MA678 homework 01

Siwei Hu

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.

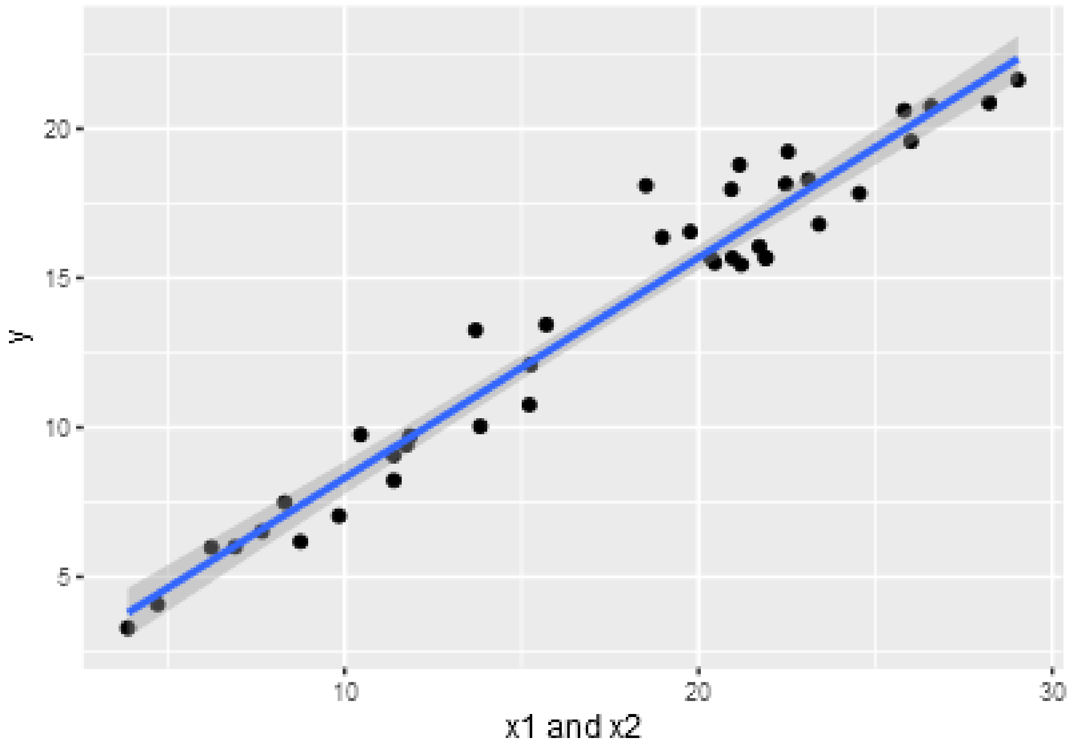
1. Use R to fit a linear regression model predicting `y` from `x1`,`x2`, using the first 40 data points in the file. Summarize the inferences and check the fit of your model.

```
fit1<- lm(y~x1+x2,data = pyth[1:40,])
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth[1:40, ])
##
## 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
```

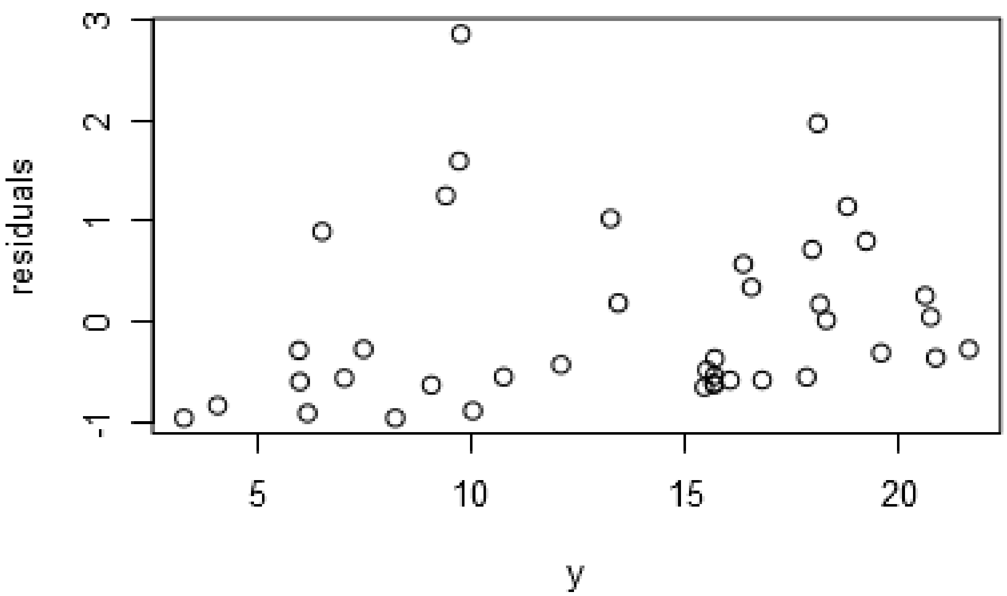
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
library(ggplot2)
ggplot(pyth[1:40,])+aes(x = x1+x2,y = y)+geom_point()+ylab("y")+xlab("x1 and x2")+stat_smooth(method =
```



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

```
plot(pyth[1:40,]$y,fit1$residuals,type = "p",xlab = "y",ylab = "residuals")
```



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

```
predict(object = fit1,newdata = pyth[41:60,2:3],interval = "prediction")
```

```
##          fit       lwr       upr
## 41 14.812484 12.916966 16.708002
## 42 19.142865 17.241520 21.044211
## 43  5.916816  3.958626  7.875005
## 44 10.530475  8.636141 12.424809
## 45 19.012485 17.118597 20.906373
## 46 13.398863 11.551815 15.245911
## 47  4.829144  2.918323  6.739965
## 48  9.145767  7.228364 11.063170
## 49  5.892489  3.979060  7.805918
## 50 12.338639 10.426349 14.250929
## 51 18.908561 17.021818 20.795303
## 52 16.064649 14.212209 17.917088
## 53  8.963122  7.084081 10.842163
## 54 14.972786 13.094194 16.851379
## 55  5.859744  3.959679  7.759808
## 56  7.374900  5.480921  9.268879
## 57  4.535267  2.616996  6.453539
## 58 15.133280 13.282467 16.984094
## 59  9.100899  7.223395 10.978403
## 60 16.084900 14.196990 17.972810
```

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.

```
beta1 <- 0.008/0.01
beta0 <- log(30000) - beta1*log(66)
beta1
```

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

```
beta0
```

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

```
sd.r = log(1.1)/2
sd.r
```

```
## [1] 0.04765509
```

$\log(earning) = \beta 0 + \beta 1 * log(heigth)$

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?

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{sd.r^2}{0.05^2} = 0.0915969$$

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.

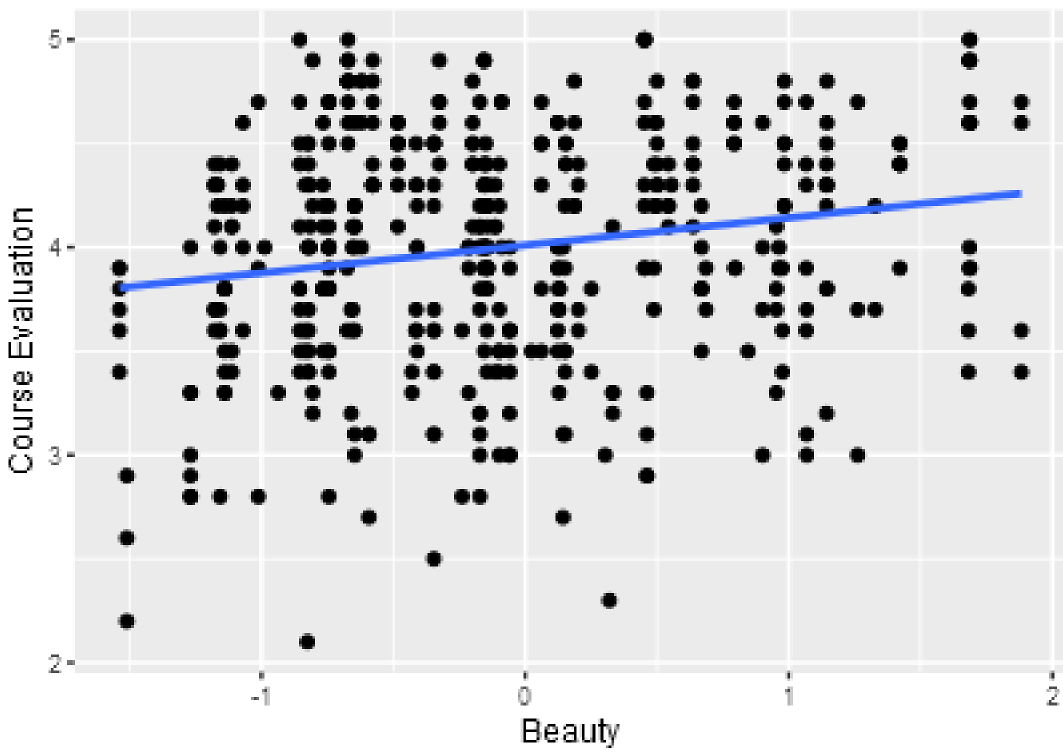
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, s
```

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.

```
beauty.fit = lm(beauty.data$courseevaluation~beauty.data$btystdave)
coef(beauty.fit)
```

```
##           (Intercept) beauty.data$btystdave
##             4.0100227             0.1330014
```

```
ggplot(beauty.data)+aes(x = btystdave,y = courseevaluation)+geom_point()+ylab("Course Evaluation")+xlab
```

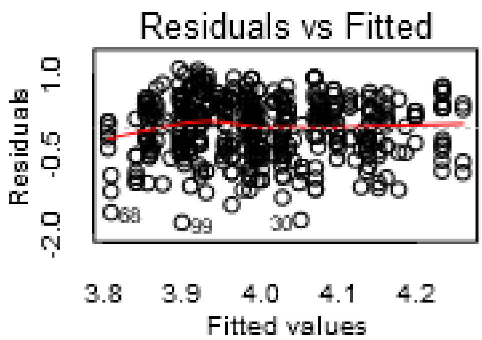



```r
summary(beauty.fit)
```

```
##
## Call:
## lm(formula = beauty.data$courseevaluation ~ beauty.data$btystdave)
##
## 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 ***
## beauty.data$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
par(mfrow=c(2,2))
par (mar=c(4,4,2,1), mgp=c(2,1,0), tck=-.01)

plot(beauty.fit, which =1)
```



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.

```
beauty.fit2 = lm(courseevaluation~btystdave+age+btystdave*age,data=beauty.data)
coef(beauty.fit2)
```

```
##   (Intercept)     btystdave           age btystdave:age
##   3.960512959  -0.339162560   0.001540134   0.010149755
```

```
beauty.fit3 = lm(courseevaluation~btystdave+btystdfl,data=beauty.data)
coef(beauty.fit3)
```

```
## (Intercept)   btystdave    btystdfl
##  4.01008588  0.18206960 -0.04592389
```

The beauty.fit2 has new input "age" and "btystdave*age" and predictors are btystdave and age and correlation between btystdave and age. The coefficents show that btystdave ,age and their correlation have influence on the course evaluation. The beauty.fit2 has new input "btystdfl" and predictors are btystdave and btystdfl. The coefficents show that btystdfl and btystdave have influence on the course evaluation.

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
```

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:

```
set.seed(1111)
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]
}
count.z = z.scores[z.scores>=2]
count.z
total.z = sum(z.scores>=2)
total.z

mean.z = mean(count.z)
mean.z
```

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

There are 5 z-cores are statistically significant. There exist 5 z-scores of 100 are away from 2 standard deviation. So these 5 coefficients estimators will keep in the model.

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"))
```

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

```
Prestige.fit = lm(prestige~income+women+census,data=Prestige)

Prestige.fit2 = lm(education~income+women+census,data=Prestige)

Prestige.education = lm(residuals(Prestige.fit)~residuals(Prestige.fit2))

Prestige.normal.fit = lm(prestige~education+income+women+census,data=Prestige)
coefficients(Prestige.normal.fit)
```

```
##   (Intercept)     education        income         women        census
## -14.949440307   4.657158047   0.001289224  -0.002086820   0.000568421
```

The residuals of $Y \sim X_2 + X_3 + X_4$ means that the number that Y dont affect by X2,X3 and X4. It's same as the residuals of $X_1 \sim X_1 + X_2 + X_3 + X_4$. From the linear regression of $Y \sim X_1 + X_2 + X_3 + X_4$, we can see that B_1 has same value as the B_1 of the linear regression of Resid($Y \sim X_1 + X_2 + X_3 + X_4$) and Resid($X_1 \sim X_1 + X_2 + X_3 + X_4$). So there almost no other influence to Y except all predictors.

(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, it is. Becasue when remove other all varibles' influence from X_1 and Y means the residual is the influence from the rest variable and error. So the B_1 is the coefficient of X_1 to Y.

(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?

This can help to analyze the relationship between outcome and only one single varible. For example, whe we do research about rabbit life environment, and we do lm(life environment ~ humdity+temperature+ food). If we do this procedure, it can analyze the relationship between life environment and each of these variables.

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.

```
prestige.r <- resid(lm(prestige~women+income,data = Prestige))
education.r <- resid(lm(education~women+income,data=Prestige))
cor <- cor(prestige.r,education.r)
cor
```

```
## [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?

If controlling for other three varibles, the B_1 shows the relationship between outcom Y and varibelX_1. So if coefficient B_1 is 0 means education has no influence on prestige. So the correlation between these two will be 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

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

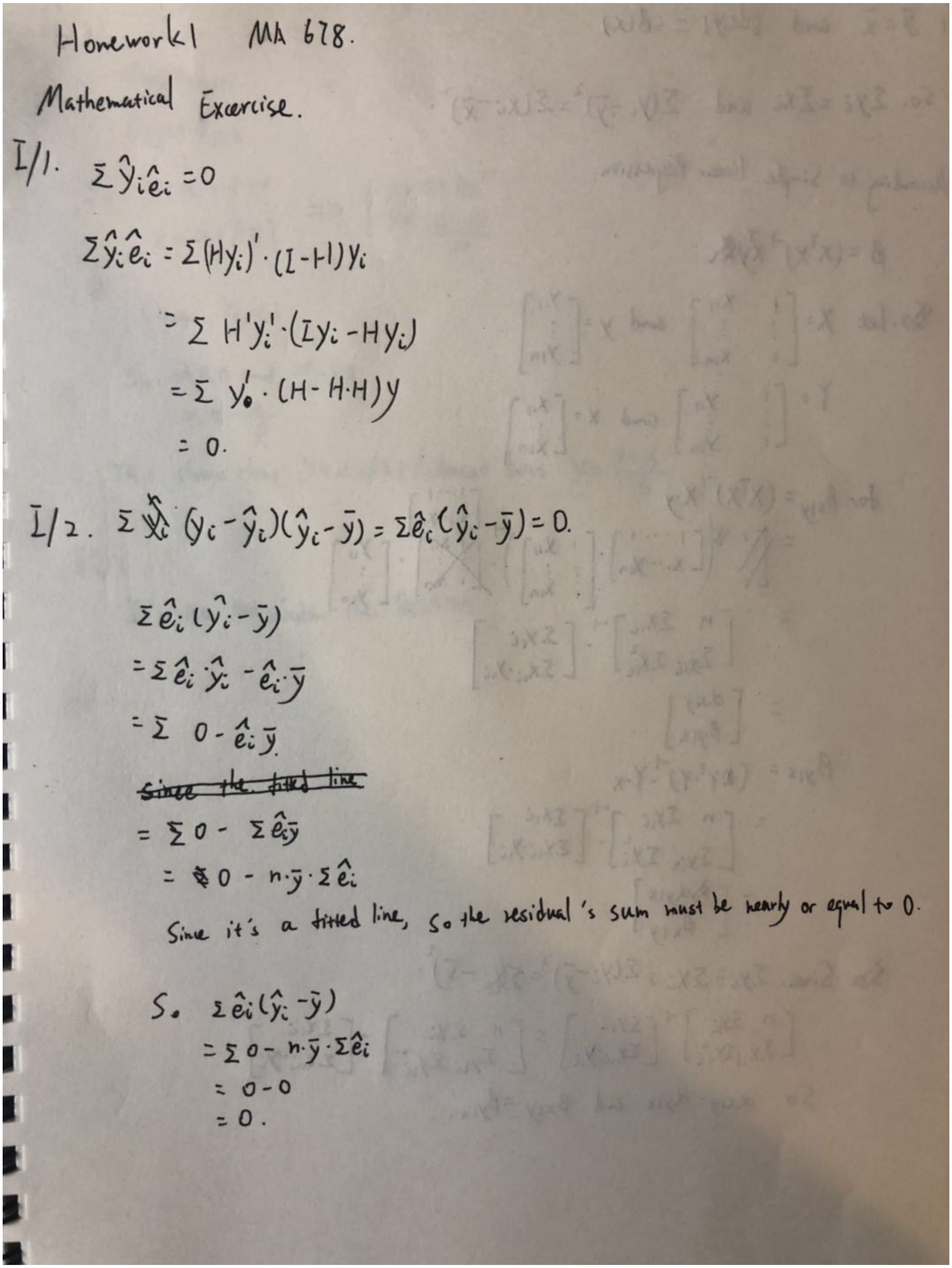
   where $\beta_{y|x}$ is the least-squares slope for the simple regression of $\boldsymbol{y}$ on $\boldsymbol{x}$, $\beta_{x|y}$ is the least-squares slope for the simple regression of $\boldsymbol{x}$ on $\boldsymbol{y}$, and r_{xy} is the correlation between the two variables. Show that the intercepts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

Answer: Researchers choose a rup of children who are reading substantially below grade level and theon only test this group. This is not a good idea. They need another group which children are reading above grade level. This can help this research to be more comprehensive. They also can do comparision for groups to see this new program is better to which group of children. # 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.



Homework1 MA 678.

Mathematical Excercise.

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

$$\sum \hat{y}_i \hat{e}_i = \sum (Hy_i)' \cdot (I - H) y_i$$
$$= \sum H' y_i' \cdot (I y_i - H y_i)$$
$$= \sum y_0' \cdot (H - H \cdot H) y$$
$$= 0.$$

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

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

~~Since the fitted line~~

$$= \sum 0 - \sum \hat{e}_i \bar{y}$$
$$= \cancel{\sum} 0 - n \cdot \bar{y} \cdot \sum \hat{e}_i$$

Since it's a fitted line, so the residual's sum must be nearly or equal to 0.

So. $\sum \hat{e}_i (\hat{y}_i - \bar{y})$
$$= \sum 0 - n \cdot \bar{y} \cdot \sum \hat{e}_i$$
$$= 0 - 0$$
$$= 0.$$

Figure 1: A caption

II/1 $\bar{y} = \bar{x}$ and $sd(y) = sd(x)$

So. $\sum y_i = \sum x_i$ and $\sum (y_i - \bar{y})^2 = \sum (x_i - \bar{x})^2$.

According to Simple linear Regression.

$$\beta = (X^T X)^{-1} X^T y$$

So. Let $X = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix}$ and $y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \end{bmatrix}$

$Y = \begin{bmatrix} 1 & y_{11} \\ \vdots & \vdots \\ 1 & y_{1n} \end{bmatrix}$ and $x = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1n} \end{bmatrix}$

for $\beta_{x|y} = (X^T X)^{-1} \cdot X \cdot y$

$$= \left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{1n} \end{bmatrix} \cdot \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_{1i} \\ \sum x_{1i} & \sum x_{1i}^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum y_{1i} \\ \sum x_{1i} \cdot y_{1i} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_{x|y} \\ \beta_{y|x} \end{bmatrix}$$

$$\beta_{y|x} = (Y^T \cdot Y)^{-1} \cdot Y \cdot x$$

$$= \begin{bmatrix} n & \sum y_{1i} \\ \sum y_{1i} & \sum y_{1i}^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum x_{1i} \\ \sum x_{1i} \cdot y_{1i} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_{y|x} \\ \beta_{x|y} \end{bmatrix}$$

So. Since $\sum y_i = \sum x_i$, $\sum (y_i - \bar{y})^2 = \sum (x_i - \bar{x})^2$.

$$\begin{bmatrix} n & \sum x_{1i} \\ \sum x_{1i} & \sum x_{1i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_{1i} \\ \sum x_{1i} \cdot y_{1i} \end{bmatrix} = \begin{bmatrix} n & \sum y_{1i} \\ \sum y_{1i} & \sum y_{1i}^2 \end{bmatrix} \cdot \begin{bmatrix} \sum x_{1i} \\ \sum x_{1i} \cdot y_{1i} \end{bmatrix}$$

So $\alpha_{x|y} = \alpha_{y|x}$ and $\beta_{x|y} = \beta_{y|x}$.

Figure 2: A caption

II/2.

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

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

So $y = \alpha + \beta x$
$x = \alpha + \beta y$ $\Rightarrow \begin{cases} y = \alpha + \beta x. \\ y = \dfrac{x - \alpha}{\beta}. \end{cases}$

Since $r_{xy} < 1$

So $\alpha \neq 0$ and $-\dfrac{\alpha}{\beta} \neq 0$

$\alpha \neq -\dfrac{\alpha}{\beta}.$

This shows that $y = \alpha + \beta x$ is different from $y = \dfrac{x - \alpha}{\beta}.$

II/3.

I write under the Question.

Figure 3: A caption