

STAT406- Methods of Statistical Learning Lecture 1

Matias Salibian-Barrera

UBC - Sep / Dec 2017

About me

- Matias Salibian-Barrera
- `matias@stat.ubc.ca`
- `http://www.stat.ubc.ca`
- `http://github.com/msalibian`
- `@msalibian`
- Professor, Department of Statistics
- Undergrad in Math, PhD in Stats

Prerequisites

- STAT306 or ECON326 or linear regression
- You are comfortable **working independently**
- You are **motivated** and enjoy being challenged
- You **want to be here** and are **interested in learning** the material

Philosophy of the class

- We're here to **help you learn** (vs. teaching you)
- We'll encourage **engagement**, **curiosity** and **generosity**
- We'll have **zero tolerance for plagiarism**
- We favour **steady work** through the Term (vs. sleeping until finals)

Lectures



- **Bring your laptop**
- **Prepare for class**
- Ask, doubt, question, discuss

Lectures / Labs / Office hours

- Two weekly lectures, one weekly lab
- Ongoing evaluation - you are expected to attend **all** course meetings
- **Pre**-lecture readings and activities
- **Office hour**: TBA, check Connect next week
- It is a **4th year course** – expectations are high

Grades

- Assignments: 10%,
 - Quizzes (Webwork): 20%
 - Lab activities: 10%
 - In-Class midterm: 20%
 - Final exam: 40%.
-
- There will be **no make-up** activities, quizzes, labs, assignments or exams. Anything you **miss** (with official documentation) will be assigned to your **final exam weight**.

Textbook?

- No textbook
- Lecture slides + verbose scripts
- Several reference books – all available on-line @ UBC Library or publisher
- **Most used: [JWHT13]** - *An Introduction to Statistical Learning*, James, Witten, Hastie, Tibshirani, R., 2013, Springer, New York.

Computer



- We will use R
 - Open source and free
 - Very flexible, relatively powerful
 - “Standard” in Statistics community, some industry
- Webwork quizzes rely on R

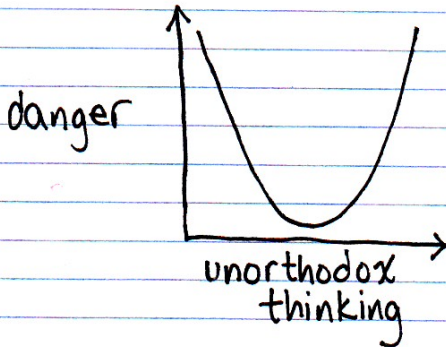
Computer

- **YOU ARE RESPONSIBLE** for learning
- We **CAN HELP** with \mathbb{R}
- We **WON'T teach all of** \mathbb{R}
- There are tons of on-line resources

Other resources

- **Connect:** announcements, e-mail blasts, general info
- **Scripts to reproduce class examples on**
`github.com/msalibian/STAT406-2017`
- **PIAZZA: student led & driven forum**
`piazza.com/ubc.ca/winterterm12017/stat406`

Lectures?



thisisindexed.com

Discussion

Statistical learning

Discussion

Models versus “predictive algorithms”

Review...

- Y is the response variable
- \mathbf{X} is a vector of auxiliary variables

$$Y = f(\mathbf{X}) + \varepsilon$$

- $f : \mathbb{R}^p \rightarrow \mathbb{R}$, unknown
- If $E[\varepsilon] = 0$

$$E[Y | \mathbf{X}] = f(\mathbf{X})$$

Review...

- In a linear model, f is a linear function

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- If $E[\varepsilon] = 0$

$$E[Y|X_1, X_2, \dots, X_p] = \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Review...

- Why would we want to estimate the coefficients of the linear model?
- What's the connection with prediction?

Review...

- Why would we want to estimate the coefficients of the linear model?
- What's the connection with prediction?
- “Best predictor”

$$\arg \min_{\mathbf{h}} E \left[(Y - \mathbf{h}(\mathbf{X}))^2 \right] = E[Y | \mathbf{X}]$$

Review...

- Best predictor is the regression function
- We need to estimate $E[Y|\mathbf{X}]$
- We propose a model (e.g. linear) for $E[Y|\mathbf{X}]$ and estimate it
- E.g. in a linear model, to estimate $f(\mathbf{X})$ we need to estimate $\beta_0, \beta_1, \dots, \beta_p$

Review...

- Data $(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \dots, (Y_n, \mathbf{X}_n),$
- Least squares estimator

$$\hat{\beta} = \arg \min_{\beta_0, \beta} \sum_{i=1}^n (Y_i - \beta_0 - \beta' \mathbf{X}_i)^2$$

Review...

- There is a closed form for $\hat{\beta}$

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{Y}$$

where $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ and

$$X = \begin{pmatrix} 1 & \dots & \mathbf{X}_1 & \dots \\ 1 & \dots & \mathbf{X}_2 & \dots \\ 1 & \dots & \mathbf{X}_3 & \dots \\ \vdots & \dots & \dots & \dots \\ 1 & \dots & \mathbf{X}_n & \dots \end{pmatrix}$$

Review...

- As long as $E[\varepsilon] = E[\varepsilon|\mathbf{X}] = 0$ we have

$$E[\hat{\beta}] = \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- The LS estimator is consistent and unbiased
- Do we need any other assumption?

Review...

- Consider the air pollution data
- $n = 60$ observations
- $p = 16$, response variable: MORT
- A linear model:

$$\text{MORT} = \beta_0 + \beta_1 \text{PREC} + \beta_2 \text{JANT} + \dots + \epsilon$$

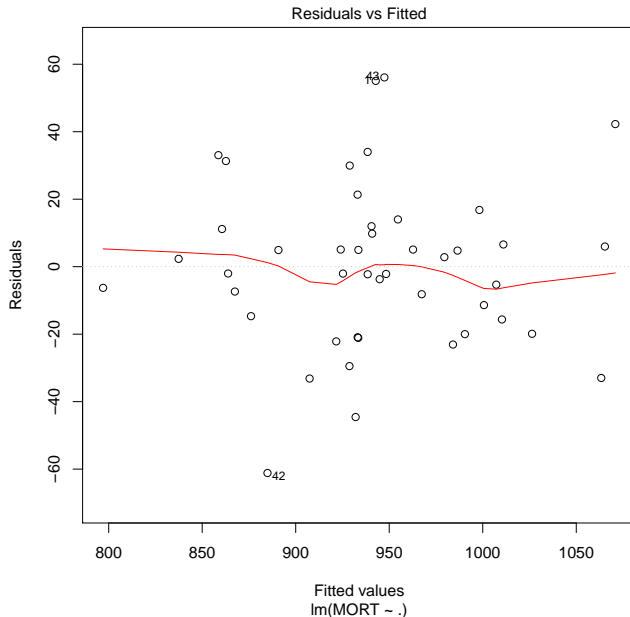
or equivalently

$$\begin{aligned} E \left(\text{MORT} \mid \text{PREC}, \text{JANT}, \dots \right) \\ = \beta_0 + \beta_1 \text{PREC} + \beta_2 \text{JANT} + \dots \end{aligned}$$

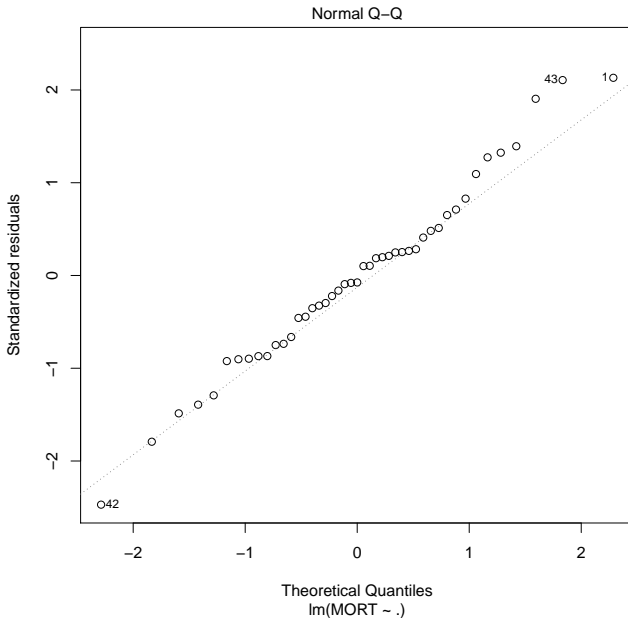
Review...

- Randomly split into a training ($n=45$) and a test set ($n=15$)
- Use training set to fit a model
- Read data into object `x.tr`
- Fit the “full” model
- “Look” at the fit

Diagnostics



Diagnostics



Diagnostics

```
> full <- lm(MORT ~ ., data=x.tr)
> summary(full)

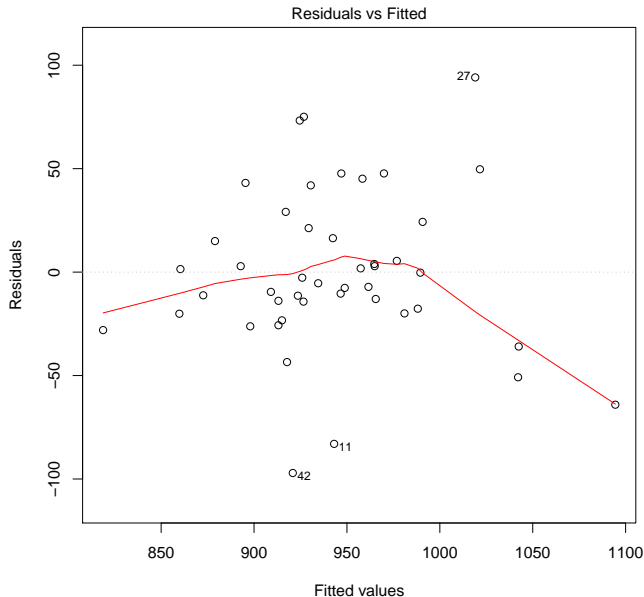
> sum( resid(full)^2 )
[1] 25898.8
```

Fit a reduced model

```
> reduced <- lm(MORT ~ POOR + HC +  
NOX + HOUS + NONW, data=x.tr)
```

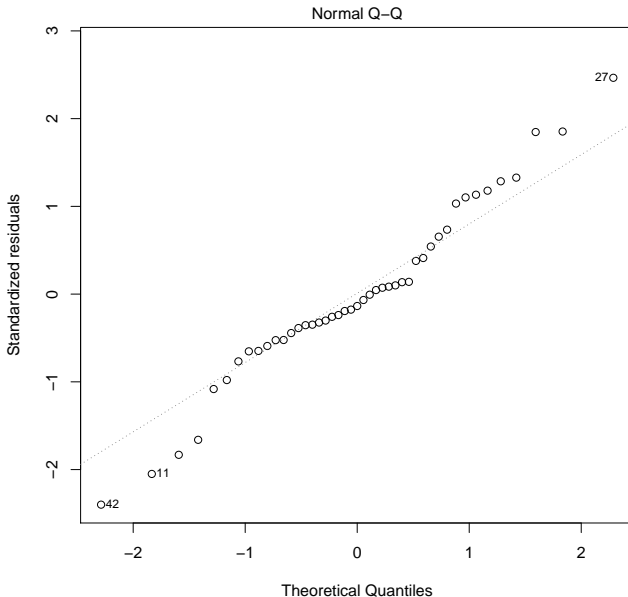
```
> sum( resid(reduced)^2 )  
[1] 66135.29
```

Diagnostics for reduced model



lm(MORT ~ POOR + HC + NOX + HOUS + NONW)

Diagnostics for reduced model



Discussion

Goodness of fit versus
prediction power

Predictions

$$Y \longleftrightarrow \hat{f}(\mathbf{X})$$

$$\left(Y - \hat{f}(\mathbf{X})\right)^2 \quad ?$$

$$E \left[\left(Y - \hat{f}(\mathbf{X})\right)^2 \right] \quad ?$$

What are we “averaging” over?
What is random?

Predictions

$$E_{(Y^*, \mathbf{X}^*)} \left[\left(Y^* - \hat{f}(\mathbf{X}^*) \right)^2 \right]$$

where (Y^*, \mathbf{X}^*) are new, future observations, not used when computing \hat{f} .

Predictions

If we assume that $Y = f(X) + \epsilon$, then

$$E_{(Y^*, \mathbf{X}^*)} \left[\left(Y^* - \hat{f}(\mathbf{X}^*) \right)^2 \right] =$$

$$E_{(Y^*, \mathbf{X}^*)} \left[\left(f(\mathbf{X}^*) - \hat{f}(\mathbf{X}^*) \right)^2 \right] + V(\epsilon)$$

- what assumptions are needed for this to be true?
- is it still true if I look at predictions for a single & fixed \mathbf{X}_0 ?

Predictions

- What we want

$$E_{(Y^*, \mathbf{X}^*)} \left[\left(Y^* - \hat{f}(\mathbf{X}^*) \right)^2 \right]$$

is very difficult to estimate

- Something similar

$$E_{\{(Y^*, \mathbf{X}^*), \text{data}\}} \left[\left(Y^* - \hat{f}(\mathbf{X}^*) \right)^2 \right]$$

is easier to estimate

Predictions

- Read the test set
- Use both models to predict `MORT`
- Compare both sets of predictions

Predictions

```
> x.te <- read.table('pollution-test.dat', ...  
>  
> x.te$pr.full <- predict(full, newdata=x.te)  
> x.te$pr.reduced <- predict(reduced,  
                             newdata=x.te)  
>  
> with(x.te, mean( (MORT - pr.full)^2 ))  
[1] 4677.45  
>  
> with(x.te, mean( (MORT - pr.reduced)^2 ))  
[1] 1401.571
```

Discuss

Discussion points

- Goodness of fit vs. prediction power
- How do we estimate prediction MSE?

$$E_{(Y^*, \mathbf{X}^*)} \left[\left(Y^* - \hat{f}(\mathbf{X}^*) \right)^2 \right]$$

- Can it be done without a test set?

Ready, set, go!

- Quiz 0 (Review quizz) on Webwork
- Practice Quiz on Connect
- Check `connect.ubc.ca` often
- Visit
`github.com/msalibian/STAT406-2017`
- Use PIAZZA
- Read before class
- Attend the lab