hw1_ma678

Yifu Dong

September 16, 2018

Part1: Pyth!

Situation

```
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.

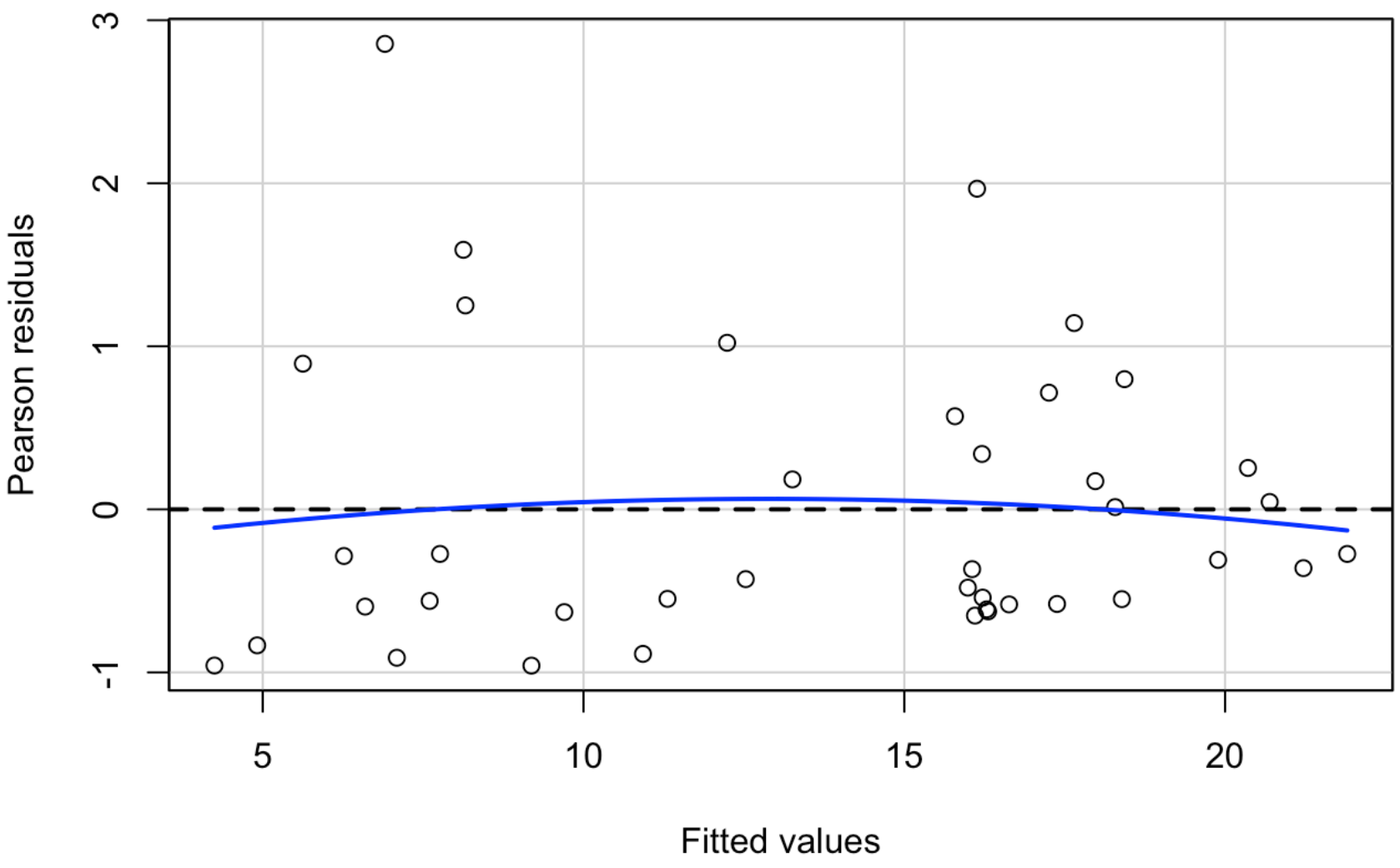
Solution

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.

```
##we have y,x1,x2 data now. lm(data=pyth)
x1 <- pyth$x1[1:40]
x2 <- pyth$x2[1:40]
y <- pyth$y[1:40]
regress1 <- lm(y~x1+x2)
summary(regress1)
```

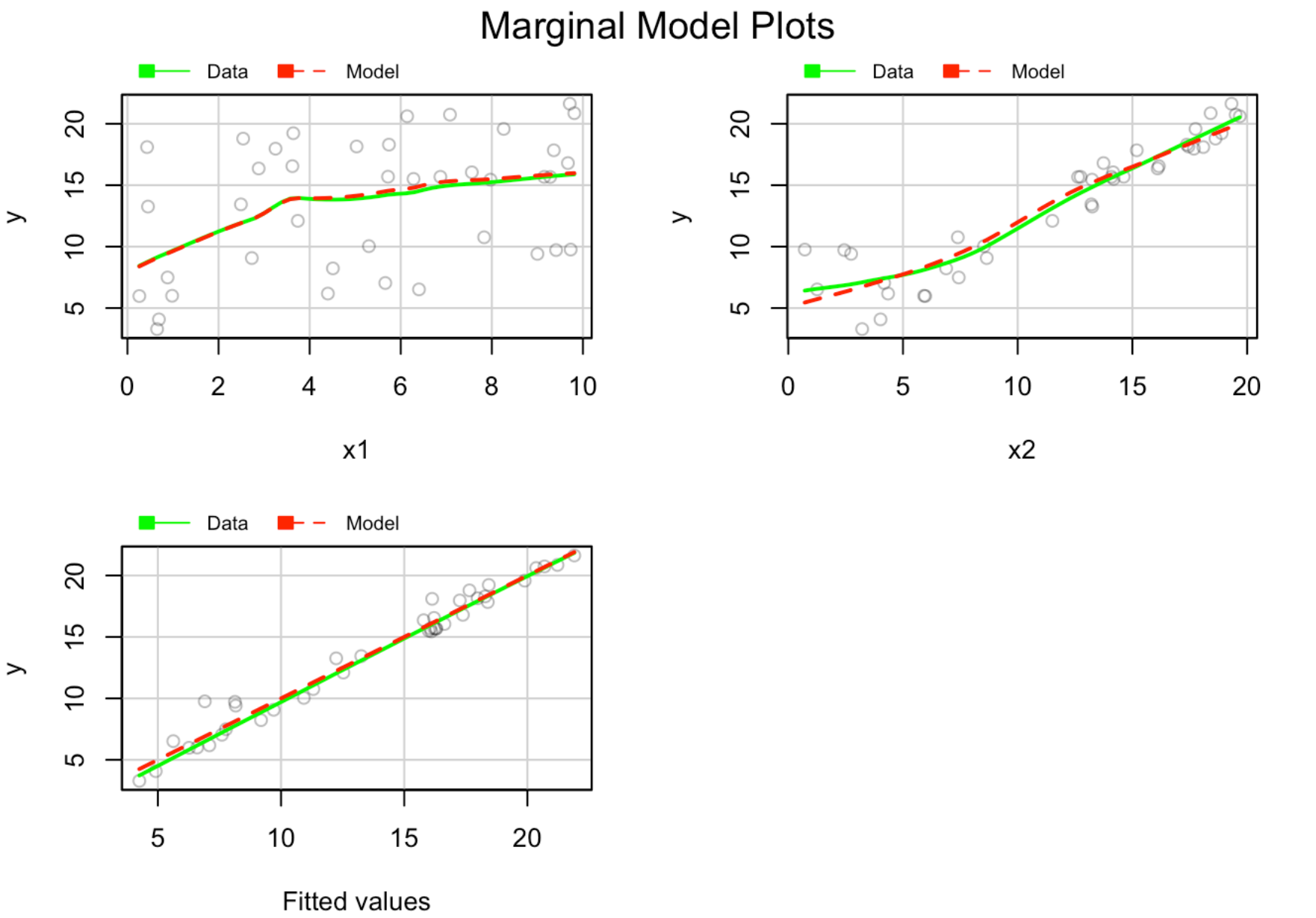
```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           0.51481  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

```
##check fit
##Residual vs Fitted
residualPlots(regress1, terms= ~ 1, fitted=TRUE)
```



```
##             Test stat Pr(>|Test stat|)
## Tukey test    -0.3576           0.7207
```

```
##overall fit
marginalModelPlots(regress1,col=rgb(0,0,0,alpha=0.3),col.line = c("green","red"))
```



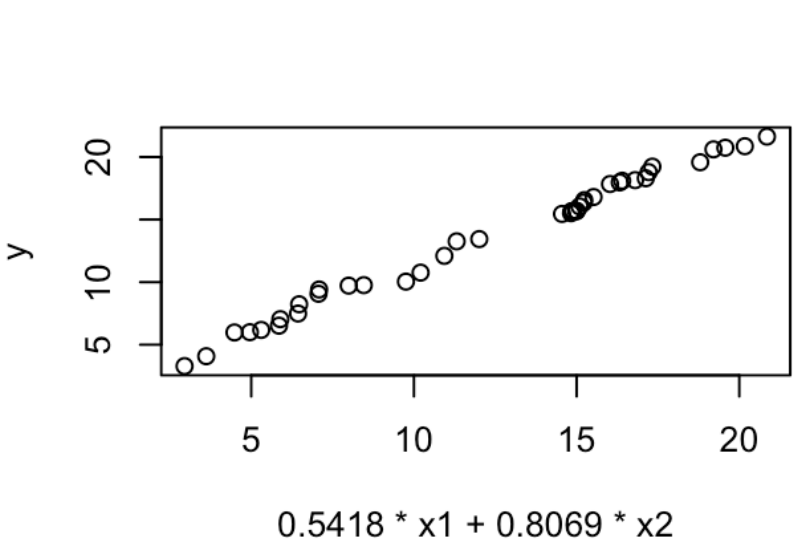

```
#p_value
#detecting heteroscedasticity
bptest(regress1)
```

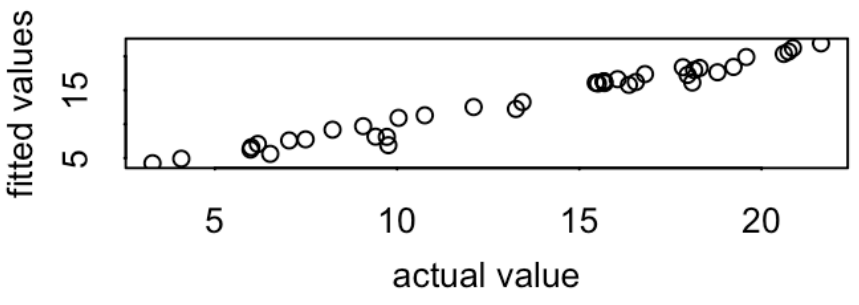
```
##
##  studentized Breusch-Pagan test
##
## data:  regress1
## BP = 6.0448, df = 2, p-value = 0.04868
```

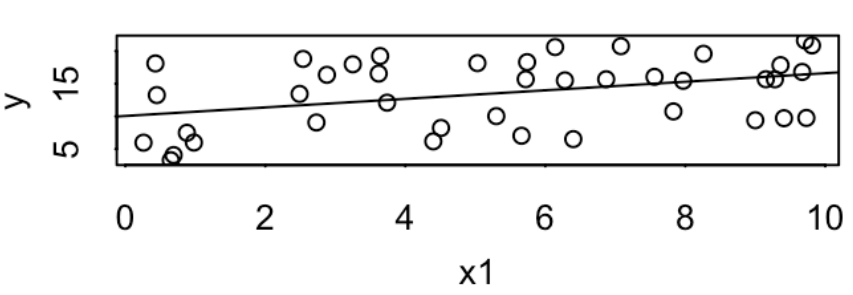
Notice that bp=6.0448, p_value=0.048<0.05,as well as the graphs, we can say this model is fitted well.

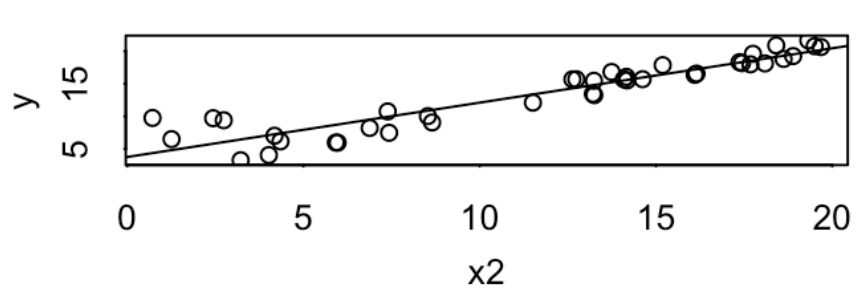
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
regress1 <- lm(y~x1+x2,data=pyth) ##regress
par(mfrow = c(2, 2))
##qqplot
qqplot(x=0.5418*x1+0.8069*x2,y=y)
##Fitted vs Actual
par (mar=c(10,3,2,1), mgp=c(2,.7,0), tck=-.01)
plot(pyth$y[1:40],fitted(regress1),xlab="actual value",ylab="fitted values")
##abline
plot(y~x1);abline(lm(y~x1))
plot(y~x2);abline(lm(y~x2))
```



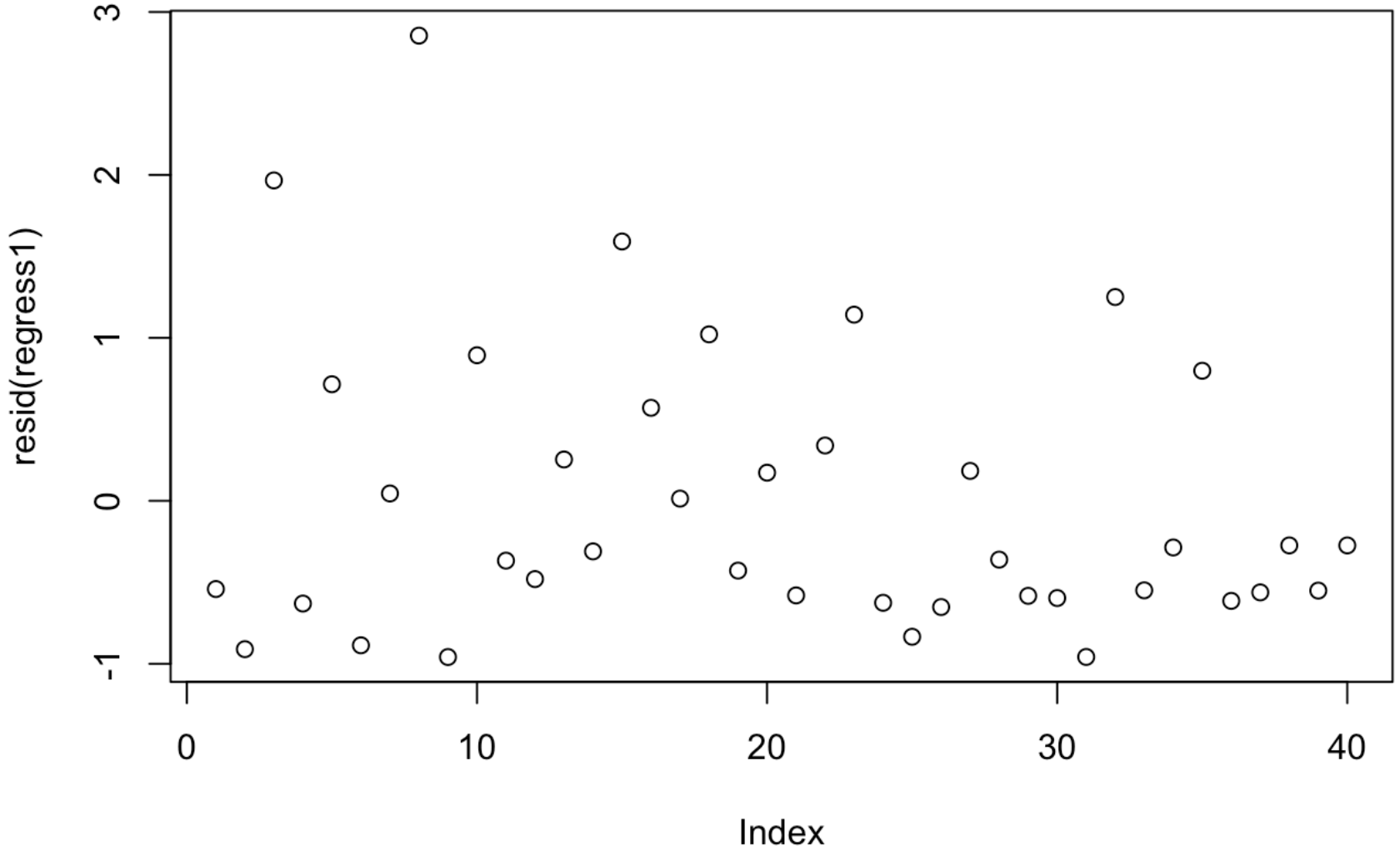




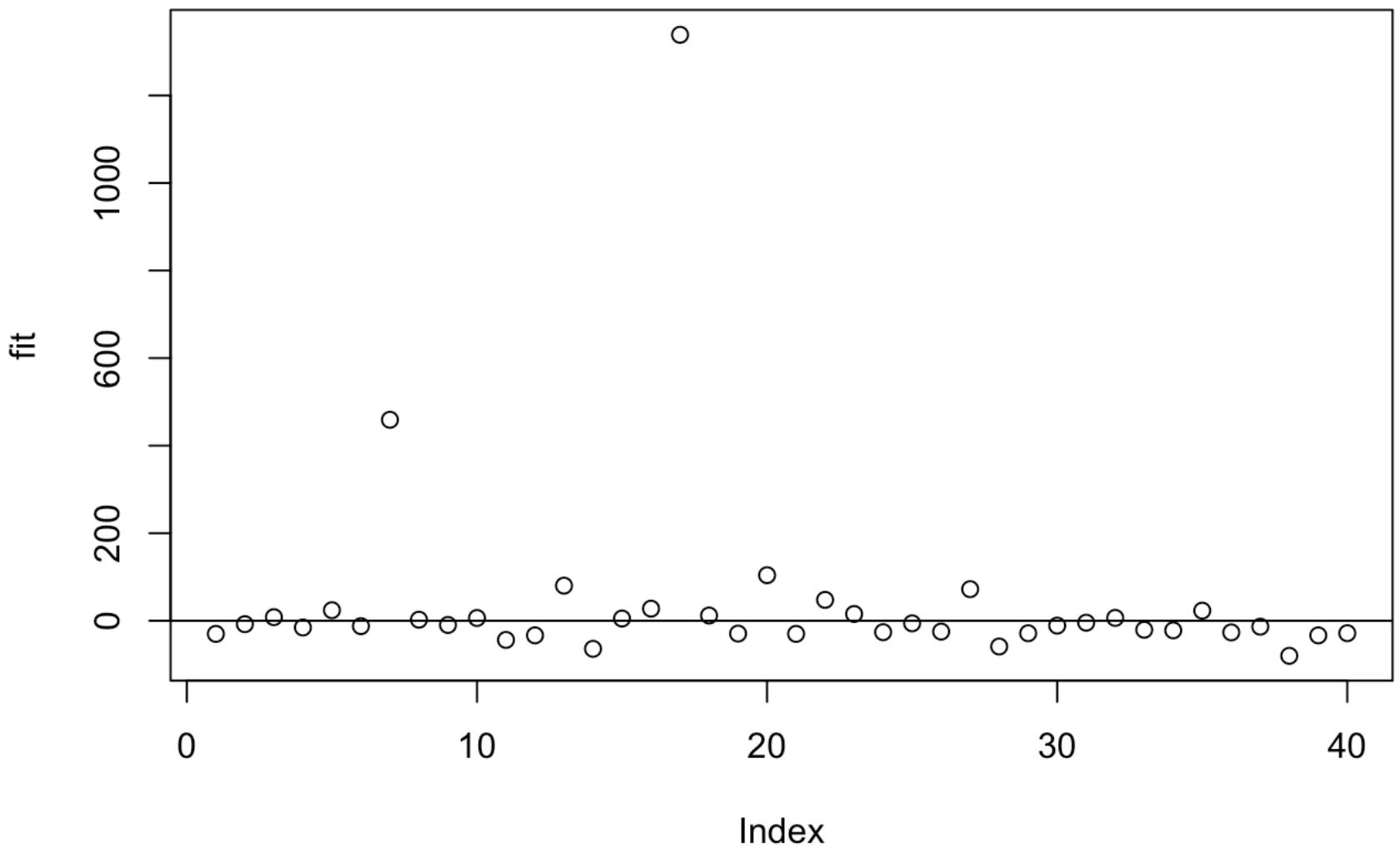


3. Make a residual plot for this model. Do the assumptions appear to be met?

```
plot(resid(regress1))
```




```
fit=fitted(regress1) / resid(regress1)
plot(fit)
abline(h=0)
```



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

```
confprediction <- predict(regress1, data.frame(x1=pyth$x1[41:60],x2=pyth$x2[41:60]
),level=0.95,interval='confidence')
confprediction
```

```
##           fit       lwr       upr
## 1  14.812484 14.295452 15.329516
## 2  19.142865 18.604860 19.680871
## 3   5.916816  5.203484  6.630147
## 4  10.530475 10.017798 11.043152
## 5  19.012485 18.501461 19.523509
## 6  13.398863 13.105741 13.691985
## 7   4.829144  4.258555  5.399733
## 8   9.145767  8.553508  9.738026
## 9   5.892489  5.313225  6.471752
## 10 12.338639 11.763150 12.914129
## 11 18.908561 18.424689 19.392433
## 12 16.064649 15.739276 16.390022
## 13  8.963122  8.510209  9.416036
## 14 14.972786 14.521738 15.423835
## 15  5.859744  5.326283  6.393204
## 16  7.374900  6.863539  7.886262
## 17  4.535267  3.940205  5.130330
## 18 15.133280 14.817297 15.449264
## 19  9.100899  8.654405  9.547393
## 20 16.084900 15.596495 16.573306
```

Part2: Earning and height

Situation

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.

Solution

1. Give the equation of the regression line and the residual standard deviation of the regression.

From what we know above, we can suppose that the equation is: log(earning)=a+b*log(height) So we can code as below:

```
# find the intercept----a
alpha = log(30000) - (0.008/0.01) * log(66)
height.example = 66
log.earnings = alpha + (0.008/0.01) * log(height.example)

#equation
log.earnings = log(30000) - (0.008/0.01) * log(66) + (0.008/0.01) * log(height.exa
mple)
# std
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?

Now we have sd=0.05/0.95. Since R2=SSR/SST=1-SSE/SST=1-sd^2/SST, so we have:

```
R2 <- 1 - (sd^2 / 0.05^2)
R2
```

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

Part3: Beauty and student evaluation

Situation

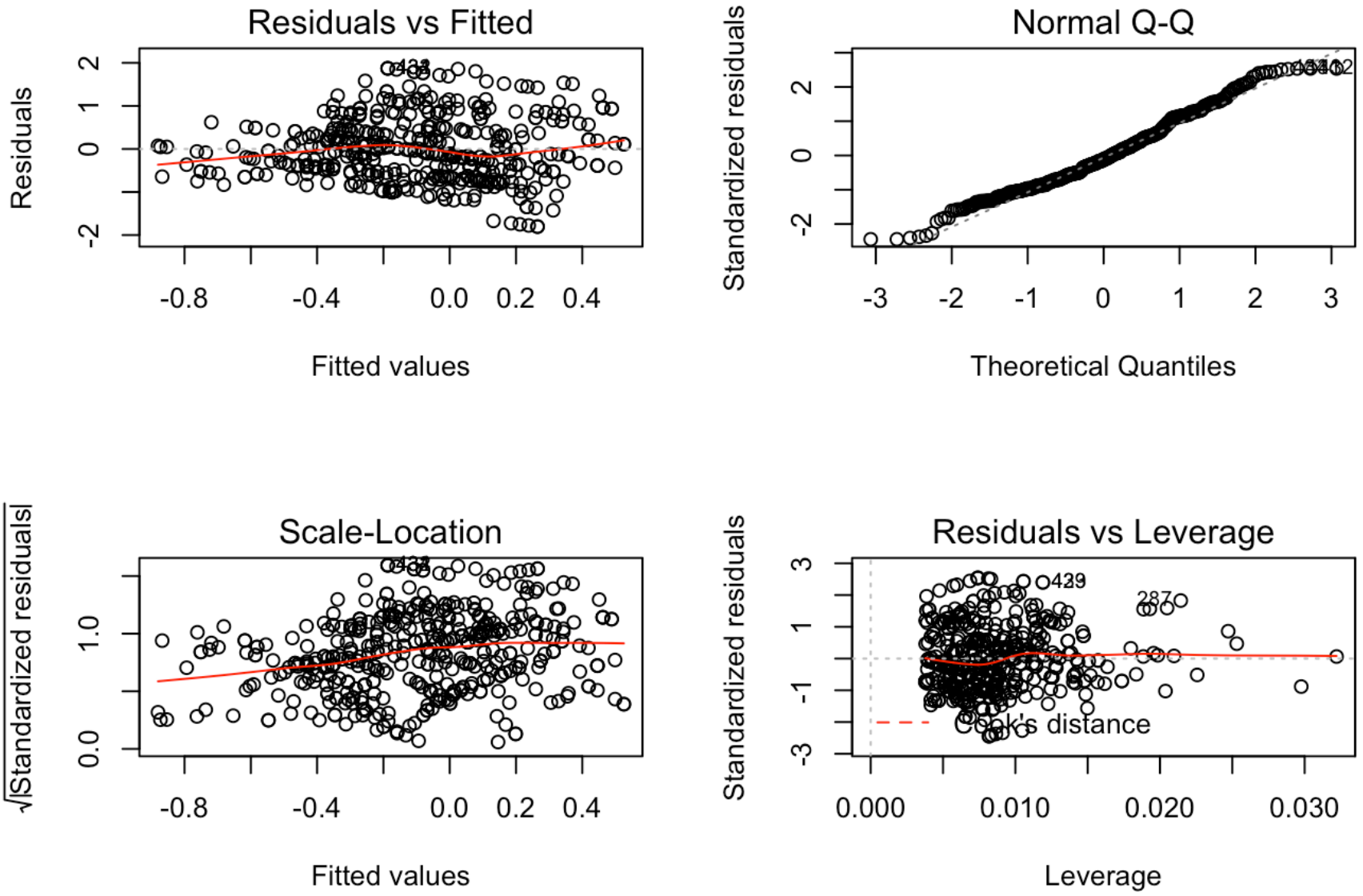
The folder beauty contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPubl
ic.csv"), header=T, sep=",")
```

Solution

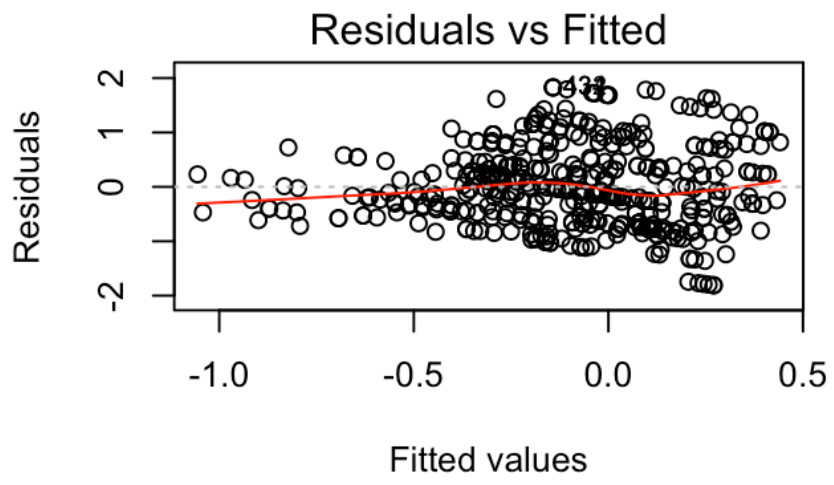
1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals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

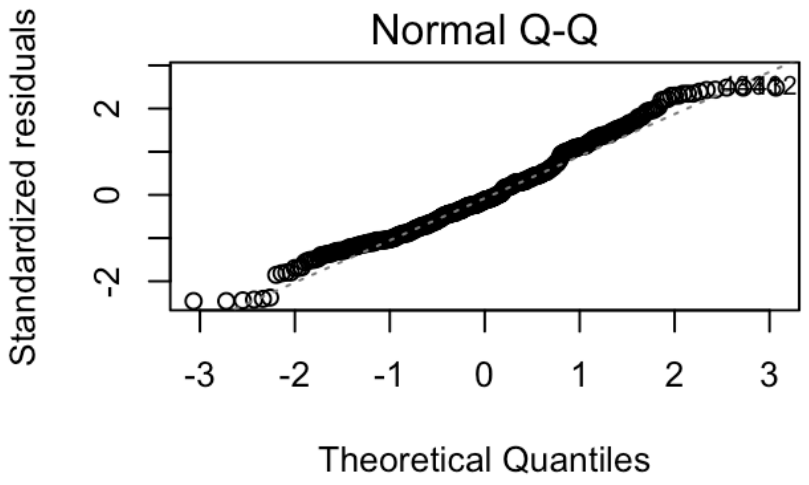
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPubl
ic.csv"), header=T, sep=",") #import
beauty1 <- lm(beauty.data$btystdave ~ beauty.data$courseevaluation + beauty.data$f
emale + beauty.data$age, data=beauty.data)
par(mfrow=c(2,2))
plot(beauty1)
```

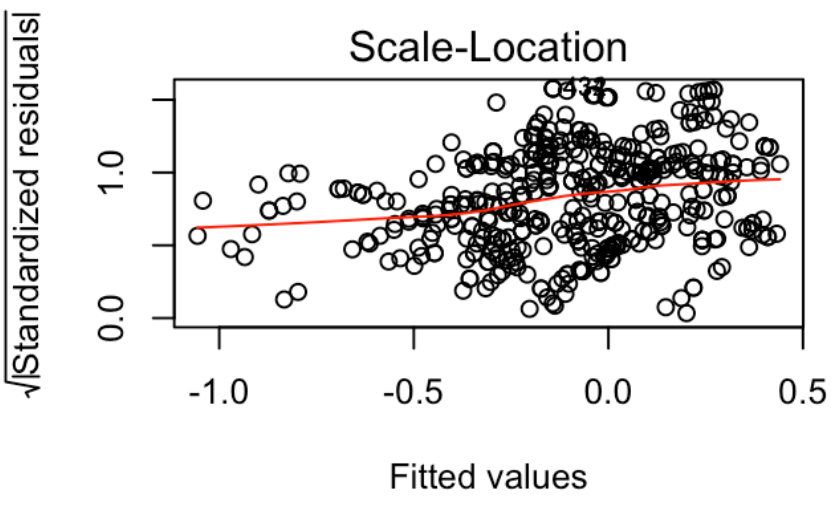


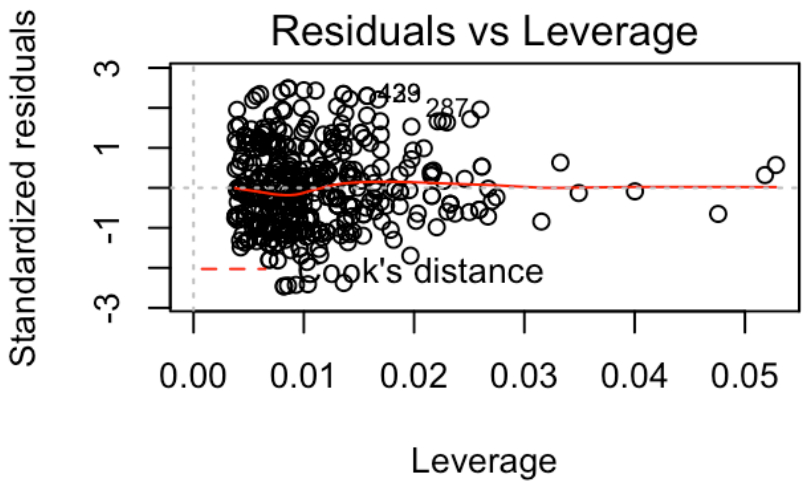
2. Fit some other models, including beauty and also other input variables. Consider at least one model with interactions. For each model, state what the predictors are, and what the inputs are, and explain the meaning of each of its coefficients.

```
beauty2 <- lm(beauty.data$btystdave ~ beauty.data$courseevaluation*beauty.data$fem
ale + beauty.data$age, data=beauty.data)
par(mfrow=c(2,2))
plot(beauty2)
```









Part4:Conceptula excercises on statistical significance

Situation

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.

Solution

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
```

-0.426<2, so we can conclude that the slope coef 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]
}
```

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

```
length(which(z.scores>2))
```

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

It means there are 1 out of there 100 z-scores are statistically significant, which means under 95% confidence level. So we believe that the slope is statistically significant at 5% level.

Part 5: Fit regression removing the effect of other variables

Situation

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.

Solution

a. Apply this procedure to the multiple regression of prestige on education, income, and percentage of women in the Canadian occupational prestige data, 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"))
fifthregress <- lm(Prestige$prestige~Prestige$education+Prestige$income+Prestige$women,data=Prestige)
summary(fifthregress)
```

```
##
## Call:
## lm(formula = Prestige$prestige ~ Prestige$education + Prestige$income +
##     Prestige$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 *
## Prestige$education  4.1866373  0.3887013  10.771  < 2e-16 ***
## Prestige$income     0.0013136  0.0002778   4.729 7.58e-06 ***
## Prestige$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
```

From the summary we can notice that the coeffcient of education is 4.187,at the same time the standard deviation is 0.3887, t_value is 10.78, which is relatively high, p_value is less than 2e-16. So we can say that the coefficient of education is properly recovered.

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

For this question, the first process which Regress Y on X2 through Xk will get the residual. We can know that the residual of Y will be a list of data consist of (Y-Yhat) The third process is to regress the residual Ey on residual Ex1. When X1 equals 0, where we can get the intercept, the value of Y is the exception the the residual. In this case, the residual Ey= 0, so the intercept for the simple regression in step 3 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”?

I think this statement is true. Because the first step and the second step is actually remove the influence of X2,⋯,Xk on X1 and Y, because the residual means (Y-Yhat), where Yhat is equal to B2 X2 + ⋯ + Bk Xk So, we can say like this.

(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 in a situation where there are many factors influencing the result, we can use this procedure to find the specific influence of the factor we want to research on the result. For example, we want to know whether there is a gender discrimination on people’s income. But there are many factors influencing the

income of male and female, not only we should consider the gender, but other factors should be included in this model,like the type of jobs female and male dominate, the different age group of male and female, etc. So we can simply use this way to find how much influence does “gender” have.

Part 6: Partial correlation

Situation

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

Solution

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

```
#(1)from the hint, we firstly try to find residuals EY|2,...,k. So we regress pres
tige on income and percentage women.
education <- Prestige$education
income <- Prestige$income
women <- Prestige$women
prestige <- Prestige$prestige
Y <- lm(prestige~income+women)
EY <- prestige-0.003334*income-0.132623*women
#then we regress education on income and percentage women.
X1 <- lm(education~income+women)
E1 <- education-0.0004826*income-0.0338047*women
r <- cor(EY,E1)
r
```

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

```
#we can calculate that r is equal to 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?

When r=0, it means Y and X1 are not related, which means Y and X1 are independant. So if we want ry1|2,…,k = 0, we have to change B1 so that X1 and Y are independant. So ry1|2,…,k = 0 if and only if B1 is 0.