MA213 Basic Statistics and Probability - Lab6

Lab 6: Linear Regression

In this lab session, we will explore practical part of linear regression with a single predictor using R.

Learning Objectives

- **Compute and Interpret Correlation and R^2**: Compute and interpret correlation coefficients and R^2 values, while recognizing that correlation does not imply causation.
- **Describe and Assess Relationships Between Two Variables**: Describe the association between two numerical variables in a scatter plot in terms of direction, shape (linear or nonlinear), and strength, and assess whether linear regression is an appropriate model.
- **Fit and Interpret Linear Models Using Least Squares**: Fit the intercept and slope of a linear model to data using the least squares method, interpret the fitted values, and use the model to predict responses to new inputs.
- **Perform Inference for Regression Coefficients**: Use fit summary and parameter estimates (e.g., $\hat{\beta}_1$ and $\hat{\sigma^2}$) to perform hypothesis tests or construct confidence intervals for the slope, and interpret the results.

Pre-lab activities

- We are going to work in groups. This time, each group have 2 people. You can discuss the exercise questions and post-lab activities with your members.

```
num_ppl_each_group = 2
student_list <- read.csv("StudentList.csv")
N <- nrow(student_list)

# Sort by last name
student_list <- student_list %>% arrange(Last)
student_list$number <- 1:N

# Shuffle the student numbers randomly
shuffled <- sample(student_list$number)

# Create group
group_ids <- rep(1:ceiling(N / num_ppl_each_group), num_ppl_each_group)[1:N]

# Assign data frame
group_df <- data.frame(number = shuffled, group = group_ids)
grouped_students <- left_join(group_df, student_list, by = "number")

grouped_students <- grouped_students %>% arrange(group)

# Output
grouped_students %>% select(!number)
```

```
##    group First Last
```

```
## 1      1     S   SS
## 2      1     U   UU
## 3      2     Y   YY
## 4      2     V   VV
## 5      3     E   EE
## 6      3     G   GG
## 7      4     A   AA
## 8      4     P   PP
## 9      5    A2  AA2
## 10     5     N   NN
## 11     6     F   FF
## 12     6     Q   QQ
## 13     7     I   II
## 14     7     D   DD
## 15     8     W   WW
## 16     8     R   RR
## 17     9    C2  CC2
## 18     9     B   BB
## 19    10     M   MM
## 20    10     K   KK
## 21    11     Z   ZZ
## 22    11     J   JJ
## 23    12     L   LL
## 24    12     C   CC
## 25    13     O   OO
## 26    13     H   HH
## 27    14     X   XX
## 28    14    B2  BB2
## 29    15     T   TT
```

Lab activities

Beauty in the class

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. (The data was originally considered by Hamermesh and Parker (2005).)

Variables in the data frame :

Name	Details
minority	is the professor from a non-Caucasian ethnic minority?
age	the professor's age.
gender	a factor indicating the professor's gender.
credits	a factor indicating whether the course is a single-credit elective (e.g. scuba diving or ballroom dancing, coded "single") or an academic course (coded "more").

Name	Details
beauty	a rating of the professor's physical attractiveness, as judged by a panel of six students. (The score was averaged across all six panelists, and shifted to have a mean of zero)
eval	the professor's average teaching evaluation for courses in the sample, on a scale of 1 to 5.
division	whether the course is an upper or lower division course.
native	whether the professor is a native English speaker.
tenure	whether the professor is tenured/tenure-track, or not.
students	the number of students that participated in the evaluation.
allstudents	the number of students enrolled in the course.
prof	a unique numerical identifier for the professor being rated.

1. Let's import the data. What are your interests? What could be the response and explanatory variables?

```
library(dplyr)
library(ggplot2)

df <- read.csv("beauty.csv")

glimpse(df)
```

```
## Rows: 463
## Columns: 12
## $ minority    <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no~
## $ age         <int> 36, 59, 51, 40, 31, 62, 33, 51, 33, 47, 35, 37, 42, 49, 37~
## $ gender      <chr> "female", "male", "male", "female", "female", "male", "fem~
## $ credits     <chr> "more", "more", "more", "more", "more", "more", "more", "m~
## $ beauty      <dbl> 0.2899157, -0.7377322, -0.5719836, -0.6779634, 1.5097940, ~
## $ eval        <dbl> 4.3, 4.5, 3.7, 4.3, 4.4, 4.2, 4.0, 3.4, 4.5, 3.9, 3.1, 4.0~
## $ division    <chr> "upper", "upper", "upper", "upper", "upper", "upper", "upp~
## $ native      <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "y~
## $ tenure      <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "y~
## $ students    <int> 24, 17, 55, 40, 42, 182, 33, 25, 48, 16, 18, 30, 28, 30, 2~
## $ allstudents <int> 43, 20, 55, 46, 48, 282, 41, 41, 60, 19, 25, 34, 40, 36, 2~
## $ prof        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
```

```
head(df)
```

```
##   minority age gender credits     beauty eval division native tenure students
## 1      yes  36 female    more  0.2899157  4.3    upper    yes    yes       24
## 2       no  59   male    more -0.7377322  4.5    upper    yes    yes       17
## 3       no  51   male    more -0.5719836  3.7    upper    yes    yes       55
## 4       no  40 female    more -0.6779634  4.3    upper    yes    yes       40
## 5       no  31 female    more  1.5097940  4.4    upper    yes    yes       42
## 6       no  62   male    more  0.5885687  4.2    upper    yes    yes      182
##   allstudents prof
## 1          43    1
## 2          20    2
```

```
## 3          55     3
## 4          46     4
## 5          48     5
## 6         282     6
```

Here are some statistics of the `beauty` and `eval`.

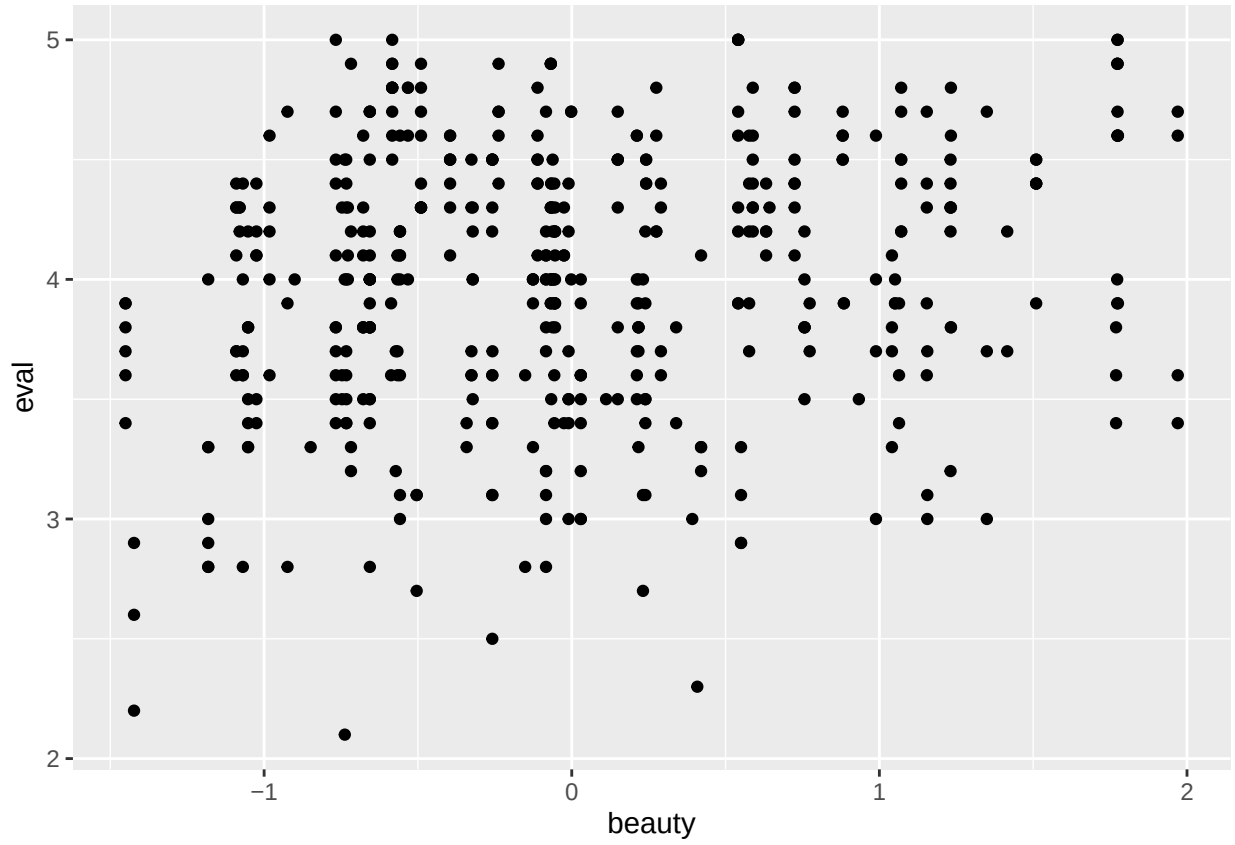
Table 2: Descriptive Statistics of beauty and eval

	beauty	eval
Min	-1.450	2.100
Median	-0.068	4.000
Mean	0.000	3.998
Max	1.970	5.000
Standard Deviation	0.788	0.554

2. Let's first see how response variable and explanatory variable are related.

We can look at scatter plot to see how `beauty` and `eval` (course ratings) are related.

```
ggplot(data=df, mapping = aes(x=beauty, y=eval)) + geom_point()
```




```
#
#
#
```

Let's also check correlation between `beauty` and `eval`.

Correlation

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

where s_x is the standard deviation of explanatory variable x and s_y is the standard deviation of response variable.

Exercise Q1 let's obtain the correlation between `beauty` and `eval`, R, using R.

```
#
#
#
```

Based on the correlation, what does it tell you?

3. Let's obtain least squares line where the model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

The estimates are

$$b_1 = \hat{\beta}_1 = (\frac{s_y}{s_x})R,$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x}$$

Exercise Q2 Let's obtain the least squares line. Obtain `b0` and `b1`.

```
#
#
#
```

4. We can check some of the assumptions (conditions) for the least square lines.

Exercise Q3 Obtain residuals and show plot of residuals against explanatory variable

```
#
#
#
```

We can check how residuals are distributed.

How do you check if residuals are normally distributed?

Exercise Q4 Show histogram plot of residuals

```
#
#
#
```

We can also check QQ plot, quantile-quantile plot. It is a graphical tool to assess if data comes from theoretical distribution (normal distribution in this case). We are expected to see If the points are on the diagonal line.

```
# qqnorm(Resid)
# qqline(Resid) #this will draw 45 degree line for reference purposes.
```

Exercise Q5 Let's calculate R^2 to describe the strength of a fit

```
#
#
#
```

What does it mean? from R^2 value you calculated?

Now, let us walk through the output from `lm()` function in R.

```
#
# model = lm(Y~X)
# summary(model)
```

Lab activities 2

Exercise Q6 Choose one dataset and perform data analysis using simple linear regression. Your group can choose a data from

Survey of Duke students on GPA, studying, and more

US Crime Rates

Normal Body Temperature

State-level data

SAT and GPA data

Pierce County House Sales Data for 2020

0. Import the data

```
#
#
#
```

1. Check the relationship between response variable and explanatory variable.

```
#
#
#
```

2. Obtain the least square line and get the residuals.

```
#
#
#
```

3. Show some plots to check model assumption.

```
#
#
#
```

Post-lab activities

The Toluca Company manufactures refrigeration equipments as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. The production of this part involves setting up the production process (which must be done no matter what is the lot size) and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labour hours required to produce the lot. To determine this relationship, data on lot size and work hours for 25 recent production runs were utilized. The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted.

Please name your submission as `lab6.R`

Data here is given in `toluca.csv` .

1. Calculate the correlation coefficient and save it to object `r2.`
2. Obtain standard deviations of `X` and `Y`, `sx` and `sy`, respectively.
3. Obtain the sum of `X` and `Y`, `sumx` and `sumy`, respectively.
4. Obtain `b0` estimate for β_0.
5. Obtain the slope estimate `b1` .
6. Obtain `R2` .
7. What is the 5th residual value? Assign it to object `resid5` .

Hamermesh, Daniel S, and Amy Parker. 2005. "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity." *Economics of Education Review* 24 (4): 369–76.