MA678 homework 01

Diptansu Singh (dips@bu.edu)

Septemeber 10, 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.

```
#Understanding the data
summary(pyth)
```

```
##        y                x1               x2
##  Min.   : 3.290   Min.   :0.190   Min.   : 0.35
##  1st Qu.: 9.325   1st Qu.:2.527   1st Qu.: 5.76
##  Median :15.590   Median :5.525   Median :12.69
##  Mean   :13.590   Mean   :5.324   Mean   :10.99
##  3rd Qu.:18.003   3rd Qu.:8.293   3rd Qu.:15.74
##  Max.   :21.630   Max.   :9.990   Max.   :19.68
##  NA's   :20
```

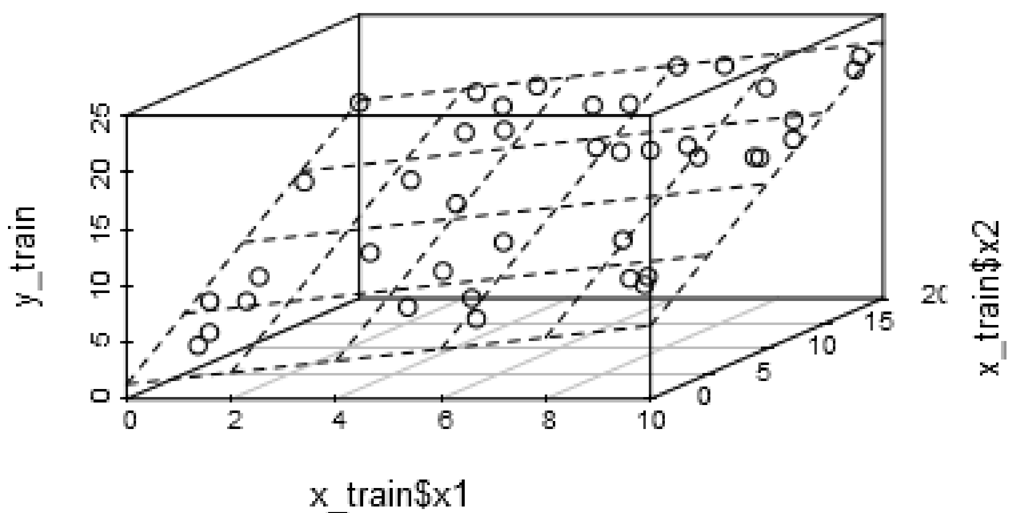
```
# x1, x2 are continuos variables. Fitting the model without centering or scaling

#Training Dataset : First 40 data points
y_train <- pyth[c(1:40),c(1)]
x_train <- pyth[c(1:40),c(-1)]
train   <- cbind(x_train, y_train)

#Test Dataste : Last 20 characters
```

```
x_test <- pyth[c(41:60),c(-1)]

#Fitting regression model
lm_1 <- lm(y_train~x_train$x1 + x_train$x2 )
lm_1
```

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

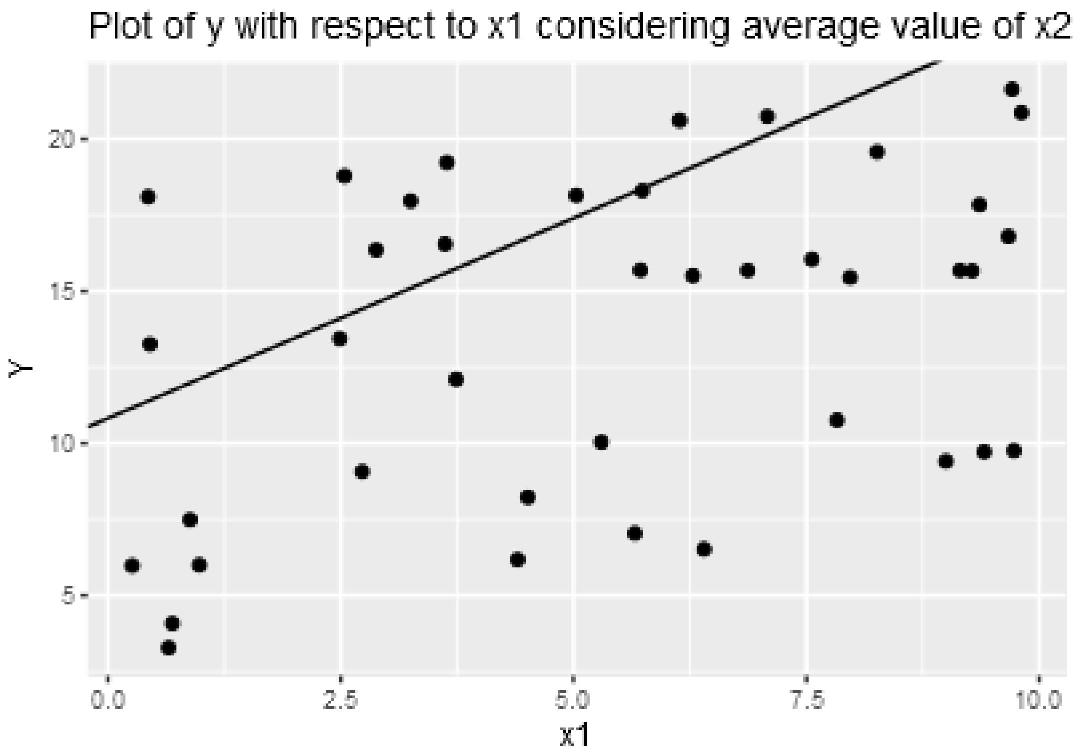
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
#Drawing a regression plane made by x1 and x2
library(scatterplot3d)
s3dplot<- scatterplot3d(x_train$x1,x_train$x2,y_train)
s3dplot$plane3d(lm_1)
```

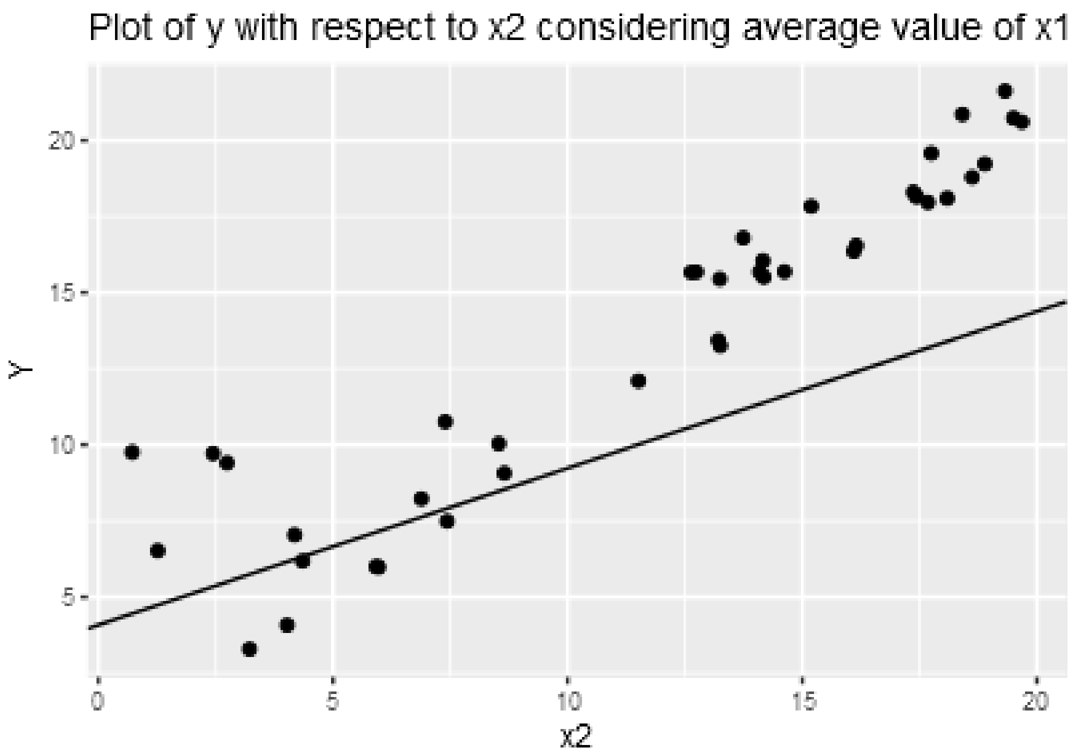



```
#This model shows the regression plane and the points in 3D space

#To understand the Variation with respect to 1 variable :

ggplot(data = train, aes(x1,y_train)) +
  geom_point() +
```

```
  geom_abline(intercept = lm_1$coefficients[1] + lm_1$coefficients[3] * mean(train$x2) , slope = lm_1$co
  ggtitle("Plot of y with respect to x1 considering average value of x2 ") + ylab("Y")
```




```
ggplot(data = train, aes(x2,y_train)) +
  geom_point() +
  geom_abline(intercept = lm_1$coefficients[1] + lm_1$coefficients[2] * mean(train$x1) , slope = lm_1$co
  ggtitle("Plot of y with respect to x2 considering average value of x1 ") + ylab("Y")
```



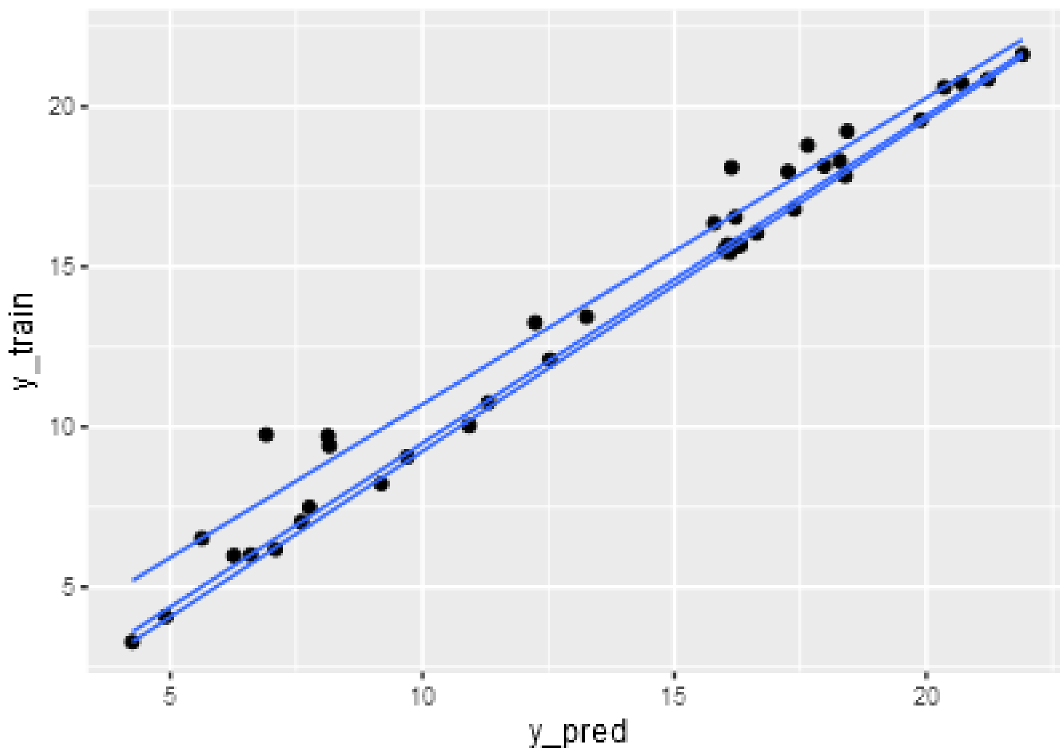

```
Y <- as.data.frame(cbind(y_train, y_pred = predict(lm_1)))
#Plot of y and y_predicted
ggplot ( Y , aes (y_pred, y_train)) +
  geom_point() +
  geom_quantile()
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

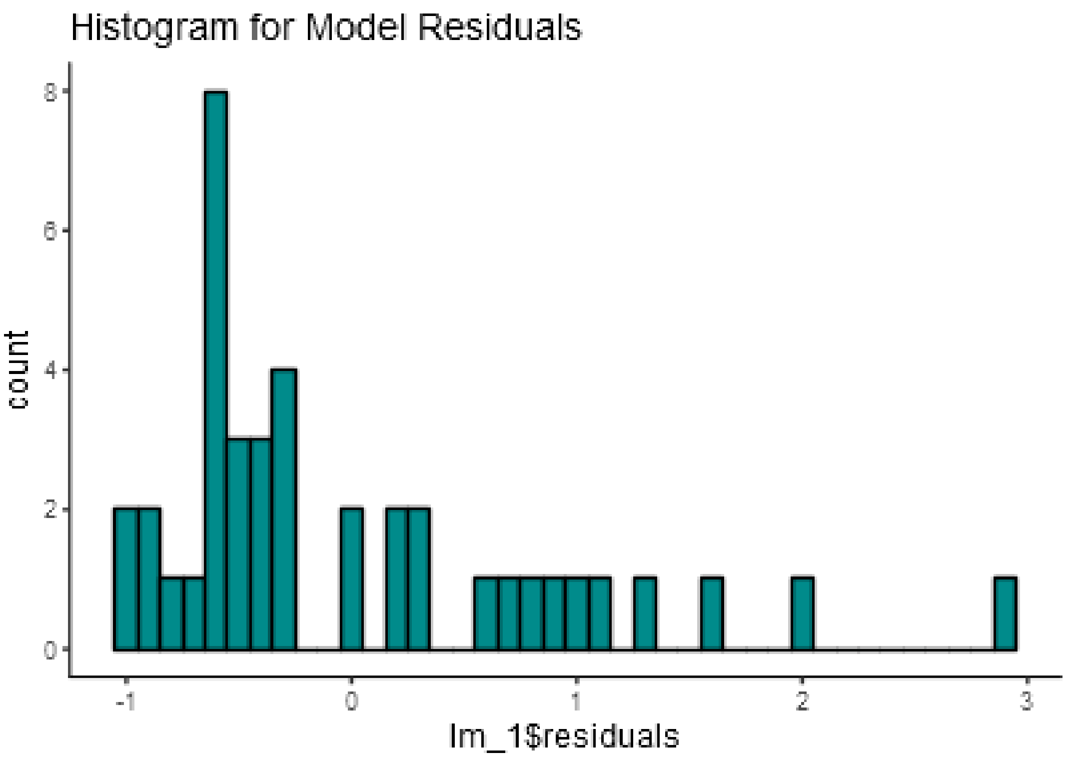
```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
## Smoothing formula not specified. Using: y ~ x
```



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

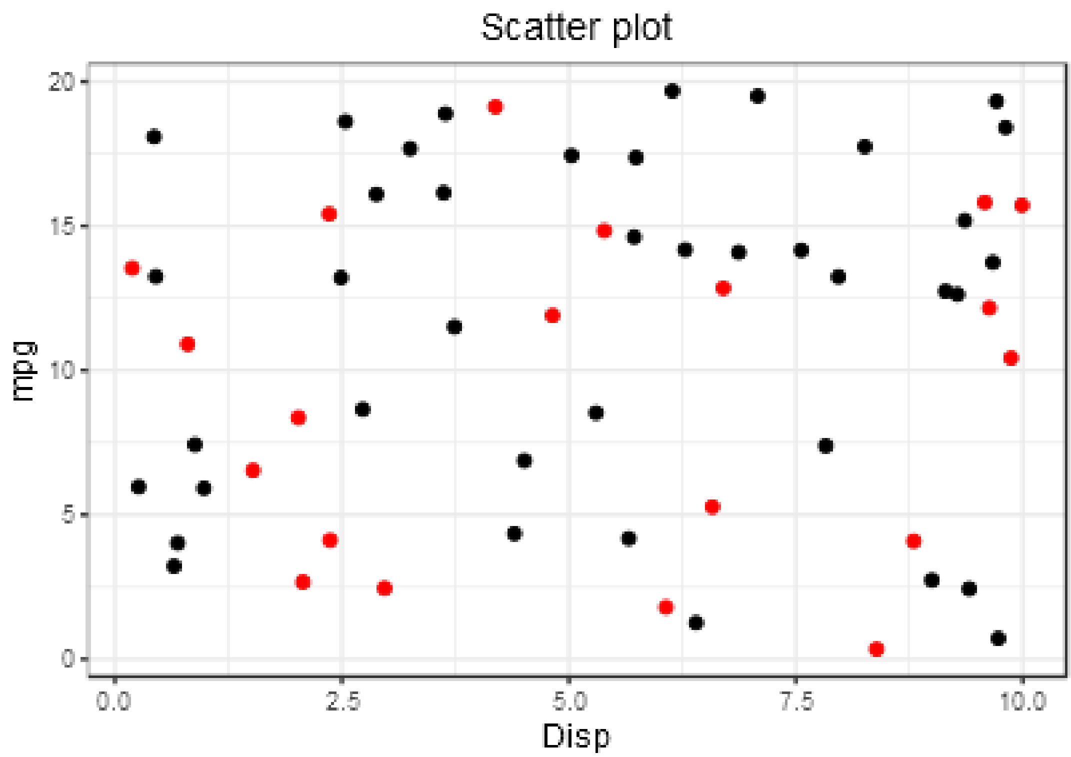
```
#Residual Plot
res <- resid(lm_1)

ggplot(data=x_train, aes(lm_1$residuals)) +
  geom_histogram(binwidth = 0.1, color = "black", fill = "cyan4") +
  theme(panel.background = element_rect(fill = "white"),
        axis.line.x=element_line(),
        axis.line.y=element_line()) +
  ggtitle("Histogram for Model Residuals")
```




```
#Residual plot is skewed towards the right
```

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

```
#Comparing distribution of x1 and x2 for test and training dataset
x <- rbind(cbind(x_train,c=1),cbind(x_test,c=2))
ggplot(data=x,aes(x=x1,y=x2)) + geom_point(col=x$c) + xlab("Disp") +
  ylab("mpg") + ggtitle("Scatter plot") + theme_bw() + theme(plot.title = element_text(hjust=0.5))
```




```
# Since distribution is similar, we predict using the formula we obtained from training model
y_test <- lm_1$coefficients[1] + lm_1$coefficients[2]*x_test$x1 + lm_1$coefficients[3]*x_test$x2
y_test
```

```
##  [1] 14.812484 19.142865  5.916816 10.530475 19.012485 13.398863  4.829144
##  [8]  9.145767  5.892489 12.338639 18.908561 16.064649  8.963122 14.972786
## [15]  5.859744  7.374900  4.535267 15.133280  9.100899 16.084900
```

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 the regression line and the residual standard deviation of the regression.

log(earning) = A + B log(height)

B = 0.008/0.01 (For every 1% increase in height, there is 0.8% increase in Y)

A = log(30000) - 0.8 log(66) = 6.957229

Calculating residual standard deviation = Standard deviation in error of Beta

(1.1 - 1) * B = SD_Error * 1.96

0.1 * 0.8 / 1.96 = SD_Error = 0.0408

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?

Var_Error = 0.0408 ^ 2

Var_Population = 0.05 ^ 2

R^2 = 1 - Var_Error / Var_Population

R^2 = 0.3341

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.

```
#Both are continuos variables
x_train <- beauty.data$btystdave
y_train <- beauty.data$courseevaluation
lm_2 <- lm(y_train~x_train)
lm_2
```

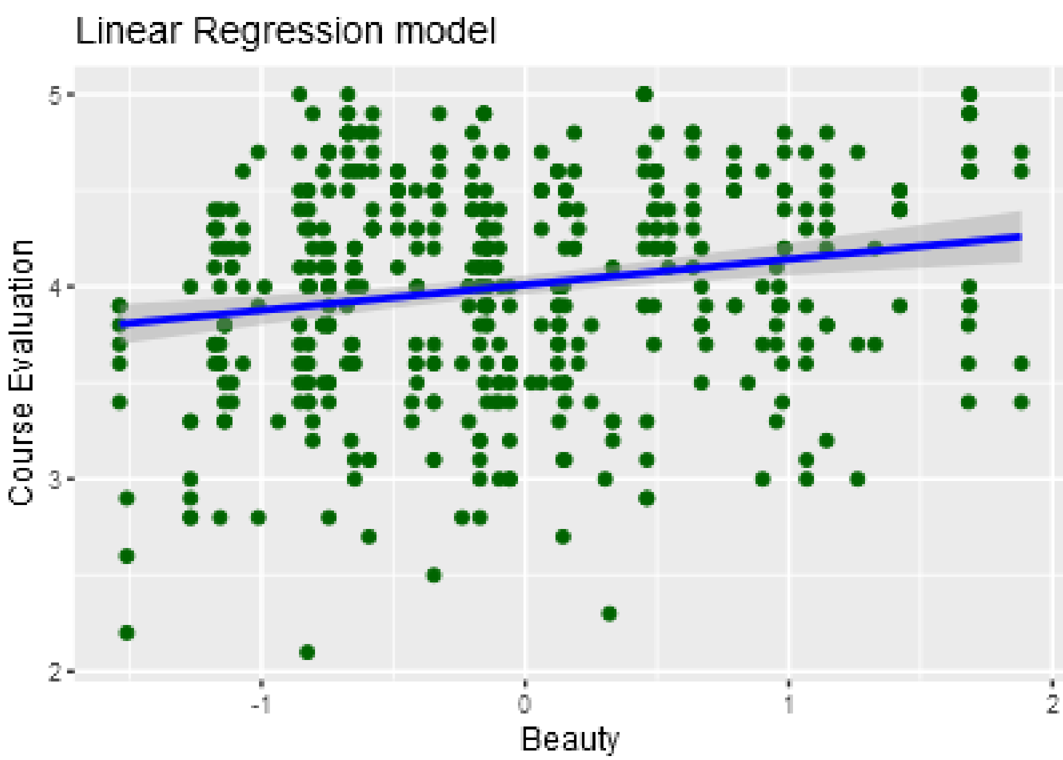
```
##
## Call:
## lm(formula = y_train ~ x_train)
##
## Coefficients:
## (Intercept)      x_train
##       4.010        0.133
```

```
summary(lm_2)
```

```
##
## Call:
## lm(formula = y_train ~ x_train)
##
## 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 ***
## x_train      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
```

```
#The R squared value is quite less. It makes sense as the data looks quite spread out
#Graphical representation
ggplot(beauty.data, aes(x = btystdave, y = courseevaluation)) +
  geom_point( color = "Dark Green") +
  stat_smooth(method = "lm", col = "blue") +
  xlab("Beauty") + ylab("Course Evaluation") + ggtitle("Linear Regression model")
```



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.

```
#In order to compare variables on the same scale: we will center and scale the variables to same value
beauty.data.sc <- as.data.frame(scale(beauty.data))
```

```r
lm_3 <- lm( courseevaluation ~ . ,beauty.data.sc )
# Stepwise regression model
step.model <- stepAIC(lm_3, direction = "both",
                      trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = courseevaluation ~ profnumber + age + beautyf2upper +
##     beautyfupperdiv + beautymupperdiv + btystdf2u + btystdmu +
##     class3 + class8 + class12 + class14 + class17 + class18 +
##     class19 + class26 + class27 + nonenglish + onecredit + percentevaluating +
##     profevaluation, data = beauty.data.sc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61495 -0.16523  0.01529  0.20435  1.02637
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.010e-11  1.511e-02   0.000  1.00000
## profnumber         3.958e-02  1.705e-02   2.322  0.02071 *
## age                7.916e-02  1.741e-02   4.546 7.06e-06 ***
## beautyf2upper      1.581e+05  8.006e+04   1.975  0.04885 *
## beautyfupperdiv    4.425e-02  2.454e-02   1.803  0.07208 .
## beautymupperdiv    2.072e+05  9.302e+04   2.228  0.02639 *
## btystdf2u         -1.581e+05  8.006e+04  -1.975  0.04885 *
## btystdmu          -2.072e+05  9.302e+04  -2.228  0.02639 *
## class3            -2.575e-02  1.575e-02  -1.635  0.10274
## class8             3.270e-02  1.539e-02   2.125  0.03418 *
## class12           -4.218e-02  1.553e-02  -2.715  0.00688 **
## class14            4.147e-02  1.566e-02   2.649  0.00837 **
## class17            2.529e-02  1.577e-02   1.603  0.10957
## class18            4.573e-02  1.563e-02   2.925  0.00362 **
## class19           -4.426e-02  1.578e-02  -2.805  0.00526 **
## class26            2.430e-02  1.544e-02   1.574  0.11621
## class27            3.733e-02  1.529e-02   2.442  0.01501 *
## nonenglish        -7.410e-02  1.623e-02  -4.566 6.46e-06 ***
## onecredit          3.231e-02  1.633e-02   1.979  0.04846 *
## percentevaluating  3.984e-02  1.634e-02   2.439  0.01512 *
## profevaluation     9.254e-01  1.655e-02  55.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3251 on 442 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8943
## F-statistic: 196.5 on 20 and 442 DF,  p-value: < 2.2e-16
```

```r
#This model has R^2 = 0.89 but there may be overfitting involved.

#Since class is an important variable as indicated by high value of effect, so we are combining the var
beauty.data$class_sum <- rowSums(beauty.data[,c('class3','class8','class12','class14','class17','class1
```

```r
#Creating a model with variables class_sum, btystdave
lm_4 <- lm(courseevaluation ~ class_sum * btystdave , beauty.data)
summary(lm_4)
```

```
##
## Call:
## lm(formula = courseevaluation ~ class_sum * btystdave, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79337 -0.36323  0.05063  0.40482  1.11076
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.01029    0.02660 150.788  < 2e-16 ***
## class_sum            -0.03446    0.09784  -0.352    0.725
## btystdave             0.14154    0.03295   4.295 2.13e-05 ***
## class_sum:btystdave -0.19460    0.15799  -1.232    0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5457 on 459 degrees of freedom
## Multiple R-squared:  0.03892,    Adjusted R-squared:  0.03264
## F-statistic: 6.196 on 3 and 459 DF,  p-value: 0.0003924
```

```r
#Interpretation of Result:
# The intercept represents the courseEval value for beauty score = 0 and class =0, This is a meaningles
# The estimate of class_sum gives the difference betwuen a students courseEval if he has taken/not take
# The estimate of btysdave gives the increase in courseEval for a unit increase in the bytsdave variabl
# The estimate of class_sum:bystvdave gives the difference in slope of regression lines for the data po
```

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
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, 1000)
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)
```

How many of these 100 z-scores are statistically significant?
6 values are statistically significant
What can you say about statistical significance of regression coefficient?
Using absolute values of z-score, in 5 times out of 100; the sample selected from the normal distribution was such that some variance in var2 was explained by var1. We can consider that we got 94 sample that actually say that there are not relationship between var1 and var2. The 95% value is also ~ same as the value of probability we get when we use sd = +-2 (95.5%). To test this out, I checked the setup with 1000 runs and z.score > 3. 2 of these were statistically significant, which ~0.2% which was expected.

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 $E_{Y|2,...,k}$.
2. Regress X_1 on X_2 through X_k, obtaining residuals $E_{1|2,...,k}$.
3. Regress the residuals $E_{Y|2,...,k}$ on the residuals $E_{1|2,...,k}$. 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"))
summary(Prestige)
```

```
##    education          income          women           prestige    
##  Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80  
##  1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23  
##  Median :10.540   Median : 5930   Median :13.600   Median :43.60  
##  Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83  
##  3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27  
##  Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20  
##      census       type   
##  Min.   :1113   bc  :44  
##  1st Qu.:3120   prof:31  
##  Median :5135   wc  :23  
##  Mean   :5402   NA's: 4  
##  3rd Qu.:8312            
##  Max.   :9517            
```

```r
#Prestige is a continuous variable which is equally distributed, hence not using log transformations
#Regression on all variables except education
lm_5 <- lm(prestige ~ income + women + census + type,Prestige)
summary(lm_5)
```

```
## 
## Call:
## lm(formula = prestige ~ income + women + census + type, data = Prestige)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20.5468  -6.6776  -0.2236   4.7902  24.9435 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.449e+01  6.141e+00   3.987 0.000134 ***
## income      1.485e-03  2.904e-04   5.115 1.71e-06 ***
## women       2.224e-02  3.545e-02   0.627 0.532059    
## census      3.315e-04  7.090e-04   0.468 0.641160    
## typeprof    2.625e+01  4.592e+00   5.718 1.33e-07 ***
## typewc      7.636e+00  3.401e+00   2.245 0.027134 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.275 on 92 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7778, Adjusted R-squared:  0.7657 
## F-statistic:  64.4 on 5 and 92 DF,  p-value: < 2.2e-16
```

```r
resi_wo_edu <- Prestige$prestige - predict(lm_5,newdata = Prestige)
# This model explains around 78% of variance in prestige
# Regression of education on variables other than prestige
lm_6 <- lm(education ~ income + women + census + type,Prestige)
resi_edu <- Prestige$education - predict(lm_6, newdata = Prestige)
#Regressing resi_wo_edu (residual of prestige variable w/o education variable)
#on resi_edu (residual of education variable obtained from model using same predictors)
lm_7 <- lm(resi_wo_edu ~ resi_edu)
summary(lm_7)
```

```
##
## Call:
## lm(formula = resi_wo_edu ~ resi_edu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9863  -4.9813   0.6983   4.8690  19.2402
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.824e-14  6.921e-01   0.000        1
## resi_edu     3.933e+00  6.362e-01   6.182 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.851 on 96 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.2847, Adjusted R-squared:  0.2773
## F-statistic: 38.21 on 1 and 96 DF,  p-value: 1.537e-08
```

```r
#Slope for this regression is 3.933

#Regression using all the variables
lm_8 <- lm(prestige ~ ., Prestige)
summary(lm_8)
```

```
##
## Call:
## lm(formula = prestige ~ ., data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9863  -4.9813   0.6983   4.8690  19.2402
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.213e+01  8.018e+00  -1.513  0.13380
## education    3.933e+00  6.535e-01   6.019 3.64e-08 ***
## income       9.946e-04  2.601e-04   3.824  0.00024 ***
## women        1.310e-02  3.018e-02   0.434  0.66524
## census       1.156e-03  6.183e-04   1.870  0.06471 .
## typeprof     1.077e+01  4.676e+00   2.303  0.02354 *
## typewc       2.877e-01  3.139e+00   0.092  0.92718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.037 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.841,  Adjusted R-squared:  0.8306
## F-statistic: 80.25 on 6 and 91 DF,  p-value: < 2.2e-16
```

```r
#While using regression of prestige on all variables; we get the same coefficient = 3.933
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case?
The intercept is zero cause the residuals for both resi_wo_edu and resi_education will have average value of zero. And regression line passes thrugh the average of the independent and dependent variables in case of a 1 continuous model

(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. This factor is helping in understanding the additional variance that only the factor X1 is exlaining

(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?
We can use it when we want to analyze if the additional of a particular variable is actually helping us explain the data better of not

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 $E_{Y|2,\dots,k}$ and $E_{1|2,\dots,k}$, given in the previous exercise. The partial correlation is denoted $r_{y1|2,\dots,k}$.

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

```
lm_9 <- lm(prestige ~ income + women ,Prestige);
resi_wo_edu <- Prestige$prestige - predict(lm_9,newdata = Prestige);
# Regression of education on variables other than prestige
lm_10 <- lm(education ~ income + women ,Prestige);
resi_edu <- Prestige$education - predict(lm_10, newdata = Prestige);

#To find correlation between the residuals
cor(cbind(resi_edu, resi_wo_edu))
```

```
##              resi_edu resi_wo_edu
## resi_edu    1.0000000   0.7362604
## resi_wo_edu 0.7362604   1.0000000
```

```
#The partial correlation between prestige and education is 0.736
```

2. In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{y1|2,\dots,k} = 0$ if and only if B_1 is 0? *If the residuals are not correlated, it means that residuals of $X_1 will not be able explain the residuals left after Y regressed onto $X_2, \dots,$X_K.*

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.

$$\Sigma \hat{y}_i \hat{e}_i = 0$$

I. Since $\hat{y}_i$ is orthogonal to $\hat{e}_i$ so

$$\Sigma \hat{y}_i \hat{e}_i = (y)^T e = 0$$ (orthogonal vectors have zero cross-prods)

II. To prove orthogonality of $\hat{y}_i$ and $\hat{e}_i$

$$\Sigma(\hat{e}_i - \bar{\hat{e}})(\hat{y} - \bar{y})$$

$$= \hat{\beta}_1 \Sigma(\hat{y}_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1^2 \Sigma(x_i - \bar{x})^2$$

$$= \frac{S_{XY}}{S_{XX}}(S_{XY}) - \left(\frac{S_{XY}}{S_{XX}}\right)^2 (S_{XX})$$

$$= 0$$

2. $\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

$$= \Sigma \hat{e}_i (\hat{y}_i - \bar{y})$$

$$= \Sigma \hat{e}_i \hat{y}_i - \Sigma e_i \bar{y}$$

$$= 0 - \Sigma \hat{e}_i \bar{y}$$ (from 1)

$$= -\bar{y} \underbrace{\Sigma \hat{e}_i} = 0$$ $\hat{e} \sim N(0, \sigma)$ so $\Sigma \hat{e} = 0$

Figure 1: Solution 1

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?

Reason 1 : Misrepresentation of Sample

Since they recruited only children reading substaintially below the grade level; the results from finding cant be generalized for all children

Solution : We can do random or stratified sampling of the students to ensure every student class is represented well

Reason 2 : The new program is not been comparing the effect to a control group

All the effect cant be attributed entirely to the new program

Solution : Form a similar control group that is not given treatment or given previous treatment to understand the incremental/decremental effect of the new proggram

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.

II

1. $\bar{y} = \bar{x}$

$sd(y) = sd(x)$

$y_i \sim N(\mu, \sigma^2) \qquad x_i \sim N(\mu, \sigma^2)$

$$\hat{y} = \alpha + \beta \hat{x} + \hat{e}$$

$$E(\hat{y}) = E(\alpha + \beta \hat{x}) + \hat{e} \qquad \hat{e} \sim N(0, \sigma^2)$$

$$\mu = \alpha + \beta \mu + 0$$

We get $\alpha_{y|x} = 0$ and $\beta = 1$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \sum (x_i - \bar{x})(y_i - \bar{y}) / \delta_x \delta_y$$

$$= 1$$

Figure 2: Solution 2