MA678 homework 01

Tingrui Huang

Septemeber 13, 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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

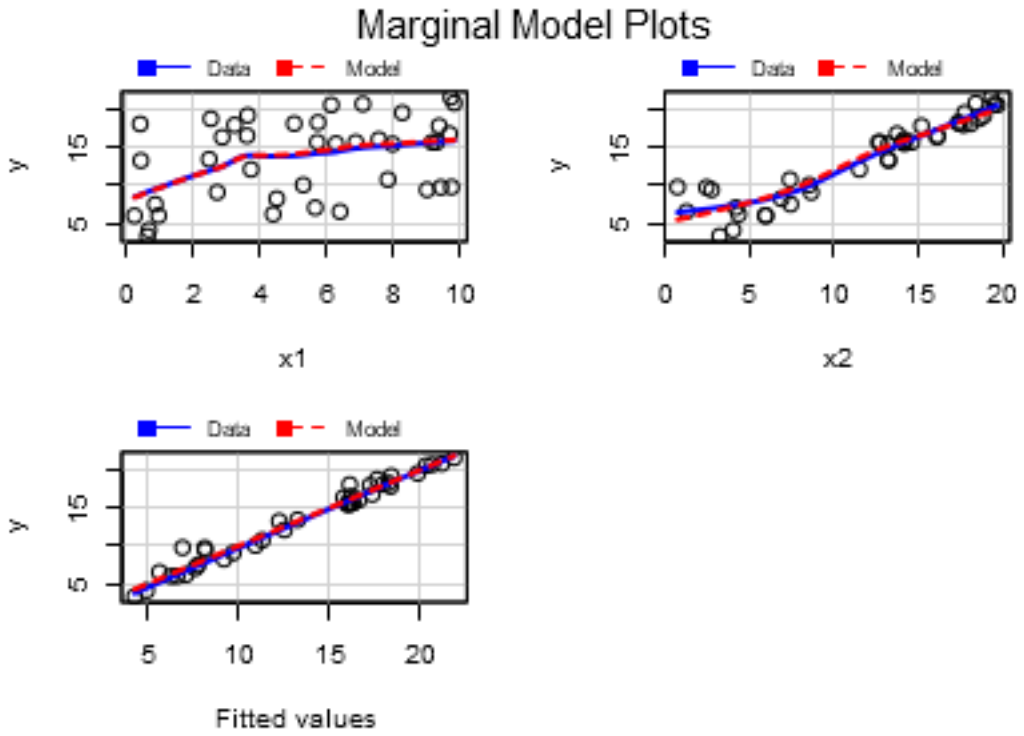
```
pyth40 <- pyth %>% head(40)
first40 <- lm(y ~ x1 + x2, data = pyth40)
summary(first40)
```

```
##
## 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
```

```
#Based on the Summary table, the P value of x1 and x2 are below 0.05, and thus I would say they are sig
#The R square is very close to 1. According to these results, I would say the fit of model is good.
```

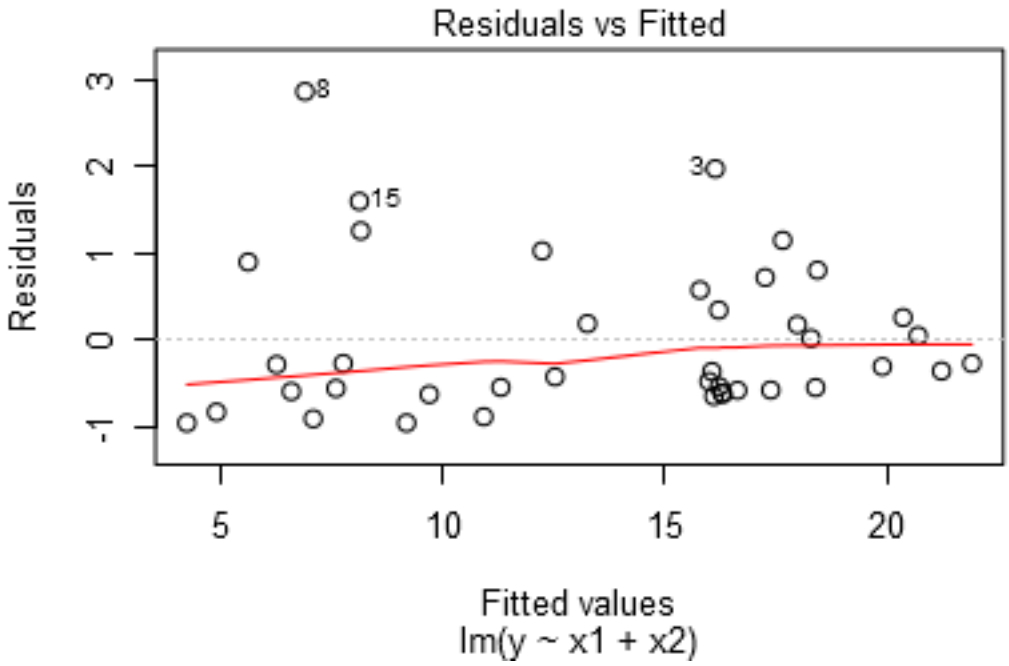
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
par(mfrow=c(2,2))
car::marginalModelPlots(first40)
```



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

```
plot(first40,which=1)
```




```
# 1.The expectation of residual should be equal to 0. As we can see in the graph below, the red line is
#from line 0, so that the residual is biased. 2.The points are distributed randomly and the range of al
#residuals is not too large.I would say the assumptions are not met well.
```

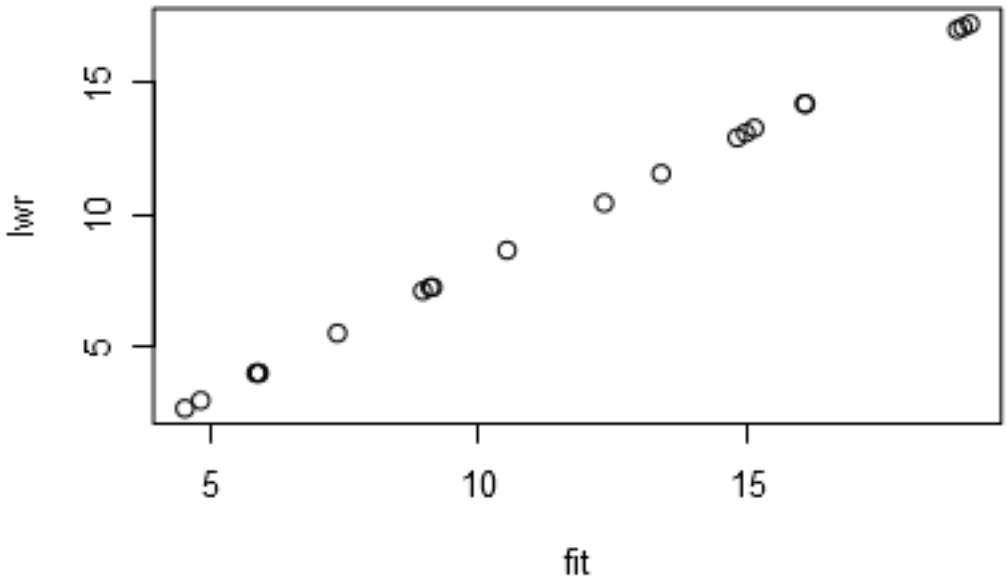
4. Make predictions for the remaining 20 data points in the file. How confident do you feel about these predictions?

```
last20 <- pyth %>% tail(20)
print(last20)
```

```
##     y   x1    x2
## 41 NA 9.87 10.43
## 42 NA 9.99 15.72
## 43 NA 8.39  0.35
## 44 NA 0.80 10.91
## 45 NA 9.58 15.82
## 46 NA 4.82 11.90
## 47 NA 2.97  2.46
## 48 NA 8.80  4.09
## 49 NA 6.07  1.80
## 50 NA 0.19 13.54
## 51 NA 4.19 19.13
## 52 NA 5.39 14.84
## 53 NA 6.58  5.28
## 54 NA 2.36 15.42
## 55 NA 2.37  4.12
## 56 NA 1.52  6.54
## 57 NA 2.07  2.67
```

```
## 58 NA 6.70 12.85
## 59 NA 2.02  8.36
## 60 NA 9.63 12.16
```

```
pl20 <- predict(first40, last20, interval = "prediction", level = 0.95)
plot(pl20)
```




```
# Since the distribution of the predicted data is linear, so I would say the predictions are good.
```

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. #regressional line: log(30000) = B + (0.008/0.01)*log(66) therefore B = 6.957229 log(earnings) = 6.957229 + 0.8*log(height) #residual standard deviation: P(log(y)+log(1.1))=0.95

```
resi_sd <- log(1.1)/2
resi_sd
```

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

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? Since $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$ and $SSE = VAR_{residual} * n^2$ and $SSR = VAR_{egression} * n^2$

```
    r_square <- 1-(resi_sd^2/0.05^2)
    r_square
```

```
## [1] 0.09159696
```

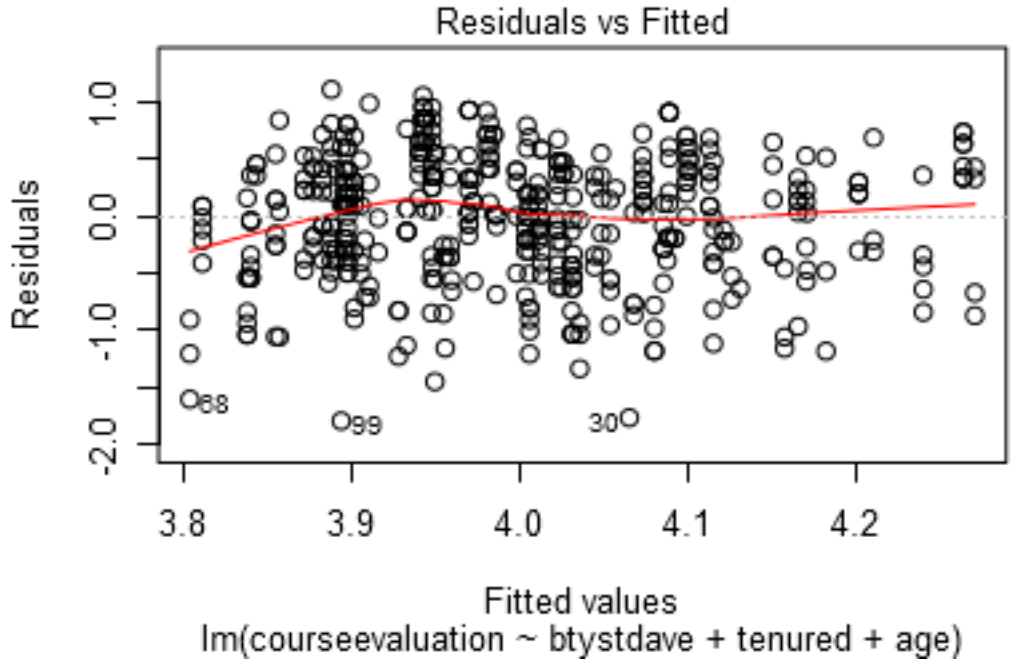
Beauty and student evaluation

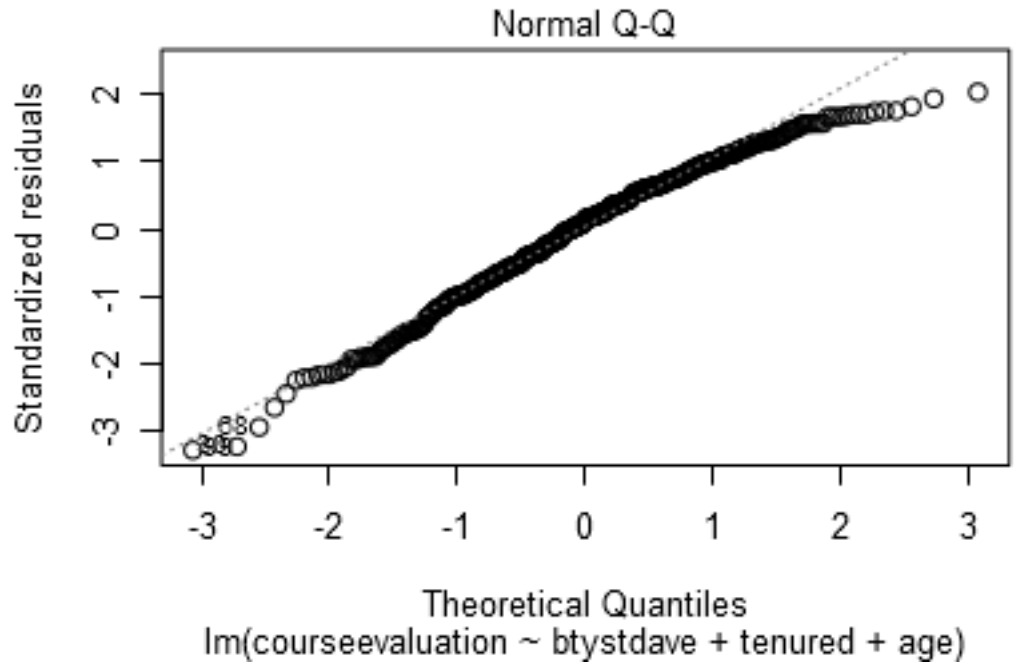
The folder beauty contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students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

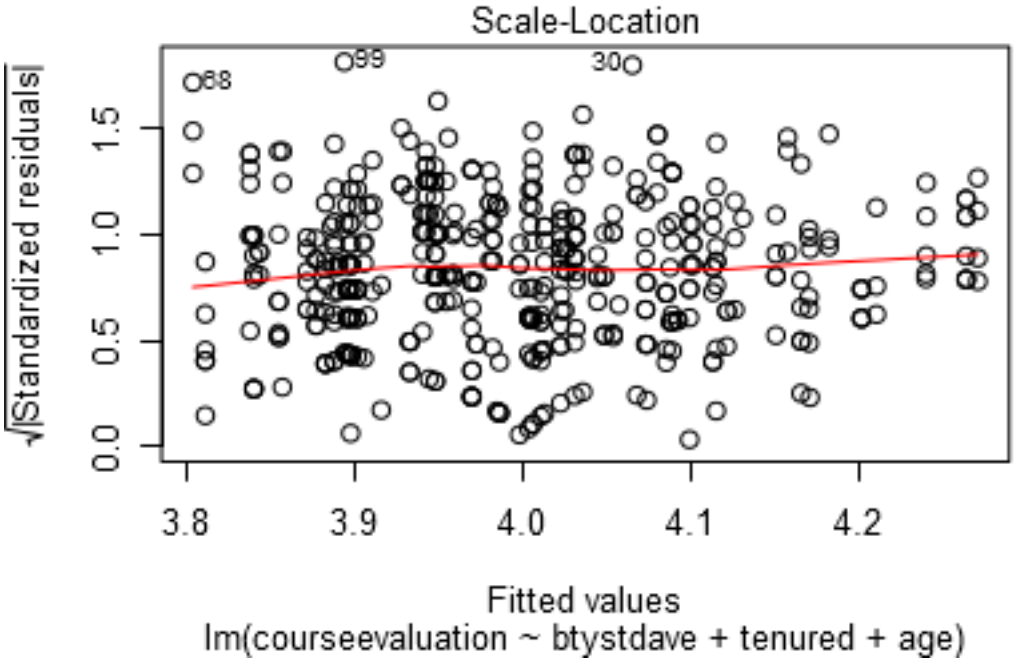
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, s
```

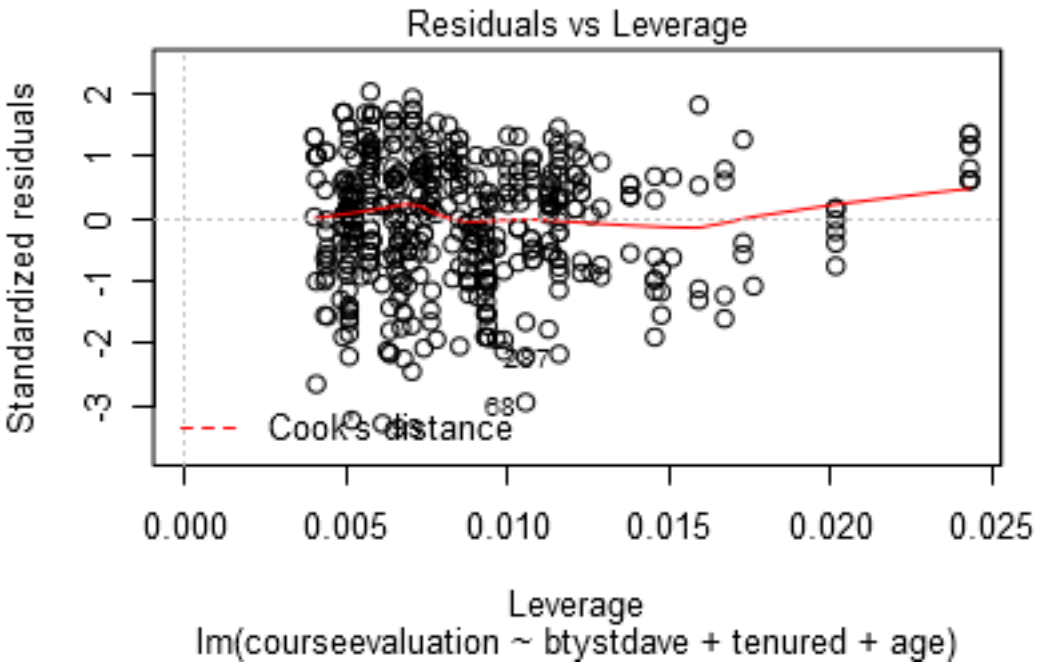
1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals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

```
breg1 <- lm(courseevaluation ~ btystdave + tenured + age, data = beauty.data)
plot(breg1)
```

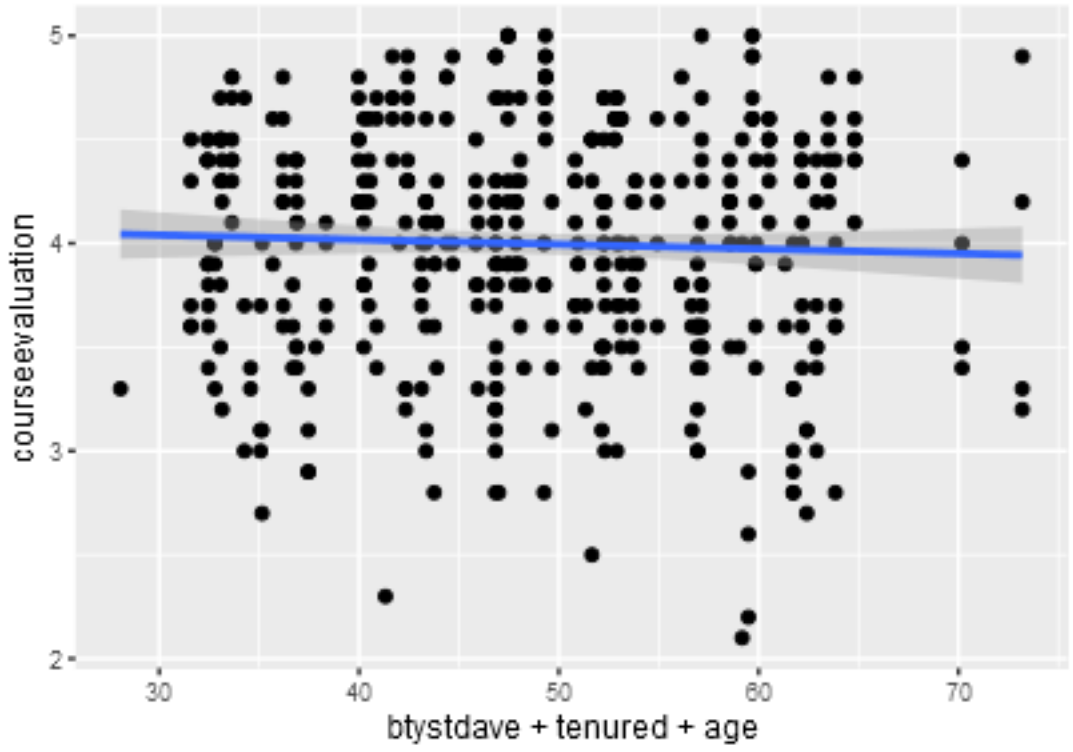



Normal Q-Q
Standardized residuals
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(courseevaluation ~ btystdave + tenured + age)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3.8
3.9
4.0
4.1
4.2
Fitted values
lm(courseevaluation ~ btystdave + tenured + age)




```
ggplot(breg1, aes(x=btystdave + tenured + age, y=courseevaluation)) + geom_point() + geom_smooth(method
```




```
summary(breg1)
```

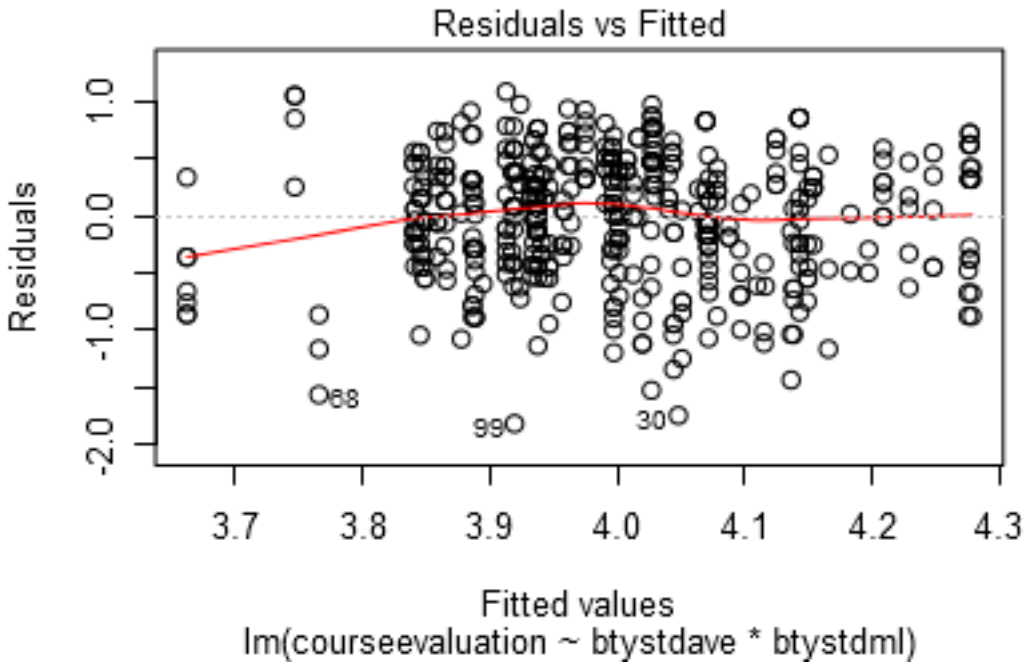
```
##
```

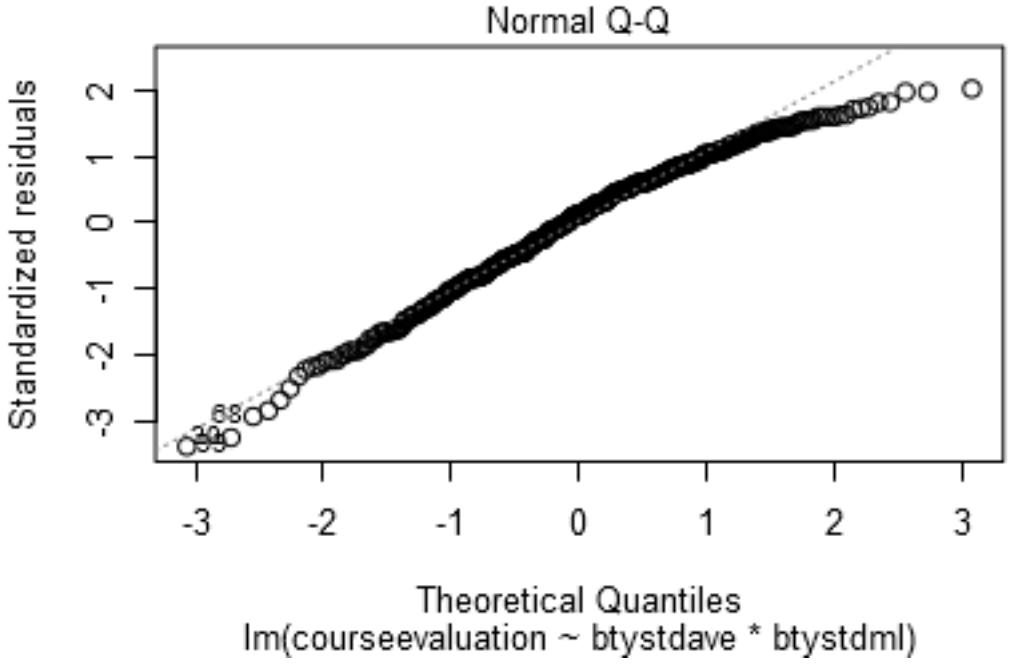
```
## Call:
## lm(formula = courseevaluation ~ btystdave + tenured + age, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79410 -0.35538  0.06695  0.40163  1.11178
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.981773   0.134720  29.556  < 2e-16 ***
## btystdave    0.132830   0.033815   3.928 9.88e-05 ***
## tenured     -0.037166   0.055603  -0.668    0.504
## age          0.001004   0.002920   0.344    0.731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5464 on 459 degrees of freedom
## Multiple R-squared:  0.0367, Adjusted R-squared:  0.0304
## F-statistic: 5.829 on 3 and 459 DF,  p-value: 0.0006492
```

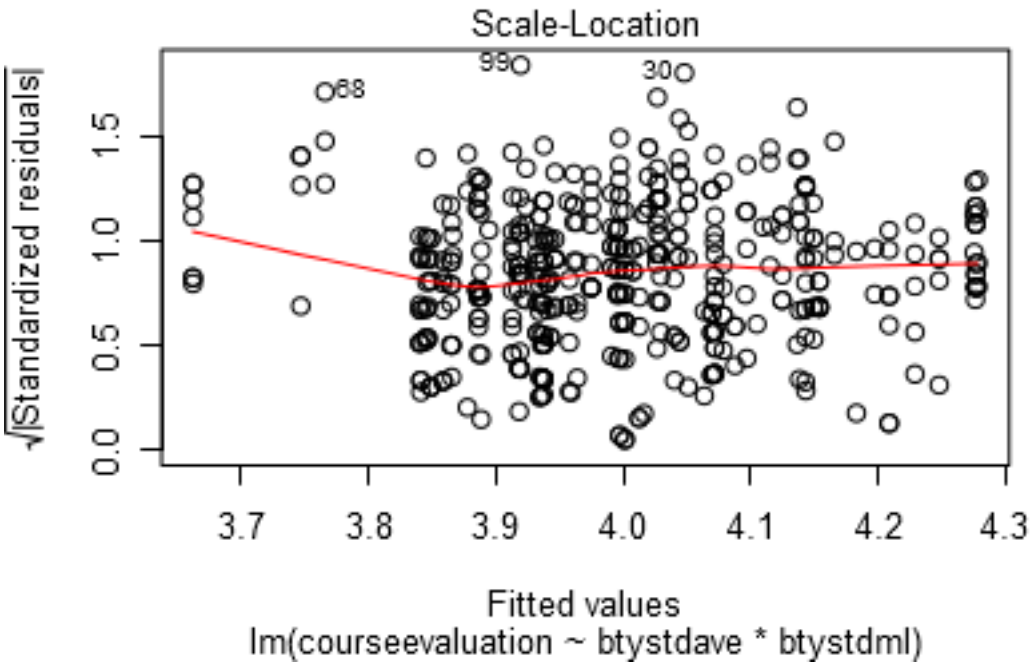
```
# From the P value we can tell that the "tenured" and "age" are not significant since their P value is
#than 0.05. While the p value of "bytstdave" is well smaller than 0.05, therefore it's significant. The
#of "btystdave" is 0.132830, which means it has positive correlation with the course evaluation, and th
#a instructor, the higher evaluation score that instructor will get.
```

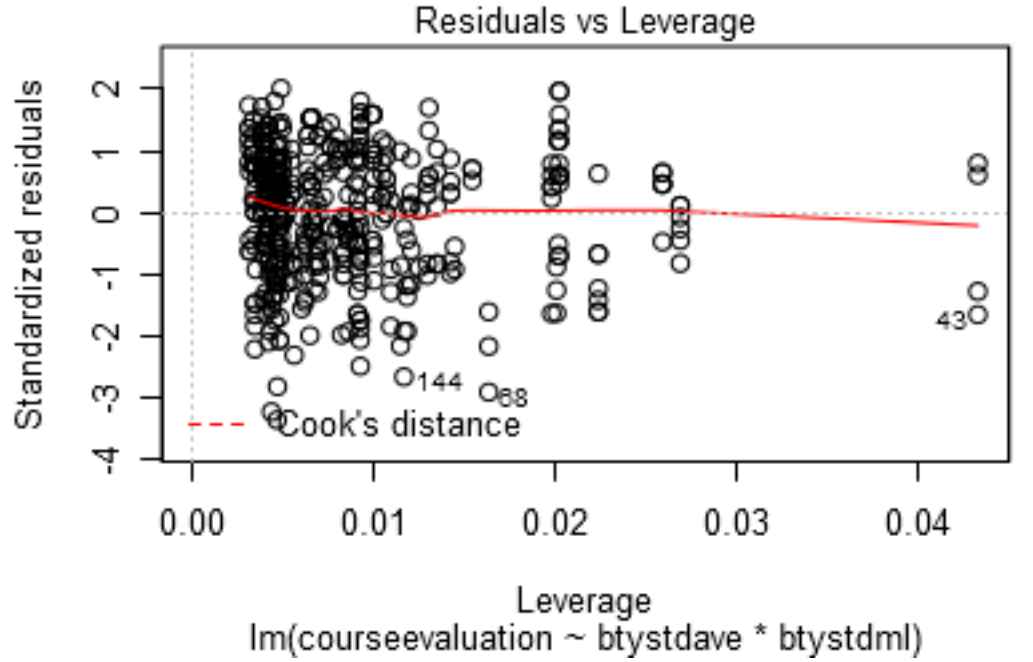
2. Fit some other models, including beauty and also other input variables. Consider at least one model with interactions. For each model, state what the predictors are, and what the inputs are, and explain the meaning of each of its coefficients.

```
breg2 <- lm(courseevaluation ~ btystdave*btystdml, data = beauty.data)
plot(breg2)
```


Residuals vs Fitted
Residuals
1.0
0.0
-1.0
-2.0
68
99
30
3.7
3.8
3.9
4.0
4.1
4.2
4.3
Fitted values
lm(courseevaluation ~ btystdave * btystdml)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
68
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(courseevaluation ~ btystdave * btystdml)






```
summary(breg2)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave * btystdml, data = beauty.data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81935 -0.36027  0.06242  0.40135  1.08719
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.00174    0.03315 120.707  < 2e-16 ***
## btystdave           0.23841    0.05305   4.495 8.84e-06 ***
## btystdml           -0.11418    0.04283  -2.666  0.00795 **
## btystdave:btystdml  0.01586    0.03387   0.468  0.63969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5424 on 459 degrees of freedom
## Multiple R-squared:  0.0508, Adjusted R-squared:  0.0446
## F-statistic: 8.189 on 3 and 459 DF,  p-value: 2.549e-05
```

```
# The inputs are "btystdave" and "btystdml", predictors include all inputs plus the interaction between
#predictors. Both predictors are significant and "btystdave" has positive correlation while "btystdml"
#negative correlation.
```

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
## 0.4882344
```

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?

```
z_100 <- abs(z.scores)>=2
sum(z_100)
```

```
## [1] 9
```

```
#There are seven of them are statistically significant.
```

What can you say about statistical significance of regression coefficient? *Since there are only 7 out of 100 z-scores are statistically significant, I would the regression coefficient is not significant.*

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"))
prreg1 <- lm(prestige ~ income+women, data = Prestige)
prreg2 <- lm(education ~ income+women, data = Prestige)
resreg1 <- lm(resid(prreg1) ~ resid(prreg2))
resreg1$coefficients
```

```
##   (Intercept) resid(prreg2)
## -2.991663e-15  4.186637e+00
```

```
prreg_tradition <- lm(prestige ~ education+income+women, data = Prestige)
prreg_tradition$coefficients
```

```
##  (Intercept)    education       income        women
## -6.794334203  4.186637275  0.001313560 -0.008905157
```

```
# as we can see, by using "resreg1$coefficients" we get the coefficient of education as 4.18663.
# By using the trditional way, the coefficient of education is the same value.
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case? *The expectation of both residuals of "prreg1" and "prreg2" are 0. When we conduct the third step, we have* $E(\epsilon_y) = E(\alpha * \epsilon_{x1} + \beta) = 0$*, so that* β *has to be zero.*

(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 would say the statement is reasonable since we have removed all other predictors from the model.*

(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. #Using the residuals from previous problem, we can get partial correlation 0.7362

```
predcor <- cor(resid(prreg1),resid(prreg2))
predcor
```

```
## [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. $\sum \hat{y}_i \hat{e}_i = 0$
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? *This is a weak research design since the researchers only picked those students that had below average reading level, and this process would result in huge sampling bias. In order to improve the research design, researchers need to go back to the very beginning and conduct a random sampling withing the population.*

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.

① $\sum \hat{y}_i \hat{e}_i = 0$

$\hat{y}_i = \hat{y}_1 \cdots \hat{y}_n \quad \hat{e}_i = \hat{e}_1, \cdots, \hat{e}_n$

$\therefore \sum \hat{y}_i \hat{e}_i = (\hat{y}_1, \hat{y}_2, \cdots \hat{y}_n)(\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n)^T = \hat{y}^T \hat{e}$

$\because \hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$

$\hat{e} = y - \hat{y}$

$\therefore \sum \hat{y}_i \hat{e}_i = [X(X^TX)^{-1}X^Ty]^T [y - X(X^TX)^{-1}X^Ty]$

$= [Hy]^T [y - Hy]$

$= [y^TH^T][y - Hy]$

$= y^TH^Ty - y^TH^THy$

$H^T = H \quad HH = H$

$\therefore \sum \hat{y}_i \hat{e}_i = y^THy - y^THy = 0$

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

$y_i - \hat{y}_i = \hat{e}_i$

$\therefore \sum \hat{e}_i (\hat{y}_i - \bar{y}) = \hat{e}_i^T (\hat{y}_i - \bar{y}) = \hat{e}_i^T \hat{y}_i - \hat{e}_i^T \cdot \bar{y}$

$\hat{e}_i^T \hat{y}_i = 0$

$\therefore \sum \hat{e}_i (\hat{y}_i - \bar{y}) = -\hat{e}^T \bar{y} \qquad \bar{y} = \frac{\sum y}{n}$

$\therefore -\hat{e}^T \bar{y} = -\bar{y} \sum \hat{e}$

$E(\hat{e}) = 0 \Rightarrow \sum \hat{e} = 0$

$\therefore -\hat{e}^T \bar{y} = 0$

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