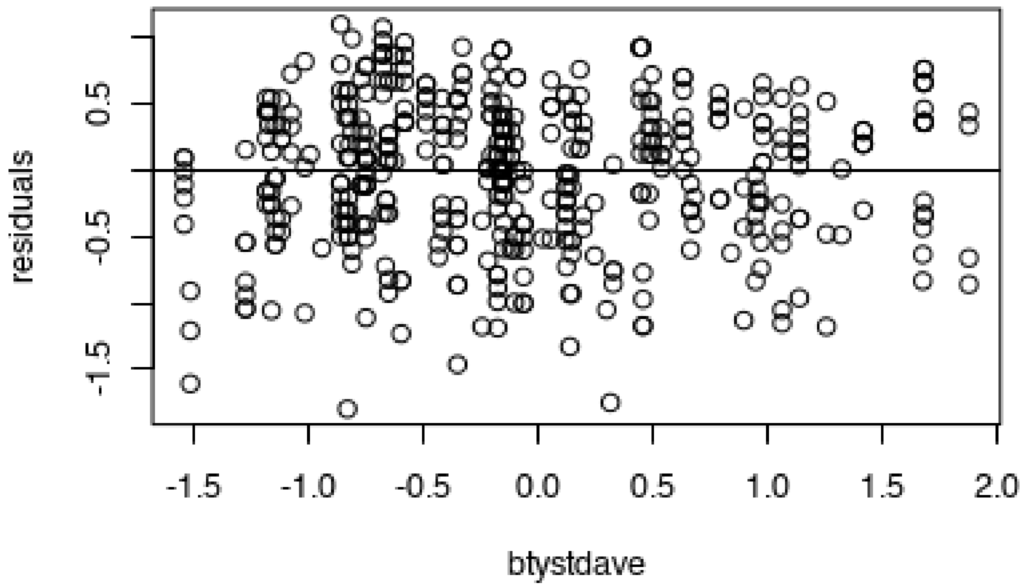
Explain the meaning of each of the coefficients:

```
coef(regout3)
```

```
## (Intercept)           x
##   4.0100227   0.1330014
```

It says the btystdave and the course evaluation is positively corelated. When the btystdave increases by 1, the course evaluation increase by 0.133. When the btystdave is 0, the course evaluation is 4.010.

```
plot(beauty.data$btystdave,regout3$residuals, xlab = "btystdave", ylab="residuals")
abline(h=0)
```



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.

The regression model uses beauty (the variable btystdave) to predict professor evaluations (profevaluation), controlling for various other inputs.

```
y = beauty.data$profevaluation
x = beauty.data$btystdave
regout4 = lm(y ~ x, beauty.data)
coef(regout4)
```

```
## (Intercept)           x
##   4.1859008   0.1264396
```

It says the btystdave and the professor evaluation is positively corelated. When the btystdave increases by 1, the professor evaluation increase by 0.126. When the btystdave is 0, the professor evaluation is 4.186.

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).

```
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)
z.scores <- coef(fit)[2]/se.coef(fit)[2]
z.scores
```

```
##      var1
## -1.471044
```

```
#The z.scores is less than 2, so I think it is not statistically significiant.
```

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]
}
```

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

```
length(z.scores[z.scores > 2]) + length(z.scores[z.scores < -2])
```

```
## [1] 9
```

Among all the values of z-scores, there are 3 of them more than 2. Therefore we can say it is statistically significant of the regression coefficient.

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 $E_{Y|2,\ldots,k}$.
2. Regress X_1 on X_2 through X_k, obtaining residuals $E_{1|2,\ldots,k}$.
3. Regress the residuals $E_{Y|2,\ldots,k}$ on the residuals $E_{1|2,\ldots,k}$. 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"))

# 0 Pick up data for the variables
y <- Prestige$prestige
x1 <- Prestige$education
x2 <- Prestige$income
x3 <- Prestige$women

# 1 Regress prestige on income and percentage of women
regout5 <- lm(y ~ x2 + x3, Prestige)
residual.1 <- regout5$residuals

# 2 Regress education on income and percentage of women
regout6 <- lm(x1 ~ x2 + x3, Prestige)
residual.2 <- regout6$residuals

# 3 Regress the residuals gotten from section 1 on the residuals gotten from section 2
regout7 <- lm(residual.1 ~ residual.2)
coef(regout7)[2]
```

```
## residual.2
##   4.186637
```

```
# 4 Regress prestige on education, income and percentage of women
regout8 <-lm(y ~ x1 + x2 + x3, Prestige)
coef(regout8)[2]
```

```
##       x1
## 4.186637
```

```
# So we can know the B1 = 4.19, the coefficient for education is properly recovered.
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case?

```
coef(regout7)[1]
```

```
##   (Intercept)
## -2.991663e-15
```

```
# As we run the simple regressions seperately with education and without education,
#the other variables like income and percantage of women have been fully controlled.
#So the intercept of the regression of residual.1 on residual.2 becomes 0.
```

(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"?

```
#I think it is reasonable to describe it in this way, because we have controlled the
#other inputs to get the reasonable outcome for the value of B1.
```

(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?

```
#I can imagine when we meet a huge set of data and we just need to come up the correlation
#of two variables, we can use this method to get the result more efficient and convenient.
```

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.

```
# 0 Pick up data for the variables
y <- Prestige$prestige
x1 <- Prestige$education
x2 <- Prestige$income
x3 <- Prestige$women

# 1 Regress prestige on income and percentage of women
regout8 <- lm(y ~ x2 + x3, Prestige)
residual.3 <- regout8$residuals

# 2 Regress education on income and percentage of women
regout9 <- lm(x1 ~ x2 + x3, Prestige)
residual.4 <- regout9$residuals

# 3 Correlation between two residuals
(Correlation <- cor(residual.3, residual.4))
```

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

```
#The partial correlation 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?

$$\epsilon_1 = E_{y|2,3}$$

$$\epsilon_2 = E_{1|2,3}$$

$$cor(\epsilon_1, \epsilon_2) = E[(\epsilon_1 - 0)(\epsilon_2 - 0)] = E(\epsilon_1 \epsilon_2)$$

$$\rho_{\epsilon_1,\epsilon_2} = \frac{cor(\epsilon_1 \epsilon_2)}{\sigma_1 \sigma_2} = \frac{E(\epsilon_1 \epsilon_2)}{\sigma_1 \sigma_2} = \frac{E(B_1 \epsilon_2^2)}{\sigma_1 \sigma_2} = \frac{B_1 E(\epsilon_2^2)}{\sigma_1 \sigma_2}$$

So there must be that

$$\beta_1 E(\epsilon_2^2) = 0$$

. Therefore B_1 is 0.

Mathematical exercises.

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

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

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, $\alpha_{y|x} = \alpha_{x|y}$.

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?

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.