Lecture 1: Course introduction and review

Donghyeon Yu

Materials in this lecture notes partially come from two lecture notes written by Jean Kyung Kim (Inha Univ.) and Pratheepa Jeganathan (Stanford Univ.).

Course introduction and review

Outline

- What is a regression model?
- Descriptive statistics – numerical
- Descriptive statistics – graphical
- Inference about a population mean
- Difference between two population means

What is course about?

- It is a course on Regression analysis.
- In class, we learn about basic concepts and theories in regression analysis.
- Taking Regression analysis Lab. is recommended.
  (In practic, we need to use a statistical software to handle data sets and fit the regression models.)
- Course notes will be R markdown.
- We will start out with a review of introductory statistics to see R in action.

What is a regression model?

A regression model is a model of the relationships between some *covariates (predictors)* and an *outcome*.

Specifically, regression is a model of the *average* outcome *given or having fixed* the covariates.

Example (Heights of mothers and daughters)

- We will consider the heights of mothers and daughters collected by Karl Pearson in the late 19th century in R package `alr4`.

```
install.packages("alr4")
```

```
library(alr4)
head(Heights)
```

```
##   mheight dheight
## 1    59.7    55.1
## 2    58.2    56.5
## 3    60.6    56.0
## 4    60.7    56.8
## 5    61.8    56.0
## 6    55.5    57.9
```

- One of our goals is to understand height of the daughter, D, knowing the height of the mother, M.
- A mathematical model might look like

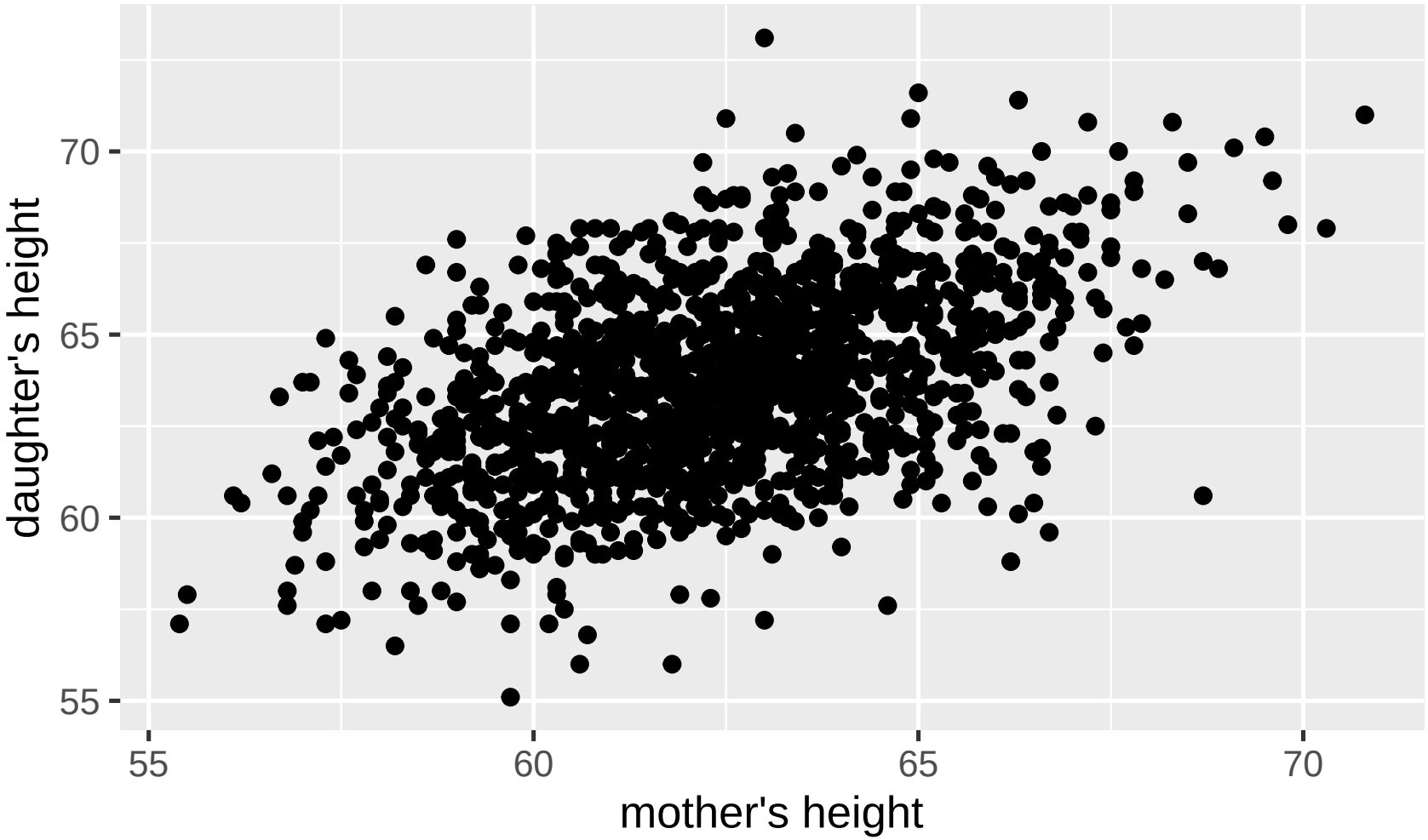
$$D = f(M) + \varepsilon,$$

  where f gives the average height of the daughter of a mother of height M and ε is *error*: not *every* daughter has the same height.
- A statistical question: is there *any* relationship between covariates and outcomes – is f just a constant?

- Let's create a plot of the heights of the mother/daughter pairs.

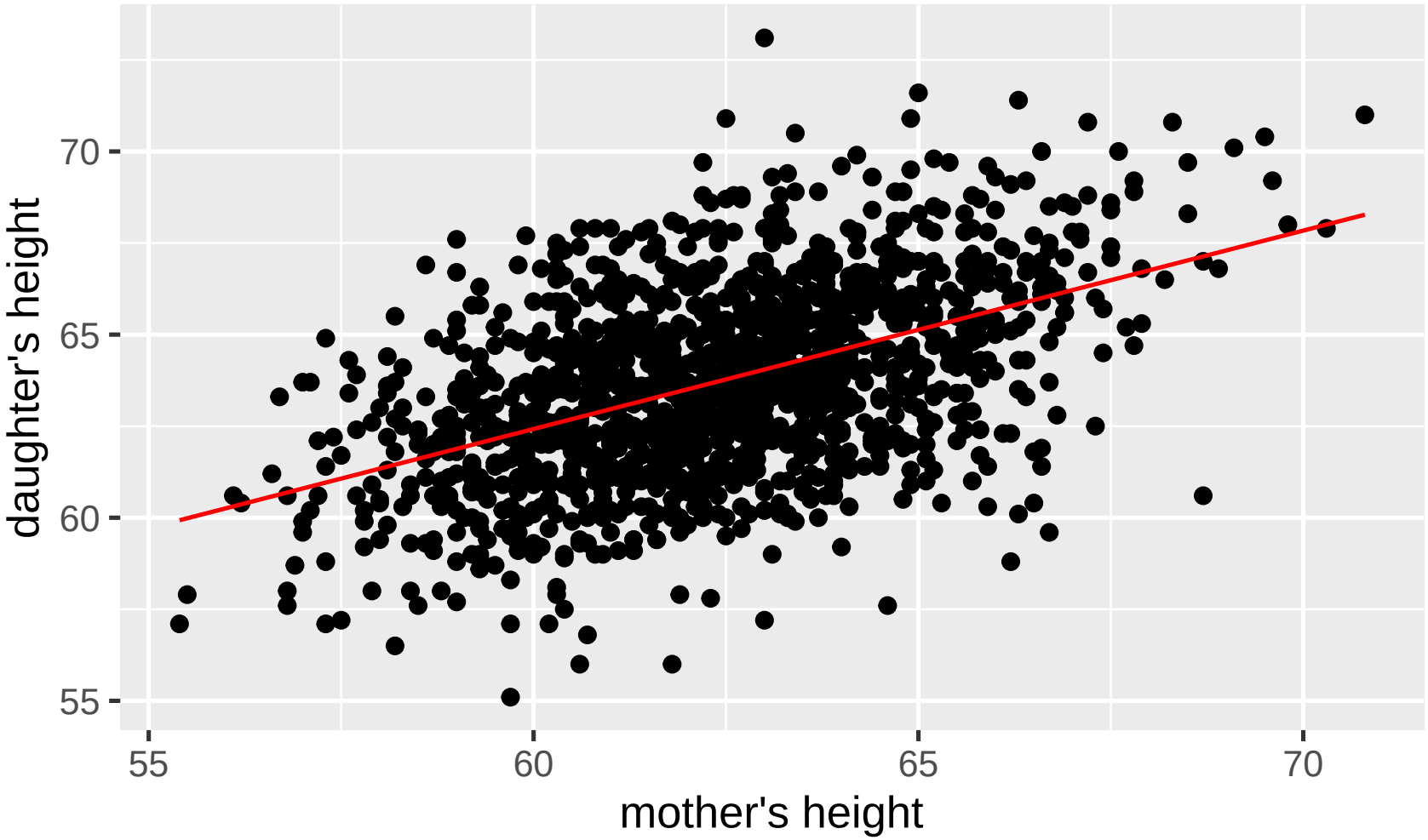
```
install.packages("ggplot2")
```

```
library(ggplot2)
p = ggplot(data = Heights) +
  geom_point(aes(x = mheight, y = dheight)) +
  xlab("mother's height") +
  ylab("daughter's height")
```


daughter's height
70
65
60
55
55
60
65
70
mother's height

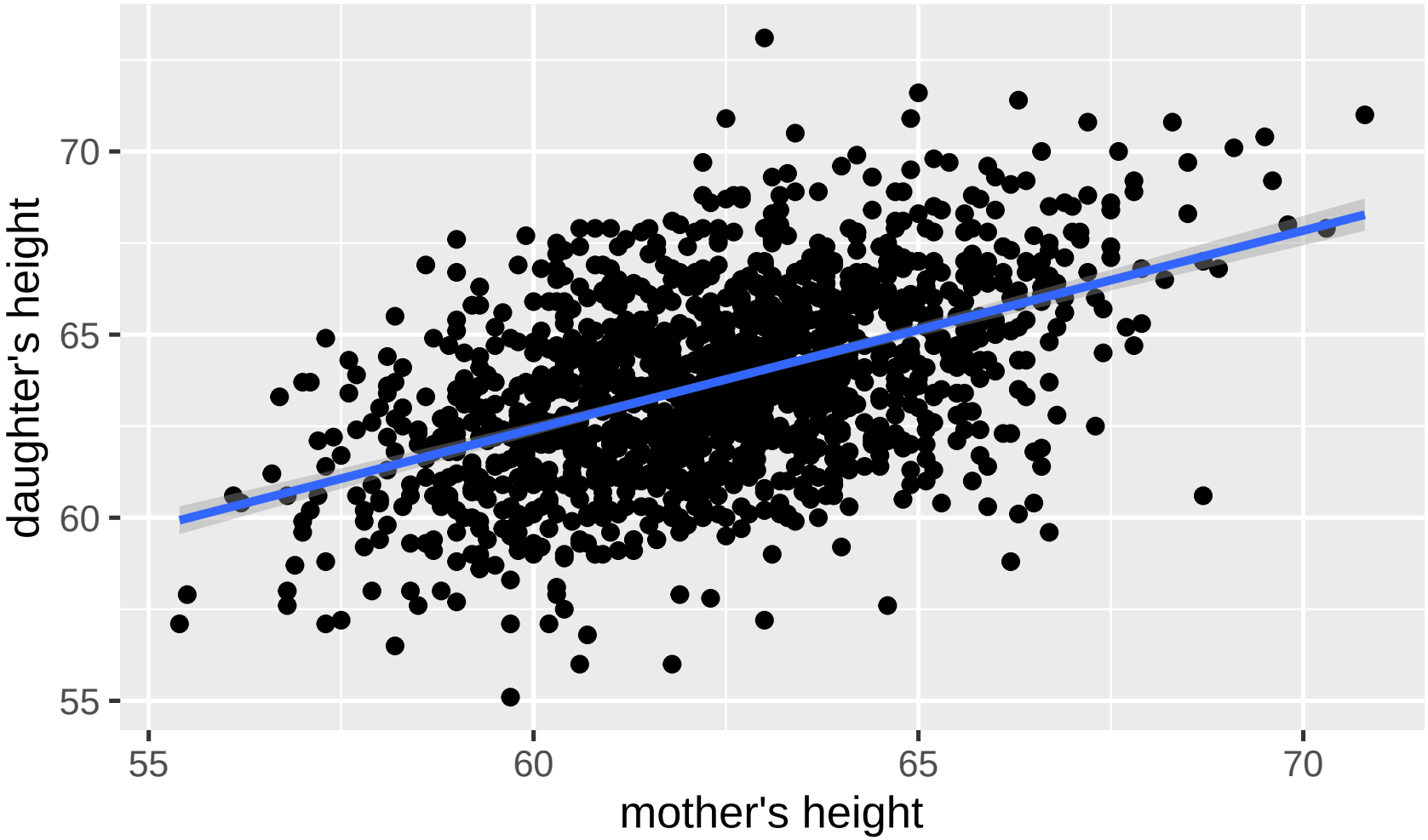
- In the first part of this course we'll talk about fitting a line to this data. Let's do that and remake the plot, including this "best fitting line".

```
fit.lm  = lm(dheight ~ mheight, data = Heights)
df = data.frame(mheight = Heights$mheight,
  dheight.fit = fitted(fit.lm))
p2 = ggplot(data = Heights) +
  geom_point(aes(x = mheight,
    y = dheight)) +
  xlab("mother's height") +
  ylab("daughter's height") +
  geom_line(data = df, aes(x = mheight,
    y = dheight.fit), color = "red")
```


daughter's height
70
65
60
55
55
60
65
70
mother's height

- We can directly call `lm` as another layer.

```
p3 = ggplot(data = Heights, aes(x = mheight,
    y = dheight)) +
  geom_point() +
  xlab("mother's height") +
  ylab("daughter's height") +
  geom_smooth(method='lm', formula = y~x)
```


daughter's height
70
65
60
55
55
60
65
70
mother's height

Linear regression model

- How do we find this line? With a model.
- We might model the data as

$$D = \beta_0 + \beta_1 M + \varepsilon.$$

- This model is *linear* in (β_0, β_1), the intercept and the coefficient of M (the mother's height), it is a *simple linear regression model*.
- Another model:

$$D = \beta_0 + \beta_1 M + \beta_2 M^2 + \beta_3 F + \varepsilon,$$

  where F is the height of the daughter's father.
- Also linear in $(\beta_0, \beta_1, \beta_2, \beta_3)$, the coefficients of $1, M, M^2, F$.
- Which model is better? We will need a tool to compare models... more to come later.

A more complex model

- Our example here was rather simple: we only had one predictor variable.
- predictor variables are sometimes called *features* or *covariates* or *independent variables*.
- In practice, we often have many more than one predictor.

Terminologies

- Regression variables
  Y : response, dependent
  X or $\mathbf{X}$: explanatory, predictor, independent, covariate, feature, regressor, factor, carrier and etc.
- Regression model

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon,$$

where ε is called error or discrepancy in the approximate and p is a number of regressors.

  - Special case: linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Applications of Regression analysis

- Agricultural Sciences
  - DHI(Diary Herd Improvement) cooperative: produce milk
  - Interest: How to develop a suitable model to predict current milk production
  - Data: samples are taken once a month
  - Variable: Table 1.1 on p.4
- Industrial and Labor Relation
  - Wagner Act and Taft-Hartley Amendments (1947)
  - Interest: Effects of laws
  - Variable: Table 1.2 on p.5
- Government
  - Interest: Building a prediction model for domestic immigration
  - Variable: Table 1.3 on p.6

- History
  - Interest: How to estimate the age of historical objects based on some age-related characteristics
  - Variable: Table 1.4 on p.6
- Environmental Science
  - Interest: How land use around a river basin contributes to the water pollution
  - Data: Table 1.5 on p.7
- Industrial Production
  - Interest: Estimating the polishing time for new products
  - Variables: Bowlm Casserolem Dish, Tray, Plate Diam, Time, Price

Steps in regression analysis

- Statement of the Problem
- Selection of Potentially relevant variables
- Data collection
- Model specification
- Choice of fitting method
- Model fitting
- Model validation and criticism

References for this lecture

- Based on the lecture notes of Pratheepa Jeganathan
- Based on the lecture notes of Jonathan Taylor .