Lecture 1: Course introduction and review

Pratheepa Jeganathan

09/23/2019

Course logistics

Course description

- Statistical tools for modern data analysis
  - regression and prediction.
  - elements of the analysis of variance.
  - bootstrap and cross-validation.
- Emphasis is on conceptual rather than theoretical understanding.
- Student assignments require use of the software package R .

Expected outcomes

By the end of the course, students should be able to:

- Enter tabular data using R .
- Plot data using R , to help in exploratory data analysis.
- Formulate regression models for the data, while understanding some of the limitations and assumptions implicit in using these models.
- Fit models using R and interpret the output.
- Test for associations in a given model.
- Use diagnostic plots and tests to assess the adequacy of a particular model.

Expected outcomes (Cont.)

- Find confidence intervals for the effects of different explanatory variables in the model.
- Use some basic model selection procedures, as found in R , to find a *best* model in a class of models.
- Fit simple ANOVA models in R, treating them as special cases of multiple regression models.
- Fit simple logistic and Poisson regression models.

General information

- Course website: Canvas @ Stanford University
- Homework will be assigned on Fridays (submit answers to gradescop)
- Midterm and finals: in-class examination.
- Instructor's office hours: Wednesday 2:30 PM - 4:30 PM in 105 Sequoia or by an email appointment.

TA's Office hours

- Benjamin Seiler
  - Zoom office hours for SCPD students: Thursday 4:30 PM - 6:30 PM.
  - Zoom meeting ID: https://stanford.zoom.us/s/793447924 .
  - All contacts about SCPD .
- Jayoon Jang
  - Office hours: Thursday 1:00 PM - 3:00 PM
  - Location: Sequoia 207 (Bowker)
- Samir Anwar Khan
  - Office hours: Tuesday 1:00 PM - 3:00 PM
  - Location: Sequoia 207 (Bowker)

Email list

The course has an email list that reaches all TAs as well as the instructor: stats191-aut1920-staff@lists.stanford.edu

As a general rule, you should send course related questions to this email list.

Questions can also be posted on Canvas Discussion .

Textbook

- Required:
  - **(CH)** Regression Analysis by Example .
    - Authors: Samprit Chatterjee, Ali S. Hadi
    - Edition: 5^{th} Edition
    - Print ISBN:978-0-470-90584-05

Textbook (Cont.)

- Comprehensive coverage of regression analysis, the assumptions underlying the methods, and examples.
- Bibliography in detail for theory.
- Recommended readings:
  - **(DH)**: Davison and Hinkley (1997). Bootstrap Method and Their Application.
  - **JSE**: Journal of Statistics Education (when I typed "regression" in the search box)
    - Find articles before 2016 in archive

Grading

The final letter grade for this course will be determined by each method of assessment weighted as follows:

- 7 weekly homework assignments (55%)
- Midterm examination (15%, Wednesday, 10/23/2019)
- Final examination (30%, according to Stanford calendar: Wednesday, 12/11/2019 @ 3:30 PM, location TBD)
- Pop quizzes: (5% Bonus points).

Homework assignments (Template)

- See the template in Canvas/Files/Templates .
  - To do the **Quiz practice**, you need to download `homework_template.Rmd`, `header.tex`, and `AppliedStat.bib`.
  - Download homework_template.Rmd
  - Download header.tex
  - Download AppliedStat.bib
- See the following link for a further outline of using R markdown for reporting .
- Write the solution for each question on a new page (use `\newpage`).
- Prepare your completed homework assignment in PDF format and submit a copy to gradescope.

- Each question in the homework assignment will be graded as follows: *scale* $\in \{0, 1, 2\}$
  - 2: submitted on time and more or less correct answer
  - 1: submitted on time and partially correct answer
  - 0: submitted with a completely incorrect answer or late submission (any day after the due date for more than one homework assignment).
- Each student can hand in only one homework late (within three days after the deadline).

Midterm examination

- In-class examination.
- 4-5 multiple-choice questions and 1-2 comprehension questions (practice exam will be posted).
- 2 single-sided pages of notes and a calculator are allowed.

Final examination

- In-class examination.
- 4-5 comprehension questions with sub parts (practice exam will be posted).
- 4 single-sided pages of notes and a calculator are allowed.

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 applied statistics.
- Hands-on: we use R , an open-source statistics software environment.
- Course notes will be R markdown.
- We will start out with a review of introductory statistics to see `R` in action.
- Main topic is *(linear) regression models*: these are the *bread and butter* of applied statistics.

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

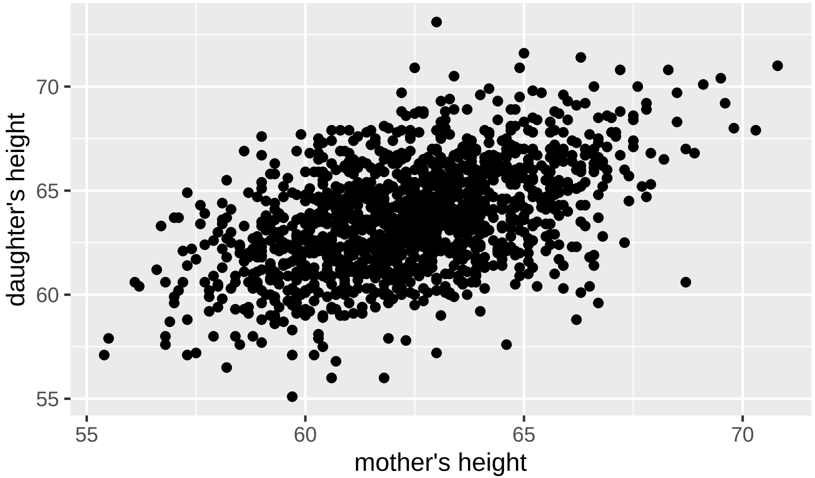
$$D = f(M) + \varepsilon,$$

  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")
```

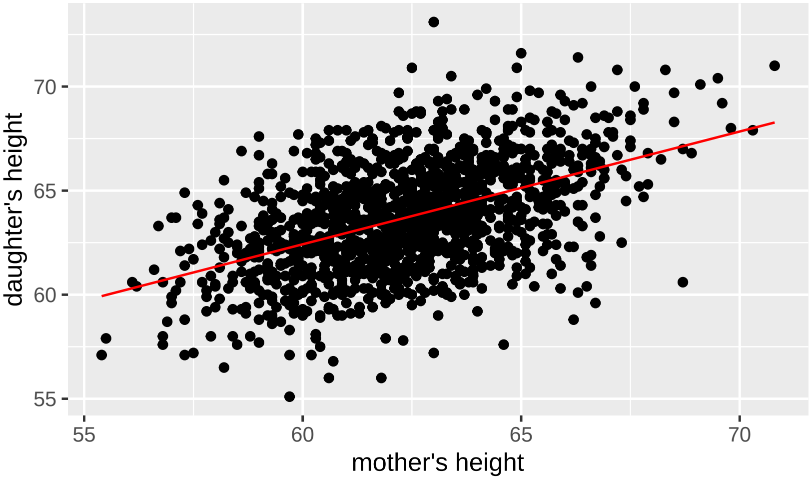
```
library(ggplot2)
p = ggplot(data = Heights) +
  geom_point(aes(x = mheight, y = dheight)) +
  xlab("mother's height") +
  ylab("daughter's height")
```



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

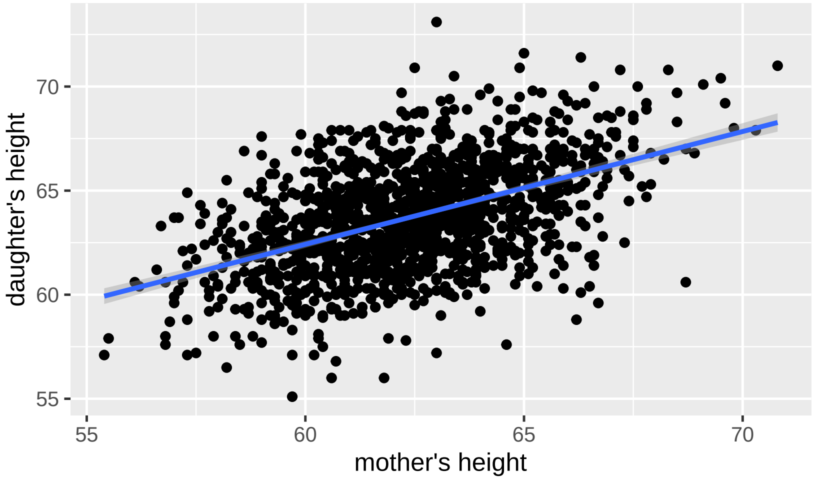
- In the first part of this course we'll talk about fitting a line to this data. Let's do that and remake the plot, including this "best fitting line".

```
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")
```


70
65
60
55
daughter's height
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.

References for this lecture

- Syllabus 191 (Autumn 2019-2020).
- Based on the lecture notes of Jonathan Taylor .