36-617: Applied Linear Models

Introduction

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Classes Will Be Recorded

- I plan to record and share these classes with Zoom with you for later viewing/review.
  - They will be available at cmu.canvas.edu
  - I'm sure the first few lectures will be pretty choppy, since I am just getting used to the tech in this classroom!

- Watching recorded lectures does not replace being in class.
  - There is a "participation" component in your grade
  - If at any time you cannot be physically in class (travel problems, illness, etc.) I can give you a zoom link to join class remotely.

Outline

- Introduction
- Course Schedule & Syllabus Stuff
- Valid vs Useful
- Quick Review of Univariate Regression
- R!
- Reading:
  - Read/skim Ch's 1-3 of Sheather
  - Look at Ch 5 of Sheather more closely
  - Supplemental:
    - G&H CH 3
    - ISLR 3.1,3.2, 3.3.1,3.3.2

Introduction – About Us

- Instructor
  Brian Junker
  brian@stat.cmu.edu

- TA
  Lorenzo Tomaselli
  ltomasel@stat.cmu.edu

- Office hours:
  - **MW Noon-1pm**
  - **132E Baker Hall**
  - ...or by appt
    (in person or Zoom)
- Office hours:
  - **TBA**
  - ...or by appt

Introduction – About The Course

- *Technical material:* The machinery of linear regression and its generalizations
- *Disposition:* When is a model adequate?
- *Translation:* ABA^{-1}:
  - A: Translate from real world to quantitative question
  - B: Answer quantitative question using Statistics
  - A^{-1}: Translate back to real word

  Lather, rinse, repeat…
- *Communication:*
  - IMRaD -> IDMRaD
  - Clear sentences, paragraphs, sections.

Introduction – Course Materials

- ***Technical material:***
  - Sheather (2009). *A Modern Approach to Regression with R.* NY: Springer *
  - James et al. (2013). *Introduction to Statistical Learning with R.* NY: Springer *
- ***Supplementary texts:***
  - Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* NY: Cambridge Univ Press.
  - Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* NY: Springer *
- ***Software:***
  - R, Rstudio, Stan (more on the next slide)
  - LaTeX (If you are new to TeX/LaTeX, try overleaf.com–free to CMU)
  - *Help: I will post some links, but get used to googling!*

* Available for free from link.springer.com if you connect from campus or using CMU's VPN service.

Introduction - Computing

- We'll mostly be working in R, RStudio and extensions of R (primarily `stan`).
  - R: https://cran.r-project.org/
  - RStudio: https://www.rstudio.com/
    - Rmarkdown: (just try it in RStudio, use google for help)
  - Stan: https://mc-stan.org/users/interfaces/rstan
- In class I'll use "raw R" more than RStudio, but you will find RStudio and Rmarkdown very convenient for hw, etc.
- It is possible to do some of the work in Python, but the lectures, technical work and project are strongly geared to facilities in R.

Introduction – Online Resources

- Online resources:
  - I will record lectures and make them available on Canvas.
  - **Canvas** (canvas.cmu.edu)
    - All course materials
    - Your grades
    - Monday quizzes
    - Submit and peer review project papers
  - **Gradescope** (within Canvas)
    - Submit weekly homework
  - **Piazza** (within Canvas)
    - Great for asking (and answering) questions outside of class
    - The TA and I will also monitor Piazza

General schedule for the semester

Week	Dates	Tentative Topics	Tentative Sources
	Aug 22–26	**Getting Ready**	Sheather, Ch 1, 2, 3
Week 1	Aug 29, 31	**Intro, Multiple Regression**	Sheather, Ch 5 ISLR 3.1, 3.2 G&H Ch 3
Week 2	Sep 5 (no class[1]), Sep 7	**Qualitative Predictors**	ISLR, 3.3.1, 3.3.2, handouts
Week 3	Sep 12, 13	**Diagnostics & Transformations**	Sheather, Ch 6 ISLR 3.3.3 G&H Ch 4
Week 4	Sep 19, 21	**Variable Selection**	Sheather Ch 7 ISLR Ch 6 G&H Ch 4
Week 5	Sep 26, 28	**Logistic Regression** *Take-Home Midterm Assigned*	Sheather Ch 8 ISLR 4.3 G&H Ch 5
Week 6	Oct 3, 5	**Generalized Linear Models** *Take-Home Midterm Due*	Sheather Ch 8 G&H Ch 6
Week 7	Oct 10, 12	**Nonparametric Regression**	ISLR Ch 7 Sheather Appx handouts
	Oct 17-21	FALL BREAK	
Week 8	Oct 24, 26	**Causal Reasoning**	G&H Ch 9, 10
Week 9	Oct 31, Nov 2	**Multilevel and Mixed Effects Models** *Project assigned*	Sheather, 10.1 Intercepts: G&H Ch 12 Slopes: G&H Ch 13
Week 10	Nov 7, 9	**Multilevel logistic regression & GLMs**	G&H parts of Ch's 14 & 15
Week 11	Nov 14, 16	**Residuals, Estimation and Model Selection**	Handouts; stuff from G&H
Week 12	Nov 21, 23 (no class)[2]	**Bayes & Shrinkage**	Lynch Ch 3, 4 G&H parts of Ch 16
Week 13	Nov 28, 30	**STAN and MLM's** *Project due*	Lynch, Ch 9 G&H parts of Ch 17
Week 14	Dec 5, 7	**MLM's with STAN**	Handouts; maybe parts of G&H Ch's 18, 21

[1]No Class Sep 5: Labor Day (US Holiday).
[2]No Class Nov 23: Thanksgiving (US Holiday) Nov 24.

Take-home and project dates are approximate.

Syllabus Stuff – Work & Rules

- 20%: 10-ish HW's
  - Please feel free to work with each other on hw; ***BUT you must list who you worked with.***
- 10%: Monday Quizzes (on weekly reading/materials)
- 25%: Take-Home Midterm
- 25%: Final Report Project
- 10% Peer review of projects
- 10% Participation (Do I remember your name? What you did in class? In office hours? On Piazza?)
- Credit where credit is due
  - Please list any *person* or any *source* you consulted in doing your work, in a list of references at the end of hw, project, take-home
- All hw will be submitted via Canvas (Gradescope)
  - *Generally we will not accept late hw or late take-homes…*

Reading & HW

- The first HW assignment will be available on Canvas later today (due next week).
- Reading:
  - You should already have read Sheather Ch's 1-3
  - For this week read Sheather Ch 5
    - ISLR 3.1 & 3.2, and G&H CH 3 are good supplemental reading
  - For next week: Sheather 5.3, and ISLR 3.3.1 & 3.3.2
- Pdf's for lecture notes etc. in the file area on Canvas
- In general, you should do the reading before each week's classes.

Introduction – Level

- Hopefully you have seen calculus-based prob & stat, matrix algebra, and a little linear regression.
  - We need to talk like statisticians!
- You have all different levels of experience with
  - Applied regression and statistical modeling
  - R
  - Writing scientific reports
- Fill in the gaps
  - Learn on your own (Google)! Help each other!
  - Ask Lorenzo and me!

Let's take a break and think about these two quotes…

- "[I]t makes sense to base inferences or conclusions only on valid models"
  – *S.J. Sheather (2007)*

- "All models are wrong but some are useful"
  – *G.E.P. Box (1978)*

Linear Regression – The Model

- Let $X_i = (x_{i0}, \ldots, x_{ip})$ and $\beta = (\beta_0, \ldots, \beta_p)^T$; then

$$\begin{aligned} y_i &= \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \\ &= X_i \beta + \epsilon_i \end{aligned}$$

- If we also stack $Y = (y_1, \ldots, y_n)^T$, $X = (X_1^T, \ldots, X_n^T)^T$, and $\epsilon = (\epsilon_1, \ldots \epsilon_n)^T$, we can write

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \epsilon_{n\times 1} \quad (k = p+1)$$

Linear Regression –The Model

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Usually $x_{i0} \equiv 1$, so we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression – The Model

- In the model

$$y_i = X_i\beta + \epsilon_i, i = 1, \ldots, n$$

it is usual to assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Recall
  - Y ~ N(0,1) iff $f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$
  - Y ~ N(μ, σ^2) iff $\frac{y-\mu}{\sigma} \sim N(0,1)$

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Digression – Multivariate Normal

- $Y = (Y_1, \ldots, Y_n)^T \sim N(0,I) = N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \right)$

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2}$$

- $Y \sim N(\mu,\Sigma)$ iff

$$\Sigma^{-1/2}(Y - \mu) \sim N(0, I)$$

…and some ugly formula for $f(y_1, \ldots, y_n)$…

Digression – Multivariate Normal

- When Y ~ N(μ,Σ), then

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

is the *mean vector*:

$$E[y_i] = \mu_i,\ i = 1, \ldots, n$$

$$\mathrm{Var}(Y) = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

is the *variance-covariance matrix*:

$$\begin{aligned} \mathrm{Var}(y_i) &= \sigma_i^2,\ i = 1, \ldots, n \\ \mathrm{Cov}(y_i, y_j) &= \sigma_{ij},\ i, j = 1, \ldots, n \end{aligned}$$

Regresssion – ML/LS Estimates

- Y = Xβ + ϵ, ϵ ~N(0,σ^2I)
- So Var(Y_i) = E[(Y_i – X_i β)^2] = E[ϵ_i^2] = Var(ϵ_i) = σ^2
- Then we can estimate (MoM!):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

- Fitting the model is basically just finding values β to minimize $\frac{1}{n}RSS$, i.e., minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)^2 = \frac{1}{n}(Y - X\beta)^T(Y - X\beta)$$

- It turns out that $\hat{\beta} = (X^TX)^{-1}X^Ty$

Regression – ML/LS Estimates

$$\boxed{y = X\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I)}$$

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$ "Hat matrix"
- $\hat{e} = y - \hat{y} = (I - H)y$

- The "residual SD" is the square root of
  $s^2 = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2 = \frac{1}{n-k} (y - X\hat{\beta})^T (y - X\hat{\beta})$

  Recall that k=p+1

- With a little more matrix algebra,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= (X^T X)^{-1}\sigma^2 \\
\mathrm{Var}(\hat{y}) &= X(X^T X)^{-1} X^T \sigma^2 = H\sigma^2 \\
\mathrm{Var}(\hat{e}) &= (I - H)\sigma^2
\end{aligned}$$

Regression – Example

- Demographic factors and income…

```
> heights <- read.dta("heights.dta")
> str(heights)
'data.frame':   2029 obs. of  9 variables:
 $ earn   : num  NA NA 50000 60000 30000 NA 50000 NA 51000 ...
 $ height1: int  5 5 6 5 5 5 5 5 5 5 ...
 $ height2: num  6 4 2 6 4 5 3 8 3 4 ...
 $ sex    : int  2 1 1 2 2 2 2 2 2 2 ...
 $ race   : int  1 2 1 1 1 1 3 2 1 1 ...
 $ hisp   : int  2 2 2 2 2 2 2 2 2 2 ...
 $ ed     : num  12 12 16 16 16 17 16 18 17 15 ...
 $ yearbn : num  53 50 45 32 61 33 99 36 51 64 ...
 $ height : num  66 64 74 66 64 65 63 68 63 64 ...
```

Regression – Example

$$\hat\beta = (X^TX)^{-1}X^Ty$$

$$SE(\hat\beta) = \sqrt{\mathrm{diag}(\widehat{\mathrm{Var}}(\hat\beta))} = \sqrt{\mathrm{diag}((X^TX)^{-1}s^2)}$$

```
> summary(lm(earn ~ height + ed, data=heights))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -106263.5     8564.7  -12.41   <2e-16 ***
height         1376.7      126.7   10.87   <2e-16 ***
ed             2590.8      197.7   13.11   <2e-16 ***

Residual standard error: 17780 on 1376 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.1915,    Adjusted R-squared:  0.1904
F-statistic:   163 on 2 and 1376 DF,  p-value: < 2.2e-16
```

$$t_j = \hat\beta_j / SE(\hat\beta_j)$$

$$\begin{aligned} \text{earn} &\approx \hat\beta_0 + \hat\beta_1(\text{height}) + \hat\beta_2(\text{ed}) \\ &= -106263.5 + 1376.7(\text{height}) + 2590.8(\text{ed}) \end{aligned}$$

$$\frac{1}{n-k}RSS = s^2 = (17780)^2 = 316128400$$

$$R^2 = \frac{\widehat{\mathrm{Var}}(\hat y)}{\widehat{\mathrm{Var}}(y)} = \frac{SS_{reg}}{SSY} = 1 - \frac{RSS}{SSY} = 0.1915$$

$$\begin{aligned} k &= p+1 = 3\ ; \quad n-k = 1376 \\ n &= 1379 \end{aligned}$$

Regression - Example

- Continuing in R…
- heights.r in the week01 folder under “Files” on Canvas…

R!

- Some people learning R for the first time; others have done extensive data analysis projects in R.
- Some references on R:
  - http://www.cookbook-r.com
  - Quick-R: https://www.statmethods.net/
  - Online course: https://www.datacamp.com/courses/free-introduction-to-r
  - **Lately I really like https://kieranhealy.org/publications/dataviz/**
- If you have not used R before…
  - http://www.cs.cmu.edu/~10702/R2/Rintro.pdf provides a good start
    - If you are new to R, type into R all the commands and examples in rintro.pdf
    - If you have worked with R before, read through rintro.pdf and try to predict what would happen with each command. If you are not sure, type in that command/example.

Summary

- Introduction
- Course Schedule & Syllabus Stuff
- Valid vs Useful
- Quick Review of Univariate Regression
- R!
- Reading:
  - Read/skim Ch’s 1-3 of Sheather
  - Look at Ch 5 of Sheather more closely
  - Supplemental:
    - G&H CH 3
    - ISLR 3.1,3.2, 3.3.1,3.3.2