MA678 homework 01

Xinyi Wang

Septemeber 6, 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.

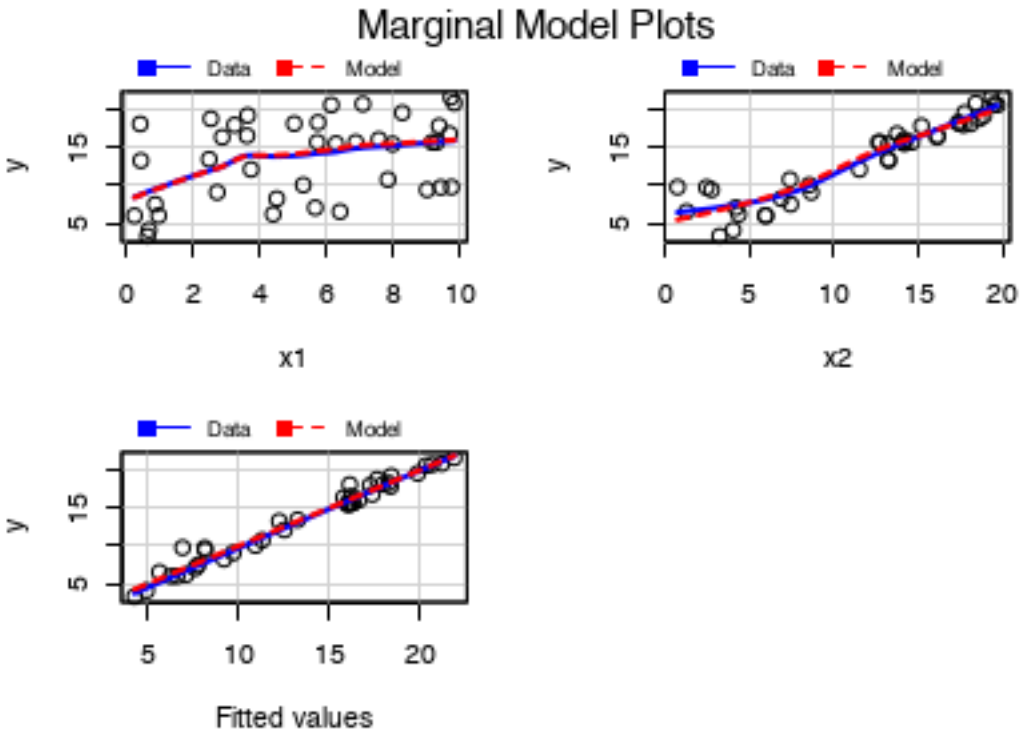
```
pyth.data=pyth[1:40,]
pyth.lm = lm(y~x1+x2, data=pyth.data)
summary(pyth.lm)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth.data)
##
## 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
```

Since R2 = 0.97 which very close to 1, we can see that roughly 97% of the variance found in the response variable(y) can be explained by the predictor variable(x1+x2), the model is fitting good to the actual data.

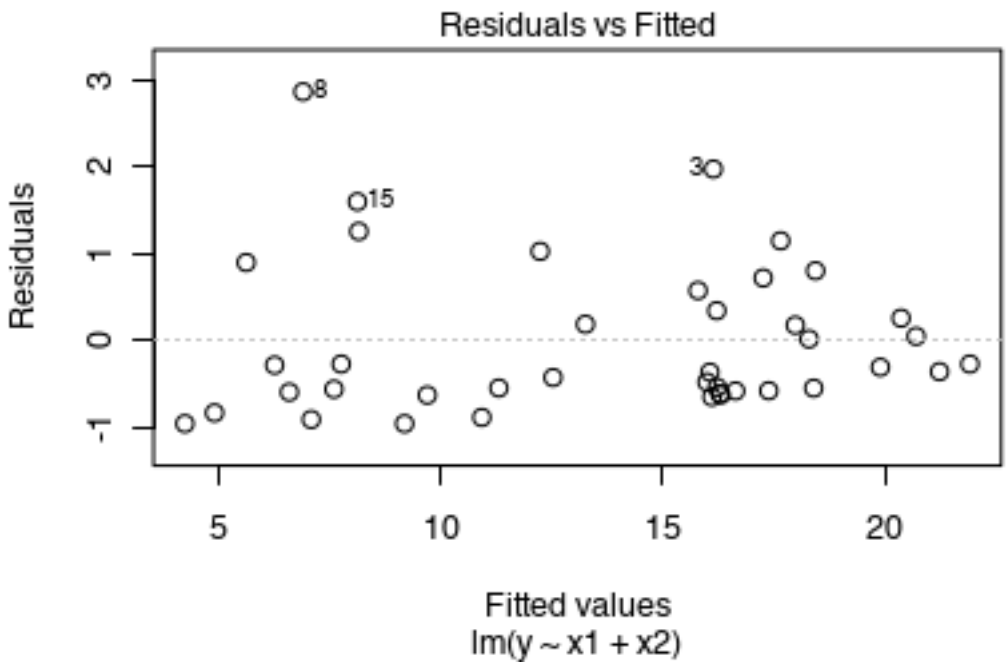
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
# library(ggplot2)
# ggplot(pyth.data) + aes(x=x1+x2,y) + geom_point() + stat_smooth(method="lm",col="red",se=FALSE)
car::marginalModelPlots(pyth.lm)
```



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

```
# pyth.res = resid(pyth.lm)
# plot(pyth.data$x1+pyth.data$x2, pyth.res, ylab="residuals", xlab="x1+x2", main="Residual Plot")
# abline(0,0)
plot(pyth.lm,which=1,add.smooth=FALSE)
```



The assumptions appears not met since those points are not distributed and tending to cluster towards the middle of the plot.

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

```
predict_data = pyth[41:60,2:3]
predict(pyth.lm,predict_data)
```

```
##        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
```

95% confident about these predictions.

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.

```r
#Using equation log(earing)=a+b*log(height)
a = log(30000) - (0.008/0.01)*log(66)
b = 0.008/0.01
a
```

```
## [1] 6.957229
```

```r
b
```

```
## [1] 0.8
```

```r
res_sd = 0.1 * 0.5/0.95
res_sd
```

```
## [1] 0.05263158
```

Equation of the regression is : log(earing) = 6.957 + 0.8*log(height) + e Residual standard deviation of the regression is 0.05 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?

```r
r2 = 1 - (res_sd^2 / 0.05^2)
r2
```

```
## [1] -0.1080332
```

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.

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

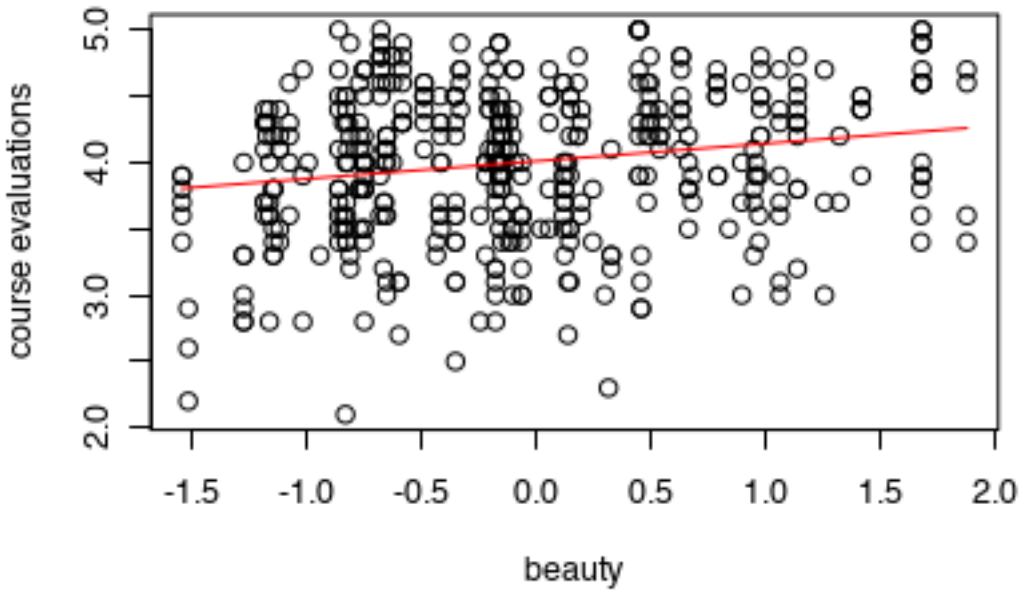
1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals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

```r
beauty.lm = lm(courseevaluation~btystdave, beauty.data)
summary(beauty.lm)
```

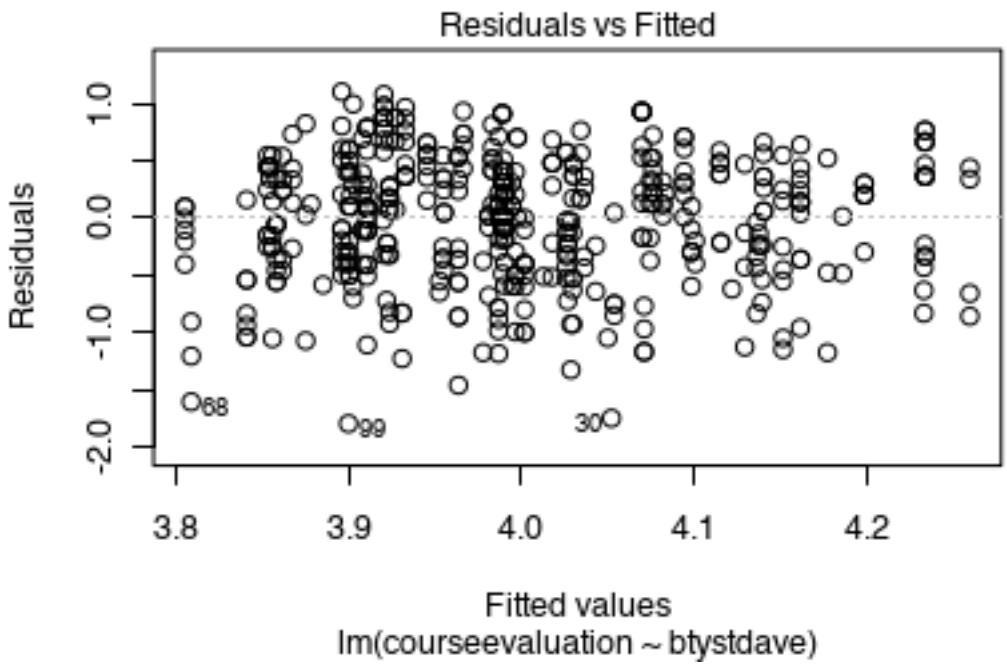
```
##
## Call:
## lm(formula = courseevaluation ~ btystdave, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## btystdave    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
library(ggplot2)
plot(beauty.data$btystdave,beauty.data$courseevaluation,xlab="beauty",ylab="course evaluations")
curve(coef(beauty.lm)[1]+coef(beauty.lm)[2]*x, add=TRUE, col="red")
```




```r
plot(beauty.lm,which=1,add.smooth=FALSE)
```



fitted model:course evaluation = 4.01 + 0.133*btystdave + e.

The intercept 4.01, is the mean of course evaluation that instructors' beauty equal to 0. The slope 0.133, means if beauty differed by 1 unit, course evaluation will differ by 0.133 units, on average. Residual standard deviation is 0.5455, which means the actual course evaluation can deviate from the true regression line by approximately 0.5455, on average.

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.

```
#model 1:
model1.lm = lm(courseevaluation ~ btystdave+age+btystdave*age, beauty.data)
summary(model1.lm)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + age + btystdave *
##     age, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74828 -0.36705  0.03469  0.41307  1.15642
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.960513   0.131992  30.006  < 2e-16 ***
## btystdave     -0.339163   0.149575  -2.268  0.02382 *
## age            0.001540   0.002715   0.567  0.57076
```

```
## btystdave:age  0.010150   0.003127   3.246  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5405 on 459 degrees of freedom
## Multiple R-squared:  0.05739,    Adjusted R-squared:  0.05123
## F-statistic: 9.316 on 3 and 459 DF,  p-value: 5.451e-06
```

Predictors:btystdave,age,btystdave*age ; Inputs:btystdave,age

fitted model:course evaluation = 3.96 - 0.339*btystdave* + *0.0015*age + 0.01*btystdave*age + e.

The first slope -0.339, the effect of beauty on course evaluation is -0.339+0.01*age. The second slope 0.0015, the effect of age on course evaluation is -0.339*beauty+0.0015. 0.01 is the slopes of the regression lines between course evaluation and beauty are different for the different age.

```
#model 2:
beauty = beauty.data$btystdave
age = beauty.data$age
model2.lm = lm(courseevaluation ~ beauty+age, beauty.data)
summary(model2.lm)
```

```
##
## Call:
## lm(formula = courseevaluation ~ beauty + age, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80242 -0.36514  0.07407  0.39913  1.10206
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.9962457  0.1328883  30.072  < 2e-16 ***
## beauty      0.1340634  0.0337441   3.973 8.24e-05 ***
## age         0.0002868  0.0027148   0.106    0.916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.546 on 460 degrees of freedom
## Multiple R-squared:  0.03576,    Adjusted R-squared:  0.03157
## F-statistic:  8.53 on 2 and 460 DF,  p-value: 0.0002305
```

Predictors:btystdave,age ; Inputs:btystdave,age

fitted model:course evaluation = 3.996 + 0.134*btystdave* + *0.00028*age +e.

The intercept 3.996, is the mean of course evaluation that instructors' beauty and age are both equal to 0. The firstslope 0.134, means if beauty differed by 1 unit and age doesn't change, course evaluation will differ by 0.134 units, on average. The firstslope 0.00028, means if age differed by 1 year and beauty doesn't change, course evaluation will differ by 0.00028 units, on average.

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
```

The result of z.scores lies bewteen -1.96 and 1.96 shows the slope coefficient is not significant at 0.05 significant level

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.table = table(z.scores)
z_scores.table
sum(z.scores>1.96 | z.scores < -1.96)
```

How many of these 100 z-scores are statistically significant?

5 of them are significant

What can you say about statistical significance of regression coefficient?

The significance of a regression coefficient in a regression model is determined by dividing the estimated coefficient over the standard deviation of this estimate. For statistical significance we expect the absolute value of the t-ratio to be greater than 2 or the P-value to be less than the significance level (0,01 or 0,05 or 0,1).

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"))
# Prestige
step1.lm = lm(prestige~income+women,Prestige)
step1.resid = resid(step1.lm)
step2.lm = lm(education~income+women,Prestige)
step2.resid = resid(step2.lm)
step3.lm = lm(step1.resid~step2.resid)
summary(step3.lm)
```

```
##
## Call:
## lm(formula = step1.resid ~ step2.resid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.992e-15  7.691e-01    0.00        1
## step2.resid  4.187e+00  3.848e-01   10.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.768 on 100 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5375
## F-statistic: 118.4 on 1 and 100 DF,  p-value: < 2.2e-16
```

```
test.lm = lm(prestige~education+income+women,Prestige)
summary(test.lm)
```

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

```
## lm(formula = prestige ~ education + income + women, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## education    4.1866373  0.3887013  10.771  < 2e-16 ***
## income       0.0013136  0.0002778   4.729 7.58e-06 ***
## women       -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

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

Because the the expected mean of step 1's residual is 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"?

It can be used to detecting joint influence on the regression coefficients.

Yes, it is.

(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.

```
# library(ppcor)
# pcor.test(Prestige$prestige,Prestige$education,Prestige[,c("income","women")])
cor(step1.resid,step2.resid)
```

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

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?

(See attached pdf)

Mathematical exercises.

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

1. $\sum \hat{y}_i \hat{e}_i = 0$ ###(See attached pdf)
2. $\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum \hat{e}_i (\hat{y}_i - \bar{y}) = 0$ ###(See attached pdf) Suppose that the means and standard deviations of $\boldsymbol{y}$ and $\boldsymbol{x}$ are the same: $\bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}$ and $sd(\boldsymbol{y}) = sd(\boldsymbol{x})$.
3. 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, $\alpha_{y|x} = \alpha_{x|y}$.

(See attached pdf)

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 $r_{xy} < 1$)?

(See attached pdf)

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?

I think this a weak research is weak because observe children's reading performance on average is not enough. To get the efiicancy of the new program, we can compare each one child's reading performance beofre the program and after the program, then we calculate the confidence interval.

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.