MA678 homework 01

Meghamala Pandit

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 <- read.table(file = "exercise2.1.man", header = T)
attach(pyth)
fx1 <- x1[1:40]
fx2 <- x2[1:40]
fy <- y[1:40]
regout <- lm(fy ~ fx1+fx2) #regression of y on x1 and x2
summary(regout)
```

```
##
## Call:
## lm(formula = fy ~ fx1 + fx2)
##
## 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 **
## fx1          0.51481    0.04590  11.216 1.84e-13 ***
## fx2          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
```

The equation of the regression line is $y = 1.3151 + 0.5148x1 + 0.8069x2$. *The y intercept is the predicted value for y when both x1 and x2 are zero.* $\beta_1 = 0.5148$ *and* $\beta_2 = 0.8069$ *are the differences in the predicted values of y for each unit difference in x1 and x2 respectively.The t-values and the p-values of the regression coefficients show that the coefficients are statistically significant.*

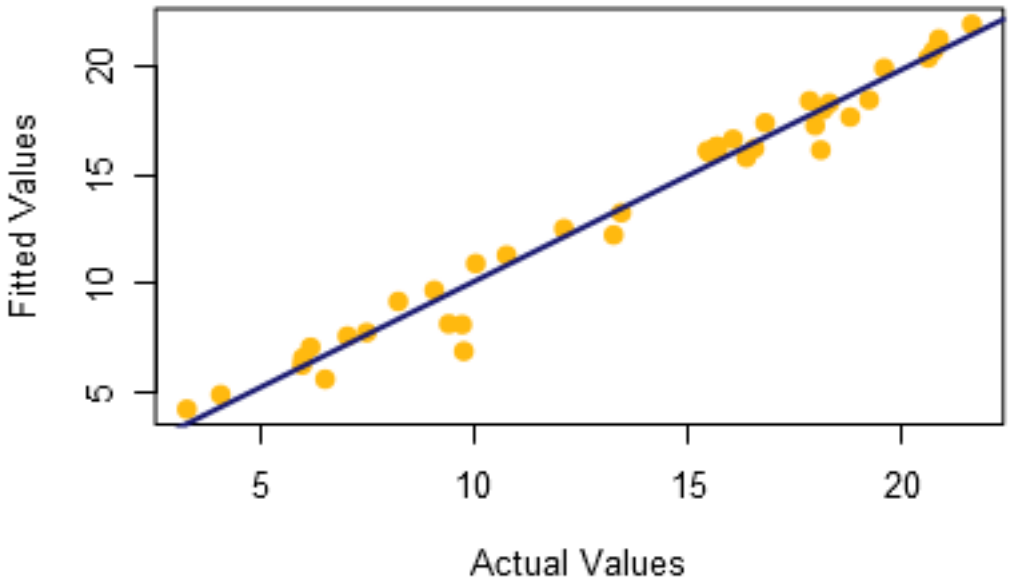
Regarding the fit of the model, the multiple and adjusted R-squared values and also the p-values and F-statistic, all suggest that the model is a very good fit. Almost all the variance of the response variable is explained by the model. However, since R-squared and p-values are not completely reliable measures in isolation, to test the fit of a model, it is good to check the residual plots of the model for any unusual patterns. The residuals are plotted in the third part of this question.

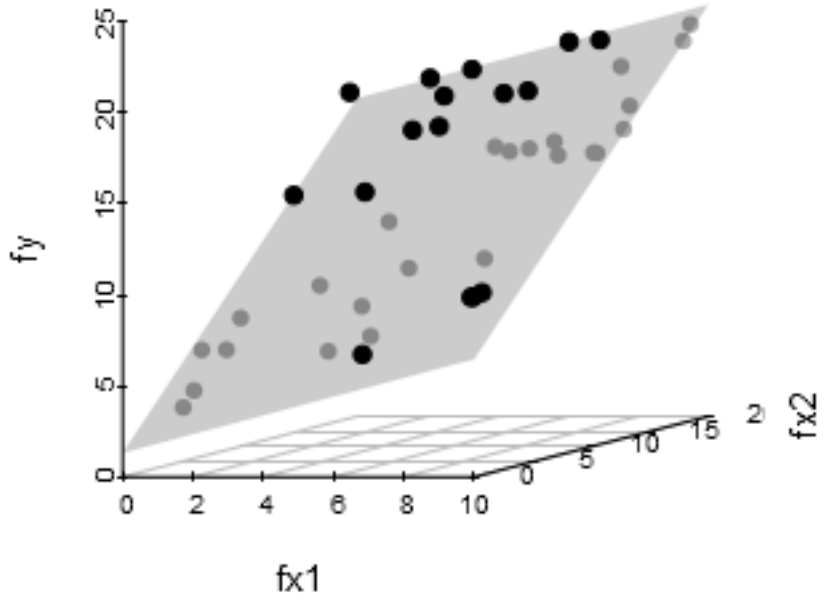
2. Display the estimated model graphically as in (GH) Figure 3.2.

```
pyth <- read.table(file = "exercise2.1.man", header = T)
attach(pyth)
```

```
## The following objects are masked from pyth (pos = 3):
##
##     x1, x2, y
```

```
fx1 <- x1[1:40]
fx2 <- x2[1:40]
fy <- y[1:40]
regout <- lm(fy ~ fx1+fx2)
fity <- regout$coef[1]+regout$coef[2]*fx1+regout$coef[3]*fx2

#plotting actual vs fitted values
plot(fy,fity, xlab = "Actual Values", ylab = "Fitted Values", pch=16,cex=1.2, col="darkgoldenrod1")
abline(lm(fity~fy), col="midnightblue",lwd=2)

#3D plot of the multiple regression
library(scatterplot3d)
spl <- scatterplot3d(x=fx1,y=fx2,z=fy,pch = 16, angle = 15,type = "p", grid = T, box = F,mar = c(5,5, 0
regl<- lm(fy~fx1+fx2)
spl$plane3d(regl,draw_polygon = TRUE, draw_lines = F)
wh <- resid(regl) > 0
spl$points3d(fx1[wh], fx2[wh], fy[wh], pch = 19)
```



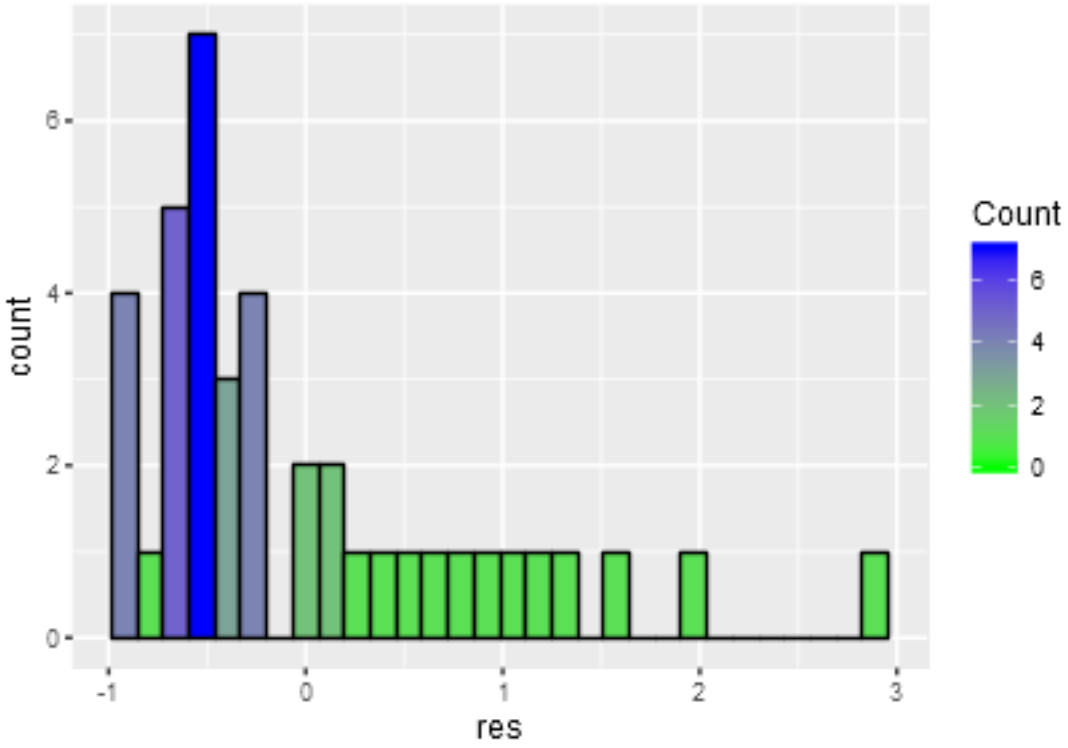


In the 3d scatterplot, the dark dots are above the regression plane and the light grey dots are below the regression plane.

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

```
#residual plot
res <- resid(regout)
df <- data.frame(fx1,fx2,fy)
library(ggplot2)
ggplot(df, aes(res))+
  geom_histogram(aes(fill= ..count..), col="black")+
  scale_fill_gradient("Count", low="green", high="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The residuals do not appear to be normally distributed as expected. The residual plot is skewed to the right.

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

```
train <- pyth[which(pyth$y > 0),  ]
test <- pyth[which(is.na(pyth$y > 0)),  ]

fit <- lm(y~x1+x2, train)
predict(fit)
```

```
##         1         2         3         4         5         6         7
## 16.221378  7.090400 16.133677  9.700421 17.254606 10.926653 20.694923
##         8         9        10        11        12        13        14
##  6.905222  9.188536  5.626640 16.057014 15.990263 20.356247 19.890290
##        15        16        17        18        19        20        21
##  8.128384 15.789193 18.286339 12.238483 12.528169 17.977308 17.380425
##        22        23        24        25        26        27        28
## 16.210499 17.647595 16.305805  4.914170 16.101788 13.256420 21.220813
##        29        30        31        32        33        34        35
## 16.633082  6.596613  4.248042  8.159388 11.309235  6.266295 18.431754
```

```
##        36        37        38        39        40
## 16.283969  7.601885 21.903629 18.390868  7.763580
```

```
predict(fit,newdata = test, type = "response")
```

```
##        41        42        43        44        45        46        47
## 14.812484 19.142865  5.916816 10.530475 19.012485 13.398863  4.829144
##        48        49        50        51        52        53        54
##  9.145767  5.892489 12.338639 18.908561 16.064649  8.963122 14.972786
##        55        56        57        58        59        60
##  5.859744  7.374900  4.535267 15.133280  9.100899 16.084900
```

```
#confidence intervals for the predictions
CI <- predict(fit,newdata = test, interval = "confidence")
library(knitr)
kable(CI)
```

	fit	lwr	
41	14.812484	14.295452	15.3
42	19.142865	18.604860	19.6
43	5.916816	5.203484	6.6
44	10.530475	10.017798	11.0
45	19.012485	18.501461	19.5
46	13.398863	13.105741	13.6
47	4.829144	4.258555	5.3
48	9.145767	8.553508	9.7
49	5.892489	5.313225	6.4
50	12.338639	11.763150	12.9
51	18.908561	18.424689	19.3
52	16.064649	15.739275	16.3
53	8.963122	8.510209	9.4
54	14.972786	14.521738	15.4
55	5.859744	5.326283	6.3
56	7.374900	6.863539	7.8
57	4.535267	3.940205	5.1
58	15.133280	14.817297	15.4
59	9.100899	8.654405	9.5
60	16.084900	15.596495	16.5
_The	confidence i	ntervals for	the predicted values of y seem to be narrow implying that the predictions may be accu

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. *y = earnings, x = height and $log(y) = \beta_0 + \beta_1 log(x)$. β_0 = log(30000) - (0.8)*log(66) = 6.9572 Therefore, the regression equation is log(y) = 6.9572 + 0.8log(x)*

*The earnings of 95% of people fall within 2 standard deviations from the mean. (Residual Standard Deviation)*2 = 0.1*0.8 Residual sd = 0.04*

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?

The R^2 for the model is $1 - (\frac{0.04^2}{0.05^2}) = 0.36$, *which implies that 36% of the variation in y can be explained by the model.*

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.

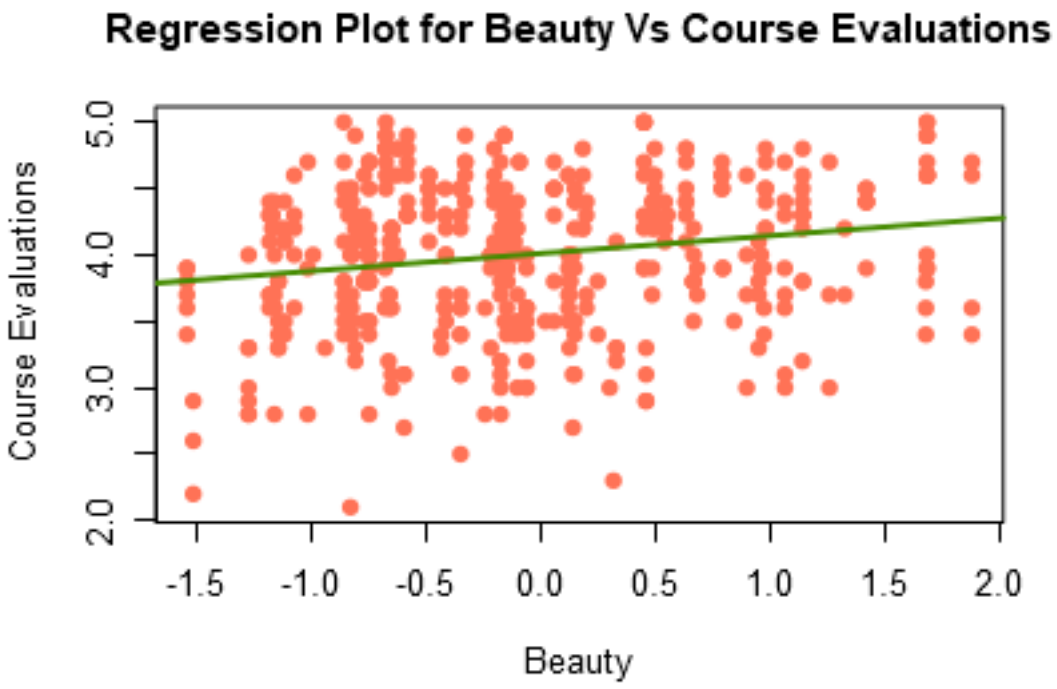
```
beauty.data <- read.table (paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, s
```

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.

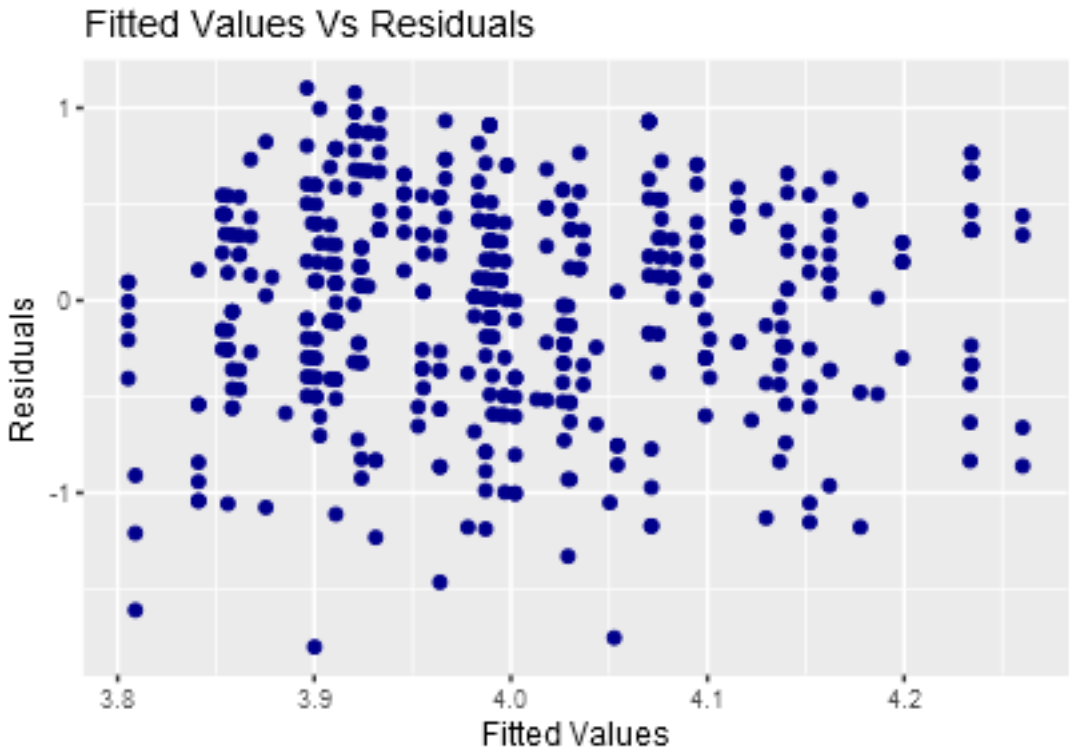
```
gelman_example_dir<-"http://www.stat.columbia.edu/~gelman/arm/examples/"
bd <- read.table(paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, sep=",")

#regression of course evaluation on beauty
fitb <- lm(bd$courseevaluation ~ bd$btystdave)
summary(fitb)
```

```
##
## Call:
## lm(formula = bd$courseevaluation ~ bd$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 ***
## bd$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(bd$btystdave, bd$courseevaluation, pch=16, col="coral1", xlab = "Beauty", ylab = "Course Evaluatio
abline(fitb, col="chartreuse4", lwd=2)
```




```
df <- data.frame(bd)
ggplot(df, aes(x=predict(fitb), y=resid(fitb)))+
  geom_point(col = "darkblue")+
  xlab("Fitted Values")+
  ylab("Residuals")+
  ggtitle("Fitted Values Vs Residuals")
```



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.

The variables are on differet scales. We need to scale them to compare them.

```
gelman_example_dir<-"http://www.stat.columbia.edu/~gelman/arm/examples/"
bd <- read.table(paste0(gelman_example_dir,"beauty/ProfEvaltnsBeautyPublic.csv"), header=T, sep=",")
bd_sc <- as.data.frame(scale(bd))
fit1 <- lm(bd$courseevaluation ~., bd_sc)
#Performing regression stepwise
stepfit <- stepAIC(fit1, direction = "both", trace = F)
summary(stepfit)
```

```
##
## Call:
## lm(formula = bd$courseevaluation ~ profnumber + age + beautyf2upper +
##     beautyfupperdiv + beautymupperdiv + btystdf2u + btystdmu +
##     class3 + class8 + class12 + class14 + class17 + class18 +
##     class19 + class26 + class27 + nonenglish + onecredit + percentevaluating +
##     profevaluation, data = bd_sc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89608 -0.09168  0.00849  0.11339  0.56950
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.998e+00  8.383e-03 476.977  < 2e-16 ***
## profnumber         2.196e-02  9.459e-03   2.322  0.02071 *
```

```
## age                4.392e-02  9.661e-03   4.546 7.06e-06 ***
## beautyf2upper      8.775e+04  4.442e+04   1.975  0.04885 *
## beautyfupperdiv    2.455e-02  1.362e-02   1.803  0.07208 .
## beautymupperdiv    1.150e+05  5.161e+04   2.228  0.02639 *
## btystdf2u         -8.775e+04  4.442e+04  -1.975  0.04885 *
## btystdmu          -1.150e+05  5.161e+04  -2.228  0.02639 *
## class3            -1.429e-02  8.738e-03  -1.635  0.10274
## class8             1.815e-02  8.541e-03   2.125  0.03418 *
## class12           -2.341e-02  8.619e-03  -2.715  0.00688 **
## class14            2.301e-02  8.688e-03   2.649  0.00837 **
## class17            1.403e-02  8.752e-03   1.603  0.10957
## class18            2.537e-02  8.675e-03   2.925  0.00362 **
## class19           -2.456e-02  8.757e-03  -2.805  0.00526 **
## class26            1.348e-02  8.566e-03   1.574  0.11621
## class27            2.071e-02  8.483e-03   2.442  0.01501 *
## nonenglish        -4.112e-02  9.006e-03  -4.566 6.46e-06 ***
## onecredit          1.793e-02  9.061e-03   1.979  0.04846 *
## percentevaluating  2.211e-02  9.064e-03   2.439  0.01512 *
## profevaluation     5.135e-01  9.181e-03  55.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1804 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
```

From the summary statistics, the model seems to be a good fit. But, these statistics are not the standalone measures of a good fit.

Combining the classes into one variable, we can check for interactions between the class variable and the beauty variable. This will be a case of interaction between one binary and one continuous variable.

```
bd$sum_class <- rowSums(bd[,c("class3", "class8", "class12","class14","class17","class18","class19","cla
fit2 <- lm(bd$courseevaluation ~ bd$sum_class*bd$btystdave)
summary(fit2)
```

```
##
## Call:
## lm(formula = bd$courseevaluation ~ bd$sum_class * bd$btystdave)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79337 -0.36323  0.05063  0.40482  1.11076
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.01029    0.02660 150.788  < 2e-16 ***
## bd$sum_class                -0.03446    0.09784  -0.352    0.725
## bd$btystdave                 0.14154    0.03295   4.295 2.13e-05 ***
## bd$sum_class:bd$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
```

The regression equation for this model is $y = 4.01 - 0.034x_1 + 0.141x_2 - 0.194x_1 : x_2$ *, where* x_1 *and* x_2 *represent the class and beauty variables respectively. Here, the calss variable is binary and the beauty variable is continuous. When both* x_1 *and* x_2 *are 0,* y *= intercept = 4.01.*

When x_1 *is 0, the difference in the predicted value of* y *for a unit change in* x_2 *is given by the coefficient of* x_2 *= 0.141*

When x_2 *is 0, the difference in the predicted value of* y *for a unit change in* x_1 *is given by the coefficient of* x_1 *= -0.034 (inverse relationship). But the high p-value of this coefficient may imply that the coefficient is not statistically significant and hence, cannot explain the variation in y very effectively.*

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).

```
set.seed(99)
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]

```
set.seed(99)
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
fit  <- lm (var2 ~ var1)
z.scores <- coef(fit)[2]/se.coef(fit)[2]
z.scores
```

The absolute value of the z-score for the slope coefficient is lesser than 2. Hence, the slope coefficient is not statistically significant.

2. Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of var1 divided by its standard error). If the absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation:

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit  <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
  z.scores[k]
```

```
}
z.scores
z_score <- z.scores[which(abs(z.scores)>2)]
z_score
summary(fit)
```

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

Since the variables are generated randomly, the z-scores for each trial are different and hence, there are a different number of statistically significant z-scores with evrry trial. However, the regression coefficient tends to remain statistically insignificant with very small t-values and large p-values. _

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"))
attach(Prestige)
```

```
## The following object is masked from package:datasets:
##
##     women
```

```
#regression of prestige on other variables excluding education
mreg <- lm(prestige ~ income + women)
yres <- resid(mreg)

#regression of education on the other variables: income and women
xreg <- lm(education ~ income + women)
xres <- resid(xreg)

#regression of the residuals of the above two regressions
ereg <- lm(yres~xres)
ereg
```

```
##
## Call:
## lm(formula = yres ~ xres)
##
## Coefficients:
## (Intercept)         xres
```

```
##  -2.992e-15    4.187e+00
```

```
summary(ereg)
```

```
##
## Call:
## lm(formula = yres ~ xres)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.992e-15  7.691e-01    0.00        1
## xres         4.187e+00  3.848e-01   10.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.768 on 100 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5375
## F-statistic: 118.4 on 1 and 100 DF,  p-value: < 2.2e-16
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case? *The sum of residuals is equal to zero and the residuals from both the regression models will have amean of zero. Since the here passes through the mean of the response and predictor variables, the intercept here is 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"? *x_1 appears to explain variability in y not explained by the other variables. Hence, we can consider describing β_1 as the effect of x_1 on y when influence of other variables is removed.*

(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.

```
fox_data_dir<-"http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/"
Pr<-read.table(paste0(fox_data_dir,"Prestige.txt"))

#Regression of prestige on income and women
reg1 <- lm(prestige ~ income + women, Pr)
res1 <- resid(reg1)
#REgression of education on income and women
reg2 <- lm(education ~ income + women, Pr)
res2 <- resid(reg2)
#Correlation between the residuals
cor(cbind(res2, res1))
```

```
##           res2      res1
```

```
## res2 1.0000000 0.7362604
## res1 0.7362604 1.0000000
```

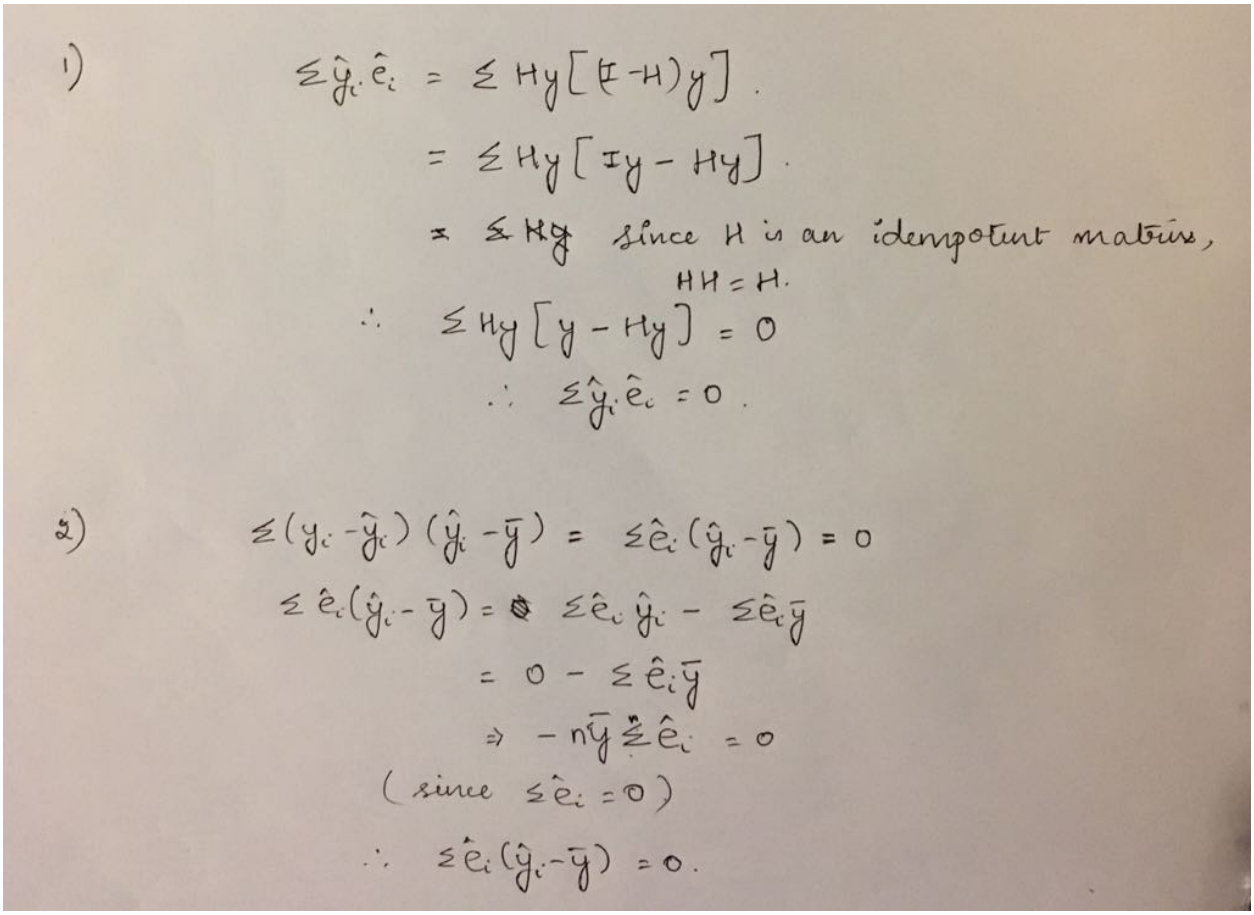
The partial correlation between x_1 and y is $r_{y1|2,\ldots,k} = 0.736$

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) $\beta_{x|y} = \frac{\sum (y-\bar{y})(x-\bar{x})}{\sum (y-\bar{y})^2}$, $\beta_{y|x} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}}$$

$$\beta_{x|y} \left(\sum (y-\bar{y})^2\right) = \sum (y-\bar{y})(x-\bar{x})$$

$$r = \frac{\beta_{x|y}\left[\sum (y-\bar{y})^2\right]}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{\beta_{x|y}[sd(y)]^2}{\sqrt{[sd(y)]^4}} = \frac{\beta_{x|y}[sd(y)]^2}{[sd(y)]^2}$$

$\Rightarrow r = \beta_{x|y}$.

Similarly; $r = \frac{\beta_{y|x}\left[\sum (x-\bar{x})^2\right]}{\sqrt{[sd(x)]^2[sd(y)]^2}} = \frac{\beta_{y|x}[sd(x)]^2}{[sd(x)]^2} = \beta_{y|x}$

$\therefore r = \beta_{y|x}$.

$\Rightarrow r = \beta_{y|x} = \beta_{x|y}$.

$y = \alpha_{y|x} + \beta_{y|x} x$ and $x = \alpha_{x|y} + \beta_{x|y} y$

$\alpha_{y|x} = \bar{y} - \beta_{y|x} \bar{x}$ and $\alpha_{x|y} = \bar{x} - \beta_{x|y} \bar{y}$

since $\bar{x} = \bar{y}$, and $\beta_{x|y} = \beta_{y|x}$.

we get $\alpha_{y|x} = \alpha_{x|y}$.

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?

The research sample would include only children who are reading below grade. The sample would not be representative of the population, which is all the children. The new program may have a different effect on children who are reading above average and children who are reading at the average level. Overall, the results would be biased.

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.