Methods of Applied Statistics I

STA2101H F LEC9101

Week 1

September 10 2020

Today

Start Recording

1. Course introduction: technical issues, people, course details, evaluation, syllabus

   Timer

2. Upcoming events of interest
3. Scholar strike
4. Review of linear regression
5. In the news: wildfire
6. Computing: RStudio, RMarkdown

Today

Royal Statistical Society
@RoyalStatSoc

Happy 100th birthday to one of our most longstanding RSS honorary fellows and formidable name in modern statistics, CR Rao! #RSSFellow rss.org.uk/news-publicati …

Technical Issues

- If **you** are having technical difficulties
    - If possible, send **me** a message in chat
    - Try leaving the class and re-joining
    - Try switching to Chrome if you are using something else
    - Don’t panic, the lecture is being recorded and both the recording and the slides will be posted

- If **Prof** is having technical difficulties
    - Check the chat to see if there’s any information there
    - If I’ve disappeared completely, give me 15 minutes before closing the call
    - Look for an announcement on Quercus
    - Don’t panic, Prof, you’ll figure it out

Course Introductions

- about me $\longrightarrow$

- TA: Sangook Kim

- Please turn on your camera to introduce yourself

roll call

- Tell us your program, current city, where you got your previous degree

STA 2101F: Methods of Applied Statistics I
Thursday, 12-3 pm Eastern September 10 – December 3 2020

Updated September 3

From the calendar:

This course will focus on principles and methods of applied statistical science. It is designed for MSc and PhD students in Statistics, and is required for the Applied Paper of the PhD comprehensive exams. The topics covered include: planning of studies, review of linear models, analysis of random and mixed effects models, model building and model selection, theory and methods for generalized linear models, and an introduction to nonparametric regression. Additional topics will be introduced as needed in the context of case studies in data analysis.
Prerequisites: ECO374H1/ECO375H1/STA302H1 (regression); STA305H1 (design of studies)

Course Delivery:

STA 2101F: Methods of Applied Statistics I
Thursday, 12-3 pm Eastern September 10 – December 3 2020

Updated September 3

From the calendar:

This course will focus on principles and methods of applied statistical science. It is designed for MSc and PhD students in Statistics, and is required for the Applied Paper of the PhD comprehensive exams. The topics covered include: planning of studies, review of linear models, analysis of random and mixed effects models, model building and model selection, theory and methods for generalized linear models, and an introduction to nonparametric regression. Additional topics will be introduced as needed in the context of case studies in data analysis.
Prerequisites: ECO374H1/ECO375H1/STA302H1 (regression); STA305H1 (design of studies)

Course Delivery:

The class will be delivered at the scheduled time (Thursdays, 12-3 pm Toronto

Course Description

- Course Delivery

  Piazza, Notifications

- Grading
- Academic Integrity
- Computing
- References

  Modules

- Contact

Use **Piazza** for course questions; email for personal questions

Today

Check Recording

1. Course introduction: technical issues, people, course details, evaluation, syllabus

   Timer

2. Upcoming events of interest
3. Scholar strike
4. Review of linear regression
5. In the news: wildfire
6. Computing: RStudio, RMarkdown

Upcoming events 1

Daniela Witten

- Data Science Applied Research and Education Seminar — Sep 14 3 pm EDT
- Register here

Upcoming events 2

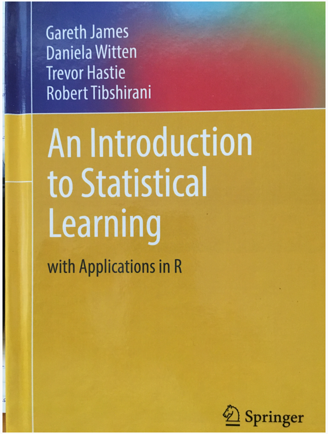
Thanks to Hidaya Ismail for the brilliant maple leaf and dinosaur hex stickers.

- Toronto Data Workshop — Thursdays 4 pm EDT
- Next one today!!

Upcoming events 3

- Methods and Theory Seminar Series Thursdays 3.30 pm EDT
- First one TBD

Upcoming events 1



`statlearning.com`

Scholar strike

Daniela Witten @daniela_witten · Jun 4

Replying to @daniela_witten

There is a very prestigious award in stats called the Fisher lecture. I am on the award committee, and I'm embarrassed to say that until a few days ago, I hadn't thought this through: like many people, when I thought Fisher, I thought "Fisher information", not eugenics. 6/

13 56 340

Daniela Witten @daniela_witten · Jun 4

But a few days ago I realized: yikes!! It is not good that a major award is named after this guy!! And so I thought — as a full professor from a good department who is a member of the award committee, I can get the name of this lecture changed! 7/

9 29 493

SSC Biostatistics Section

- Case Study Competition

- “take any COVID-19 data publicly available and create an analytic tool or develop a model that can be useful for decision makers”

- “Each team should be composed of at most four students and a professor”
- “Three awards

  will be presented for the top 3 teams.”

- “Each team is expected to provide a 10-pages report”

- October 1 2020 5 pm EDT: deadline for registration
- December 31 2020 5 pm EDT: deadline for report submission

Today

Start Recording

1. Course introduction: technical issues, people, course details, evaluation, syllabus

   Timer

2. Upcoming events of interest
3. Scholar strike
4. Review of linear regression
5. In the news: wildfire
6. Computing: RStudio, RMarkdown

- Model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

- Equivalently:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

- Standard Assumptions
    - y_i independent equivalently ϵ_i independent — y is often called response
    - $\mathbb{E}(\epsilon_i) = 0$ — why?
    - $\mathrm{var}(\epsilon_i) = \sigma^2$ — constant
    - x_i known, β to be estimated — x_i often called explanatory variables
- More concisely:

$$\mathbb{E}(Y \mid X) = X\beta, \quad \mathrm{var}(Y \mid X) = \sigma^2 I$$

I ??

Nice big equation:

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Or, if you prefer:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad \epsilon_i \text{ i.i.d. mean 0, constant variance, } i = 1, \dots, n$$

Or, if you prefer:

$$\mathbb{E}(y_i \mid x_i) = x_i^{\mathrm{T}}\beta, \qquad \mathrm{var}(y_i \mid x_i) = \sigma^2, \quad i = 1, \dots, n$$

y_i independent

- often not completely clear: X might be fixed by design, or measured on each individual

  e.g.?

- If measured, then should we consider its distribution? E.g. should our model be $(y_i, x_i^{\mathrm{T}}) \sim$??

  some $(p+1)$-dimensional distribution

- Almost always in regression settings we condition on X, as on previous slide

  ancillary statistic

- often not emphasized: interpretation of β_j
  - version 1: effect on the expected response of a unit change in jth explanatory variable, all other variables held fixed
  - version 2:

$$\beta_j = \frac{\partial \mathbb{E}(y_i \mid x_{ij})}{\partial x_{ij}} \qquad \frac{\partial \mathbb{E}(y \mid x_j)}{\partial x_j}$$

notation ambiguous, see CD §6.5.2

Aside: ancillarity

added after the lecture

- in a statistical model (i.e. likelihood function) that factorizes, in a way that separates the parameters, there are strong theoretical (and often practical) reasons for using only the relevant factor for inference
- in our example, even if (y, X) have a $(p+1)$-dimensional distribution (maybe even jointly multivariate normal), it would typically be the case that β, which by definition is $\mathbb{E}(y \mid X)$, is not a parameter of the distribution of X.
- So we would have something like

$$f(y, X; \beta, \sigma^2, \theta) = f_1(y \mid X; \beta, \sigma^2) f_2(X \mid \theta)$$

- and we base our inference for β and σ^2 (which are the parameters of interest, by assumption) on $f_1(y \mid X; \beta, \sigma^2)$
- in technical terms X is ancillary for (β, σ^2)
- If we did decide to include the variability in X as part of our analysis, our inference about β would be much less precise, and needlessly so, because we are worrying about explanatory variable values that were not seen in our data set

Least squares estimation

- Definition

$$\hat{\beta}_{LS} := \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}}\beta)^2$$

- Equivalently,

$$\hat{\beta}_{LS} := \min_{\beta} (y - X\beta)^{\mathrm{T}}(y - X\beta)$$

- Equivalently,

$$\hat{\beta}_{LS} := \min_{\beta} ||y - X\beta||_2^2$$

$L2$ distance

- Equivalently, $\hat{\beta}$ is the solution of the score equation

$$X^{\mathrm{T}}(y - X\beta) = 0$$

?how?

- Solution

$$\hat{\beta}_{LS} = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}y)$$

check dimensions

… least squares estimation

- Solution

$$\hat{\beta}_{LS} = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}y)$$

check dimensions

- Expected value

$$\mathbb{E}(\hat{\beta}_{LS}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbb{E}(y) = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}X)\beta = \beta$$

why?

- Least squares estimates are unbiased
- Variance

really variance-covariance matrix

$$\mathrm{var}(\hat{\beta}_{LS}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{var}(y)X(X^{\mathrm{T}}X)^{-1} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sigma^2 IX(X^{\mathrm{T}}X)^{-1} = \sigma^2(X^{\mathrm{T}}X)^{-1}$$

ASIDE: here and following all assume X is fixed

What about the normal distribution?

- If we further assume $\epsilon_i \sim N(0, \sigma^2)$ (and independent across i), then
- $y \mid X \sim N(X\beta, \sigma^2 I)$, and
- likelihood function is

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

- log-likelihood function is

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)$$

constants in params don't matter

- maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \hat{\beta}_{LS}$$

… what about the normal distribution?

- maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \hat{\beta}_{LS}$$

- distribution of $\hat{\beta}$ is normal

why?

$$\hat{\beta} \sim N_{p}(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1})$$

- distribution of $\hat{\beta}_j$ is

$$N(\beta_j, \sigma^2(X^{\mathrm{T}}X)^{-1}_{jj}), \quad j = 1, \ldots, p$$

- maximum likelihood estimate of σ^2 is $\frac{1}{n}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$
- but we use

$$\tilde{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

Inference

- If you really like likelihood theory, the expected Fisher information is SM §8.2.3

$$\mathcal{I}(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2} X^{\mathrm{T}} X & 0 \\ 0 & \frac{1}{2} n \sigma^{-4} \end{pmatrix}$$

$\mathcal{I}^{-1}$ gives (asymptotic) variance of MLE

- but just using previous slide we have

$$\frac{\hat{\beta}_j - \beta_j}{\sigma [\{(X^{\mathrm{T}} X)^{-1}\}_{jj}]^{1/2}} \sim N(0, 1)$$

- and

$$\frac{\hat{\beta}_j - \beta_j}{\tilde{\sigma} [\{(X^{\mathrm{T}} X)^{-1}\}_{jj}]^{1/2}} \sim t_{n-p}$$

Example


```
install.packages("faraway")
library(faraway)
data(prostate)
head(prostate)


model1 <- lm(lpsa ~ ., data = prostate)

summary(model1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
```

```
summary(model1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thread

Women in Statistics and Data Science
@WomenInStat

Today, we're going to play a game I'm calling "IT'S JUST A LINEAR MODEL" (IJALM).

It works like this: I name a model for a quantitative response Y, and then you guess whether or not IJALM.

1/

5:59 PM · Jul 23, 2020 · TweetDeck

Many special cases $\mathbb{E}(Y \mid X) = X\beta$

- $$y = X\beta + \epsilon$$

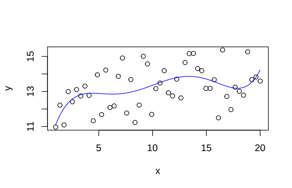
linear in β

- $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$$

1st column of X?

- $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 \epsilon_i$$

- $$y_i = \beta_0 \pm \beta_1 + \epsilon_i$$



... many special cases

- $y_i = \beta_0 + \beta_1 \sin(x_i) + \beta_2 \cos(x_i) + \epsilon_i$
- $y_i = \gamma_0 x_{1i}^{\gamma_1} x_{2i}^{\gamma_2} \eta_i, \quad \eta_i \sim$ positive r.v. — SM Example 8.5
- $y_i = \varphi_0 + \sum_{k=1}^{K} \varphi_k s_k(x_i) + \epsilon_i$ — Smoothing splines, e.g.
- mean function (expected value) $\quad \mathbb{E}(y) =$ linear in β
- measured with additive error $\quad y = \mathbb{E}(y) + \epsilon$