MA678_Homework1

Sky Liu

Septemeber 18, 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.

```
pyth40 <- pyth[1:40,]
regpyth <- lm(y ~ x1 + x2, data=pyth40)
summary(regpyth)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth40)
##
## 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
```

The summary of linear regression model predicting ‘y’ from ‘x_1’, ‘x_2’ is shown above.

From the summary we can see that $y = 1.32 + 0.51x_1 + 0.81x_2$

This infers that:

(1) When x_1 and x_2 are both 0, the value of y is 1.32.

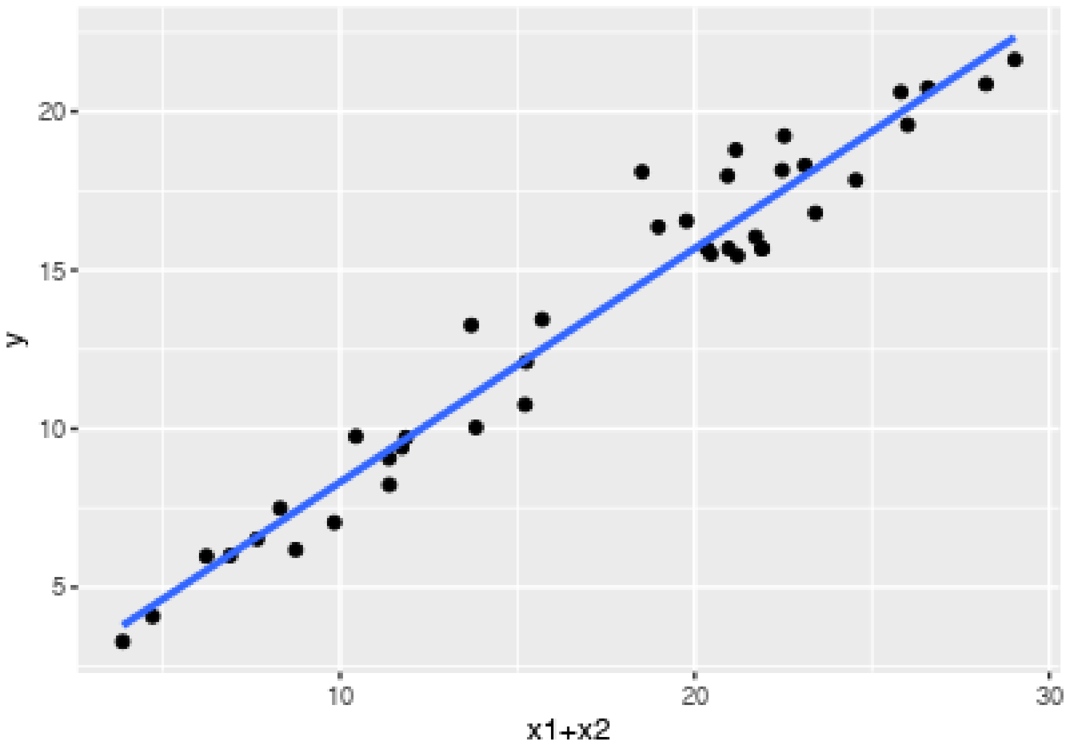
(2) Holding x_2 constant, the average value of y will increase by 0.51 if x_1 is incremented by one unit.

(3) Holding x_1 constant, the average value of y will increase by 0.81 if x_2 is incremented by one unit.

From the R square statistics we can see that 97% of varience can be explained by this model, and all the coefficients are denoted as statistically significant.

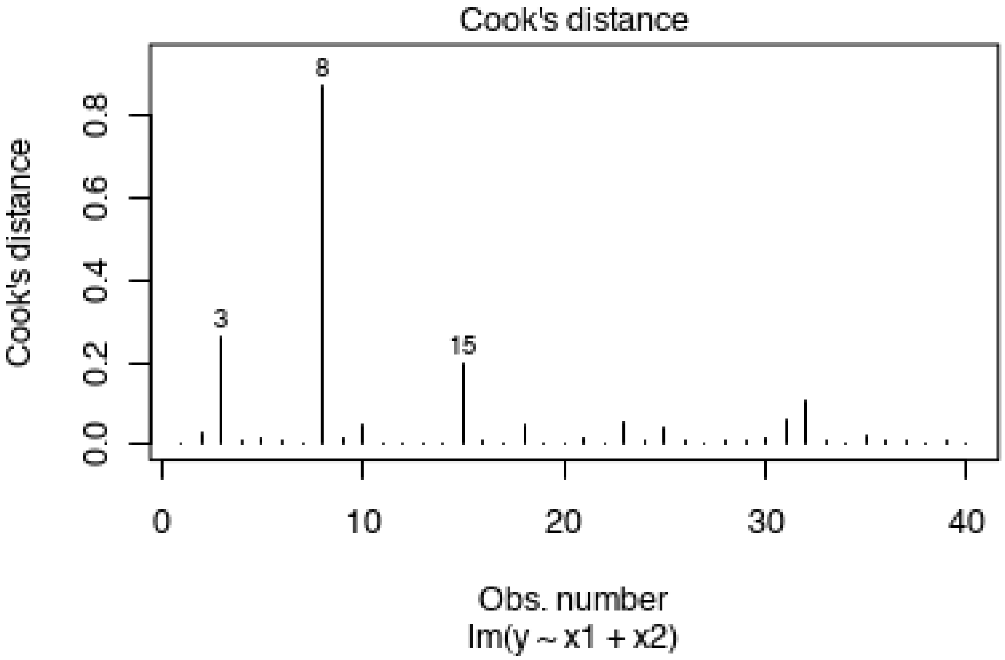
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
ggplot(regpyth)+aes(x=x1+x2,y=y)+geom_point()+ylab("y")+xlab("x1+x2")+geom_smooth(method="lm",se=FALSE)
```



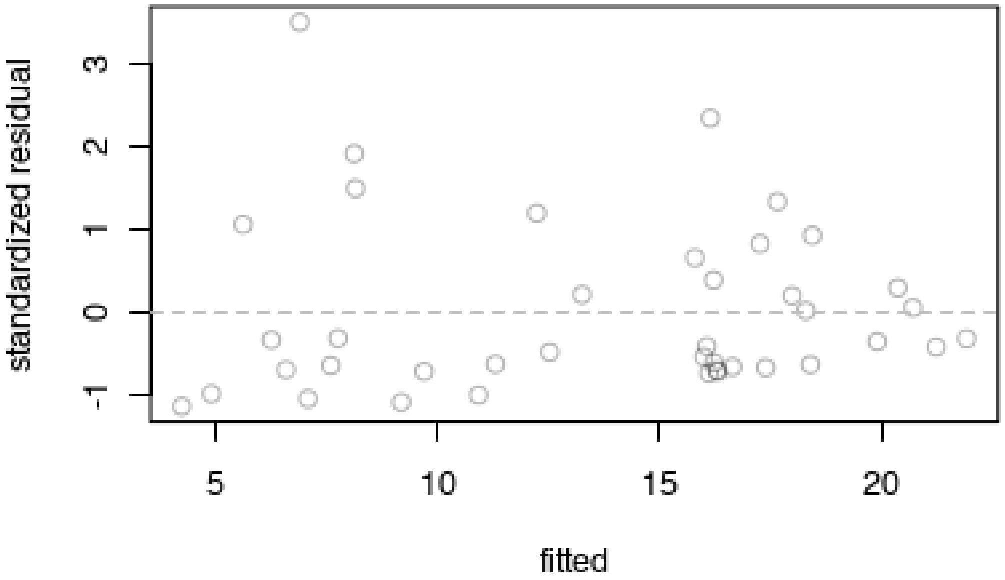
We use cook’s distant to check for outliers. From the plot we can see that there is only one point to be considered as an outlier.

```
plot(regpyth,which = 4)
```



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

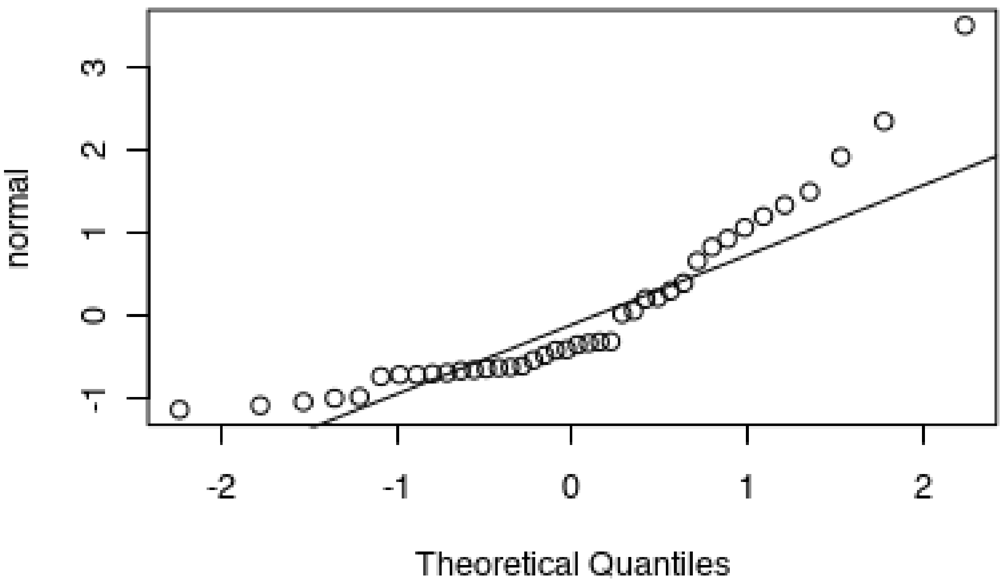
```
plot(fitted(regpyth),rstandard(regpyth),ylab="standardized residual",xlab="fitted",col=rgb(0,0,0,alpha=
```



From the standarized residual plot, we can see that residuals below 0 are more dense and the residuals to the left size are more spreading. Thus, heteroscedasticity exsits.

Also, we need to check to normality of residuals.

```
qqnorm(rstandard(regpyth),ylab="normal",main=""); qqline(rstandard(regpyth))
```



```
shapiro.test(rstandard(regpyth))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(regpyth)
## W = 0.85034, p-value = 9.168e-05
```

From the QQ plot we can see that the residuals are not quite normally distributed.

Also from the result of shapira test, we find the p-value is small enough to reject the null hypothesis of residuals being normally distributed.

Then, we need to check the correlation among residuals.

From the Durbin-Watson test we can see that the residuals are not correlated.

```
lmtest::dwtest(regpyth)
```

```
##
##  Durbin-Watson test
##
## data:  regpyth
## DW = 2.5092, p-value = 0.9511
## alternative hypothesis: true autocorrelation is greater than 0
```

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

```
pyth20 <- pyth[41:60,2:3]
predict (regpyth, pyth20, 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
```

Even though this model appears to be a good fit with significant coefficient and R^2 value, the assumption of residuals are not met.

Therefore, I am not very confident about these predictions.

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 is:

$log(earning) = 6.96 + 0.8 * log(height)$

The residual standard deviation of the regression is $log(1.1) = 0.041$.

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 is

$1 - \frac{0.041^2}{0.05^2} = 0.33$

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.

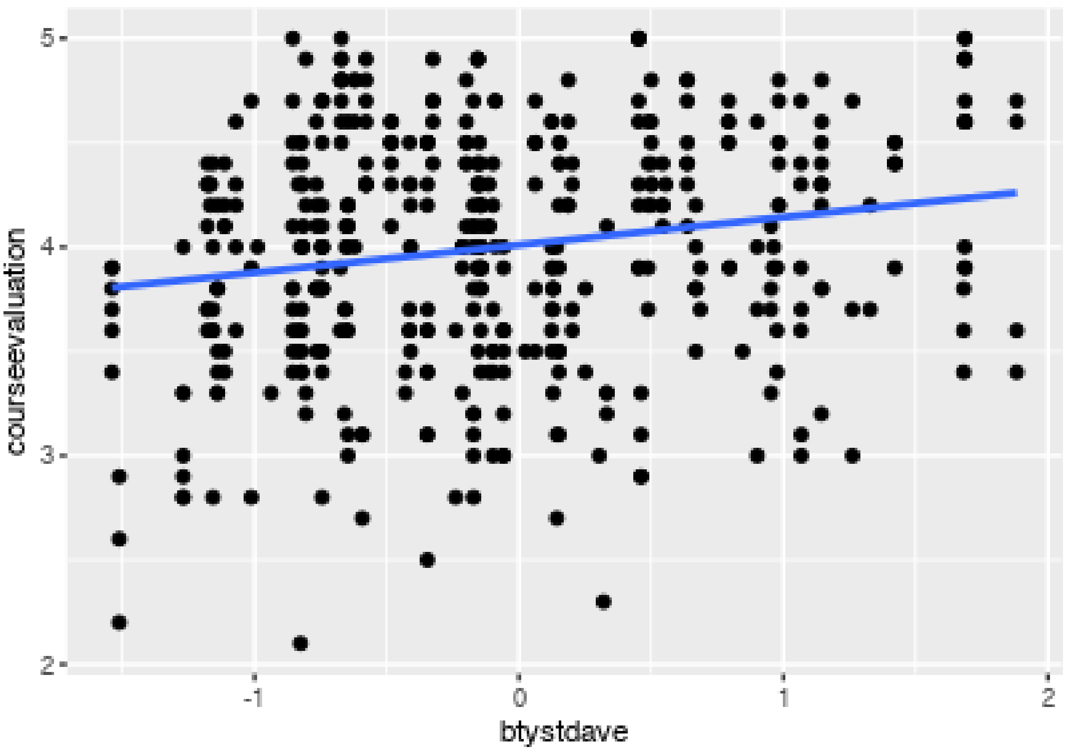
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, s
```

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.

```
regbt <- lm(courseevaluation~btystdave,data=beauty.data)
summary(regbt)
```

```
##
## 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
```

```
ggplot(regbt)+aes(x=btystdave,y=courseevaluation)+geom_point()+ylab("courseevaluation")+xlab("btystdave
```

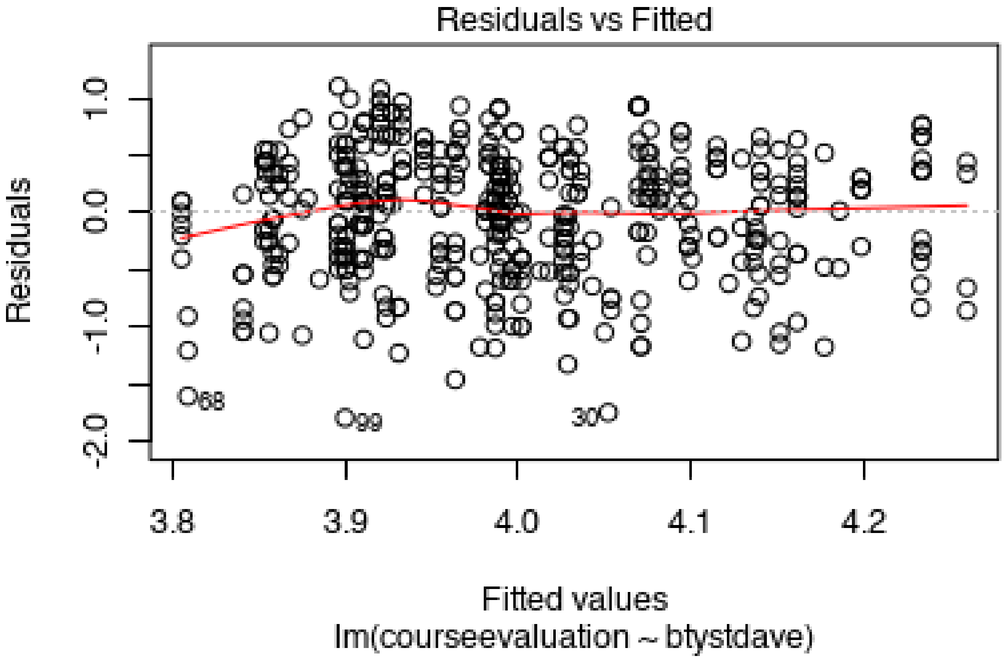


The intercept is 4.01, which means the average course evaluation is 4 if the beauty of the instructor is 0.

The beauty coefficient is 0.13, which means the average course evaluation will increase by 0.13 the in beauty of the instructor is incremented by one unit.

The plot of residuals versus fitted values is shown below:

```
plot(regbt,which=1)
```



The residual standard deviation refers to the scale of residuals.

Here the residual standard deviation is 0.55, that is to say this model can predict course evaluation to about an accuracy of 0.55 points. This model is very weak.

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.

```
regbt_1 <- lm(courseevaluation~btystdave+female+btystdave:female,data=beauty.data)
summary(regbt_1)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female + btystdave:female,
##     data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83820 -0.37387  0.04551  0.39876  1.06764
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.10364    0.03359 122.158  < 2e-16 ***
## btystdave         0.20027    0.04333   4.622 4.95e-06 ***
## female           -0.20505    0.05103  -4.018 6.85e-05 ***
## btystdave:female -0.11266    0.06398  -1.761   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5361 on 459 degrees of freedom
## Multiple R-squared:  0.07256,    Adjusted R-squared:  0.0665 
## F-statistic: 11.97 on 3 and 459 DF,  p-value: 1.471e-07
```

In this model which uses beauty, gender, interaction of gender and beauty to predict course evaluation.

The intercept is 4.1, which means the average course evaluation is 4.1 if the gender of the instructor is male and the beauty and the age of the instructor is 0.

The beauty coefficient is 0.2 and the interaction coefficient of beauty and female is -0.11, which means the average course evaluation will increase by 0.2 if the in beauty of a male instructor is incremented by one unit and the average course evaluation will not change if the in beauty of a female instructor is incremented by one unit.

The female coefficient is -0.2, which means the average course evaluation of a female instructor is 0.2 lower than a male instructor with the same beauty.

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).

```r
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]

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

No, the absolute value of this zscore is less than 2.

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:

```r
set.seed(201809)
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]
}
length(which(abs(z.scores)>2)) ## number of z-score with absolute value greater than 2
```

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

2 out of 100 z-scores are statistically significant.

What can you say about statistical significance of regression coefficient?

The term with repression coefficient is supposed to have a significant influence on the model, a term without statistical significance might be considered to be discard depending on the situatioin.

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_fit1 = lm(prestige~income+women+census,data=Prestige)
prestige_fit2 = lm(education~income+women+census,data=Prestige)
prestige_fitedu = lm(residuals(prestige_fit1)~residuals(prestige_fit2))
coefficients(prestige_fitedu)
```

```
##              (Intercept) residuals(prestige_fit2)
##             8.053058e-16             4.657158e+00
```

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

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

Here the education coefficient B_1 is 4.657

(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, in that case we just consider the errors and influence of other variables as residuals.

(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?

It helps to understand the relationship between the dependent variable and all other single 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.

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

```
## [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?

Mathematical exercises.

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

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

Proof:

$\sum \hat{y}_i \hat{e}_i$

$= \sum(\beta_0 + \beta_1 X_i)\hat{e}_i$

$= \sum \hat{e}_i \beta_0 + \sum \hat{e}_i \beta_1 X_i$

$= \beta_0 \sum \hat{e}_i + \beta_1 \sum \hat{e}_i X_i$

Because $\sum \hat{e}_i$ is 0, we obtian

$\sum \hat{y}_i \hat{e}_i$

$= \beta_0 0 + \beta_1 0$

$= 0$

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

Proof:

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 $r_{xy} < 1$)?

Solution:

The coefficient of the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ is α

The coefficient of the regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ is $-\frac{\alpha}{\beta}$

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?

Because it does not control other factors that could impact children's reading performance. For example, children's reading performance could be improved by age. It will be better if there is another control group with the same age. One group uses the new program and one uses the old one. After a period of time see if there is any difference in reading performance between two groups.

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.