STAT406- Methods of Statistical Learning Lecture 1

Matias Salibian-Barrera

UBC - Sep / Dec 2016

About me

- Matias Salibian-Barrera
- `matias@stat.ubc.ca`
- `http://www.stat.ubc.ca/~matias`
- `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**: Wed 1:30 - 2:30, ESB 3174
- It is a **4th year course** – expectations are high

Grades

- Homeworks & quizzes: 30%,
- Lab activities: 25%
- In-Class activities: 5%
- Final exam: 40%.

- There will be **no make-up** activities, quizzes, labs, homeworks or exams. Anything you **miss** (with official documentation) will be assigned to your **final exam weight**.

Textbook?

- No textbook
- Several reference books – all available on-line @ UBC Library

- **Most used**: [**JWHT13**] - *An Introduction to Statistical Learning*, James, Witten, Hastie, Tibshirani, R., 2013, Springer, New York.

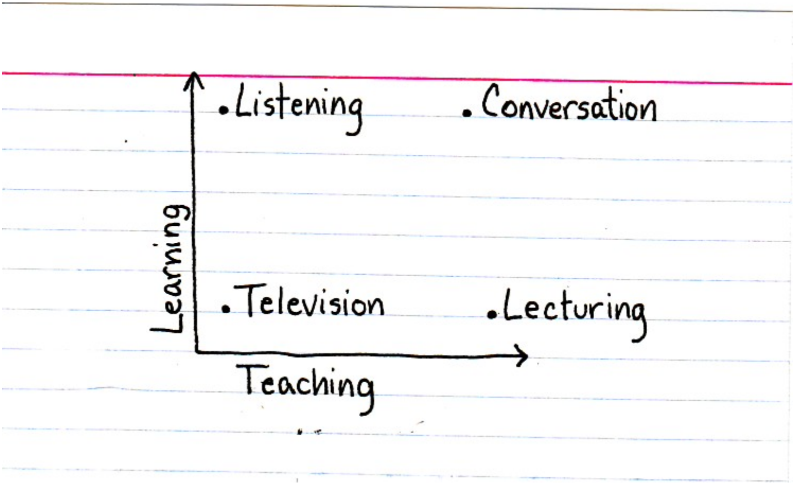
Computer

- I use R
  - Open source and free
  - Very flexible, relatively powerful
  - "Standard" in Statistics community
- Any other software is also fine

Computer

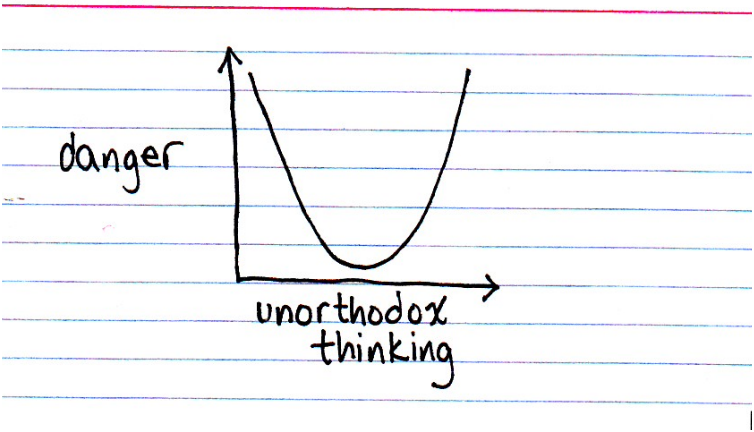
- Whichever software you use, learn it
- We can **help** with `R`
- We **won't teach all of** `R`
- **You are responsible** for learning it
- There are tons of on-line resoures
- Example: `http://swirlstats.com/`

Lectures?



thisisindexed.com

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

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

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

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

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 $\beta_0, \beta_1, \ldots, \beta_p$

Review...

- Data $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, $\ldots$, $(Y_n, \mathbf{X}_n)$,

- Least squares estimator

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

Review...

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

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

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

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

Review...

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

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

- 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:

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

or equivalently

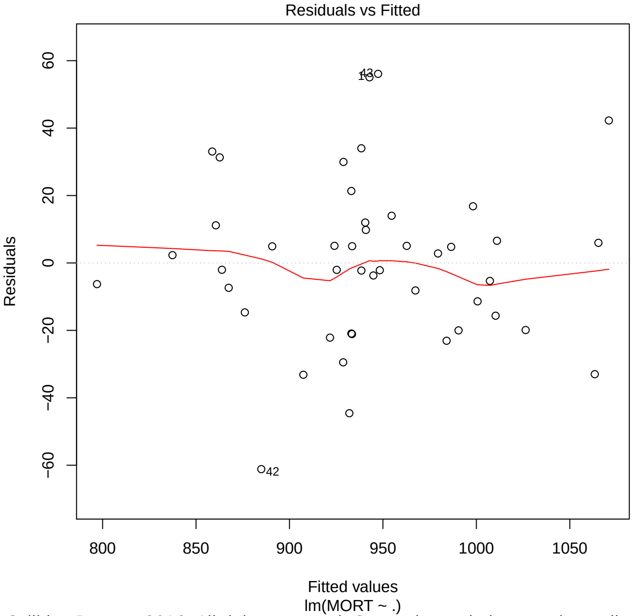
$$E\left(\text{MORT} \,\middle|\, \text{PREC}, \text{JANT}, \ldots\right) = \beta_0 + \beta_1\,\text{PREC} + \beta_2\,\text{JANT} + \ldots$$

Review...

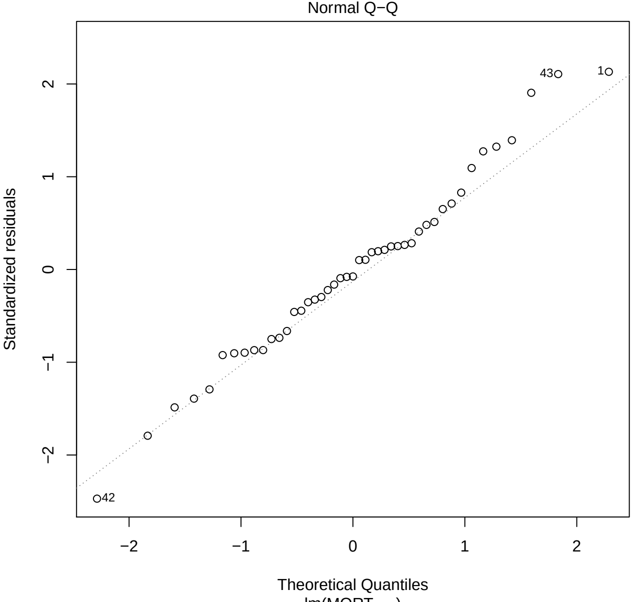
- Randomly split into a training (`n=45`) and a test set (`n=15`)
- Use training set to fit a model

- Read data into object `x.tr`
- Fit the "full" model
- "Look" at the fit

Diagnostics



Diagnostics



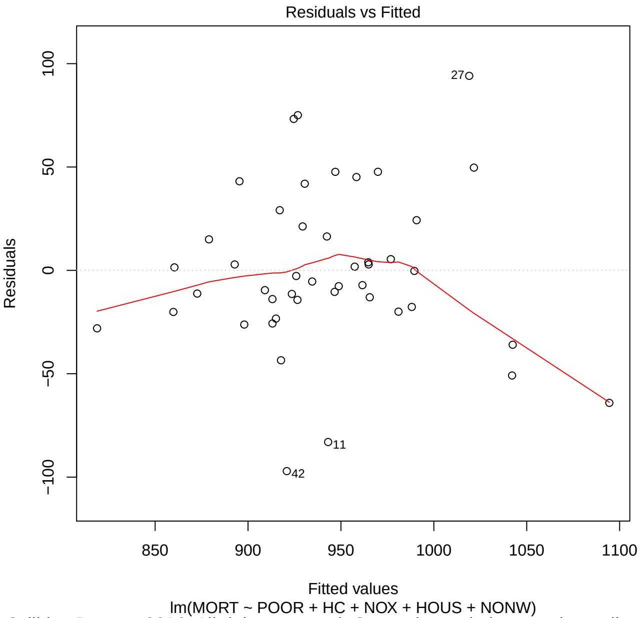
Diagnostics

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

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

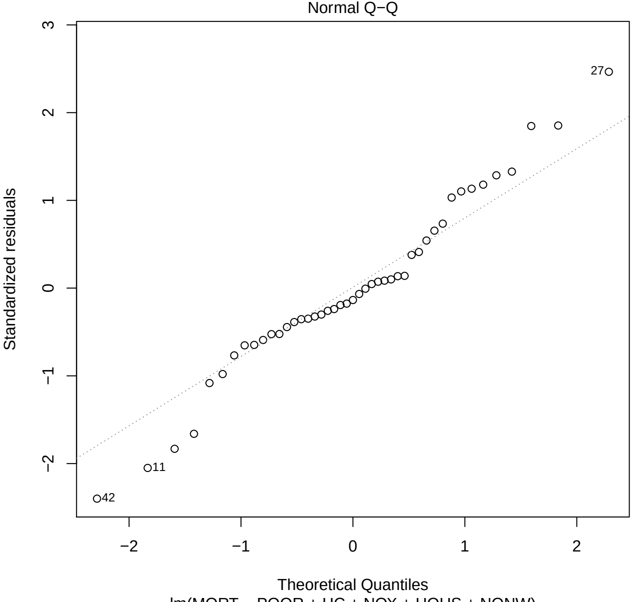
Fit a reduced model

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

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

Diagnostics for reduced model



Diagnostics for reduced model



Discussion

Goodness of fit versus prediction power

Predictions

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

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

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

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

Predictions

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

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

Predictions

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

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

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

- 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

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

is very difficult to estimate

- Something similar

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

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?

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

- Can it be done without a test set?

Next week...

- Quiz 0 (Review) is out, due next class.
- Check `connect.ubc.ca` often
- Read the suggested sections of [JWHT13]
- Attend the lab