homework_01

Sample Solution

Septemeber 12, 2017

% vectors and matrices

% vectors and matrices

Please do NOT distribute.

Data analysis

Pyth!

```
pyth <- read.table ("http://www.stat.columbia.edu/~gelman/arm/examples/pyth/exercise2.1.dat",
                    header=T, sep=" ")
```

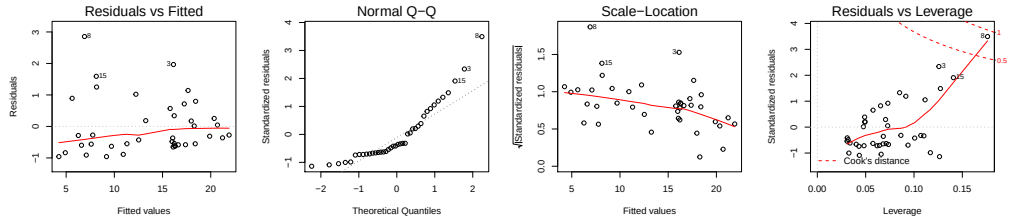
The folder pyth contains outcome `y` and inputs `x1`, `x2` for 40 data points, with a further 20 points with the inputs but no observed outcome. Save the file to your working directory and read it into R using the `read.table()` function.

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.

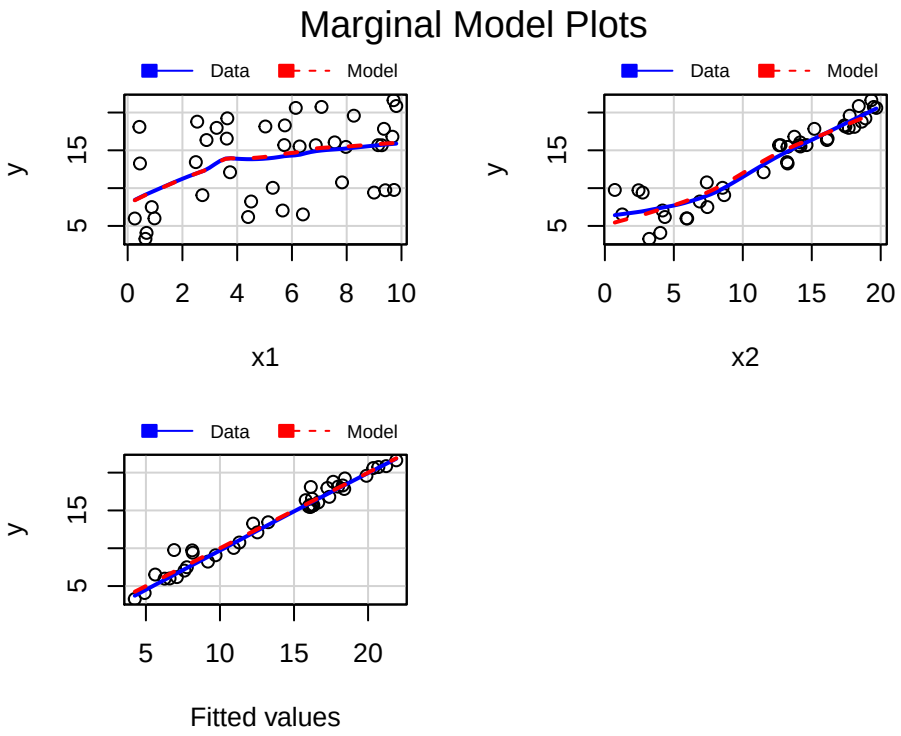
```
fit1 <- lm(y~x1+x2, data = pyth[1:40,])
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth[1:40, ])
##
## 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
```

```
par(mfrow=c(1,4))
plot(fit1)
```




```
car::marginalModelPlots(fit1)
```

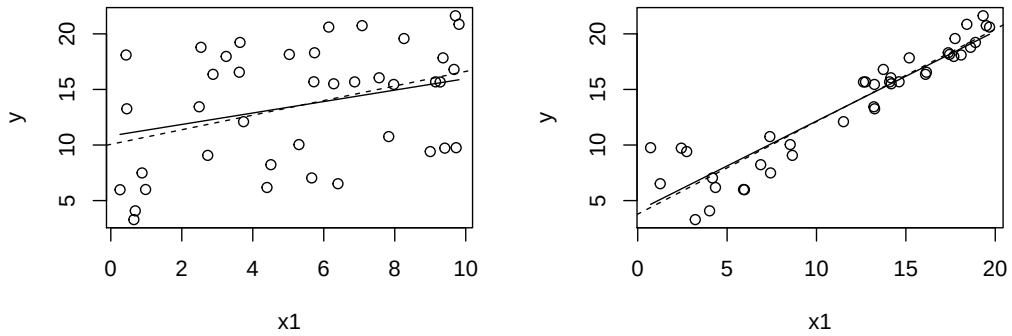


2. Display the estimated model graphically as in Figure 3.2.

There are couple of points that I'd like you to understand with this exercise. - When you are fitting the model jointly you are fitting a plane. - To draw a line, you need to fix one of the variable to a value. In the example below I'm fixing the other variable to be at the sample mean. - the lines you fit just to each predictor variable (dashed) is not the same as the lines you get from fitting jointly (solid).

```
my_surface <- function(f, n=10, ...) {
  ranges <- rgl:::.getRanges()
  x <- seq(ranges$xlim[1], ranges$xlim[2], length=n)
  y <- seq(ranges$ylim[1], ranges$ylim[2], length=n)
  z <- outer(x,y,f)
  surface3d(x, y, z, ...)
}

f <- function(x1, x2){
coef(fit1)[1] + coef(fit1)[2]*x1+ coef(fit1)[3]*x2
  }
plot3d(pyth[1:40,]$x1,pyth[1:40,]$x2,pyth[1:40,]$y, type="p",
       col="red", xlab="X1", ylab="X2", zlab="Y", site=5, lwd=15)
my_surface(f, alpha=.2 )
```

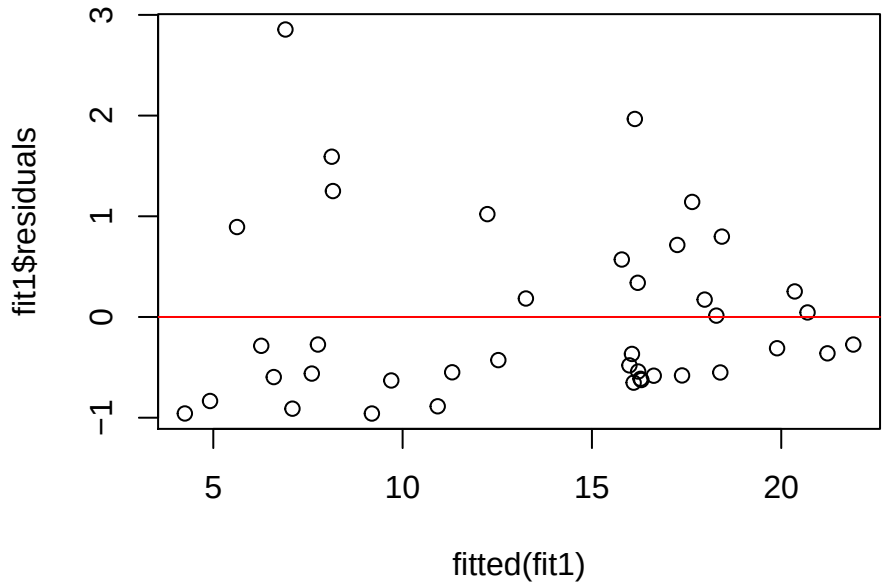
```
par(mfrow=c(1,2))
plot ( pyth[1:40,]$x1,  pyth[1:40,]$y, xlab = "x1", ylab="y")
curve (coef(fit1)[1] + coef(fit1)[2]*x+ coef(fit1)[3]*mean(pyth[1:40,]$x2), add=TRUE)
abline(lm(y~x1,data=pyth[1:40,]),lty=2)

plot ( pyth[1:40,]$x2,  pyth[1:40,]$y, xlab = "x1", ylab="y")
curve (coef(fit1)[1] + coef(fit1)[2]*mean(pyth[1:40,]$x1)+ coef(fit1)[3]*x, add=TRUE)
abline(lm(y~x2,data=pyth[1:40,]),lty=2)
```



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

The constant variance assumption does not seem to be met.

```
plot(fitted(fit1),fit1$residuals)
abline(0,0, col="red")
```



However, it also seems like we have some outliers.

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

```
predict (fit1, newdata=pyth[41:60, ], interval="confidence", level=0.95)
```

```
##          fit       lwr       upr
## 41 14.812484 14.295452 15.329516
## 42 19.142865 18.604860 19.680871
## 43  5.916816  5.203484  6.630147
```

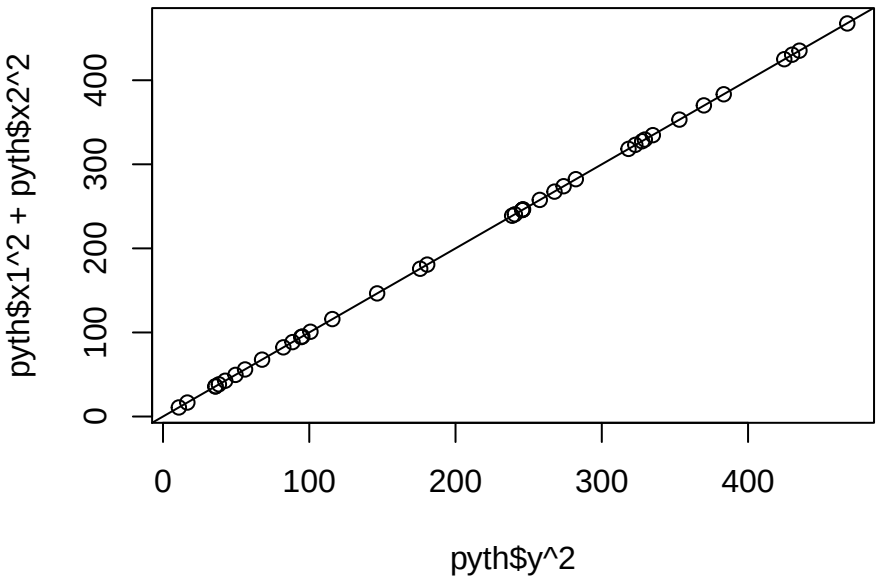
```
## 44 10.530475 10.017798 11.043152
## 45 19.012485 18.501461 19.523509
## 46 13.398863 13.105741 13.691985
## 47  4.829144  4.258555  5.399733
## 48  9.145767  8.553508  9.738026
## 49  5.892489  5.313225  6.471752
## 50 12.338639 11.763150 12.914129
## 51 18.908561 18.424689 19.392433
## 52 16.064649 15.739276 16.390022
## 53  8.963122  8.510209  9.416036
## 54 14.972786 14.521738 15.423835
## 55  5.859744  5.326283  6.393204
## 56  7.374900  6.863539  7.886262
## 57  4.535267  3.940205  5.130330
## 58 15.133280 14.817297 15.449264
## 59  9.100899  8.654405  9.547393
## 60 16.084900 15.596495 16.573306
```

After doing this exercise, take a look at Gelman and Nolan (2002, section 9.4) to see where these data came from.

When you look at Gelman and Nolan, it turns out that this data was generated from a model

$$y^2 = x1^2 + x2^2$$

```
plot(pyth$y^2,pyth$x1^2+pyth$x2^2)
abline(0,1)
```



however, ys are calculated only up to the second decimal creating small error.

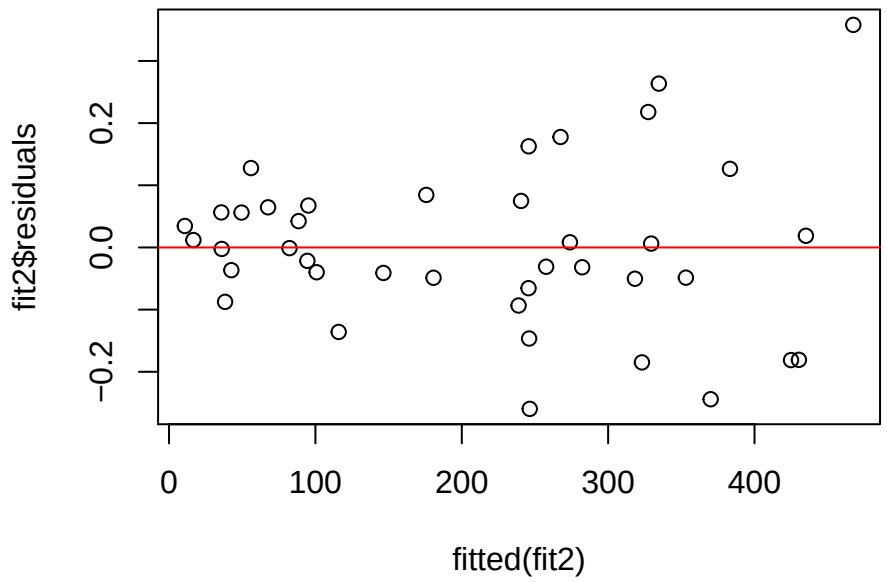
```
round(sqrt(pyth[1:40,]$x1^2+pyth[1:40,]$x2^2),2)-pyth[1:40,]$y
```

```
##  [1]  0.00  0.01  0.00  0.00  0.01  0.00  0.01  0.00  0.00  0.00  0.01
## [12]  0.00  0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.00
## [23]  0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00 -0.01  0.00  0.01
## [34]  0.00  0.01  0.00  0.00 -0.01  0.00 -0.01
```

```r
f2 <- function(x1, x2){
  sqrt(x1^2+x2^2)
}
n <- 200
plot3d(pyth[1:40,]$x1,pyth[1:40,]$x2,pyth[1:40,]$y,
       type="p", col="red",
        xlab="X1", ylab="X2", zlab="Y", site=5, lwd=15)
my_surface(f2, alpha=.2 )

fit2 <- lm(I(y^2)~I(x1^2)+I(x2^2)-1, data = pyth[1:40,])
summary(fit2)
```

```
##
## Call:
## lm(formula = I(y^2) ~ I(x1^2) + I(x2^2) - 1, data = pyth[1:40,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25966 -0.05413 -0.00175  0.06525  0.35803
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## I(x1^2) 0.9999889  0.0005356    1867   <2e-16 ***
## I(x2^2) 0.9998752  0.0001266    7900   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1326 on 38 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.171e+07 on 2 and 38 DF,  p-value: < 2.2e-16
```

```r
plot(fitted(fit2),fit2$residuals)
abline(0,0, col="red")
```



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.

$$log(earning) = a + b * log(height)$$

$$log(1.008earning) = a + b * log(1.01height)$$

$$log(1.008) + log(earning) = a + b * log(1.01) + b * log(height)$$

$$log(1.008) = b * log(1.01)$$

$$b = log(1.008)/log(1.01)$$

```
b <- log(1.008)/log(1.01)
a <- log(30000)-b*log(66)
a;b
```

```
## [1] 6.9539
```

```
## [1] 0.8007944
```

$$log(earning) = 6.9539 + 0.8 * log(height)$$

To compute the standard deviation we simply use the information given in the third bullet point. We know that the earnings of approximately 95% of people fall within a factor of 1.1 of predicted values. This means that 95% of predicted values fall within 2 standard deviations from the predicted means.

$$\log(earning) \pm \log(1.1) \approx \log(earning) \pm 2 * \sigma$$

We can reverse engineering what is the standard error.

```
sd =  log(1.1)/2
sd
```

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

2. Suppose the standard deviation of log heights is 5% in this population. What, then, is the R^2 of the regression model described here?

From P.41

$$R^2 = 1 - \hat{\sigma}^2/s_y^2$$

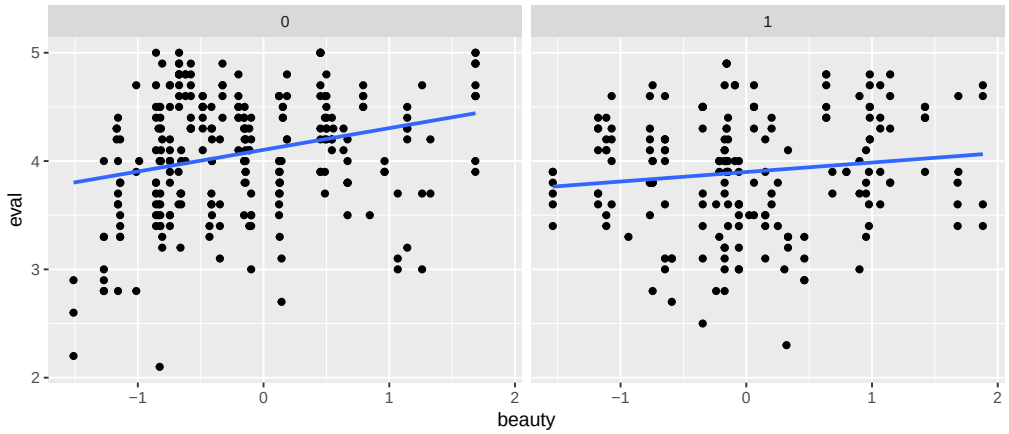
```
sd.population = 0.05
R2 <- 1 - (sd^2 / sd.population^2)
R2
```

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

Beauty and student evaluation

(From Gelman 3.5) 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.

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

beauty.data.dt <- data.table(beauty.data)
setnames(beauty.data.dt,c("btystdave","courseevaluation"),c("beauty","eval"))
n <- length (beauty.data.dt$eval)

# Make a scatterplot
ggplot(beauty.data.dt)+aes(x=beauty,y=eval)+
  geom_point()+
  facet_grid(.~female)+geom_smooth(method="lm",se=FALSE)
```



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.

```
lm_fit_beauty<-lm(eval ~ beauty+tenured+nonenglish+female,data=beauty.data.dt)
display(lm_fit_beauty)
```

```
## lm(formula = eval ~ beauty + tenured + nonenglish + female, data = beauty.data.dt)
##             coef.est coef.se
## (Intercept)  4.18     0.05
## beauty       0.14     0.03
## tenured     -0.10     0.05
## nonenglish  -0.35     0.10
## female      -0.22     0.05
## ---
## n = 463, k = 5
## residual sd = 0.53, R-Squared = 0.09
```

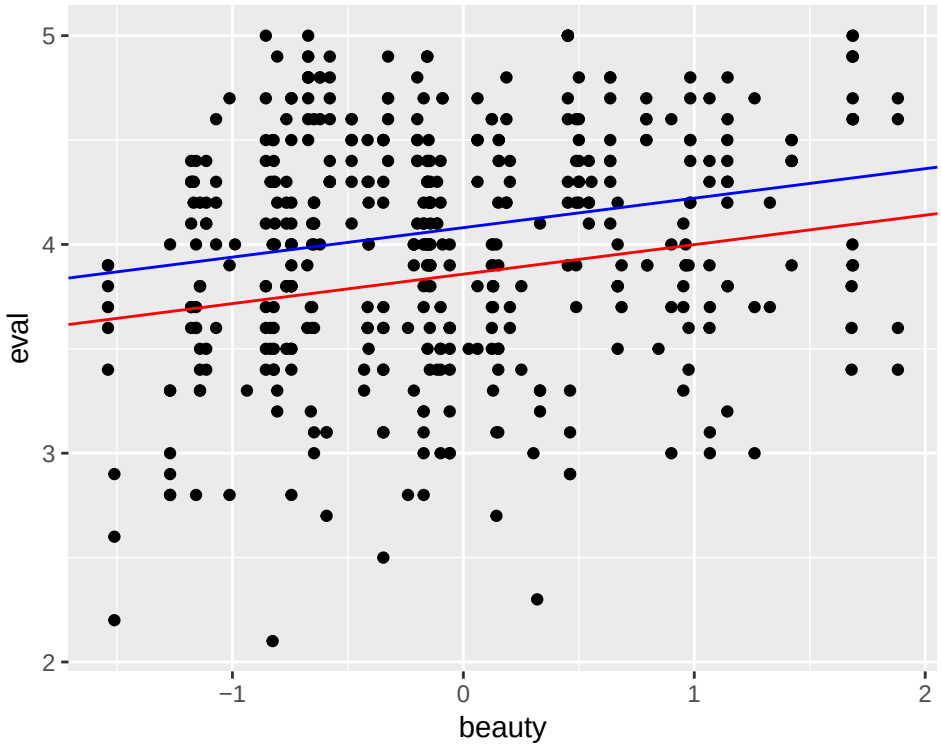
```
beauty.data$pred<-predict(lm_fit_beauty)
beauty.data$resid<-resid(lm_fit_beauty)
```

- `(Intercept)`: Expected evaluation for male, native English speaking non-tenured professor is 4.18.

- `beauty`: On average evaluation increases by 0.14 with every unit increase in beauty controlling for other variables.
- `tenured`: Difference in expected evaluation for tenured professor compared with the non-tenured professor controlling for the other variables is -0.10 .
- `nonenglish`: Difference in expected evaluation for non native English speaker compared with the native English speaker controlling for the other variables is -0.35.
- `female`: Difference in expected evaluation for female compared with male controlling for the other variables is -0.22.

Male tenured native English speaker (blue) and female tenured native English speaker (red)

```
# Make a scatterplot
ggplot(beauty.data.dt)+aes(x=beauty,y=eval)+
  geom_point()+
  geom_abline(intercept=coef(lm_fit_beauty)%*%c(1,0,1,0,0),
              slope=coef(lm_fit_beauty)["beauty"],color="blue")+
  geom_abline(intercept=coef(lm_fit_beauty)%*%c(1,0,1,0,1),
              slope=coef(lm_fit_beauty)["beauty"],color="red")
```

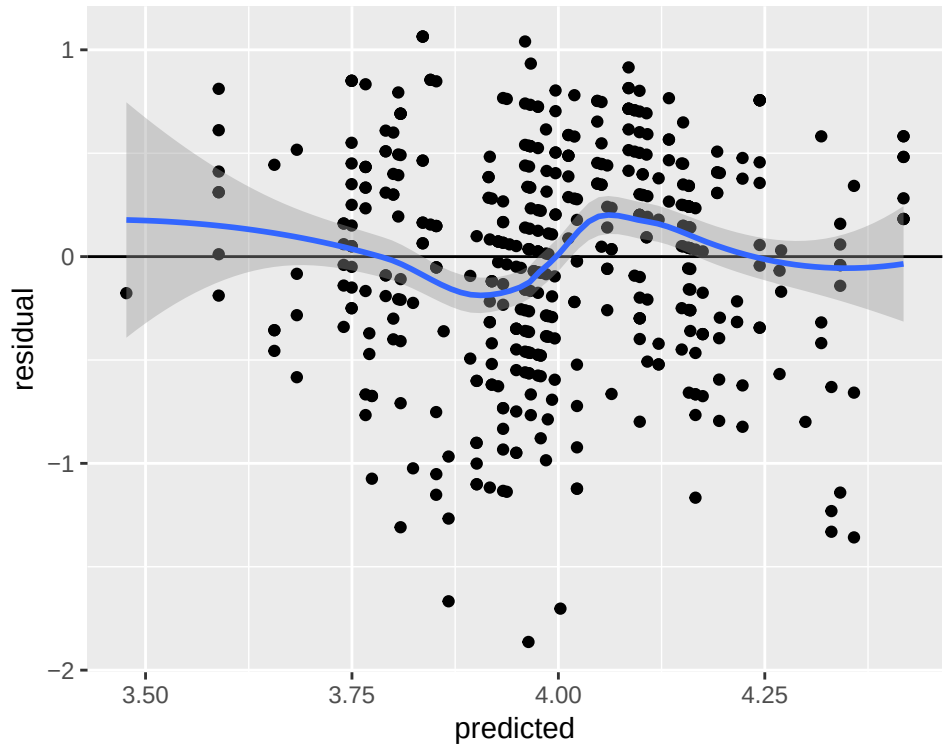


We see that on average female instructor gets lower evaluation compared to males.

- Fitted vs residual

```
ggplot(beauty.data)+aes(x=pred,y= resid)+geom_point()+
  xlab("predicted")+
  ylab("residual")+
  theme(legend.position="none")+geom_hline(yintercept=0)+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There seems to be some nonlinear effect that we are not taking into account.

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.

```
lm_fit_beauty_02<-lm(eval ~ beauty*female+tenured+nonenglish+female,data=beauty.data.dt)
display(lm_fit_beauty_02)
```

```
## lm(formula = eval ~ beauty * female + tenured + nonenglish +
##     female, data = beauty.data.dt)
##                coef.est coef.se
## (Intercept)     4.18     0.05
## beauty          0.19     0.04
## female         -0.23     0.05
## tenured        -0.08     0.05
## nonenglish     -0.35     0.10
## beauty:female -0.10     0.06
## ---
## n = 463, k = 6
## residual sd = 0.53, R-Squared = 0.10
```

I'm being lazy here and just added interaction between the beauty and female.

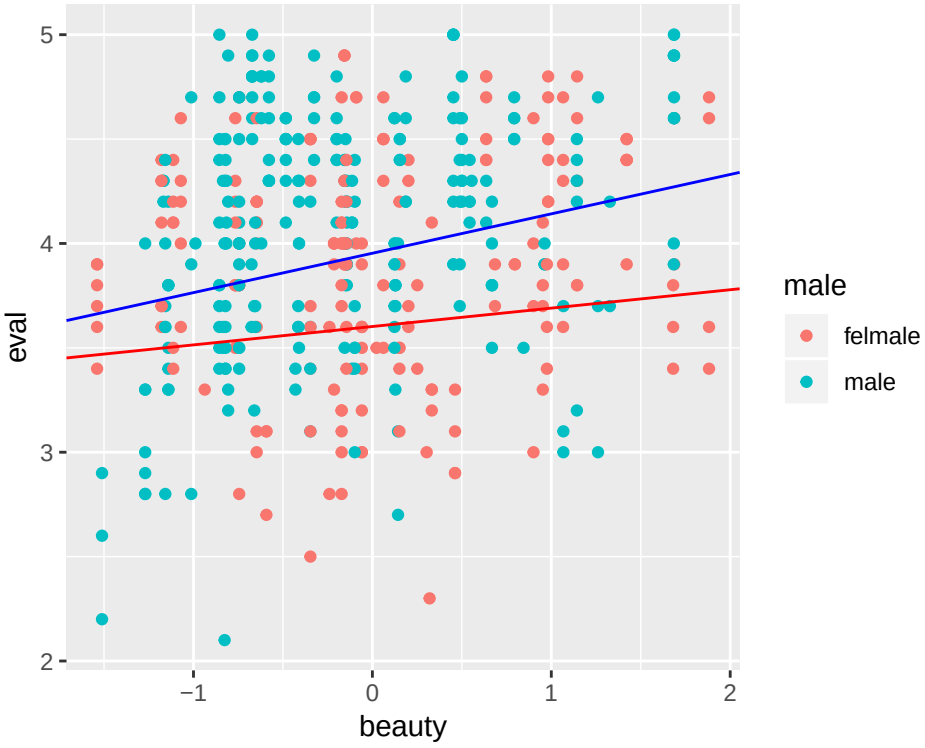
- `(Intercept)`: Expected evaluation for male, native English speaking non-tenured professor is 4.18.
- `beauty`: On average evaluation increases by 0.19 with every unit increase in beauty controlling for other variables.
- `tenured`: Difference in expected evaluation for tenured professor compared with the non-tenured professor controlling for the other variables is -0.08.
- `nonenglish`: Difference in expected evaluation for non native English speaker compared with the native English speaker controlling for the other variables is -0.35
- `female`: Difference in expected evaluation for female compared with male controlling for the other

variables is -0.23.

- `beauty:female`: Difference in expected evaluation with every unit increase in beauty for female compared with male controlling for the other variables is -0.1.

Male tenured native English speaker (blue) and female tenured native English speaker (red)

```
beauty.data.dt$male = factor(1-beauty.data.dt$female,labels = c("felmale","male"))
# Make a scatterplot
ggplot(beauty.data.dt)+aes(x=beauty,y=eval,col=male)+
  geom_point()+
  geom_abline(intercept=coef(lm_fit_beauty_02)%*%c(1,0,1,0,0,0),
              slope=coef(lm_fit_beauty_02)["beauty"],color="blue")+
  geom_abline(intercept=coef(lm_fit_beauty_02)%*%c(1,0,1,0,1,0),
              slope=coef(lm_fit_beauty_02)["beauty"]+coef(lm_fit_beauty_02)["beauty:female"],color="red
```

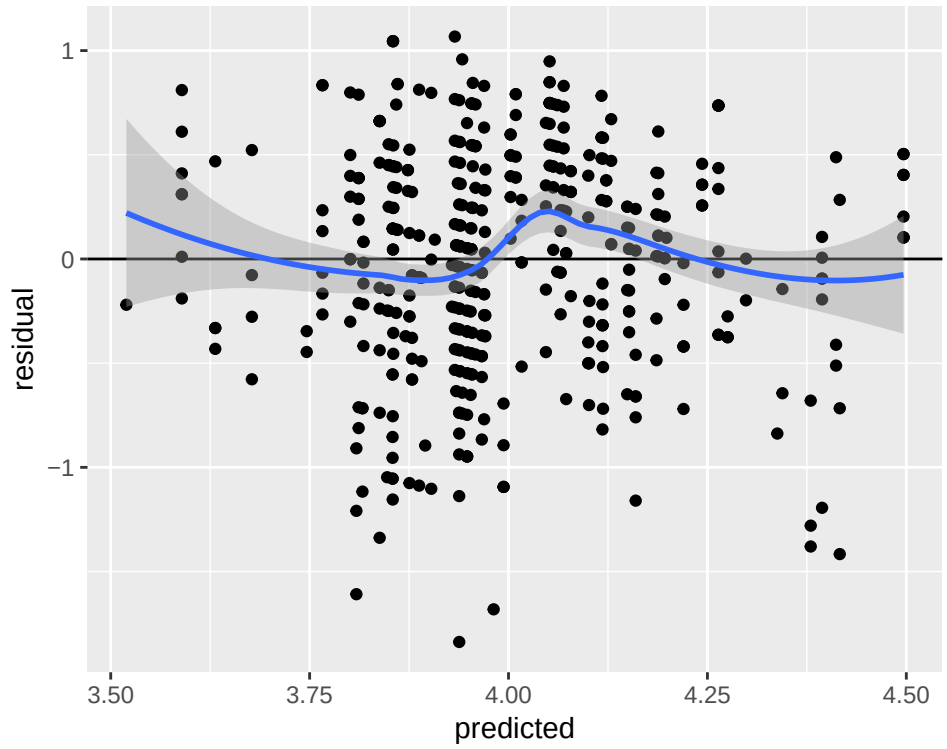


What we see is that the effect of beauty on the evaluation of the instructor is less for female than male.

- Fitted vs residual

```
beauty.data$pred2<-predict(lm_fit_beauty_02)
beauty.data$resid2<-resid(lm_fit_beauty_02)
ggplot(beauty.data)+aes(x=pred2,y= resid2)+geom_point()+
  xlab("predicted")+
  ylab("residual")+
  theme(legend.position="none")+geom_hline(yintercept=0)+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There's a weird bump in the residual indicating there might be something we are not taking into account.

See also Felton, Mitchell, and Stinson (2003) for more on this topic. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=426763

Conceptula excercises

On statistical significance.

Note: This is more like a demo to show you that you can get statistically significant result just by random chance. We haven't talked about the significance of the coefficient so we will follow Gelman and use the approximate definition, which is if the estimate is more than 2 sd away from 0 or equivalently, if the z score is bigger than 2 as being "significant".

(From Gelman 3.3) In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.

1. First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing in R. Generate another variable in the same way (call it var2).

```
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
```

Run a regression of one variable on the other. Is the slope coefficient statistically significant? [absolute value of the z-score(the estimated coefficient of var1 divided by its standard error) exceeds 2]

```
fit      <- lm (var2 ~ var1)
z.scores <- coef(fit)[2]/se.coef(fit)[2]
z.scores
```

```
##       var1
## -0.7200545
```

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:

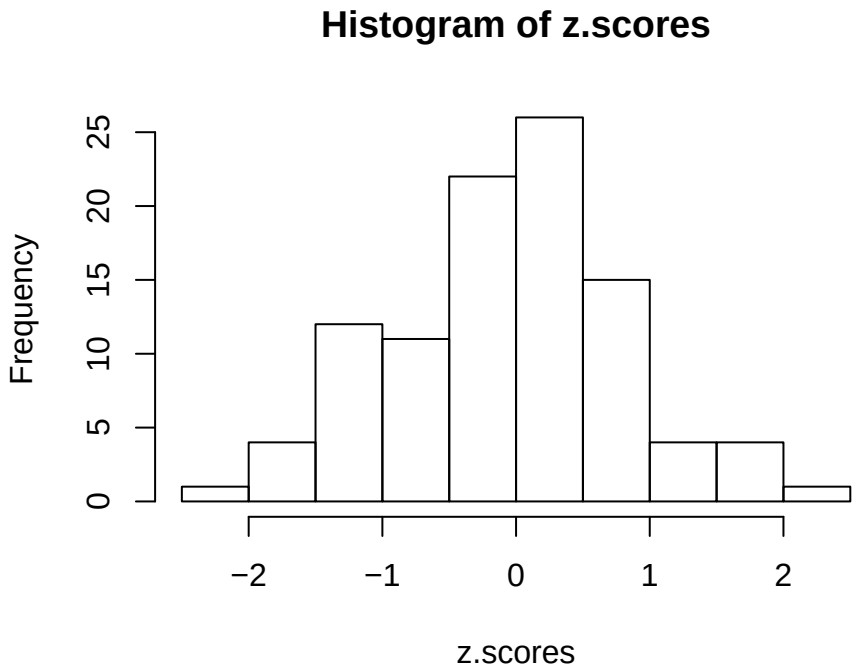
```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit  <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
```

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

There were 2 siginificant z-scores out of 100.

What can you say about statistical significance of regression coefficient?

```
hist( z.scores )
```



Since we are using approximately 5% alpha level by thresholding at zscore of 2, the result is what we would expect due to random chance.

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.

```
Prestige<-read.table("http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/Presti
Prestige_dt <- data.table(Prestige)
display(Prestige_dt[,lm(prestige~ education + income + women)])
```

```
## lm(formula = prestige ~ education + income + women)
##             coef.est coef.se
## (Intercept) -6.79     3.24
## education    4.19     0.39
## income       0.00     0.00
## women       -0.01     0.03
## ---
## n = 102, k = 4
## residual sd = 7.85, R-Squared = 0.80
```

```
Prestige_dt<- Prestige_dt[,yresid:=resid(lm(prestige ~ income + women))]
Prestige_dt<- Prestige_dt[,xresid:=resid(lm(education~  + income + women))]
display(Prestige_dt[,lm(yresid~  xresid)])
```

```
## lm(formula = yresid ~ xresid)
##             coef.est coef.se
## (Intercept) 0.00     0.77
## xresid      4.19     0.38
## ---
## n = 102, k = 2
## residual sd = 7.77, R-Squared = 0.54
```

The coefficient indeed agrees.

(b) The intercept for the simple regression in step 3 is 0. Why is this the case?

This is simple when you think about the geometry of the problem. When you project the residual of $\mathbf{Y}$ on residual of $\mathbf{X}_1$ after regressing them on the other Xs, both of the residuals are orthogonal to $\mathbf{1}$ by the definition of the residuals.

(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 at least in terms of linear influence.

(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?

When we need to update a regression model based on a new predictor variable for some reason. It's probably not good for distributed computing since you will need all the other variables.

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.

```
Prestige_dt[,cor(yresid, xresid)]
```

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

2. In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{y1|2,\ldots,k} = 0$ if and only if B_1 is 0?

If you think again in terms of projection, the answer should be obvious.

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, $\alpha_{y|x} = \alpha_{x|y}$.

Since $sd(\boldsymbol{y}) = sd(\boldsymbol{x})$, $(y_i - \bar{y})^2 = (x_i - \bar{x})^2$.

Therefore

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

Also

$$\beta_{y|x} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{Var(\boldsymbol{x})} = r_{xy} \frac{sd(\boldsymbol{y})}{sd(\boldsymbol{x})} = r_{xy}$$

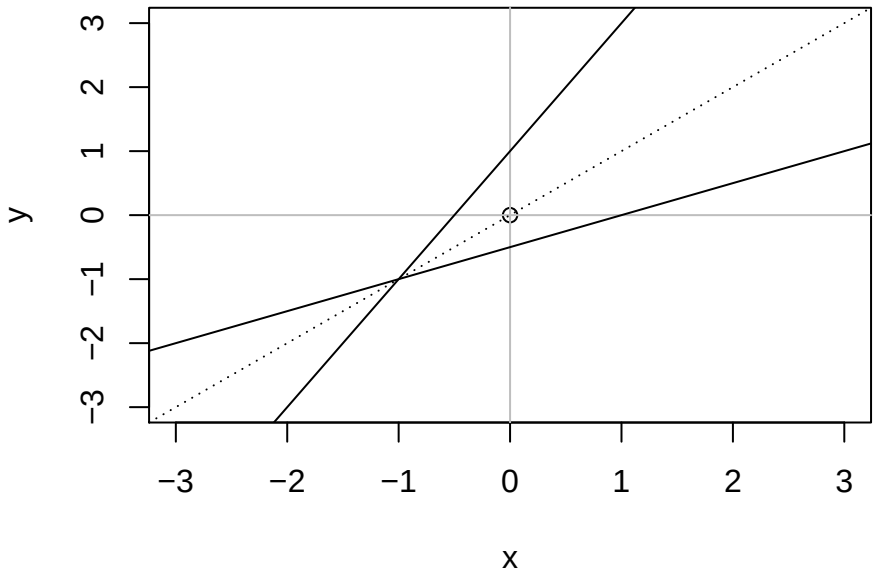
Finally,

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

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 $r_{xy} < 1$)?

Because they will be identical transposed but otherwise different. The only time they will match is when the line is slope 1 going throught the origin.

```
plot(0,0,xlim=c(-3,3),ylim=c(-3,3),xlab="x",ylab="y")
abline(h=0,col="grey"); abline(v=0,col="grey")
abline(1,2)
abline(-1/2,1/2)
abline(0,1,lty=3)
```



3. Imagine that educational researchers wish to assess the efficacy of a new program to improve he reading performance of children. To test the program, they recruit a group of children who are reading substantially vbelow grade level; after a year in the program, the researchers observe that the children, on average, have imporved their reading performance. Why is this a weak research design? How could it be improved?

This is weak because it's easy to "improve" poorly performing students by just the regression to the mean. The resercher should randomly sample students and assess their improvement.