STAT 447B: Methods for Statistical Learning

Introduction and Linear Models (Review)

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics
University of British Columbia

Sep-Dec 2014

STAT 447B Teaching Team

- Instructor: Matias Salibian (second half)
- TA: David Lee
- Instructor: Will Welch (first half)

About Me

- B.Sc in Management Sciences, Loughborough University, England
- M.Sc, PhD in Statistics, Imperial College, University of London
- Professor at Prof at UBC (Commerce) for 3 years, University of Waterloo for 16 years and UBC (Statistics) again for 11 years
- Researcher in statistical methods for nearly 30 years
  - many topics, a major one is design and analysis of experiments via computer models
  - for 10+ years classification, particularly applied to drug discovery
- In recent years, taught STAT 305 (many times), STAT 404, STAT 447K (similar to this course)
- My hobbies include photography

About You

- What is your program?
  Clicker question 1.
- Have you taken and passed STAT 305 Introduction to Statistical Inference, or an equivalent course?
  Clicker question 2.
- Have you taken and passed STAT 306 Finding Relationships in Data, or an equivalent course (least-squares regression, logistic regression)?
  Clicker question 3.
- If you do not have STAT 306 background you need to talk to me.

Prerequisites

- So you should have some basic statistic courses.
- You are an undergrad student that has taken some basic Statistic courses including STAT306 or equivalent.
- Or you are a graduate student, say in a Science program other than Statistics, and have used regression models.
- You are comfortable working independently (i.e., finding by yourself details of a method, its computer implementation, etc.)
- You are motivated to learn some fairly novel statistical methods and enjoy being challenged.

Evaluation

- 2 in-class quizzes (20% of the course grade)
  - Dates: Sep 23 and ???
  - If you miss a quiz for a valid reason the 10% weight will be reassigned to the final and other in-class quiz
- 4 assignments (20% of the course grade)
  - Due dates: ??
  - If you miss an assignment for a valid reason the 5% weight will be reassigned to the other assignments
- Lab activities (5% of the course grade)
- In-class activities, mainly based on clickers (5% of the course grade)
- Final exam (50% of the course grade)
- To pass the course you must get a passing grade (35/70) from the 70% available for the final exam and in-class quizzes
- There will not be any make-up quizzes or assignments, **NONE, ZERO**

Tentative Outline

- Prediction models, linear, non-linear and non-parametric
- Classification (supervised learning), parametric and non-parametric models
- Classification and regression trees, boosting, neural networks
- Additive and Generalized Additive Models
- Principal components analysis (PCA)
- Clustering (unsupervised learning), model-based and otherwise
- Support vector machines (SVMs) for classification and regression
- Throughout
  - Variable (feature) selection via cross validation, etc.
  - Model misspecification, bias-variance trade-off

Lectures / Labs / Office hours

- Two weekly lectures
- One weekly lab (Wednesday or Thursday, 4:00–5:00 pm)
- Attending both is strongly recommended
- Pre-lecture readings, lecture activities
- Office hours
  - Will: tentatively Tuesday, 2:00–3:00, Thursday, 3:00–4:00
  - David: TBA
- This is a **4th year course** – expectations are high

Lectures

- Bring your laptop
- Short in-class activities
- Doubt, question, challenge your instructors and TA!

Labs / Tutorials / Office hours

- Computer lab will have scheduled learning activities
- Sometimes it will work as a Tutorial session
- Attendance is strongly recommended

Website and Email

- Course website: `www.stat.ubc.ca/~will/STAT447B`
- Email: `will@stat.ubc.ca`
- Above will be updated when Professor Salibian takes over
- I will be posting:
  - Lecture slides, which include the short in-class activities
  - Clicker question answers
  - Computer code and datasets to reproduce lecture examples
  - News, important information

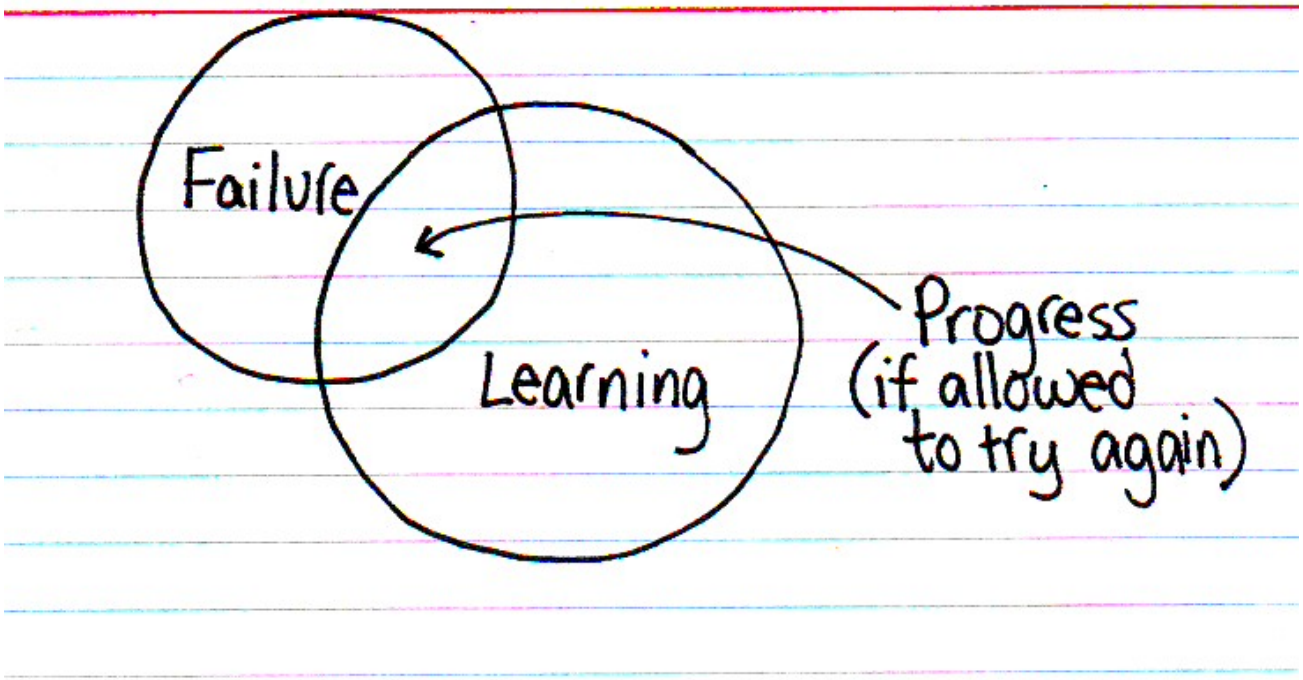
Textbook?

- No required textbook
- Several reference books – see Course Outline
  - Most useful is probably “An Introduction to Statistical Learning”, James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013, Springer, New York
  - All are available online

R Software

- We will use `R`
  - `http://www.r-project.org`
  - Open source and free
  - Very flexible, relatively powerful
  - "Standard" in statistics community
- Use R's help, e.g., `help(lm)`

Questions?



thisisindexed.com

Linear Models (Review)

- Consider the "union membership" data
- `wage` (wages, in dollars / hr)
- `years.experience`
- `age`
- `years.educ` (years of education)
- `female` ("0": male, "1": female)
- Why would I want to study the relationship between these variables (if any exists)?

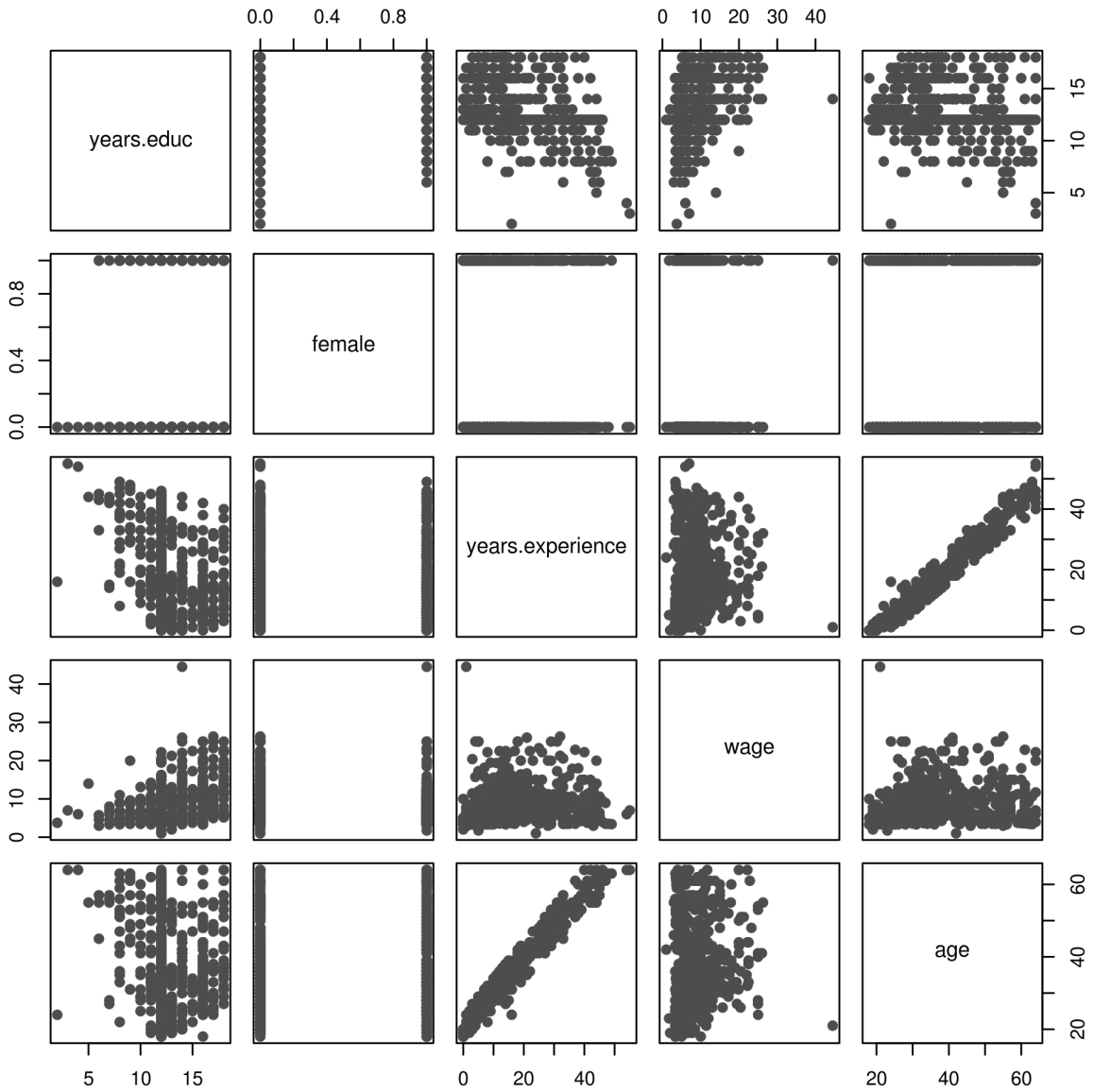
Trade Union Data

```
> # read data
> library(SemiPar)
> data(trade.union)
> x <- trade.union
>
> # look at read object
> str(x)
'data.frame': 534 obs. of  11 variables:
 $ years.educ      : int  8 9 12 12 12 13 10 12 16 12 ...
 $ south           : int  0 0 0 0 0 0 1 0 0 0 ...
 $ female          : int  1 1 0 0 0 0 0 0 0 0 ...
 $ years.experience: int  21 42 1 4 17 9 27 9 11 9 ...
 $ union.member    : int  0 0 0 0 0 1 0 0 0 0 ...
 $ wage            : num  5.1 4.95 6.67 4 7.5 ...
 $ age             : int  35 57 19 22 35 28 43 27 33 27 ...
 $ race            : int  2 3 3 3 3 3 3 3 3 3 ...
 $ occupation      : int  6 6 6 6 6 6 6 6 6 6 ...
 $ sector          : int  1 1 1 0 0 0 0 0 1 0 ...
 $ married         : int  1 1 0 0 1 0 0 0 1 0 ...
>
> # scatter plot of some variables
> pairs(x[,c(1,3,4,6,7)])
```

Review...

- Look at the data
- Objective: **predict wages**
  - Should we use other variables?
  - Regression?
  - What does regression mean?
  - Why would regression give good predictions?
  - Could there be a better way to predict wages with the available information?

Look at the data



Linear Regression

- Y is the response variable
- $x_1, x_2, \ldots, x_p$ are potential auxiliary variables
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
- $E[\varepsilon] = 0$
- $E[Y | x_1, x_2, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated.

Linear Regression: Trade Union Data

- Y is wages
- One x_j variable, years.experience, called x
- $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
- $E[\varepsilon] = 0$
- To a statistician, is this a linear model?
  Clicker question 4.