MA678 homework 01

WEILING LI

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=" ")
write.table(pyth, "~/Desktop/MSSP\ MA\ 678/MA678/pyth.txt",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.

```
pyth = read.table(file = "pyth.txt",sep = " ")
```

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 = head(pyth,40)
#hist(pyth40$x1)
#hist(log(pyth40$y))
#hist(log(pyth40$x2))
pythfit = lm(y ~ x1 + x2 , data = pyth40 )
summary(pythfit)
```

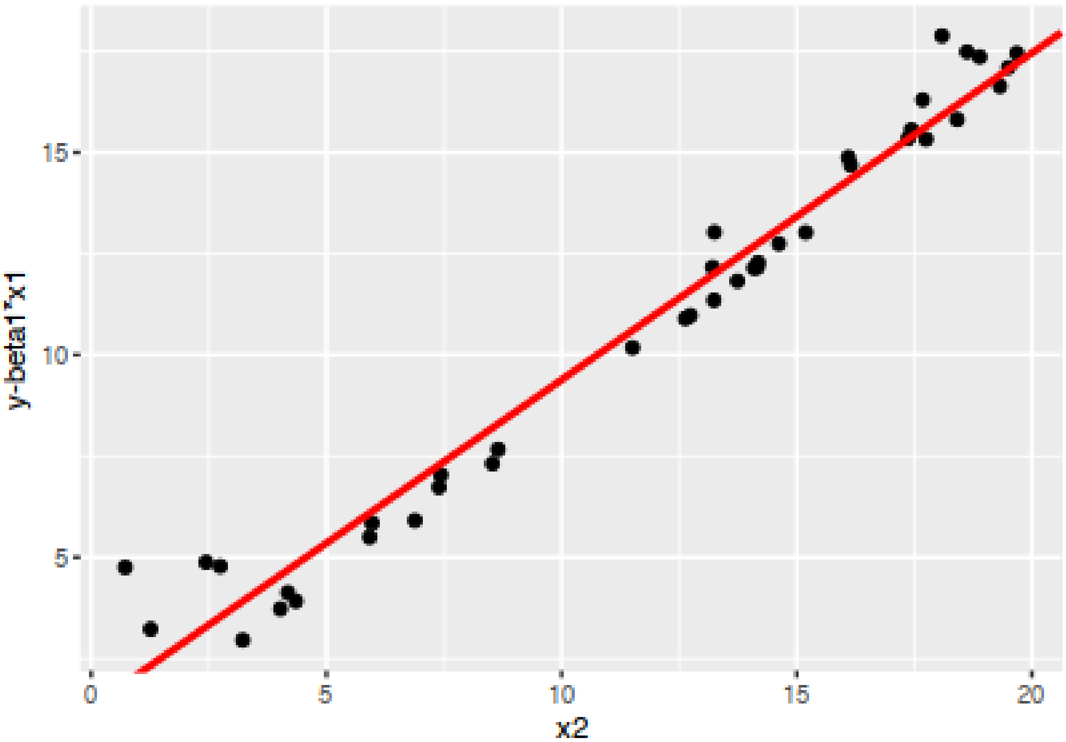
```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth40)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           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
```

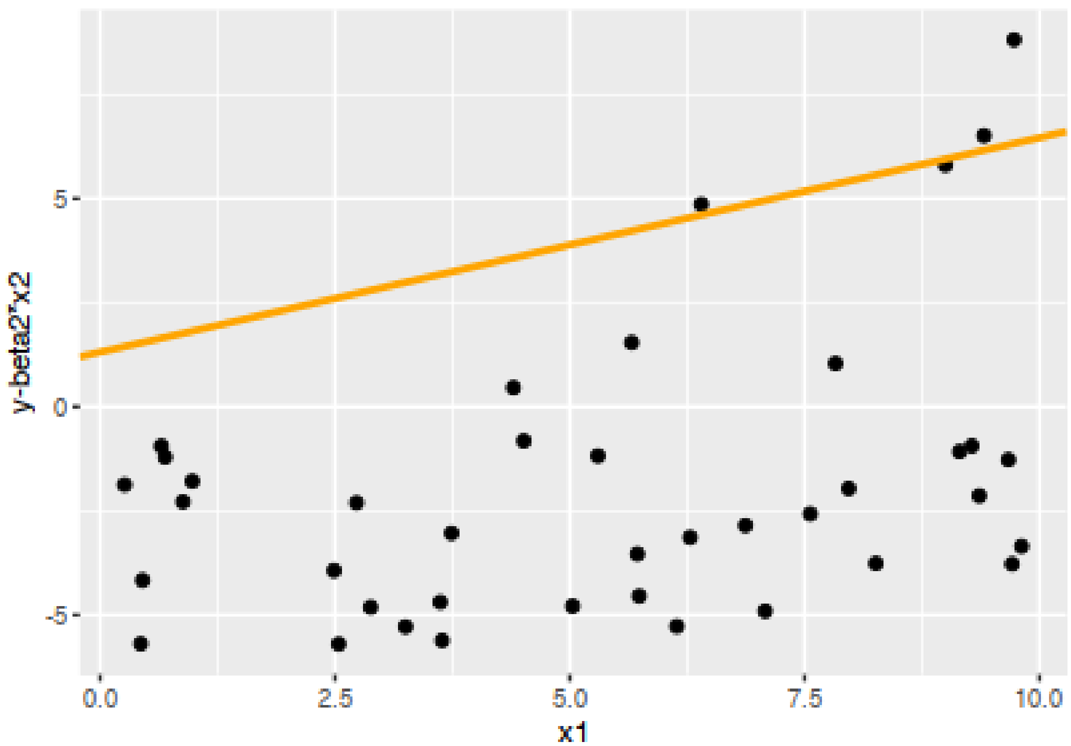
```
#plot(y ~ x1 + x2, data = pyth40)
#display(pythfit)
#fitpoints = predict(pythfit)
```

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

```
ggplot(data = pyth40)+
  geom_point() + aes(x = x2, y = y - coef(pythfit)[2]*x1)+geom_abline(intercept = coef(pythfit)[1], slop
```

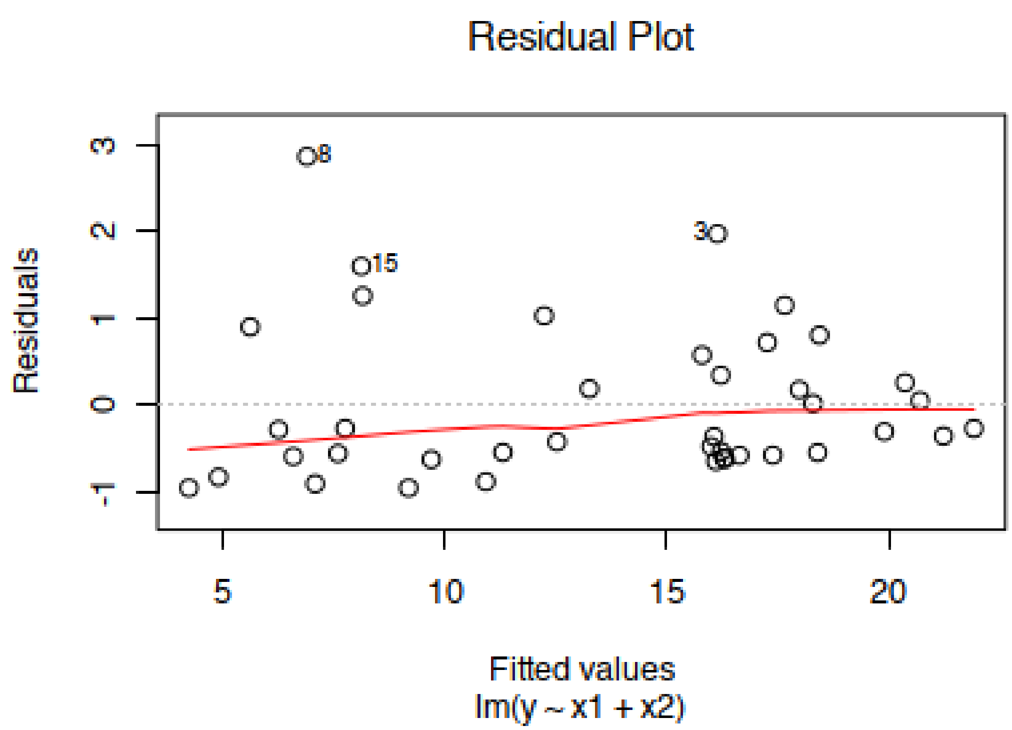



```
ggplot(data = pyth40)+
  geom_point() + aes(x = x1, y = y - coef(pythfit)[1]*x2)+geom_abline(intercept = coef(pythfit)[1], slop
```

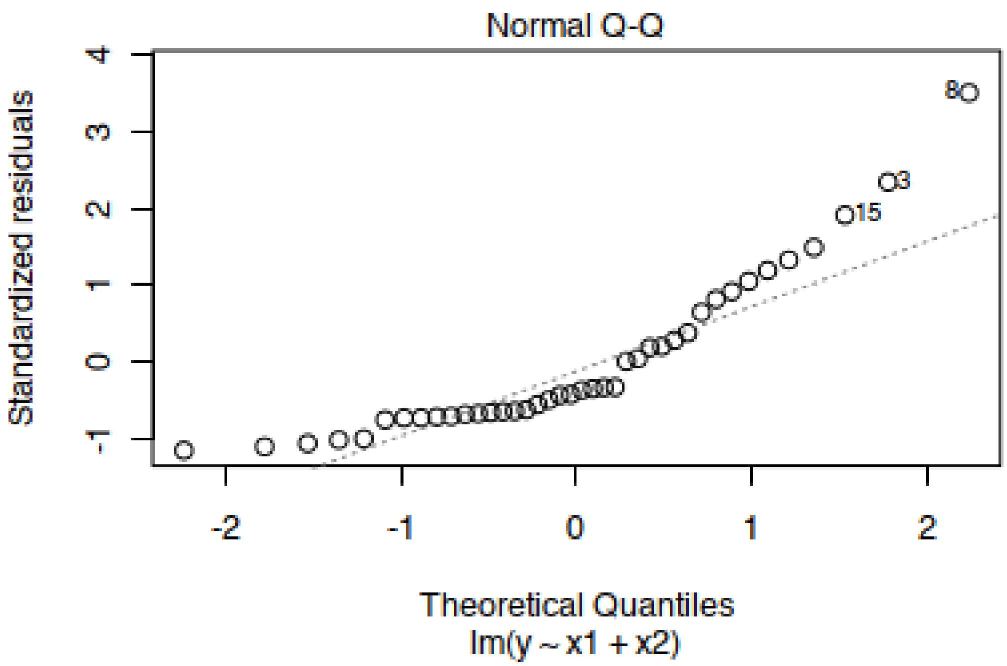


3. Make a residual plot for this model. Do the assumptions appear to be met?

```
#Residual Plot
plot(pythfit,which =1,title("Residual Plot"))
```



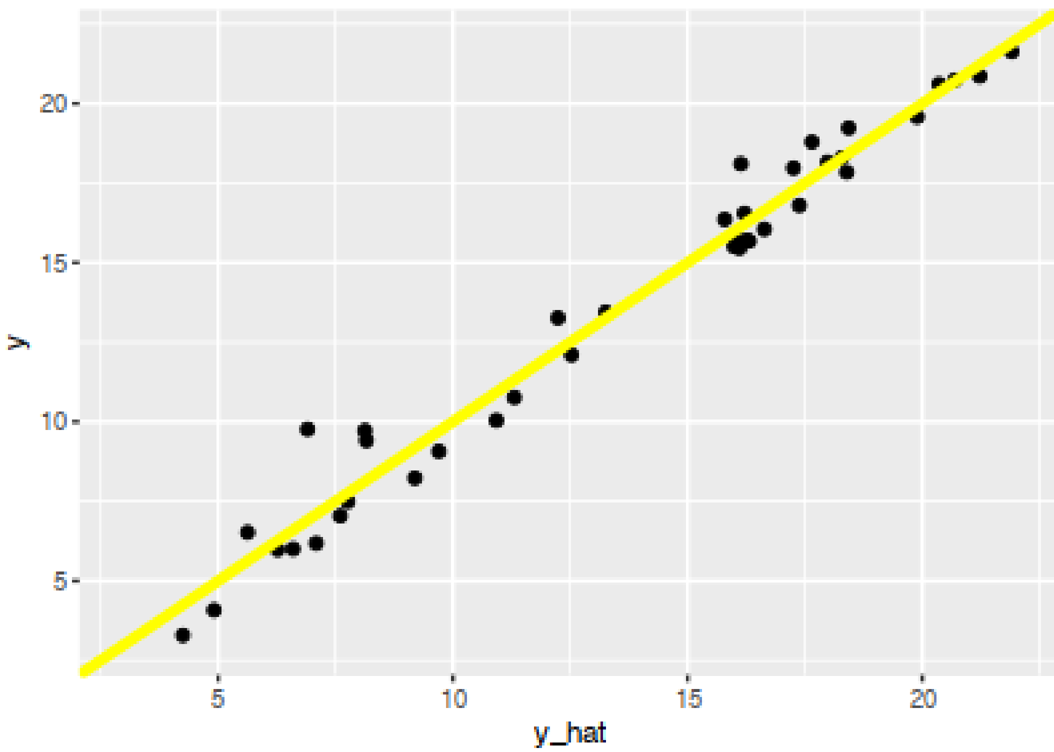
```r
#QQ plot
plot(pythfit,which =2)
```



Judging from the plot, the assumptions for linear regression for this particular model is not met. the variance of errors is not equal. Especially, the variance of x1 is really large compared to the variance of x2.

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

```r
pyth.tail = tail(pyth, 20)
pyth.pred = predict(pythfit,newdata = pyth.tail)
#pyth.pred1 = predict(pythfit,newdata = pyth.tail[-1])
ggplot()+geom_point()+aes(x=predict(pythfit),y = pyth40$y)+geom_abline(intercept = 0, slope = 1,color =
```



Judging from the outcome from the plot above and the residual plot. the prediction is actually a okay estimate for larger predicted y values(maybe over 15?).

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 regression equation can be written as:
   $ln(earnings) = \beta_0 + \beta_1 ln(heights)$

We know that for 66inches height, the predicted earnings is $30,000 and every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings. These two constraints will translate to:
$ln(30000) = \beta_0 + \beta_1 \times ln(66)$ (1)
$ln(1.008 \times earnings) = \beta_0 + \beta_1 \times ln(1.01 \times height)$ (2)
take exponential of equation (2) we get:
$1.008 \times e^{\beta_0} \times e^{beta_1 \times ln(height)} = e^{\beta_0} \times e^{beta_1 \times ln(1.01 \times height)}$
$1.01^{\beta_1} = 1.008$
$\beta_1 = log_{1.01}(1.008)$ (3) By substituing eq.(3) to eq.(1), we can get:
$\beta_0 = ln(30000) - log_{1.01}1.008 \times ln(66)$

```
## [1] "beta1="            "0.80079444259282"
```

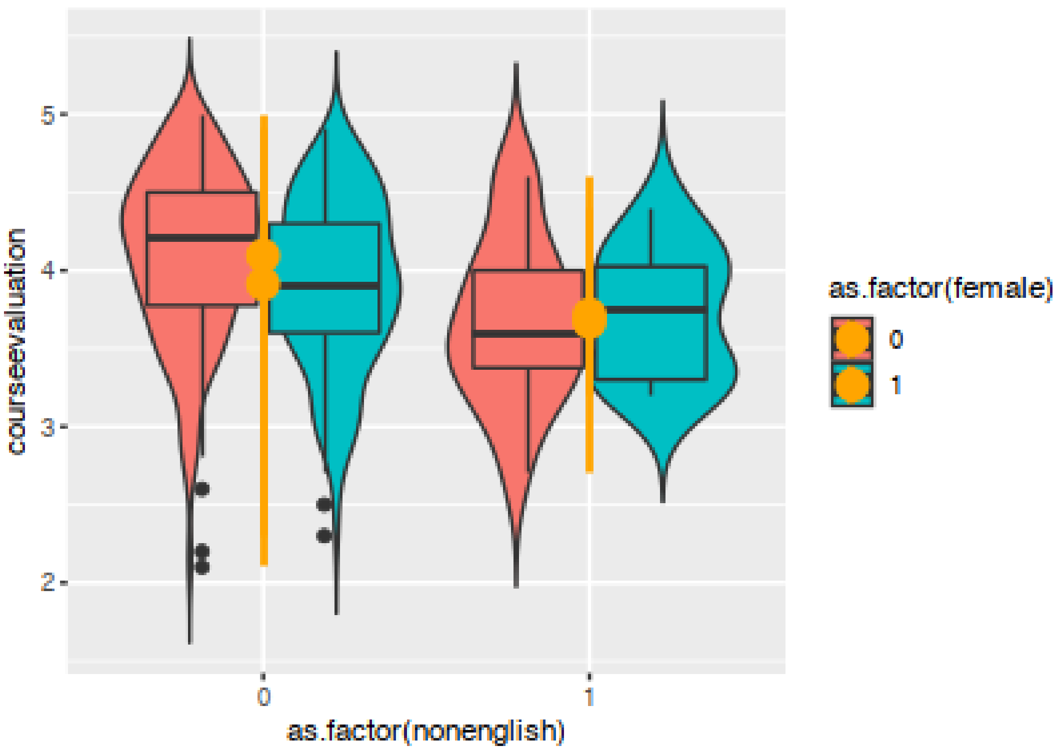
```
## [1] "beta0="            "6.95390042684688"
```

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?

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.

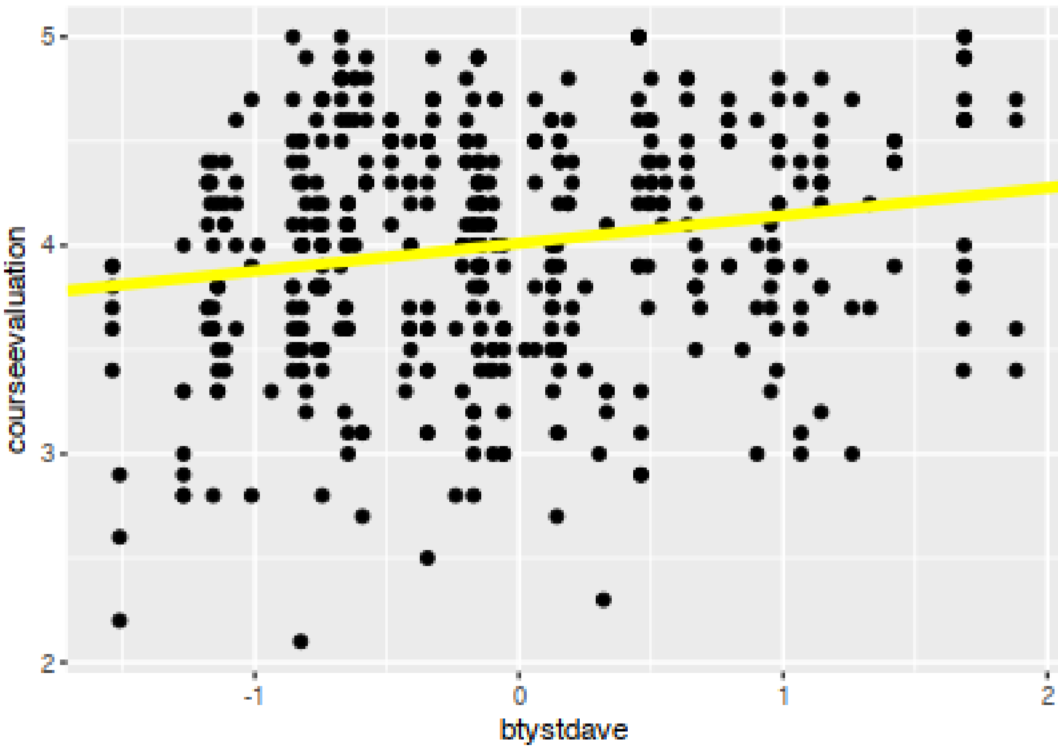
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, se

ggplot(data = beauty.data,aes(x = as.factor(nonenglish), y = courseevaluation,fill = as.factor(female))
```

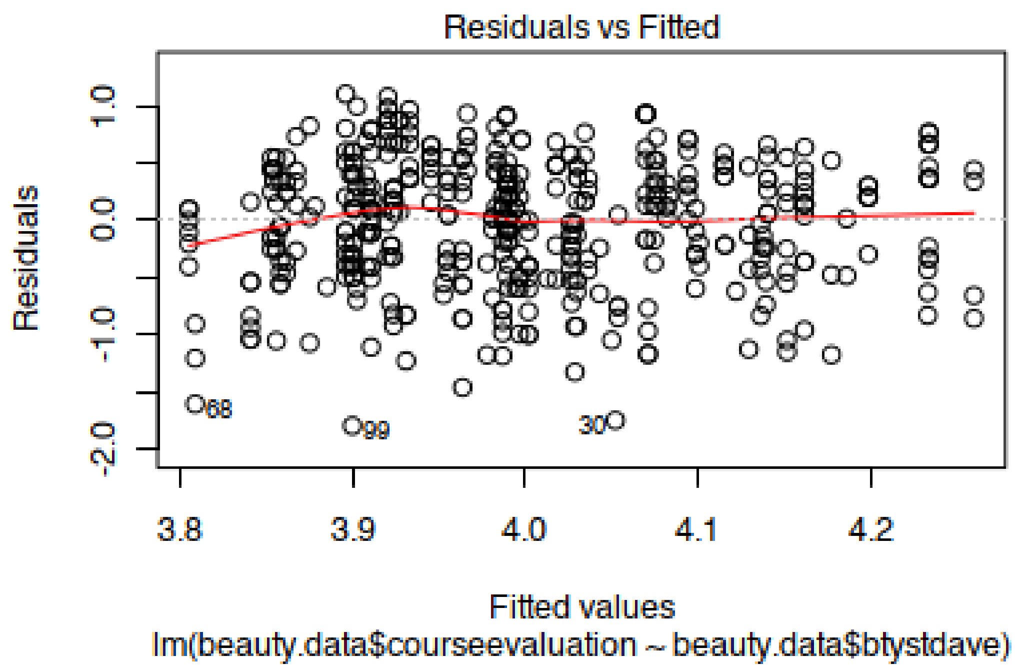


1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals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

```
##
## Call:
## lm(formula = beauty.data$courseevaluation ~ beauty.data$btystdave)
##
## 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.data$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
```





By observing the bystdave value, this is the Z-score transformed value of the beauty rating, so the intercept 4.01 means the average course evaluation score for average instructor's beauty rating.

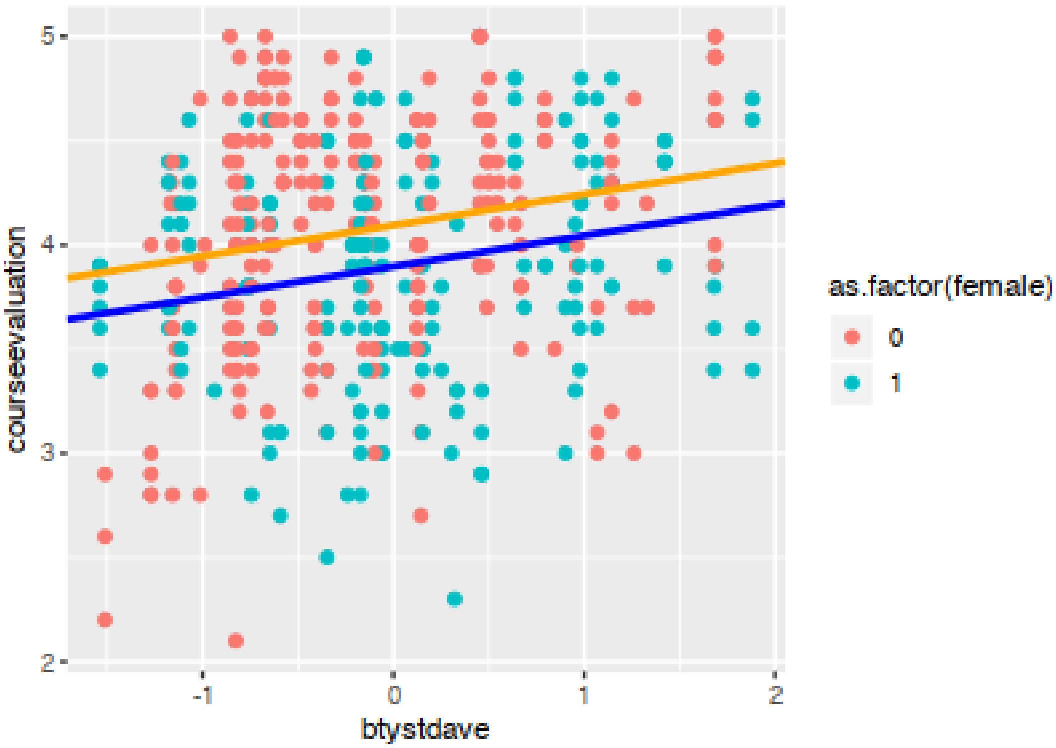
The slope means for every σ increase in beauty rating of instructor, the course evaluation score will increase 0.133

The residual standard error shows the variability of actual observations spead around the predicted values. The residual plot actually shows that the variability is high but roughly evenly spreaded around the predicted value.

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.

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87196 -0.36913  0.03493  0.39919  1.03237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.09471    0.03328  123.03  < 2e-16 ***
## btystdave    0.14859    0.03195    4.65 4.34e-06 ***
## female      -0.19781    0.05098   -3.88  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5373 on 460 degrees of freedom
```

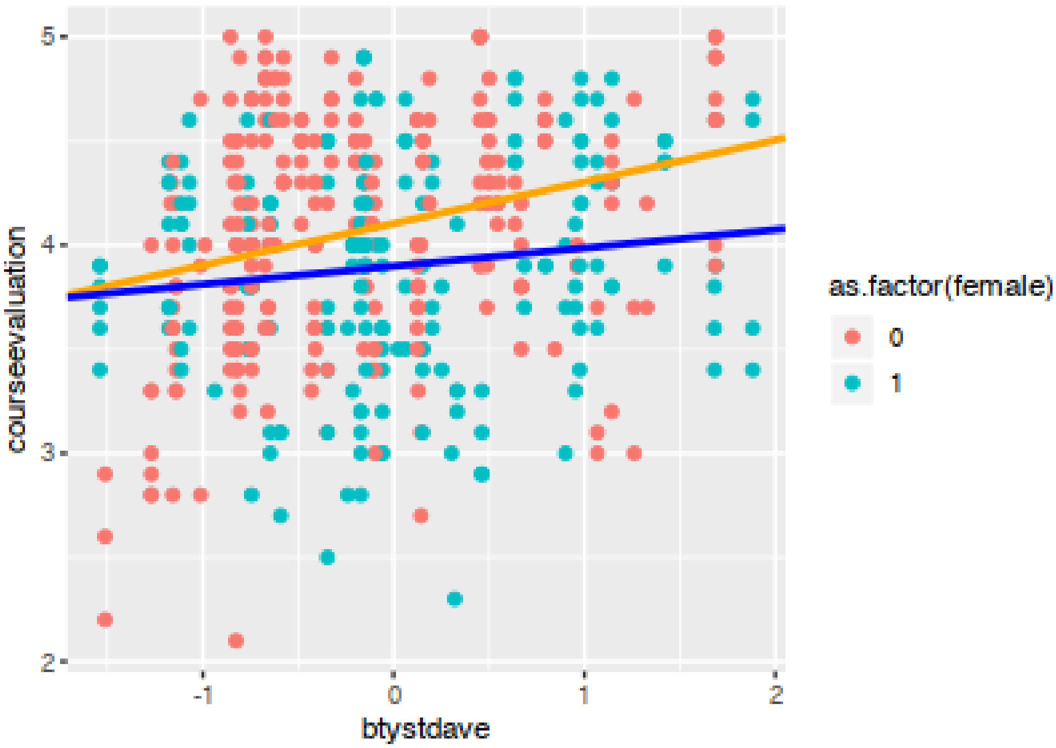
```
## Multiple R-squared:  0.0663, Adjusted R-squared:  0.06224
## F-statistic: 16.33 on 2 and 460 DF,  p-value: 1.407e-07
```



The model shown above shows a model with 1 binary predictor:female and 1 continuous predictor: btystdave with no interaction term. The model can be written as follows:
$courseeval = \beta_0 + \beta_1 \times btystdave + \beta_2 \times female$ with the binary components taking 0 or 1, the model can be describes as two regression lines $courseeval = \beta_0 + \beta_1 \times btystdave$ is the model for men
$courseeval = (\beta_0 + \beta_2) + \beta_1 \times btystdave$ is the model for women
Coef β_0 is the average course eval score 4.09 for male instructor with mean beauty eval.
Coef β_1 describes that for 1 σ increase in beauty rating(regardless male of female) the course eval will increase 0.2.
Coef β_2 is the difference of the mean course eval between two genders, averagely, female instructors are graded 0.11 less than male instructors

```
##
## 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
```



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
```

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?

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?

(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"?

(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.
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?

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.