MA678 homework 01

Jinfei Xue

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.

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

```
##
## 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
```

From the summary above, we can say the model fits the observation very well. The reasons are as follows:

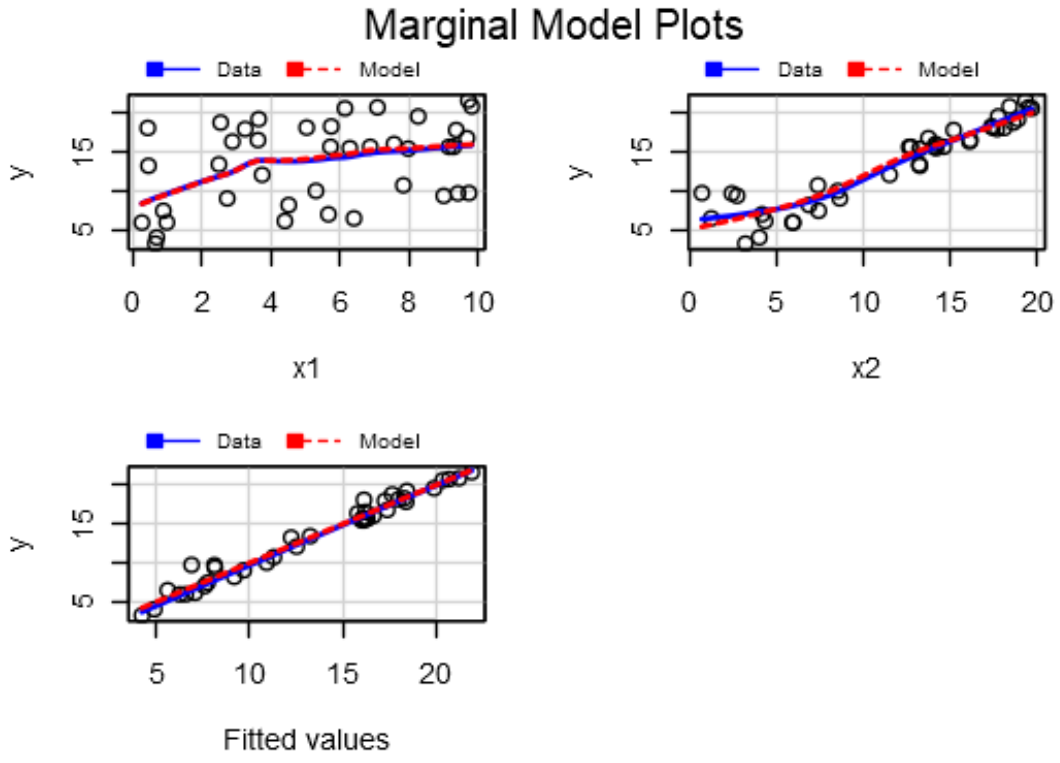
1> The value of adjusted R-squared is near 1;

2> Because all the p-values of coefficients are very small, the coefficients in the regression are statistically significant;

3> Because the p-value of F Statistics is very small, the linear relation in the model is statistically significant, which means all the independent variables can explain the dependent variable very well.

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

```
car::marginalModelPlots(r)
```

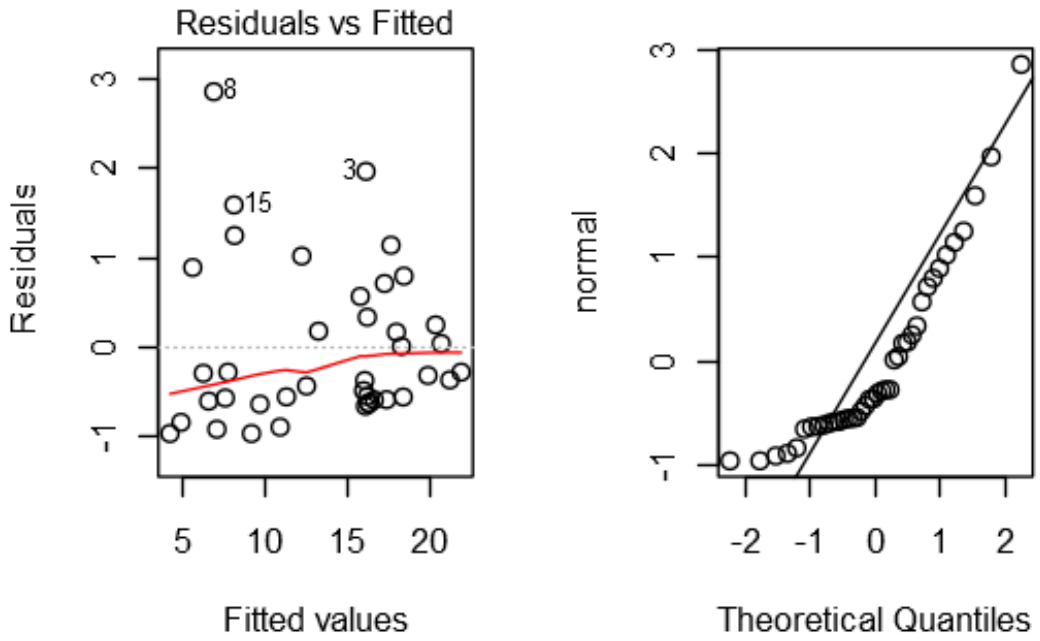


There is one marginal model plot for each independent variable and one additional plot that displays the predicted values on the horizontal axis.

According to the third graph, we can say the model fit the data well.

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

```
par(mfrow=c(1,2))
plot(r,which=1)
resid<-resid(r)
qqnorm(resid,ylab="normal",main="");qqline(rnorm(40))
```



The assumptions are not met. The reasons are as follows:

1> The value of residuals are not evenly distributed around the dotted line.

2> The red line does not closely coincide with the dotted line.

3> The qqplot shows the residuals are not normally distributed.

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

```
predict(r,pyth[41:60,],interval="prediction",level=0.95)
```

```
##          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
```

These prediction intervals are 95% statistically significant. The second and third columns represent the interval prediction of data.

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.

```
beta_0=log(30000)-(0.008/0.01)*log(66)
beta_0
```

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

So the equation of the regression line is

$$log(earnigs) = 6.957229 + 0.8log(height) + \epsilon$$

```
sd=0.1*0.5/0.95
sd
```

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

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?

```
sd.log_heights=0.05
R2<- 1-(sd^2/sd.log_heights^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.

```
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.

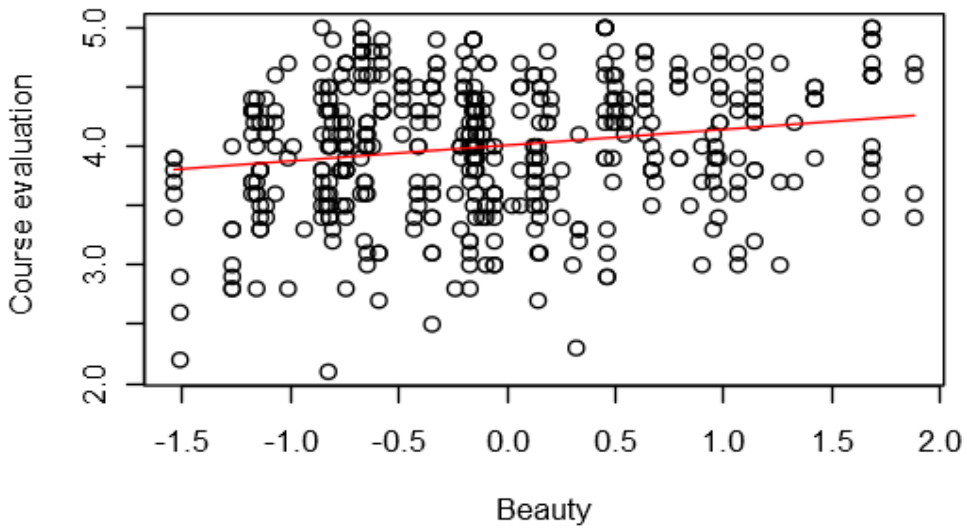
```
beauty<-beauty.data$btystdave
eval<-beauty.data$courseevaluation
r_1<-lm(eval~beauty)
summary(r_1)
```

```
##
## Call:
## lm(formula = eval ~ beauty)
##
## 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       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
```

So the fitted model is:

$$\hat{eval} = 4.01002 + 0.13300 \times beauty$$

```
plot(beauty,eval,xlab="Beauty", ylab="Course evaluation")
curve(coef(r_1)[1]+coef(r_1)[2]*x, add=TRUE, col="red")
```



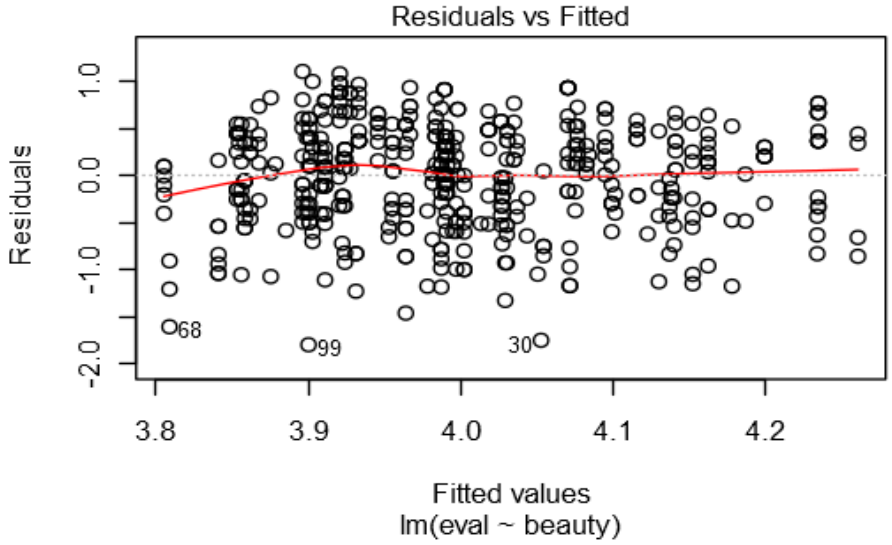
$\beta_0 = 4.01002$ *means when beauty=0, courseevaluation is 4.01002;*

$\beta_1 = 0.13300$ *means when beauty inceases 1 unit, the expected value of courseevaluation will increase 0.13300 unit.*

Residual standard error is to measure the degree of dispersion of the observation value from the fiited value of dependent variable.

the residuals versus fitted values plot is as follows:

```
plot(r_1,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.

```
#Model 1
age<-beauty.data$age
beauty<-beauty.data$btystdave
gender<-beauty.data$female
eval<-beauty.data$courseevaluation
fit_1<-lm(eval~age+beauty+gender)
summary(fit_1)

##
## Call:
## lm(formula = eval ~ age + beauty + gender)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85612 -0.35831  0.04697  0.39308  1.04276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.225242   0.142820  29.584  < 2e-16 ***
## age         -0.002602   0.002768  -0.940    0.348
## beauty       0.139978   0.033243   4.211 3.06e-05 ***
## gender      -0.210792   0.052824  -3.990 7.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5374 on 459 degrees of freedom
## Multiple R-squared:  0.06809,    Adjusted R-squared:  0.062
## F-statistic: 11.18 on 3 and 459 DF,  p-value: 4.305e-07
```

So the fitted model is:

$$\hat{eval} = 4.225242 - 0.002602 \times age + 0.13300 \times beauty - 0.210792 \times gender$$

The predictors in model 1 are age, beauty and gender. The inputs are the same as predictors.

$\beta_0 = 4.225242$ *means when age=0, beauty=0 and gender is male, courseevaluation is 4.225242;*

$\beta_1 = -0.002602$ *means when age inceases by 1 unit, the expected value of courseevaluation will decrease by 0.002602 unit.*

$\beta_2 = 0.13300$ *means when beauty inceases by 1 unit, the expected value of courseevaluation will increase by 0.13300 unit.*

$\beta_3 = -0.210792$ *means the expected value of courseevaluation of female is 0.210792 unit less than that of male.*

```
#Model2
tenured<-beauty.data$tenured
eval_prof<-beauty.data$profevaluation
eval<-beauty.data$courseevaluation
fit_2<-lm(eval~tenured+eval_prof+tenured*eval_prof)
summary(fit_2)

##
## Call:
## lm(formula = eval ~ tenured + eval_prof + tenured * eval_prof)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01496 -0.11496  0.00564  0.12113  0.78242
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.16674    0.10801  -1.544   0.1233
## tenured            0.31723    0.14305   2.218   0.0271 *
## eval_prof          0.99584    0.02541  39.196   <2e-16 ***
## tenured:eval_prof -0.07326    0.03391  -2.160   0.0313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1962 on 459 degrees of freedom
## Multiple R-squared:  0.8758, Adjusted R-squared:  0.8749
## F-statistic:  1079 on 3 and 459 DF,  p-value: < 2.2e-16
```

So the fitted model is:

$\hat{eval} = -0.16674 + 0.31723 \times tenured + 0.99584 \times evalprof - 0.07326 \times tenured \times evalprof$

The predictors in model 2 are tenured, profevaluation and tenured multiply profevaluation. The inputs include tenured and eval_prof.

$\beta_0 = -0.16674$ *means when a professor is not tenured and evalprof=0, courseevaluation is -0.16674;*

$\beta_1 = 0.31723$ *means the courseevaluaiton of tenured professor(with profevaluation is 0) is 0.31723 unit more than that of non-tenured professor(with profevaluation is 0).*

$\beta_2 = 0.99584$ *means when a professor is non-tenured, if profevaluation inceases by 1 unit, the expected value of courseevaluation will increase by 0.99584 unit.*

$\beta_3 = -0.07326$ *represents the difference in the slope for profevaluation, comparing professors who did and did not tenure.*

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
## -0.7952306
```

Because the absolute value of z.score is smaller than 2, the slope coefficient is 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
```

```
##   [1] -0.95110248  0.03655228 -0.51188662 -0.67615226  0.03243308
##   [6] -1.59991528 -1.04949525 -1.56696038  0.97461540  0.39402659
##  [11]  0.13470677  1.24619882  0.98706794  0.78883590 -0.70763344
##  [16]  0.04616834 -0.50108779  0.38718848  1.26251777  2.12333158
##  [21]  0.14251116  0.59279703 -0.70453056  0.45097029 -1.98771234
##  [26]  1.34490629  0.02482558  0.43738301 -0.97382713  1.98391805
##  [31]  0.14636712 -1.54656584 -0.79488268 -2.40513377  0.10270541
##  [36]  0.91673913 -1.36909199 -0.02310948 -0.33029788 -0.01407498
##  [41] -1.12545290 -0.09907172  0.42199809 -0.12087443 -1.71363124
##  [46] -0.14964013 -0.96868004  1.30760417 -0.42994941  0.40810363
##  [51] -0.29491047  0.30861353  0.92208603 -0.47620834 -0.82319438
##  [56] -0.90343200 -0.46274340  0.32901593 -0.48051917  0.97236881
##  [61]  0.09379524  0.16544010  0.61198146 -0.73426176 -0.48507732
##  [66] -0.17911107  0.07518543  0.58172671 -1.02363053  0.99028932
##  [71] -0.31481911  0.03443187 -0.36039645 -2.31595429 -0.52054093
##  [76]  1.67055790  1.24939865  0.44226656  0.59266879  2.56284317
##  [81]  2.02158803 -0.89282384  1.11696381 -0.02956308  2.54789784
##  [86] -0.10388274 -1.85068524 -0.17538100  0.01260801 -1.40472497
##  [91] -0.93082359  0.32402618 -3.10943232  0.61616261 -1.71659341
##  [96] -0.66451578 -0.28429225  0.94799403  0.76797810 -0.18760055
```

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

```
length(z.scores[abs(z.scores)>2])
```

```
## [1] 7
```

So there are 7 z-scores which are statistically significant in these 100 z-scores.

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"))
```

First, make a multiple regression of prestige on education, income and percentage of women

```
edu<-Prestige$education
inc<-Prestige$income
women<-Prestige$women
prestige<-Prestige$prestige
r_0<-lm(prestige~edu+inc+women)
summary(r_0)
```

```
##
## Call:
## lm(formula = prestige ~ edu + inc + women)
##
## 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 *
## edu          4.1866373  0.3887013  10.771  < 2e-16 ***
## inc          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
```

```
coefficients(r_0)
```

```
##  (Intercept)          edu          inc        women
## -6.794334203  4.186637275  0.001313560 -0.008905157
```

The coefficient of education is 4.186637275

Next, follow three steps mentioned previously.

```
#Step1:
r_1<-lm(prestige~inc+women)
summary(r_1)
```

```
##
## Call:
## lm(formula = prestige ~ inc + women)
##
## 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 ***
## inc         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
```

```
resid_1<-residuals(r_1)

#Step2:
r_2<-lm(edu~inc+women)
summary(r_2)
```

```
##
## Call:
## lm(formula = edu ~ inc + women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8421 -1.3210 -0.0429  1.2760  5.6208
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.478e+00  5.268e-01  12.297  < 2e-16 ***
## inc         4.826e-04  5.298e-05   9.109 9.60e-15 ***
## women       3.380e-02  7.090e-03   4.768 6.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.029 on 99 degrees of freedom
## Multiple R-squared:  0.458,  Adjusted R-squared:  0.4471
## F-statistic: 41.84 on 2 and 99 DF,  p-value: 6.78e-14
```

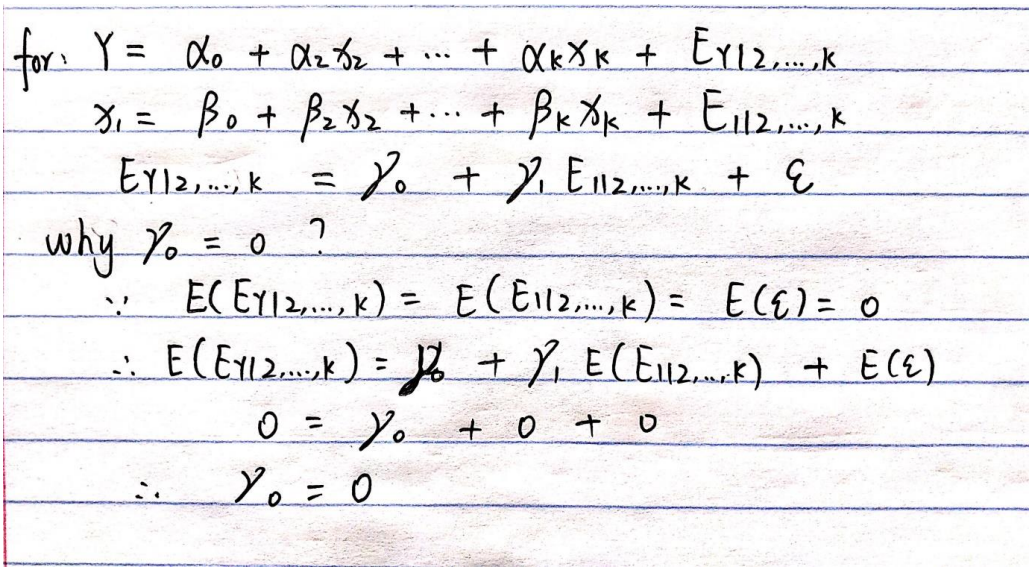
```
resid_2<-residuals(r_2)

#Step3:
r_3<-lm(resid_1~resid_2)
summary(r_3)
```

```
##
## Call:
## lm(formula = resid_1 ~ resid_2)
##
## 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
## resid_2      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
```

The coefficient of resid_2 is 4.187e+00, which is same as the coefficient of education in the first regression. Therefore, we can confirm that the coefficient for education is properly recovered.

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



(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 is 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. According to the previous procedure, $E_{Y|2,\ldots,k}$ and $E_{1|2,\ldots,k}$ can be seen as Y and X_1 from which the influence of $X_2, \cdots, X_k$ is removed respectively. Therefore, it's reasonable.

(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?

When the independents have multicolinearity, this method can remove the influence of other independents. Thus, the equation which removes the influence of other related independents can fit the data very well.

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 $r_{y1|2,\ldots,k}$.

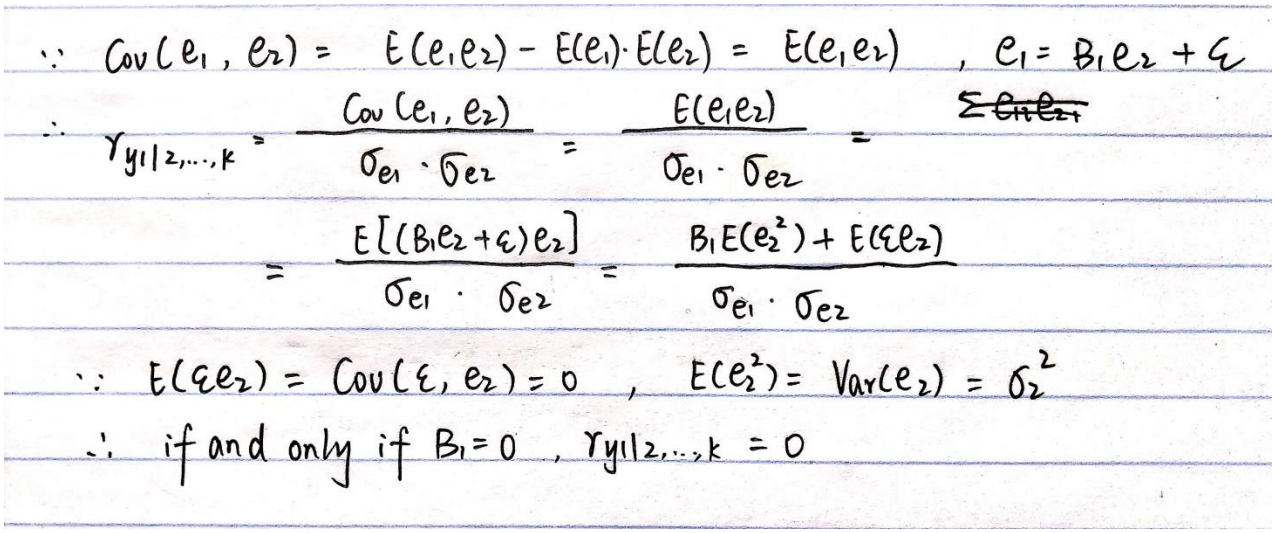
1. Using the Canadian occupational prestige data, calculate the partial correlation between prestige and education, controlling for income and percentage women.

```
cor(resid_1,resid_2)
```

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

Therefore, the partial correlation between prestige and education is 0.7362604, which is same as the simple correlation between resid_1 and resid_2.

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?



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

Mathematical exercises

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

Proof: $\sum \hat{y}_i e_i = (\hat{y}_1\ \hat{y}_2 \cdots \hat{y}_n)\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \hat{Y}^T e$

$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$, where $X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}$

$e = Y - \hat{Y} = IY - X(X^TX)^{-1}X^TY = [I - (X(X^TX)^{-1}X^T)]Y$

So, $\hat{Y}^T e = Y^TX(X^TX)^{-1}X^T \cdot [I - X(X^TX)^{-1}X^T]Y$

$\because X(X^TX)^{-1}X^T = H$ and $H \cdot H = H$

$\therefore \hat{Y}^T e = Y^T H \cdot (I - H)Y = Y^T(H - H \cdot H)Y$

$= Y^T(H - H)Y = 0$

$\therefore \sum \hat{y}_i e_i = 0$

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

Proof: $\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = (Y - \hat{Y})^T(\hat{Y} - \bar{Y})$

$= e^T(\hat{Y} - \bar{Y}) = e^T\hat{Y} - e^T\bar{Y}$

where $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\bar{Y} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}$

$\because e^T\hat{Y} = \hat{Y}^T e = 0$, $e^T\bar{Y} = \bar{y} \cdot \sum e_i = \bar{y} \cdot 0 = 0$ ($\because e_i \sim N(0, \sigma^2)$, $\therefore \sum e_i = 0$)

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

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

1. Show that, under these circumstances

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

II.

1. $\beta_{Y|X} = (X^TX)^{-1}X^TY_1$, $\beta_{X|Y} = (Y^TY)^{-1}Y^TX_1$

where $X_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ $Y_1 = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ $Y = \begin{pmatrix} 1 & y_1 \\ 1 & y_2 \\ \vdots & \vdots \\ 1 & y_n \end{pmatrix}$

$$\beta_{Y|X} = c \cdot \begin{pmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{pmatrix} \cdot \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix}, \text{ where } c = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= c \cdot \begin{pmatrix} \Sigma x_i^2 \Sigma y_i - \Sigma x_i \cdot \Sigma x_i y_i \\ -\Sigma x_i \Sigma y_i + n\Sigma x_i y_i \end{pmatrix} = \begin{pmatrix} \alpha_{y|x} \\ \beta_{y|x} \end{pmatrix}, \text{ where } d = \frac{1}{n\Sigma y_i^2 - (\Sigma y_i)^2}$$

the same: $\beta_{X|Y} = d \cdot \begin{pmatrix} \Sigma y_i^2 \Sigma x_i - \Sigma y_i \cdot \Sigma x_i y_i \\ -\Sigma y_i \Sigma x_i + n\Sigma x_i y_i \end{pmatrix} = \begin{pmatrix} \alpha_{x|y} \\ \beta_{x|y} \end{pmatrix}$

$\because \bar{y} = \bar{x}$, $sd(y) = sd(x)$

$\therefore \frac{1}{n-1}\Sigma(y_i - \bar{y})^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$ $\therefore \Sigma x_i^2 = \Sigma y_i^2$, $\Sigma x_i = \Sigma y_i$

$\therefore \alpha_{y|x} = \alpha_{x|y}$, $\beta_{y|x} = \beta_{x|y}$

$$\left(r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)} \cdot \sqrt{Var(y)}} = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sqrt{\Sigma(x_i - \bar{x})^2} \cdot \sqrt{\Sigma(y_i - \bar{y})^2}}\right.$$

$$= \frac{\Sigma x_i y_i - \bar{y}\Sigma x_i - \bar{x}\Sigma y_i + n\bar{x}\bar{y}}{\Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2} \xrightarrow[\bar{x} = \bar{y} = a]{\text{let}} \frac{\Sigma x_i y_i - 2na^2 + na^2}{\Sigma x_i^2 - 2na^2 + na^2}$$

$$= \frac{\Sigma x_i y_i - na^2}{\Sigma x_i^2 - na^2}$$

$$\because \beta_{y|x} = \beta_{x|y} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{n\Sigma x_i y_i - n^2a^2}{n\Sigma x_i^2 - n^2a^2} = \frac{\Sigma x_i y_i - na^2}{\Sigma x_i^2 - na^2}$$

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

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

The conclusion has been proved when prove the equation in the first question.

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

$y = \alpha_{y|x} + \beta_{y|x} \cdot x$

$x = \alpha_{x|y} + \beta_{x|y} \cdot y \Rightarrow y = \frac{-\alpha_{x|y}}{\beta_{x|y}} + \frac{1}{\beta_{x|y}} \cdot x$

$\because \alpha_{y|x} = \alpha_{x|y}, \quad \beta_{y|x} = \beta_{xy} = r_{xy} < 1$

$\therefore \alpha_{y|x} \neq \frac{-\alpha_{x|y}}{\beta_{x|y}}, \quad \beta_{y|x} \neq \frac{1}{\beta_{x|y}}$

$\therefore$ the two regression lines are different

4. 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 research is designed because it can obtain a good result when assessing the new program. It can be improved by randomly obtain identities in the sample and then compare their performance before and after the program.

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.

Some contents in the homework are not included in the tutorial so that I am confused about them.