MA678 homework 01

Shiyu Zhang

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.

```
pyth[(1:40),]
```

```
##         y   x1    x2
## 1   15.68 6.87 14.09
## 2    6.18 4.40  4.35
## 3   18.10 0.43 18.09
## 4    9.07 2.73  8.65
## 5   17.97 3.25 17.68
## 6   10.04 5.30  8.53
## 7   20.74 7.08 19.50
## 8    9.76 9.73  0.72
## 9    8.23 4.51  6.88
## 10   6.52 6.40  1.26
## 11  15.69 5.72 14.62
## 12  15.51 6.28 14.18
## 13  20.61 6.14 19.68
## 14  19.58 8.26 17.75
## 15   9.72 9.41  2.44
## 16  16.36 2.88 16.10
## 17  18.30 5.74 17.37
## 18  13.26 0.45 13.25
## 19  12.10 3.74 11.51
## 20  18.15 5.03 17.44
## 21  16.80 9.67 13.74
```

```
## 22 16.55 3.62 16.15
## 23 18.79 2.54 18.62
## 24 15.68 9.15 12.74
## 25  4.08 0.69  4.02
## 26 15.45 7.97 13.24
## 27 13.44 2.49 13.21
## 28 20.86 9.81 18.41
## 29 16.05 7.56 14.16
## 30  6.00 0.98  5.92
## 31  3.29 0.65  3.22
## 32  9.41 9.00  2.74
## 33 10.76 7.83  7.39
## 34  5.98 0.26  5.97
## 35 19.23 3.64 18.89
## 36 15.67 9.28 12.63
## 37  7.04 5.66  4.18
## 38 21.63 9.71 19.32
## 39 17.84 9.36 15.19
## 40  7.49 0.88  7.43
```

```r
y<-pyth$y
x1<-pyth$x1
x2<-pyth$x2
lm(y~x1+x2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1            x2
##      1.3151       0.5148        0.8069
```

```r
summary(lm(y~x1+x2))
```

```
##
## 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
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

"as the summary outpout shows, beta is 1.315, 0.514, 0.806, respectively which matched the linear
 regression model i just created ,all coefficient are statistically significant (p-value < 0.01),
 what's more, the adjusted r square is 0.97 which is close to 1, shows that the model is
 appropraite. we can say that the model fits well."

[1] "as the summary outpout shows, beta is 1.315, 0.514, 0.806, respectively which matched the linea

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

```
library(ggplot2)
library(Cairo)
ggplot(pyth)+aes(x=x1+x2,y=y)+geom_point()+ylab("y")+xlab("x1+x2")+geom_smooth(method="lm",se=FALSE)
```

Warning: Removed 20 rows containing non-finite values (stat_smooth).

Warning: Removed 20 rows containing missing values (geom_point).
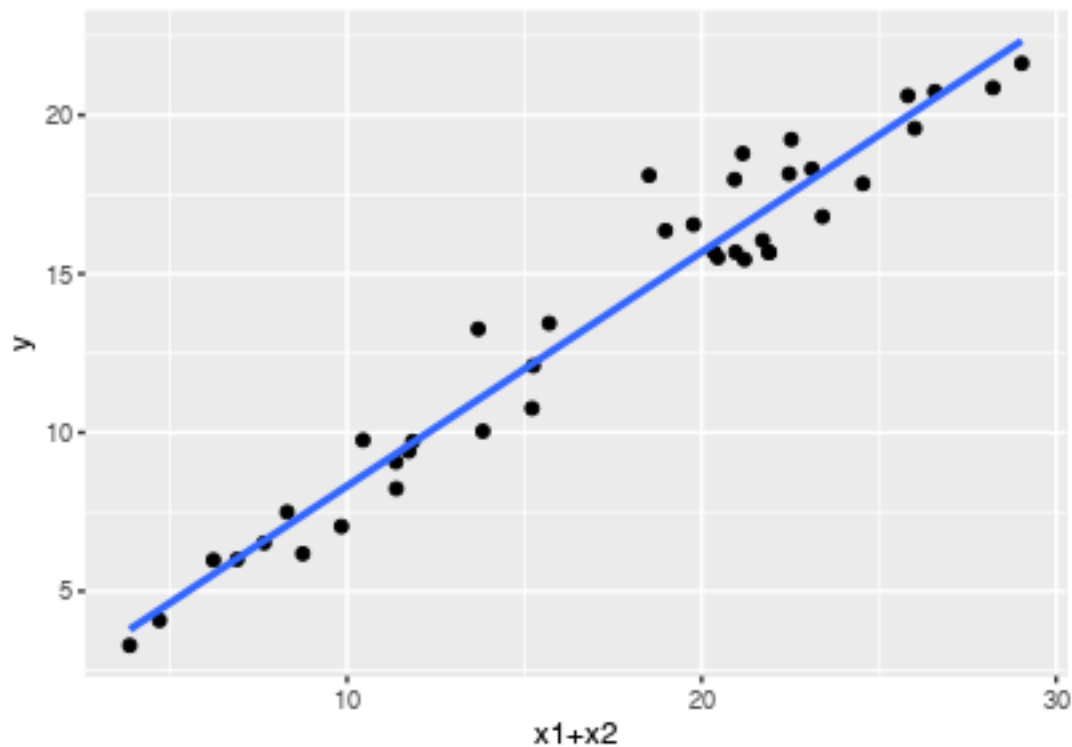

3. Make a residual plot for this model. Do the assumptions appear to be met?

```
a<-lm(y~x1+x2)
residual<-resid(a)
residual
```

```
##           1           2           3           4           5           6
## -0.54137824 -0.91040037  1.96632254 -0.63042100  0.71539425 -0.88665335
##           7           8           9          10          11          12
##  0.04507693  2.85477816 -0.95853592  0.89336021 -0.36701364 -0.48026286
##          13          14          15          16          17          18
##  0.25375318 -0.31029015  1.59161590  0.57080695  0.01366145  1.02151685
##          19          20          21          22          23          24
## -0.42816934  0.17269245 -0.58042548  0.33950129  1.14240527 -0.62580455
##          25          26          27          28          29          30
```

3

```
## -0.83417039 -0.65178806  0.18358044 -0.36081314 -0.58308177 -0.59661250
##          31           32           33           34           35           36
## -0.95804235  1.25061230 -0.54923536 -0.28629500  0.79824557 -0.61396875
##          37           38           39           40
## -0.56188513 -0.27362887 -0.55086758 -0.27357995
```

```
plot(residual)
```
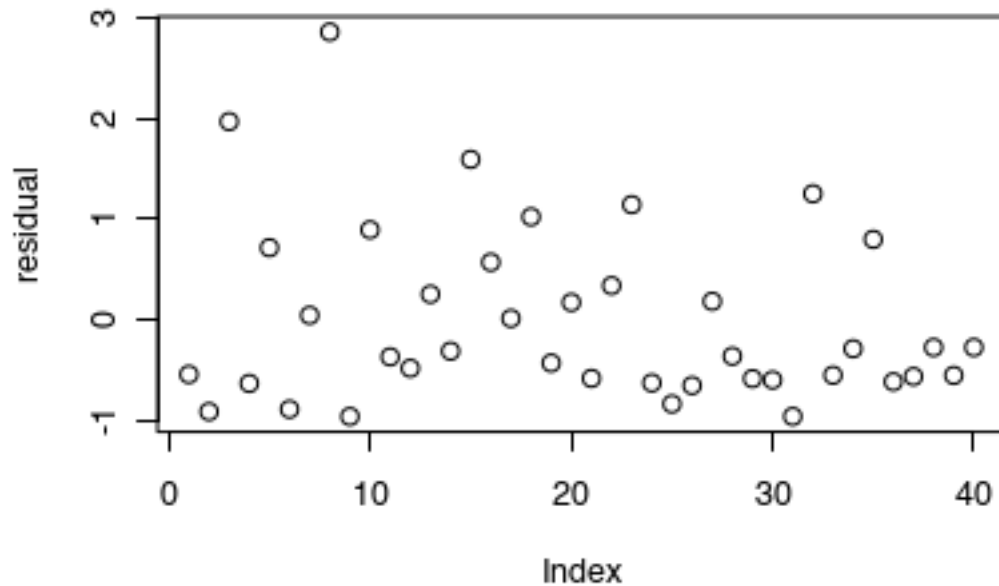


```
"the assumption doesn't meet the requirement since residual expectation does not equal
to zero - doesn't meet the unbiased assumption."
```

```
## [1] "the assumption doesn't meet the requirement since residual expectation does not equal \nto zero
```

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

```
pyth2<-pyth[41:60,]
pyth2
```

```
##     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
```

```r
predict(a,newdata=pyth2,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
```

```
## based on the upper level and lower level output, the result can be seen.
```

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.

answer:

alpha = log(30000) - (0.008/0.01) * log(66) # for the y-intercept alpha # alpha = 6.957729 log_y = alpha + (0.008/0.01) * log(66) log_y # log_y=10.30895 exp(log.y) # the equation is: log_earnings=6.957229+(0.008/0.01)*log(height)

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?

answer:

sd=0.1*0.5/0.95 # sd_population=0.05263158 R2=(sd$^{2/0.05}$2) R2=1.10

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.

```
btystdave=beauty.data$btystdave
courseevaluation=beauty.data$courseevaluation
b=lm(courseevaluation~btystdave, data=beauty.data)
summary(b)
```

```
##
## 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
```

```
plot(btystdave, courseevaluation,xlab="Beauty", ylab="Course evaluation")
```
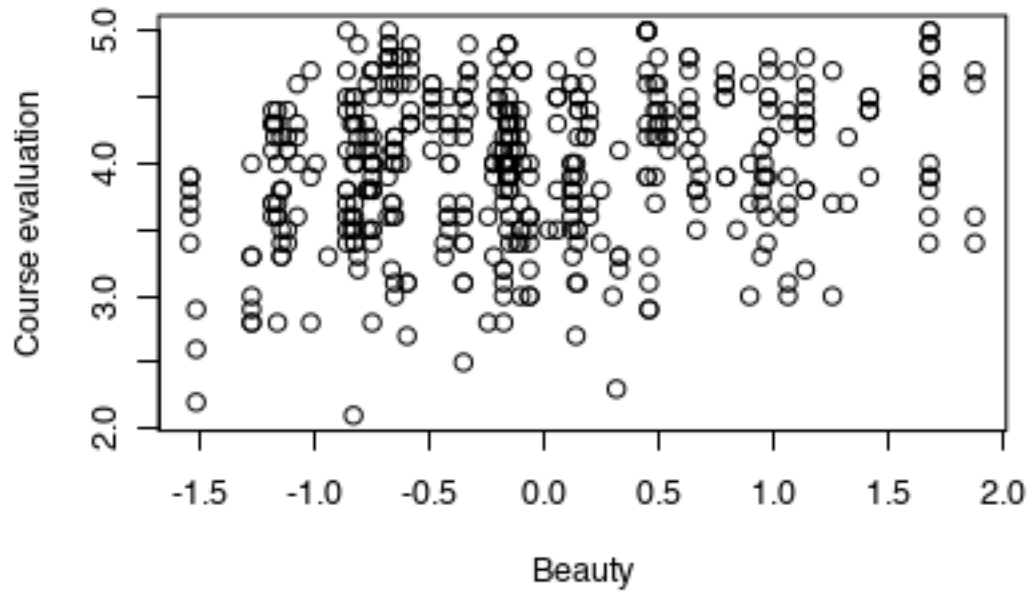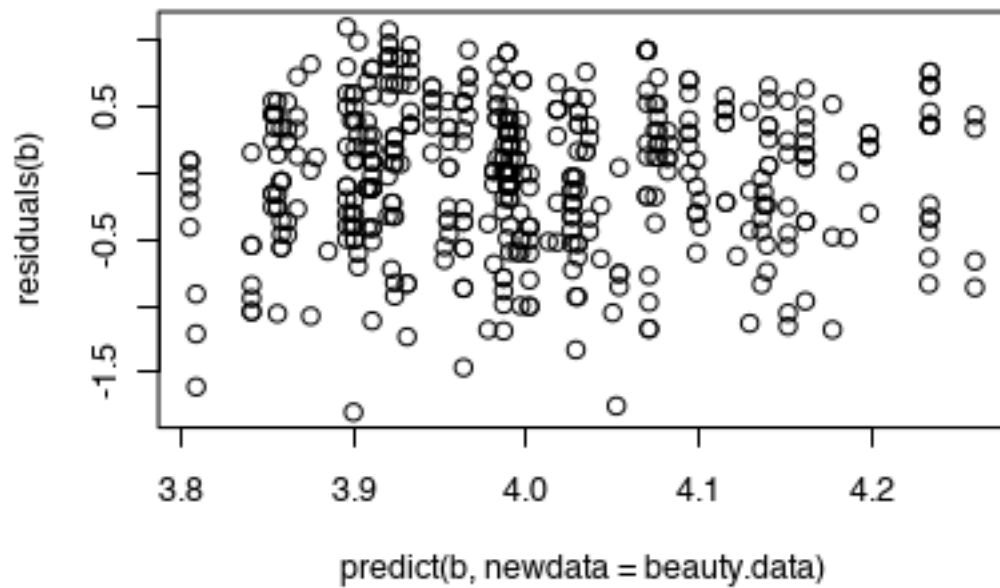


```
coef(b)
```

```
## (Intercept)    btystdave
##   4.0100227    0.1330014
```
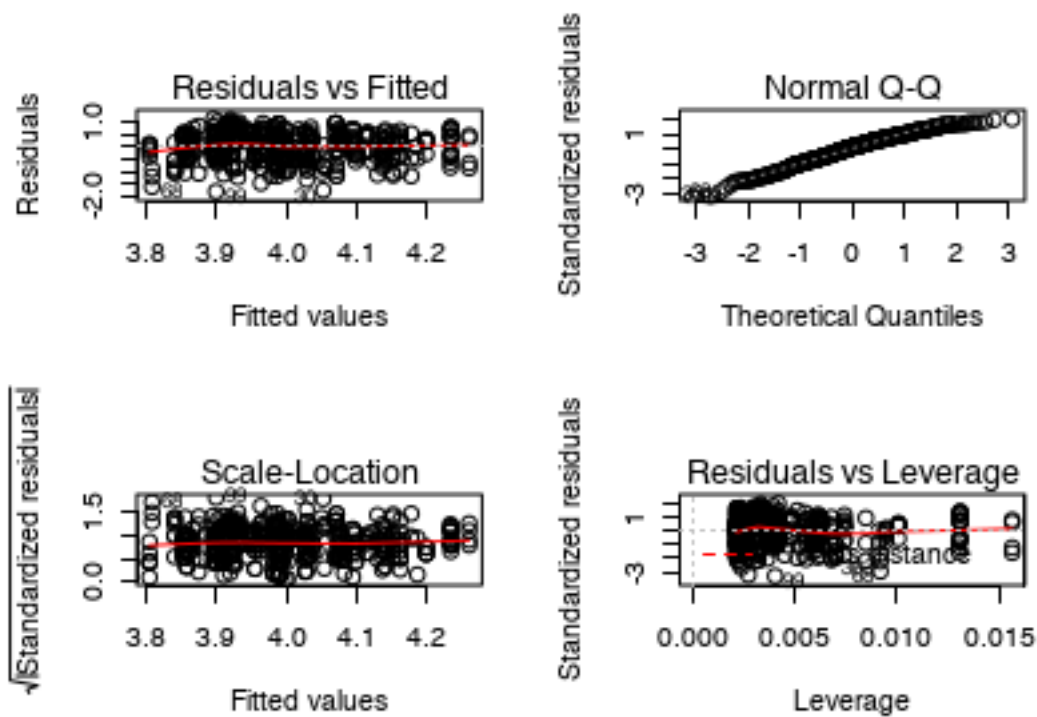
```
b1=4.0100227
b2=0.1330014
plot(predict(b, newdata= beauty.data), residuals(b))
```
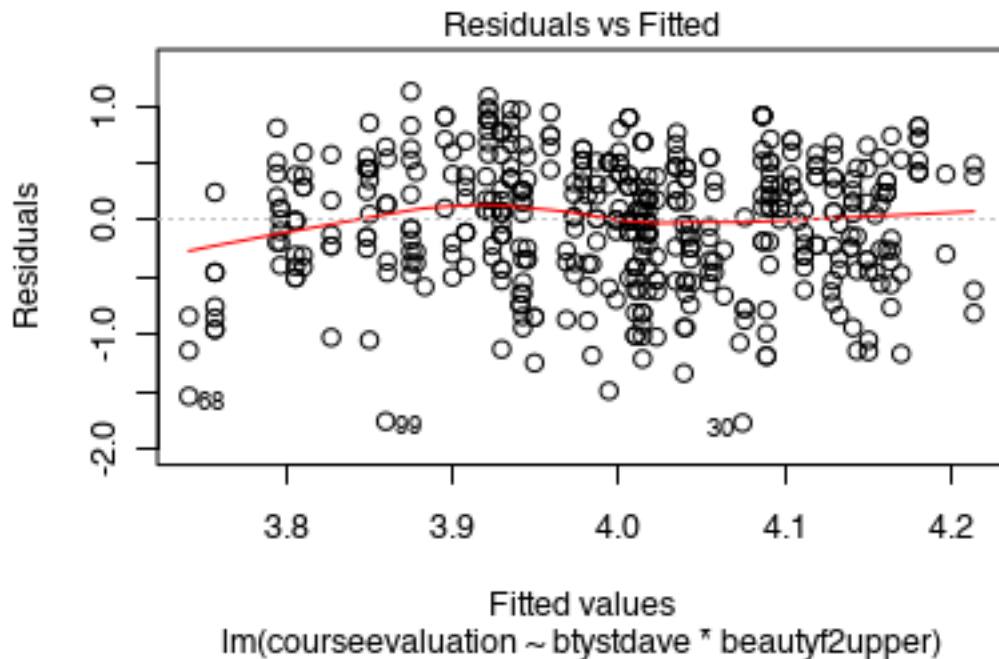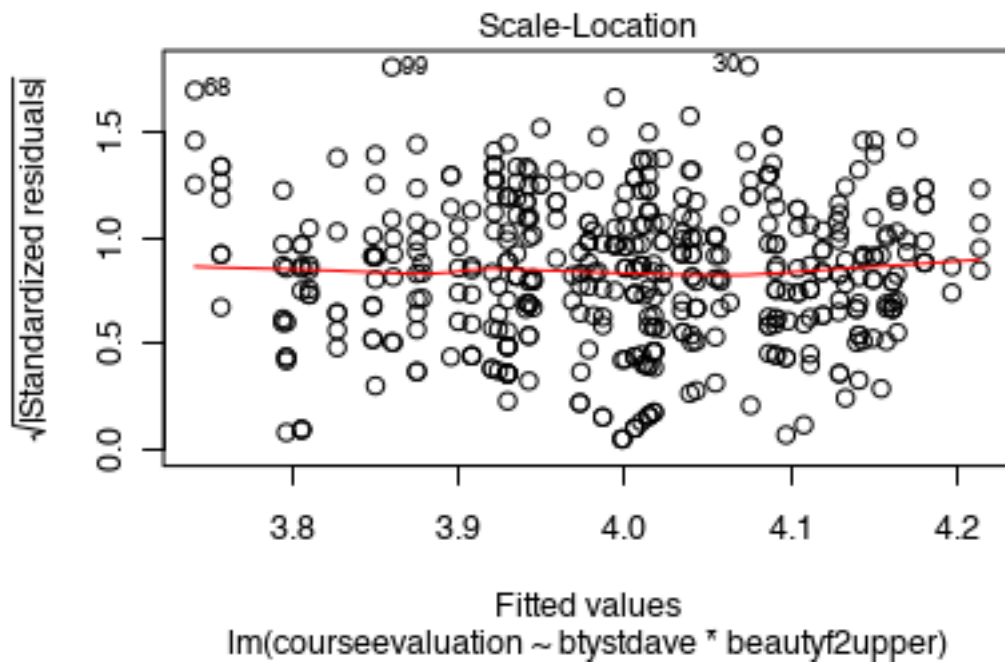
```
# or
par(mfrow=c(2,2))
plot(b)
```

```
#"the coefficients for this model is 4.0100227 and 0.1330014, respectively." about btystdave's
#coefficient, when other predictors in the model remain unchanged, adding 1 to the predictor
#btystdave will lead the whole output increase by 0.1330.
```

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.

```
btystdave=beauty.data$btystdave
beautyf2upper=beauty.data$beautyf2upper
age=beauty.data$age
model_2=lm(courseevaluation~btystdave*beautyf2upper,data=beauty.data)
plot(model_2)
```


Residuals vs Fitted

lm(courseevaluation ~ btystdave * beautyf2upper)

Normal Q-Q

Standardized residuals

Theoretical Quantiles
lm(courseevaluation ~ btystdave * beautyf2upper)

Scale-Location

√|Standardized residuals|

Fitted values
lm(courseevaluation ~ btystdave * beautyf2upper)

Residuals vs Leverage

lm(courseevaluation ~ btystdave * beautyf2upper)
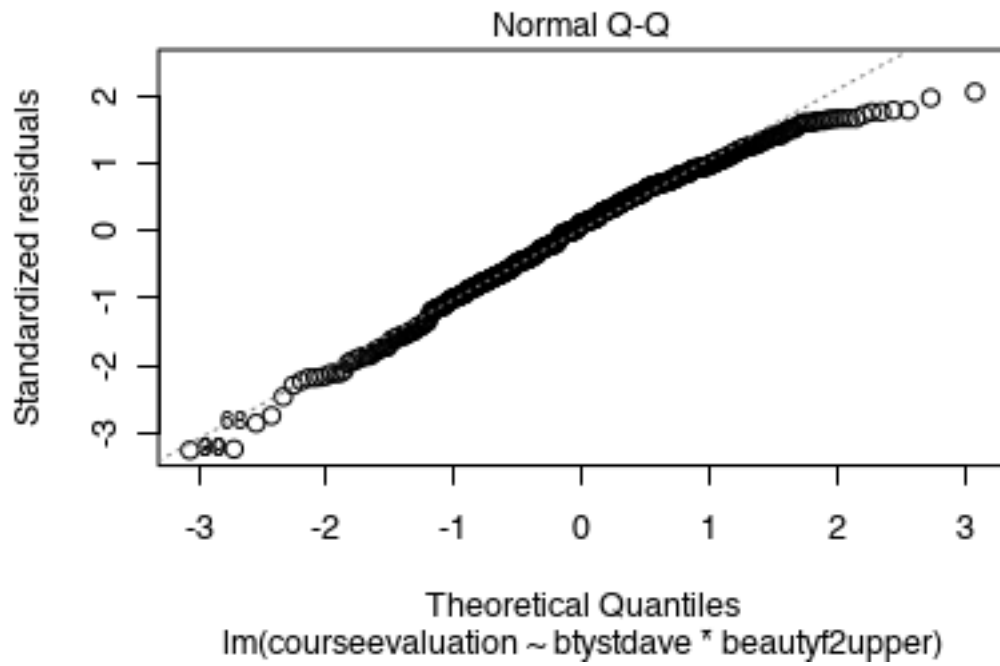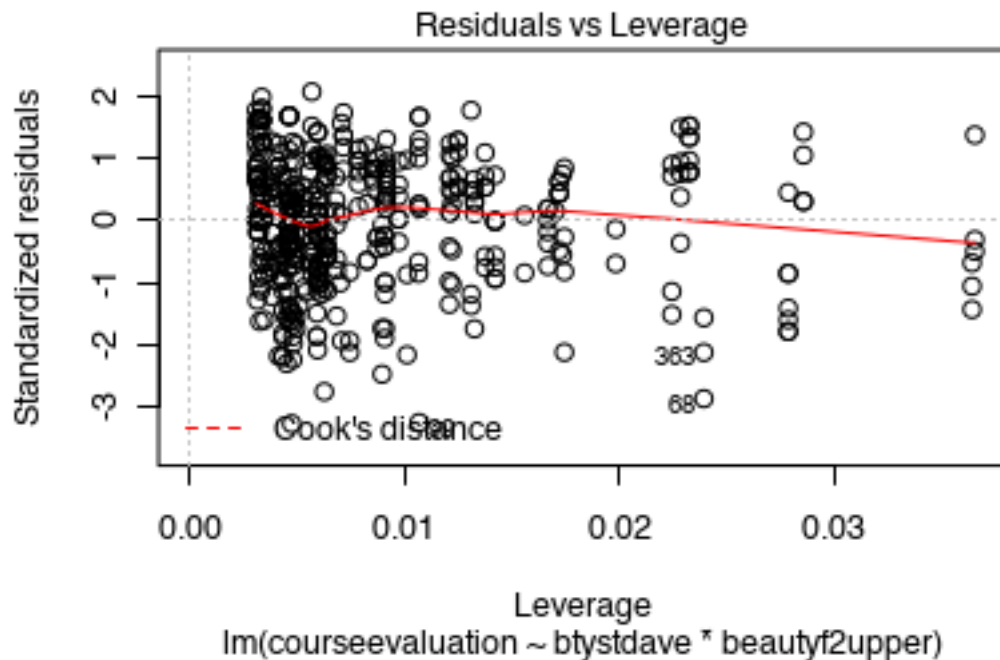
```
summary(model_2)
```
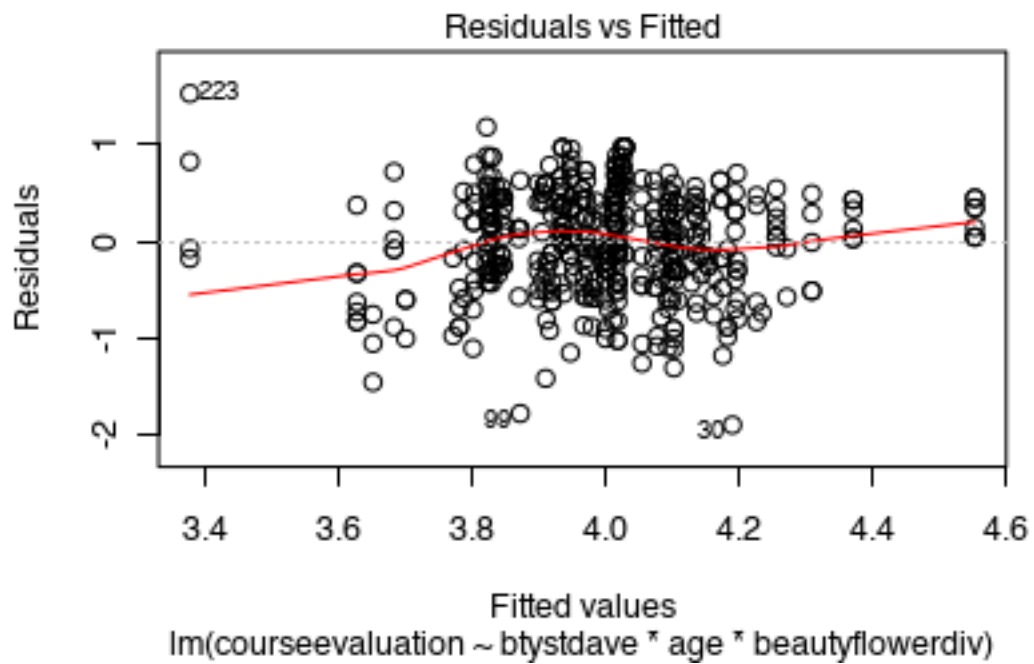
```
## 
## Call:
## lm(formula = courseevaluation ~ btystdave * beautyf2upper, data = beauty.data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7745 -0.3622  0.0706  0.4000  1.1253
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.986913   0.121593  32.789   <2e-16 ***
## btystdave             0.201812   0.103518   1.950   0.0518 .
## beautyf2upper         0.007876   0.021561   0.365   0.7151
## btystdave:beautyf2upper -0.014338 0.014401  -0.996   0.3199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5459 on 459 degrees of freedom
## Multiple R-squared:  0.03817,    Adjusted R-squared:  0.03188
## F-statistic: 6.072 on 3 and 459 DF,  p-value: 0.0004652
```

in this model, i used the beauty and beautyf2upper as input. predictor is beauty:beautyf2upper. fr

```
beautyflowerdiv=beauty.data$beautyflowerdiv
model_3=lm(courseevaluation~ btystdave*age*beautyflowerdiv, data=beauty.data)
model_3
```

```
##
```

```
## Call:
## lm(formula = courseevaluation ~ btystdave * age * beautyflowerdiv,
##     data = beauty.data)
##
## Coefficients:
##              (Intercept)                       btystdave
##                 6.206496                        0.208374
##                      age                  beautyflowerdiv
##                -0.040721                       -0.583109
##            btystdave:age       btystdave:beautyflowerdiv
##                -0.001782                        0.101496
##      age:beautyflowerdiv  btystdave:age:beautyflowerdiv
##                 0.010739                       -0.001474
```

```
plot(model_3)
```

Normal Q-Q

Standardized residuals

Theoretical Quantiles
lm(courseevaluation ~ btystdave * age * beautyflowerdiv)

Scale-Location

√|Standardized residuals|

Fitted values
lm(courseevaluation ~ btystdave * age * beautyflowerdiv)

Residuals vs Leverage

lm(courseevaluation ~ btystdave * age * beautyflowerdiv)
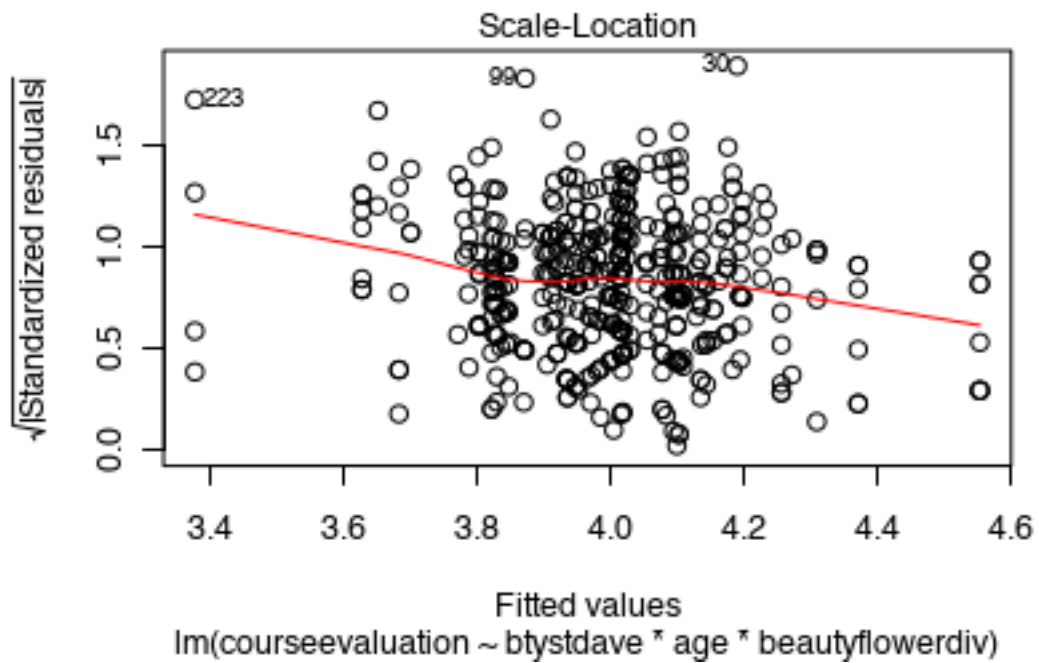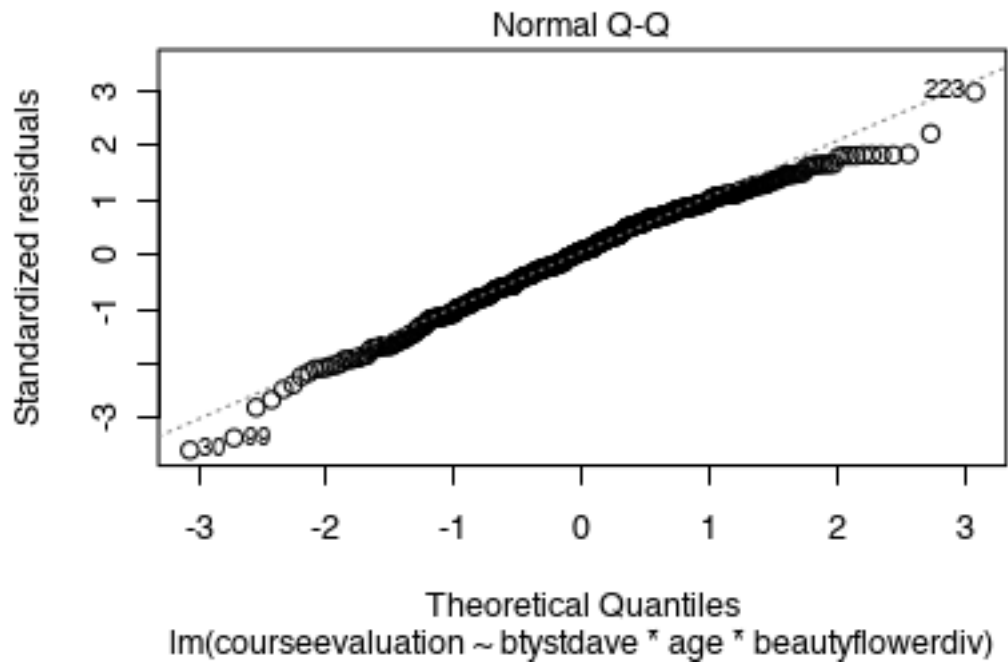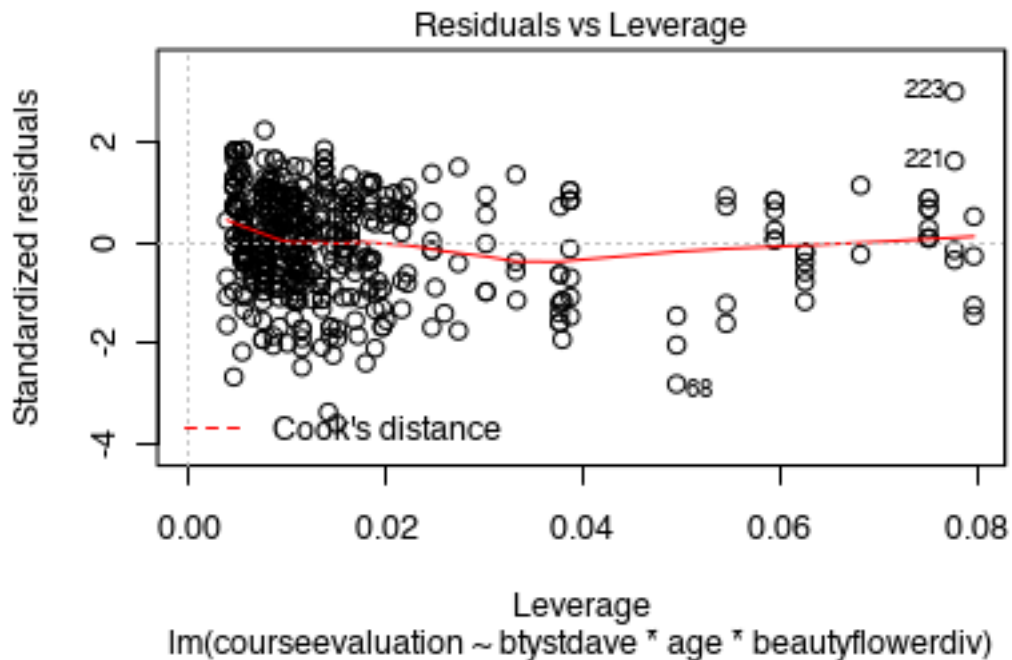
```r
summary(model_3)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave * age * beautyflowerdiv,
##     data = beauty.data)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.89114 -0.33475  0.04281  0.39402  1.52346
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.206496   0.595406  10.424  < 2e-16 ***
## btystdave                   0.208374   0.464310   0.449 0.653802
## age                        -0.040721   0.012405  -3.283 0.001107 **
## beautyflowerdiv            -0.583109   0.139269  -4.187 3.4e-05 ***
## btystdave:age              -0.001782   0.009920  -0.180 0.857520
## btystdave:beautyflowerdiv   0.101496   0.079506   1.277 0.202399
## age:beautyflowerdiv         0.010739   0.002747   3.909 0.000107 ***
## btystdave:age:beautyflowerdiv -0.001474 0.001594  -0.925 0.355486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.532 on 455 degrees of freedom
## Multiple R-squared:  0.09473,    Adjusted R-squared:  0.0808
## F-statistic: 6.802 on 7 and 455 DF,  p-value: 1.08e-07
```

```
### for this model, the inputs are beauty, age and beautyflowerdiv. the predictors are beauty:age, beau
```

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
## -1.793637
```

```
(abs(z.scores)>2)
```

```
##  var1
## FALSE
```

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]
}
sum((abs(z.scores)>2))
```

```
## [1] 4
```

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

sum((abs(z.scores)>2)) -there are 2 of them

What can you say about statistical significance of regression coefficient?

when coefficient is more than 2 standard error from the true value is considered statistically significant, this variable is useful in this model.

Fit regression removing the effect of other variables

Consider the general multiple-regression equation

$$Y = A + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + E$$

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"))
prestige=Prestige$prestige
education=Prestige$education
income=Prestige$income
women=Prestige$women
c=lm(prestige~education+income+women, data=Prestige)
coef(c)
```

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

```
m1<-lm(prestige~income+women)
m2<-lm(education~income+women)
e1<-prestige-fitted(m1)
e2<-education-fitted(m2)
m3<-lm(e1~e2)
coef(m3)[2]
```

```
##       e2
## 4.186637
```

(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"?

yes,it is 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?

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(e1,e2)
```

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

```
#or
data.frame(prestige, education)
```

```
##      prestige education
## 1       68.8      13.11
## 2       69.1      12.26
## 3       63.4      12.77
## 4       56.8      11.42
## 5       73.5      14.62
## 6       77.6      15.64
## 7       72.6      15.09
## 8       78.1      15.44
## 9       73.1      14.52
## 10      68.8      14.64
## 11      62.0      12.39
## 12      60.0      12.30
## 13      53.8      13.83
## 14      62.2      14.44
## 15      74.9      14.36
## 16      55.1      14.21
## 17      82.3      15.77
## 18      58.1      14.15
## 19      58.3      15.22
## 20      72.8      14.50
## 21      84.6      15.97
## 22      59.6      13.62
## 23      66.1      15.08
## 24      87.2      15.96
## 25      66.7      15.94
## 26      68.4      14.71
## 27      64.7      12.46
## 28      34.9       9.45
## 29      72.1      13.62
## 30      69.3      15.21
## 31      67.5      12.79
## 32      57.2      11.09
## 33      57.6      12.71
## 34      54.1      11.44
## 35      46.0      11.59
## 36      41.9      11.49
## 37      49.4      11.32
## 38      42.3      10.64
## 39      47.7      11.36
## 40      30.9       9.17
## 41      32.7      12.09
## 42      38.7      11.04
## 43      36.1       9.22
## 44      37.2      10.07
```

```
## 45       38.1      10.51
## 46       29.4      11.20
## 47       51.1      11.13
## 48       35.7      11.43
## 49       35.6      11.00
## 50       41.5       9.84
## 51       40.2      11.13
## 52       26.5      10.05
## 53       14.8       9.62
## 54       23.3       9.93
## 55       47.3      11.60
## 56       47.1      11.09
## 57       51.1      11.03
## 58       43.5       9.47
## 59       51.6      10.93
## 60       29.7       7.74
## 61       20.2       8.50
## 62       54.9      10.57
## 63       25.9       9.46
## 64       20.8       7.33
## 65       17.3       7.11
## 66       20.1       7.58
## 67       44.1       6.84
## 68       21.5       8.60
## 69       35.3       8.88
## 70       38.9       7.54
## 71       25.2       7.64
## 72       34.8       7.64
## 73       23.2       7.42
## 74       33.3       6.69
## 75       28.8       6.74
## 76       42.5      10.09
## 77       44.2       8.81
## 78       35.9       8.40
## 79       41.8       7.92
## 80       35.9       8.43
## 81       43.7       8.78
## 82       50.8       8.76
## 83       37.2      10.29
## 84       28.2       6.38
## 85       38.1       8.10
## 86       50.3      10.10
## 87       27.3       6.67
## 88       40.9       9.05
## 89       50.2       9.93
## 90       51.1       8.24
## 91       38.9       6.92
## 92       36.2       6.60
## 93       29.9       7.81
## 94       42.9       8.33
## 95       26.5       7.52
## 96       66.1      12.27
## 97       48.9       8.49
## 98       35.9       7.58
```

```
## 99        25.1      7.93
## 100       26.1      8.37
## 101       42.2     10.00
## 102       35.2      8.55
```

```
cor(data.frame(prestige, education))
```

```
##             prestige education
## prestige  1.0000000 0.8501769
## education 0.8501769 1.0000000
```

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

$$\hat{y}_i = X\hat{\beta} = X(X^TX)^{-1}X^Ty_i = Hy_i$$

$$\hat{e} = y_i - \hat{y}_i = y_i - Hy_i = (I - H)y_i$$

$$\sum \hat{y}_i \hat{e}_i = \sum Hy_i * (I - H)y_i = y_i * H * (Iy_i - Hy_i) = \sum y_i * y_i * (H - HH) = \sum y_i * y_i * 0 = 0$$

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?

Feedback comments etc.

me and several of my classmmates whose undergrad major was not statistics all think that this homework is a little bit challenging for us. i will work harder and hope i will spend less time on r project next time.