Take Home Exam

Wenduo Wang

August 1, 2016

Problem 1: Beauty Pays!

After an initial exploration

First load the necessary libraries for logistic regression and plotting.

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 366723 19.6     592000 31.7   460000 24.6
## Vcells 567050  4.4    1308461 10.0   786432  6.0
```

```r
library(dplyr)
library(ggplot2)
# library(reshape2)
library(glmnet)
# library(randomForest)
```

(1) estimate the effect of "beauty" into course ratings

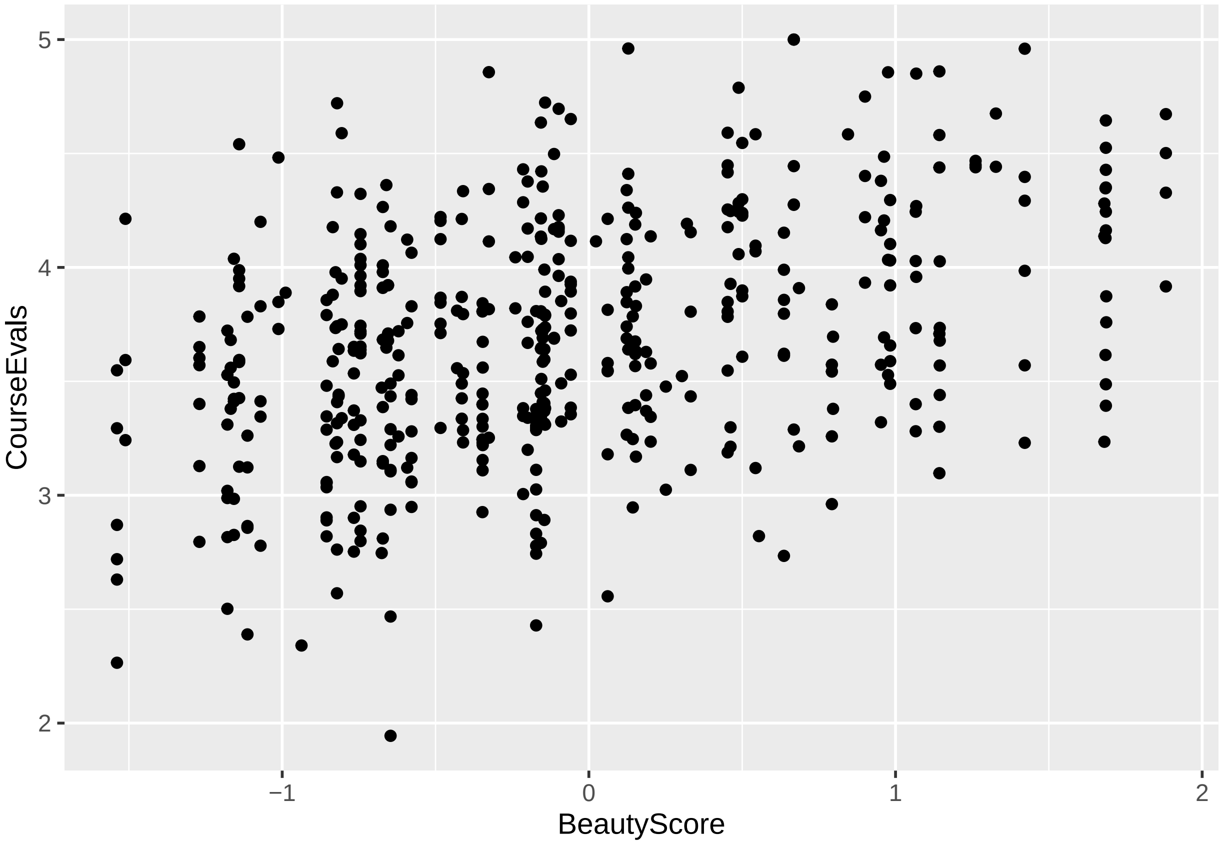
Read in the data and convert `female`, `nonenglish`, `lower` and `tenuretrack` into binary variables whose values are either `1` or `0`.

```r
col_class <- c("numeric", "numeric", rep("factor", 4))
Beauty <- read.csv("BeautyData.csv", colClasses=col_class, header=TRUE)
summary(Beauty)
```

```
>>>   CourseEvals     BeautyScore        female   lower    nonenglish tenuretrack
>>>  Min.   :1.944   Min.   :-1.53884   0:268    0:306    0:435      0:102
>>>  1st Qu.:3.326   1st Qu.:-0.74462   1:195    1:157    1: 28      1:361
>>>  Median :3.682   Median :-0.15636
>>>  Mean   :3.689   Mean   :-0.08835
>>>  3rd Qu.:4.067   3rd Qu.: 0.45725
>>>  Max.   :5.000   Max.   : 1.88167
```

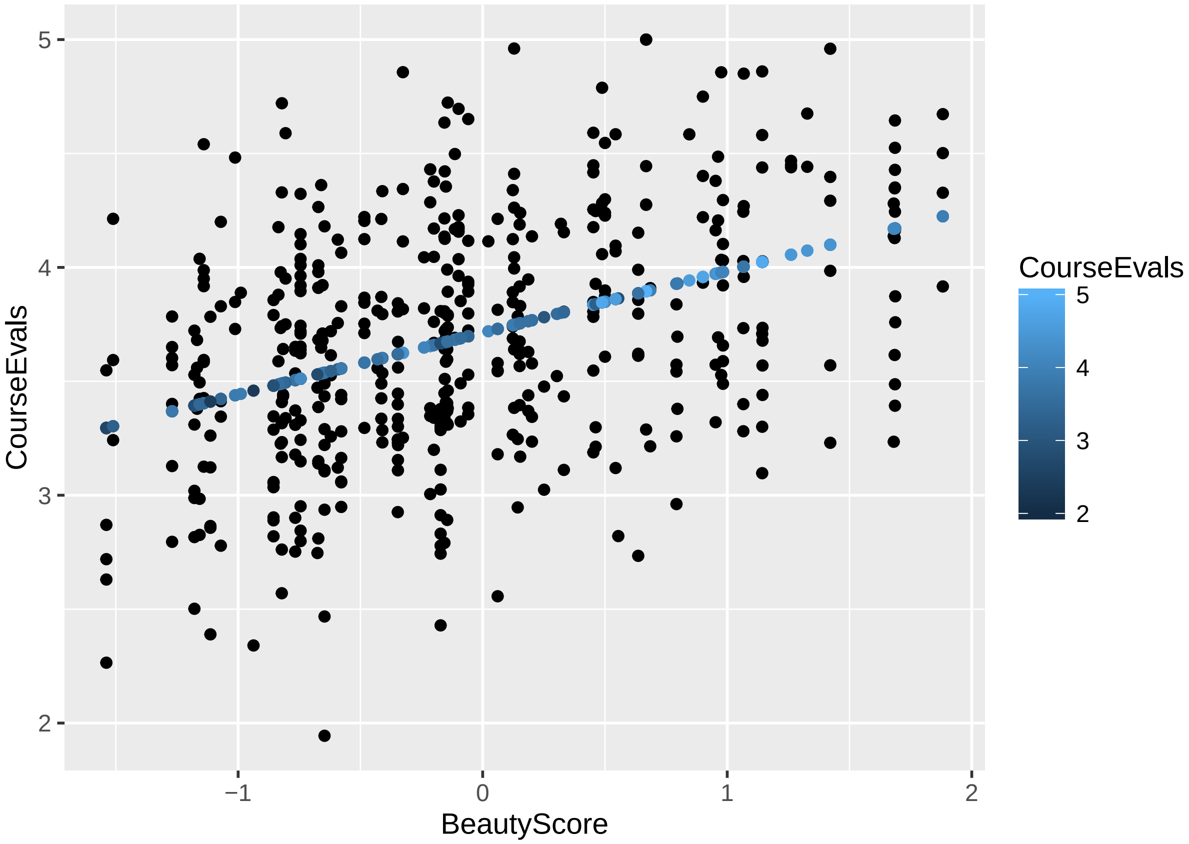
Before delving into quantitative analysis, let's make a plot and see if there is a pattern between course score and beauty score.

```r
p <- ggplot(data=Beauty, aes(x=BeautyScore, y=CourseEvals))
p + geom_point()
```



It appears that course score is positively related with beauty score, though the variation is quite big. To show this, let fit a linear model between them, and show the fitted values on the graph.

```
lm_fit <- lm(CourseEvals~BeautyScore, data=Beauty)
prediction_lm <- predict(lm_fit)
p + geom_point() + geom_point(aes(y=prediction_lm, color=CourseEvals))
```



Here the black points represent the original data, whereas blue points are predicted values using simple linear regression on `BeautyScore`. With this illustration, the correlation between the two variables is clearer. Let's print out the confidence interval of linear model coefficients to quantitatively inspect this question.

```
confint(lm_fit)
```

```
>>>                  2.5 %    97.5 %
>>> (Intercept) 3.6692053 3.7575937
>>> BeautyScore 0.2157292 0.3272273
```

Since the confidence interval of `BeautyScore` does not include zero, it is likely that their correlation is true. But is it because of other variables? Let's verify this by fitting a logistic model on all variables, and print out the confidence intervals of the coefficients.

```
logistic_fit <- glm(CourseEvals~., data=Beauty, family="gaussian")
confint(logistic_fit)
```

```
>>> Waiting for profiling to be done...
```

```
>>>                    2.5 %       97.5 %
>>> (Intercept)   3.9645800  4.166251734
>>> BeautyScore   0.2542984  0.353993240
>>> female1      -0.4118699 -0.252118878
>>> lower1       -0.4264810 -0.258620180
>>> nonenglish1  -0.4242491 -0.091915040
>>> tenuretrack1 -0.1952444 -0.003656772
```

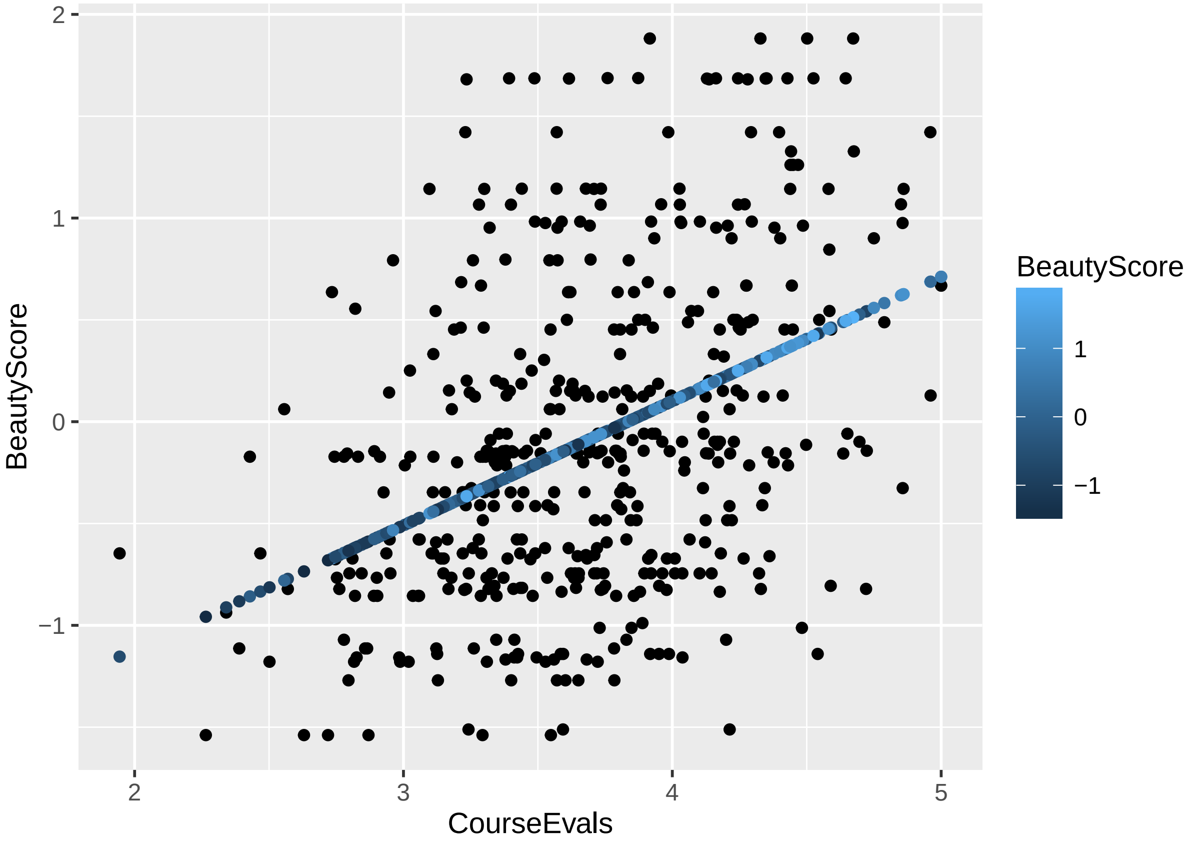
The reason of regressing on all the variables is that, if the correlation is false, then by adding more predictors, the truthfully relevant information will be drawn out of `BeautyScore`, and thereby its coefficient will drop to zero. However, the output further proves a positive correlation between course score and beauty score, which is our conclusion given the available information.

(2) How to understand Dr. Hamermesh's sentence

To understand Dr. Hamermesh's quote "Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible" we need to understand the concept of causality.

If the positive correlation is "productive", then instructors with better looks could actually provide extra benefits to the students, who then will achieve better scores and give higher ratings for the class. Another way to understand this point is, by giving better lectures the instructors have improved their image in front of the students, so the beauty score is mixed with a certain bias. To clarify this point, let's reverse the model and plot `BeautyScore` against `CourseEvals`, and it can be seen that the result also shows a significant positive correlation.

```
lm_fit_rev <- lm(BeautyScore~CourseEvals, data=Beauty)
prediction_lm_rev <- predict(lm_fit_rev)
p2 <- ggplot(data=Beauty, aes(x=CourseEvals, y=BeautyScore))
p2 + geom_point() + geom_point(aes(y=prediction_lm_rev, color=BeautyScore))
```



```
confint(lm_fit_rev)
```

```
>>>                   2.5 %     97.5 %
>>> (Intercept) -2.8077055 -1.8733814
>>> CourseEvals  0.4850899  0.7358054
```

On the other side, if the results indicates "discrimination", it could be said that the students have prejudice on the looks of the instructors and therefore adjust their ratings of the class accordingly. Or, the term "discrimination" can be perceived as an interplay between beauty score and other determinants.

But either way, with the available data it is difficult to differentiate the two interpretations, which is largely because of the lack of information to determine the cause and result direction.

Problem 2: Housing Price Structure

Load the data and make `Home`, `Nbhd`, `Brick` factor variables.

```r
col_class <- c("factor", "factor", "integer", "numeric", "factor", "integer", "integer", "numeric")
MidCity <- read.csv("MidCity.csv", header=TRUE, colClasses=col_class)
```

(1) Is there a premium for brick houses everything else being equal?

(2) Is there a premium for houses in neighborhood 3?

Since there are a couple categorical variables as said about, we first try logistic regression on the dataset to qualitatively inspect the significance of house material and neighbourhood.

```r
logistic_fit <- glm(Price~.-Home, data=MidCity)
confint(logistic_fit)
```

```
>>> Waiting for profiling to be done...

>>>                    2.5 %      97.5 %
>>> (Intercept) -15240.68903 19559.68541
>>> Nbhd2        -6258.15303  3136.99479
>>> Nbhd3        14509.20133 26852.87337
>>> Offers      -10393.61178 -6141.36486
>>> SqFt            41.75484    64.23265
>>> BrickYes     13413.45281 21181.24625
>>> Bedrooms      1114.94618  7378.64160
>>> Bathrooms     3733.96534 12032.59164
```

The coefficient confidence intervals indicate that `Brick` and `Nbhd3` have positive correlation with `Price`, so the answer to question 1 and 2 is yes.

(3) Is there an extra premium for brick houses in neighborhood 3?

To further determine the correlation of a house being made of brick and located in neighbourhood 3 to price, let's create an interaction term `Nbhd*Brick` in the previous logistic regression model.

```r
logistic_fit_2 <- glm(Price~.-Home+Brick*Nbhd, data=MidCity)
confint(logistic_fit_2)
```

```
>>> Waiting for profiling to be done...

>>>                       2.5 %      97.5 %
>>> (Intercept)    -13609.75880 21000.78156
>>> Nbhd2           -6570.06401  3934.75171
>>> Nbhd3           10243.36354 23718.23106
>>> Offers         -10475.49817 -6288.04272
>>> SqFt               42.60158    64.88903
>>> BrickYes         4092.15473 20093.95744
>>> Bedrooms         1667.93494  7886.49676
>>> Bathrooms        2222.06519 10692.50973
>>> Nbhd2:BrickYes  -7266.39867 12603.29725
>>> Nbhd3:BrickYes   1464.97598 22401.41775
```

Similar to last question, here in the confidence intervals of all variables, it is clear that `Nbhd3:BrickYes` has a significant positive correlation with price, so the answer to question is again yes.

(4) For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

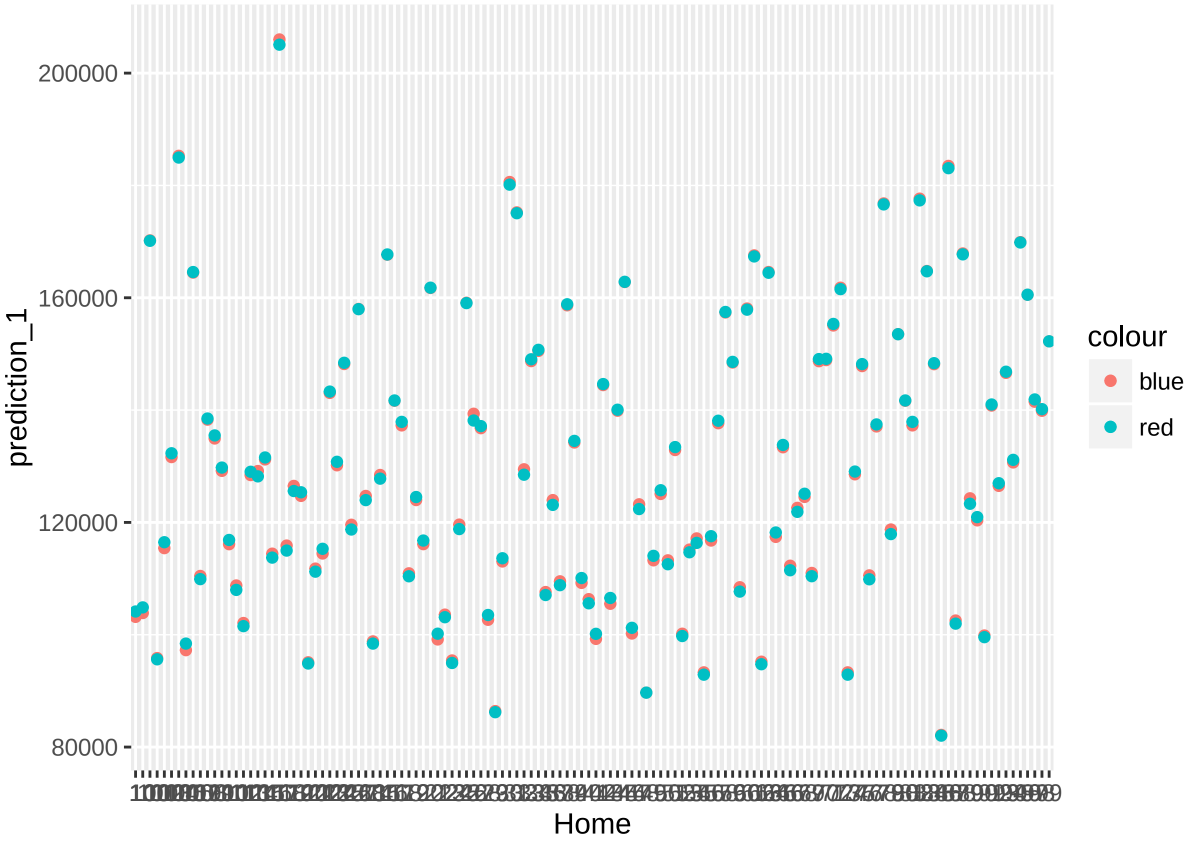
To assess the difference in terms of prediction after merging neighbourhood 1 and 2, we create a new column in such a way and compare the prediction results of the new and original datasets. Im-sample RMSE is used as a measure.

```
prediction_1 <- predict(logistic_fit)

Nbhd_mod  <- MidCity$Nbhd
levels(Nbhd_mod) <- c(levels(Nbhd_mod), "older")
Nbhd_mod[Nbhd_mod!= 3] <- factor("older")
MidCity_mod <- cbind(MidCity, Nbhd_mod)

logistic_fit_2 <- glm(Price~.-Home-Nbhd, data=MidCity_mod)
prediction_2 <- predict(logistic_fit_2)

p3 <- ggplot(data=MidCity_mod, aes(x=Home))
p3 + geom_point(aes(y=prediction_1, color="blue")) + geom_point(aes(y=prediction_2, color="red"))
```



```r
RMSE_1 <- mean((prediction_1 - MidCity_mod$Price)^2)^.5
RMSE_2 <- mean((prediction_2 - MidCity_mod$Price)^2)^.5
cat("RMSE without merging older neighbourhoods is:", RMSE_1)
cat("RMSE after merging older neighbourhoods is:", RMSE_2)
```

```
>>> RMSE without merging older neighbourhoods is: 9700.801RMSE after merging older neighbourhoods is: 9717.9
```

The predicted prices using original `Nbhd` and merged `Nbhd` are plotted on the same graph, and from the large overlap it is intuitive that the two sets of prices are highly similar. This is re-emphasized by the in-sample RMSE comparison. So for prediction purposes it is fine to combine neighbourhoods 1 and 2 into a single "older" neighbourhood.

Problem 3: What causes what??

(1) It is difficult to draw decisive conclusions on the causal relationships between crime rate and police force, because they could affect each other in both directions. For example in places where the crime rate is high, the government tends to hire more policemen, but that doesn't mean more policemen drive up crime rate.

(2) The researchers took advantage of the "terrorism alert system". The system enabled the researchers to observe the effect of number of policemen on crime rate in one direction, since when the alert escalates more policemen are marshalled in DC, which is an independent event of crime rate change.

Table 2 shows a significant negative correlation between the presence of high alert and crime rate, which means when there is a high terrorism alert, the crime rate will go down. Given the rise of policemen number on such occasions, that is equivalent to a negative correlation between policemen number and crime rate. Another point in this table is the positive correlation between midday ridership and crime rate, i.e. the more people on the street the higher the crime rate will be.

(3) Metro ridership is an important alternative factor in this case, because without controlling this factor, the possibility cannot be ruled out that the crime rate dropped actually due to fewer people are on the street. By measuring the ridership volume, the researchers are able to determine the alternative hypothesis is false, because on the high alert days the ridership volume did not diminish, while crime rate went down, even though in table 2 it was proven that the two variables are positively correlated.

(4) From the first column in table 4, it is likely the researchers built a regression model of crime rate on high alert presence in 2 specific districts (1 and others) and midday metro ridership. It is difficult to determine if this model is linear regression or not, but the coefficient levels are nonetheless indicating that in district 1 a high alert will cause a drop in crime rates, whereas that is not necessarily true in other districts.

so we can conclude that for district 1, the presence of a high alert and therefore an increased number of policemen will lead to a lower crime rate, and this correlation is statistically significant.