# STAT 540: Data Analysis and Regression

Wen Zhou

http://www.stat.colostate.edu/~riczw/
Email: riczw@stat.colostate.edu

Department of Statistics
Colorado State University

**Fall 2015**

# Class organization

- Prerequisites and assumed knowledge:
  - Two previous upper-division statistics classes
  - Two or three semesters of calculus
  - Linear algebra (matrices: operations, inversion, eigenvalues, etc.)
- KNNL book: Needed for reading and homework (expensive though).
- R: Install and learn–mostly self-taught–during the semester. Abundance of online resources. Start immediately!
- Homework: Start early.
- Reading: Self-directed. Follow along in the book.
- Come to class with printed slides (current section), R file, and having already read the book chapter(s)

# What is Statistics

- Dictionary definitions:
  - Statistic – a single term or datum; a quantity that is computed from a sample
  - Statistics – a branch of mathematics dealing with collection, analysis, interpretation and presentation of numerical data
  - Statistics - art and science of drawing justifiable conclusions from incomplete data

# Statistics as a Mathematical Science

- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$,

  or

  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \ N(0, \sigma^2 I)$.

- The model defines a set of random variables: $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)'$
- The unknown parameters are $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and $\sigma^2$

Can derive some of the following results (details later):

- The least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\hat{\beta}_0, \hat{\beta}_1)'$$

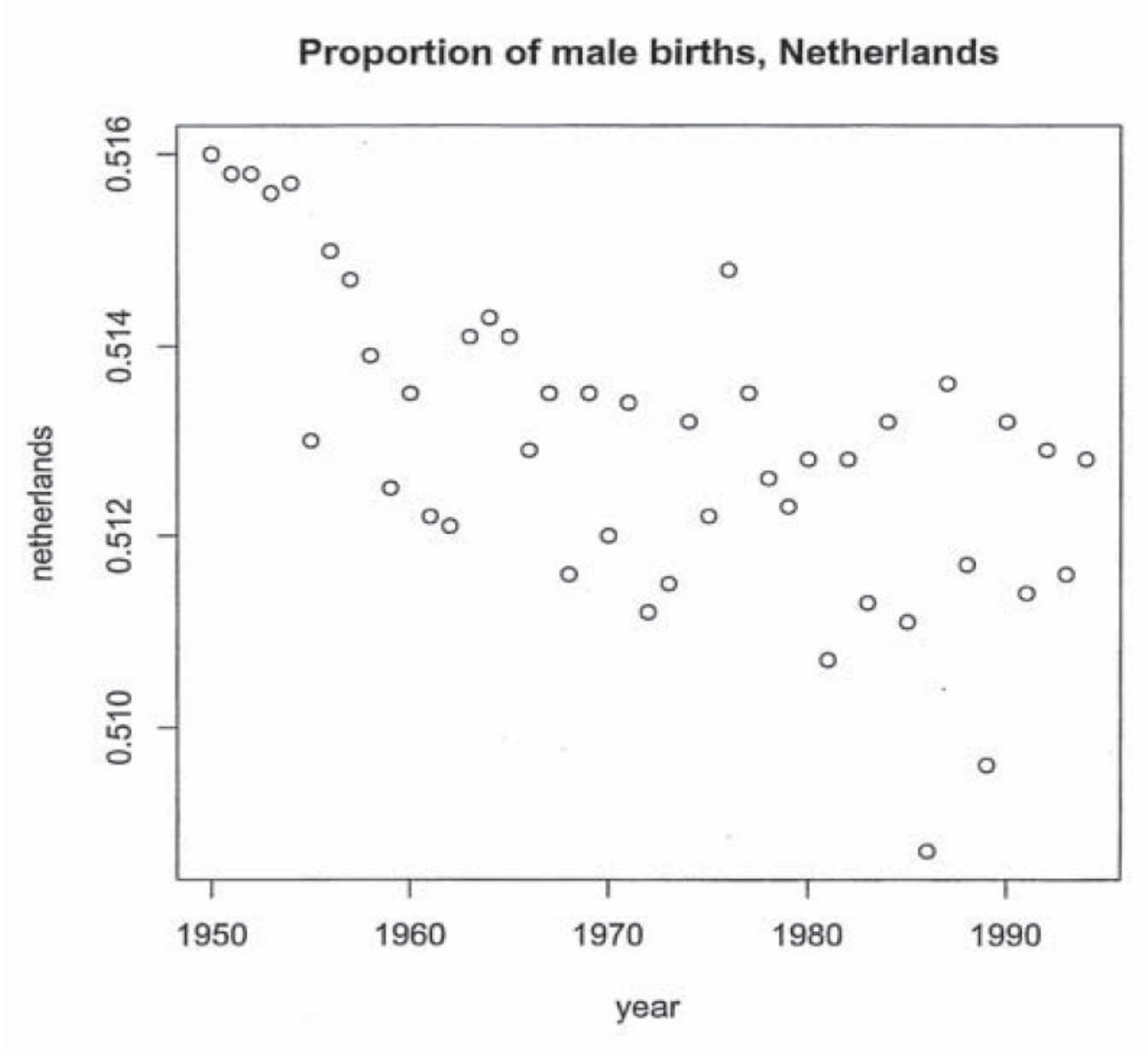
  is the minimum variance linear unbiased estimator
- $\text{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{V} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- $\boldsymbol{c}'\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{c}'\boldsymbol{\beta}, \boldsymbol{c}'\boldsymbol{V}\boldsymbol{c})$
- Test $H_0 : \boldsymbol{c}'\boldsymbol{\beta} = 0$ using $t = \frac{|\boldsymbol{c}'\hat{\boldsymbol{\beta}} - 0|}{\sqrt{\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}}}$

# Statistics as Art

**Friend, researcher, client says to you:** *I have data on percentages of male births in the Netherlands from 1950 to 2009. They seems to be declining. I need to estimate the expected change in the next ten years (2011 - 2020). You are a statistician; can you help me?*

Natural to use the linear model described above. Some issues:

- What type of model can be used to address the questions?
- Is the model reasonable? How can one tell?


Proportion of male births, Netherlands
netherlands
0.516
0.514
0.512
0.510
1950
1960
1970
1980
1990
year

- Are conclusions reasonable when
  - Observations (errors) are not normally distributed
  - Error variances may not be constant
  - Errors may not be independent
  - Is the relationship linear
- No absolutely optimal solution, more than one reasonable approach
- Most appropriate model?

# Statistics as Science

Statistics is the science of using information to make decisions and quantify uncertainty inherent to those decisions.

There are four basic steps in the process:

- Define the questions to be answered (Plan)
- Gather appropriate information (Do)
- Analyze the information and make decisions (Study)
- Implement the decisions (Act)

# Summary

- Statistics is a discipline where "relatively" simple models are applied to approximately describe "random" phenomena observed in the real world and inference/prediction are made.
- Probability theory provides the mathematical foundations (17th and 18th centuries).
- The method of least squares was invented around the turn of the 19th century.
- Modern computers have expedited large-scale statistical computation, making new methods computationally feasible.

# Essential Components of Practicing Statistics

- Study design.
- Data collection.
- Data analysis.
- Conclusion and interpretation.

# Study Design

- You are part of a team.
  - Subject-matter experts, database managers, programmers, etc.
  - Understand the background of a study.
  - Learn something about the subject area.
- What is the goal?
  - Understand the objectives.
  - Understand and help shape the expectation.
  - Suggest new avenues for research.
- Don't forget the design.
  - Maximize amount and relevance of information in the data.
  - Survey or observational studies vs. experiment.
  - Sample size constraints.

# Observational Studies

- Observe the population or process as it naturally occurs
- Often involves some type of sample
  - Simple random sample of size $n$: sampling without replacement or with replacement
  - Other sampling methods: cluster sampling, stratification, opportunistic samples
- **Sampling procedure dictates method of analysis**
- Cause and effect inferences are not possible

# Observational Studies – Examples

- Retrospective study of potential effects of smoking on lung cancer
  - Simple random sample of patients diagnosed with lung cancer
  - Independent simple random sample of non-lung cancer patients
  - Compare smoking histories for the two samples
- Prospective study: Nesting success of pheasants
  - Random sample of $N$ locations
  - Find nests and implant transmitters in chicks
  - Relate survival probability to local habitat

# Experiments

- Basic component of experiments
  - Treatment: combinations of levels of factors, and applied to experimental units
  - Experimental Units (observational units): receives the assigned treatment, and provides the measured response
  - Runs: operation of a process under a specific treatment
  - Replications: allows for estimation of error variance
  - Randomization: randomly assign experimental units to treatment groups
- Basic principles in design of experiments
  - Control of extraneous variation
  - Comparison with a control group
  - Blocking / matching
  - Blinding
- Extending inferences beyond the units used in the experiment
  - Were experimental units randomly selected from some population?

# Data Collection

- Time consuming and labor intensive.
- It may be difficult to follow the plan for the experiment or survey.
  - Nonresponse.
  - Missing values.
  - Adjustments to study design.
- Data management.
  - Data entry and coding.
  - Data checks.
  - Organizing data for analysis
    - ⋆ Depends on the planned analysis, therefore...
    - ⋆ Often left to the statistician
    - ⋆ Can be a huge task!
  - Data confidentiality.

# Data Analysis

- Perform exploratory data analysis (EDA).
  - Graphical methods. Critical. R is very powerful.
  - Summary statistics.
- Formulate the problem in statistical terms.
- Develop statistical models.
- Evaluate model appropriateness and goodness of fit.
- Carry out statistical inference.
  - Hypothesis testing.
  - Estimation.
  - Prediction.

# Conclusion and Interpretation

- Report conclusions.
  - Summarize data using best parts of the EDA
  - Summarize model
  - Provide estimates
  - All conclusions include a measure of uncertainty:
    - ★ Confidence intervals (preferred)
    - ★ p-values
  - Explain implications in plain language
  - Caveats
- Help team make a decision.
  - Change beliefs; alter the status quo.
  - Collect more information.
  - Take no further action.

# Exploratory Data Analysis (EDA)

- EDA is one of the most important parts of any data analysis
- Some goals of EDA are:
  - Understand the data.
  - Find unusual aspects or 'errors' in the data.
  - Anticipate results of complicated analysis.
  - Plan how to explain data and results to the consumer of your analysis
  - Doublecheck and debug computer code
- A picture is worth a thousand words. Try to show your results in a plot.

# Approaches to EDA

- 1-D exploration
  - Goal: Describe each variable separately.
  - Numerical summary: Mean, median, sd, etc.
  - Graphical summary: Histogram, stem-and-leaf diagram, etc.
- 2-D exploration
  - Goal: Study the relationship between pairs of variables.
  - Numerical summary: Variance matrix, correlation matrix, etc.
  - Graphical summary: Scatter plots, all-pairs plots, etc.
- $n$-D exploration
  - Goal: Simultaneously explore the relationships between $n$ variables.
  - For example, 3-D plots, Chernoff faces, etc.

# R

- Interactive, not compiled
- You type in code, or paste blocks of code. Hit enter.
- Output printed immediately. Usually output should be assigned to a variable to examine further details and save results for later
- Components of working in R
- Working directory: Default path where R looks for files. (File→Change Dir or `getwd()` and `setwd()`; backward slashes)
- Workspace: Binary file with all R objects (i.e, variables and functions you create).
- Script file: Block of code saved for future use.

# R Workspace

- Workspace: Binary file with all R objects (i.e, data, variables and functions you create).
  - Called .RData
  - File→Load Workspace
  - Recommended: one .RData per class or research project.
  - The loaded workspace comes from the directory where you started R, or the default directory if no .RData exists there
  - Can double-click on .RData file to open R with it, in that directory
  - Save it each session: must say "yes" to save upon exit, or `save.image()` anytime
  - To initialize, copy from somewhere else. Or change directory and then `save.image()`. Restart in proper directory, and (could) delete all variables using `rm(list = ls())`

# R Script File

- Script File: your programming code
  - Coding will be iterative. Lots of re-executing as you develop and debug the code
- A fundamental choice: your editor or theirs. Plain text!
- Theirs (path of least resistance):
  - File→New/Open Script
  - Terrible text editor. (Slow and PITA)
  - Some handy integrated commands (e.g., F5 = quick execute)
- Yours:
  - (Potentially) powerful text editor. (Fast)
  - Must cut/paste into R. My 5-button mouse has dedicated cut/paste buttons. . .fast
  - Watch out for non-ascii symbols, e.g., quotation marks (“ ”) in Word. (Use " )
- R Studio: Integrated development environment
  - Not recommended: ornate, with clumsy editor.

# Data Example: `lake`

- lake = ID number of the lake (10–40).
- site = site number (1–8).
- build = number of buildings per km of shoreline per lake.
- ow1822 = relative exposure of site to winds over 18 m/s when the water is open (April to October) per site.
- lnls = ln(density of logs +1) per site (response).
- lnll = ln(density of logs +1) per lake.
- sqrtts = square root of tree density per site (#/hectare).
- sqrtbas = square root of total basal area per site ($cm^2$/hectare).
- sqrtbal = total basal area per lake (averaged over sites) ($cm^2$/hectare).
- sqrtttl = square root of tree density per lake (averaged over sites) (#/hectare).

# Data Example: `eire`

- towns: towns/unit area.
- ROADACC: arterial road network accessibility in 1961.
- OWNCONS: percentage in value terms of gross agricultural output of each county consumed by itself.
- POPCHG: 1961 population as percentage of 1926.
- RETSALE: value of retail sales (£1000).
- INCOME: total personal income (£1000).
- names: county names