Lecture 1

Course Introduction
Scientific method and the role of regression

Course Description

- 140.653 Linear regression for continuous outcomes
- 140.654 Regression for discrete outcomes plus some survival analysis
- 140.654 Introduction to machinic learning approaches (classification/regression trees and random forests)
- You will be learning the underlying theory behind how linear regression works with an emphasis on developing, fitting, interpreting and evaluating models to address specific scientific questions.

Course Objectives

Upon successfully completing this course, students will be able to:

1. Formulate a scientific question about the relationship of a continuous response variable Y and predictor variables X in terms of the appropriate linear regression model.

2. Interpret the meaning of regression coefficients in scientific terms as if for a substantive journal. Explicitly define the epidemiologic terms “confounding” and “effect modification” in terms of multiple regression coefficients

3. Develop graphical and/or tabular displays of the data to display the evidence relevant to describing the relationship of Y with one X controlling for others. Use an adjusted variable plot to explain the meaning of a multiple regression coefficient.

Course Objectives

Upon successfully completing this course, students will be able to:

4. Estimate the model using a modern statistical package such as STATA or R and interpret the results for substantive colleagues. Derive the least squares estimators for the linear model and the distribution of coefficients, predicted values, residuals and linear functions of them.

5. Check the major assumptions of the model including independence and model form (mean, variance and distribution of residuals) and make changes to the model or method of estimation and inference to appropriately handle violations of standard assumptions. Use weighted least squares for situations with unequal variances. Use robust variance estimates for violations of independence or variance or distributional assumptions. Use regression diagnostics to prevent a small fraction of observations from having undue influence on the results

6. Write a methods and results section for a substantive journal, correctly describing the regression model in scientific terms and the method used to specify and estimate the model. Correctly interpret the regression results to answer the specific substantive questions posed in scientific terms that can be understood by substantive experts

7. Critique the methods and results from the perspective of the statistical methods chosen and alternative approaches that might have been

Course Objectives: Highlights

I want you to know the statistical theory AND very importantly how to:

1. **Formulate a scientific question about the relationship of a continuous response variable Y and predictor variables X in terms of the appropriate linear regression model.**
   - **Real-estate and retail: location, location, location**
   - **Statistical science: question, question, question**

6. **Write a methods and results section for a substantive journal, correctly describing the regression model in scientific terms and the method used to specify and estimate the model. Correctly interpret the regression results to answer the specific substantive questions posed in scientific terms that can be understood by substantive experts**
   - **This is new for many of you; my red ink story for inspiration**

7. **Critique the methods and results from the perspective of the statistical methods chosen and alternative approaches that might have been**
   - **Think sensitivity analysis/limitations/discussion section of all scientific papers!**

Teaching Team

- Elizabeth Colantuoni, ejohnso2@jhu.edu
- Erjia Cui, ecui1@jhmi.edu
- Jingning Zhang, jzhang218@jhu.edu
- Jiawei Bai, Jiawei.bai@jhu.edu

Office hours:

- Thursday 10:30-11:50am EST or by appointment

Course Format

Lectures:

- Synchronous Lecture Tuesdays: 7:30-8:50am EST OR 10:30-11:50am EST
  - Combination of me talking (lecture slides) and you doing (hands on in-class exercises)
- Asynchronous Lecture: Will post this on Wednesday morning, should be viewed prior to next synchronous lecture.

Labs:

- Synchronous Tuesdays: 3:30-4:20pm EST OR 9:30 – 10:20pm EST
  - These are designed to compliment the material you are learning in lecture and provide additional guidance for tools needed to complete the homework assignments
  - See syllabus for schedule
  - Will often cover material not in the lecture
  - Structure: Presentation by TA following by short hands-on exercise linked with problem sets

Recordings of all sessions will be available!

Course Evaluations

- Three problem sets (10% each)
- Three on-line quizzes (20-30 minutes) (10% each)
- Participation (10% each): do you attend lectures, participate in in-class exercises and discussion, do you attend lab sessions, participate in lab exercises and discussion
- Fourth problem set: final project
- Across the 4 problem sets, we provide less and less guidance so that the fourth problem set is a reflection of your accumulating knowledge and ability to work from the start (formulate the statistical analysis based on the scientific question through written/graphical presentation of results)

Key Dates

- Math Certification Quiz: Complete by February 5th
- Problem Set 1: February 11th
- On-line Quiz 1: February 15th
- Problem Set 2: February 25th
- On-line Quiz 2: March 1st
- Problem Set 3: March 11th
- In-class Quiz 3: March 12th
- Problem Set 4: March 19th

Course Communication

- Direct email from course faculty and announcement via Courseplus in the event of major changes to due dates or important messages
- Slack workspace
  - Please subscribe to this forum
  - We have set up topic categories to help keep things organized
  - No question is too big/small
- Questions about grading:
  - Please send questions about grading to Elizabeth only.
- General guidelines on personal emails to course faculty
  - Please use the Slack workspace for questions relating to course content and problem sets. There will be other students in the course who have similar questions as you. So by posting questions in a public forum, we gain efficiency
  - Please send personal communications (e.g. problem set due date extension request) to Elizabeth

What should you be doing in the first two weeks of class?

- Reviewing the materials in “Review of Linear Algebra” pdf linked with today’s class session
- Take the math certification quiz
- Problem set 1

Now for the fun stuff.....

- Icebreaker
  - What are your personal objectives for the course?
  - How many inches of snow do you prefer?

- 5-minute break

Science / Scientific Method / Statistics

- What is science?
  - Search for parsimonious hypotheses (laws) that explain observations of nature
  - Search for “truth”
- What is the scientific method?
  - Methodology for generating and interpreting data as evidence to support some competing hypotheses about nature more than others
- What is statistics?
  - The science of data
  - Principles and quantitative methods for implementing the scientific method

Statistical Model

Description of how the data are generated by nature; precise statement of *a set* of hypotheses

- "Statistical" – part deterministic, part stochastic (involving probability)
  - Deterministic part – combinations of simple functions
  - Stochastic part –realizations of random variables
  - Data = deterministic model (signal) + stochastic deviation (noise)
  - Example:

    BP for person i = population mean + person i's deviation from mean;

    $BP_i = mu + e_i$

- "Model":
  - An approximation to reality; simplification of reality; cartoon that captures key characteristics of a problem
  - **Models are (almost) never true**; they are more or less useful for a given problem

A “useful” statistical model

- Translates a scientific question into a statistical one to learn from data
- Captures the key aspects of the process (biological, social, physical,…) in a small set of unknown parameters or functions
- Adequately consistent with the observations
- Facilitates quantification of the strength of evidence in the data for each of the competing hypotheses
- Statistical models are like prisms; view the data from different perspectives

See talk Dr. Scott Zeger did on “right model” or “best model” or “correct model” here:

www.biostat.jhsph.edu/newEvent/event/lecture.php

What is regression?

Some answers:

1) Methods for describing the dependence of a response (Y) on predictor variables (X) using a sample of data (X_i, Y_i), i=1,…,n.

2) Average Y at each value of X; E(Y | X) = f(X) where f is some function

The name “regression” originated from Galton’s study of the heights of fathers and sons

www.biostat.wisc.edu/~kbroman/talks/regression_ho.pdf

Example 1:

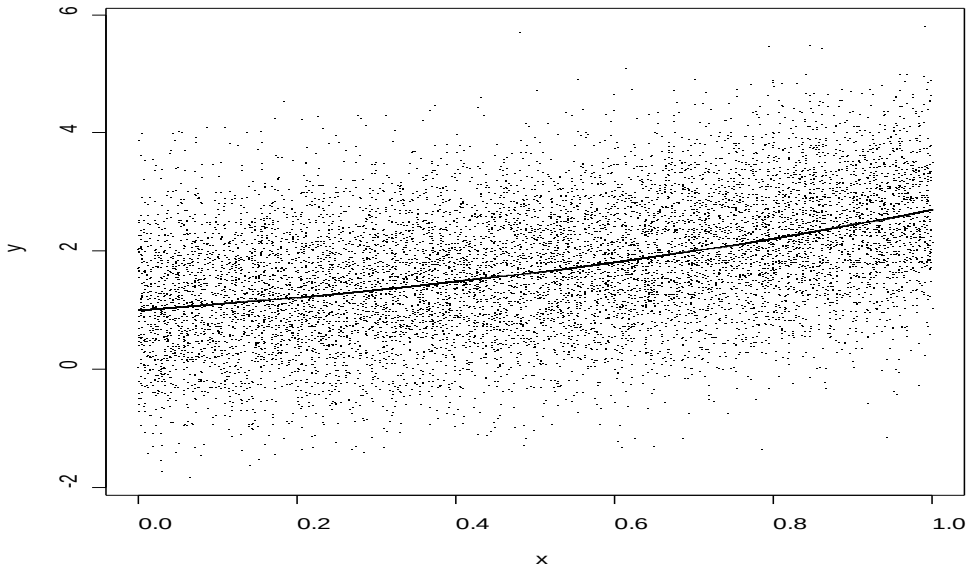
Calculate the regression of Y on X for the population below.

i.e. calculate the average Y at each value of X

X	Y=0	Y=1	Total
0	100	100	200
1	50	150	200
Total	150	250	400

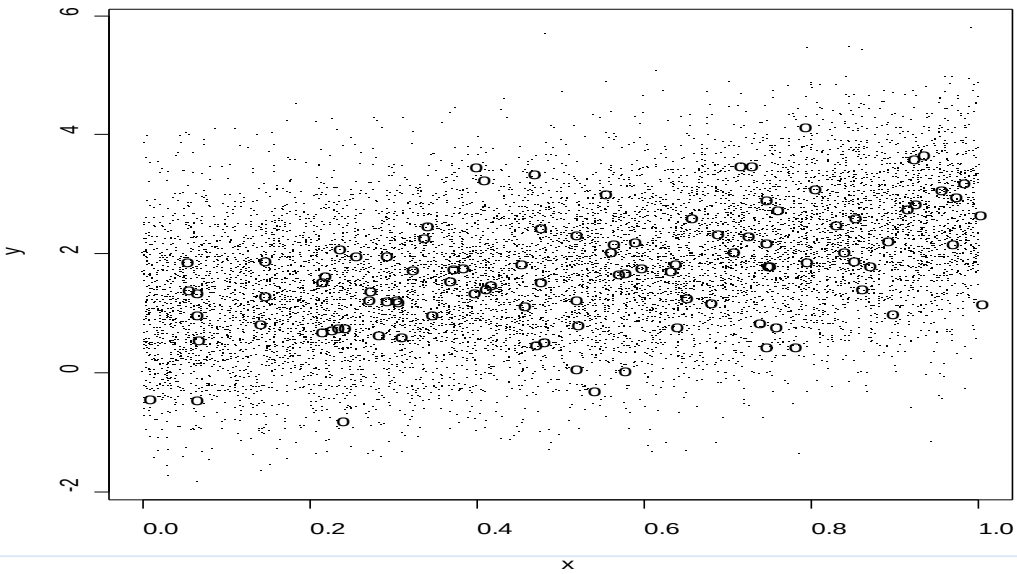
Example 2: Set up

- The figure below displays values of Y and X for a population of interest. The solid line is the E(Y | X). What do you notice about this line?



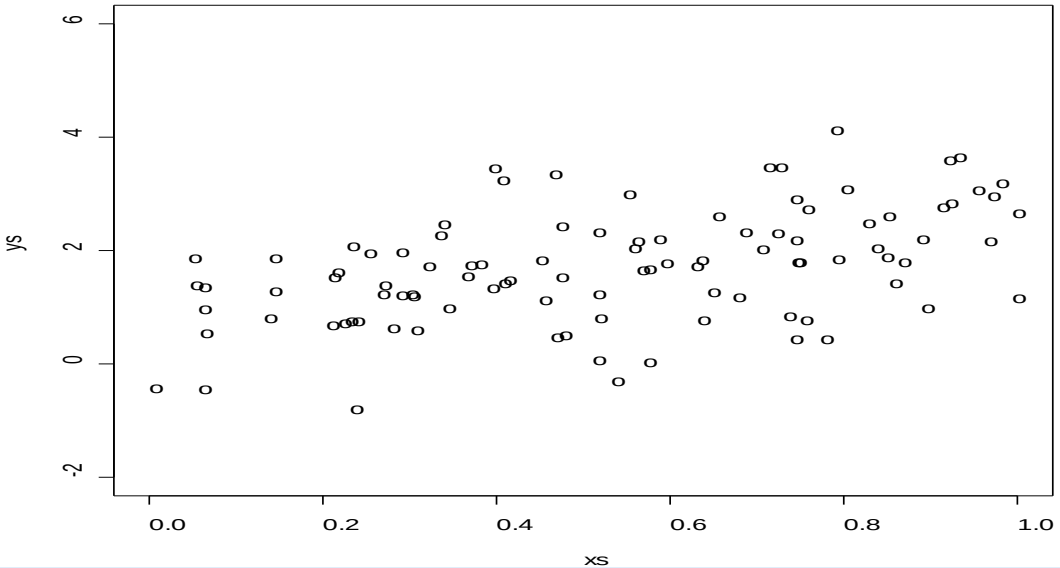
Example 2: Statistical Problem

- We never see the whole population (all the sets of (X,Y)); only see a sample. Then we try to say what the conditional mean of Y given X looks like.



Example 2: Regression estimation

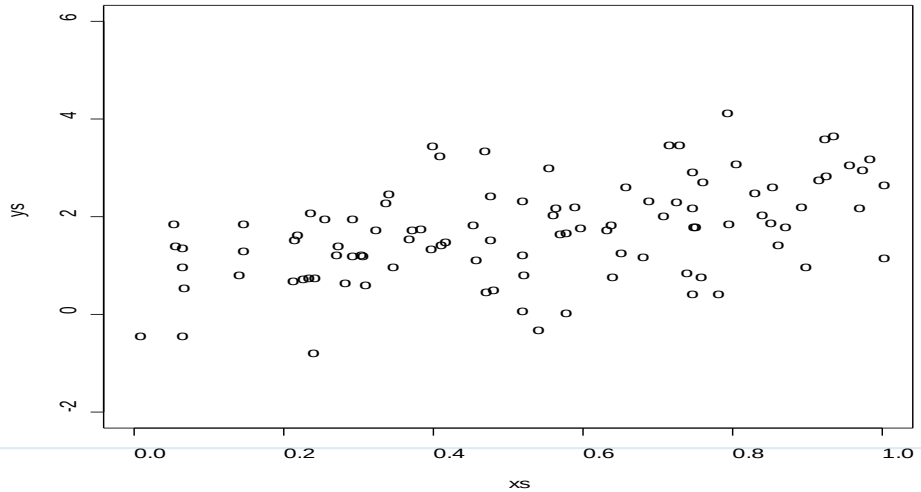
- How would you estimate the population average Y at each value of X given only the sample of data?



Example 2: Regression estimation

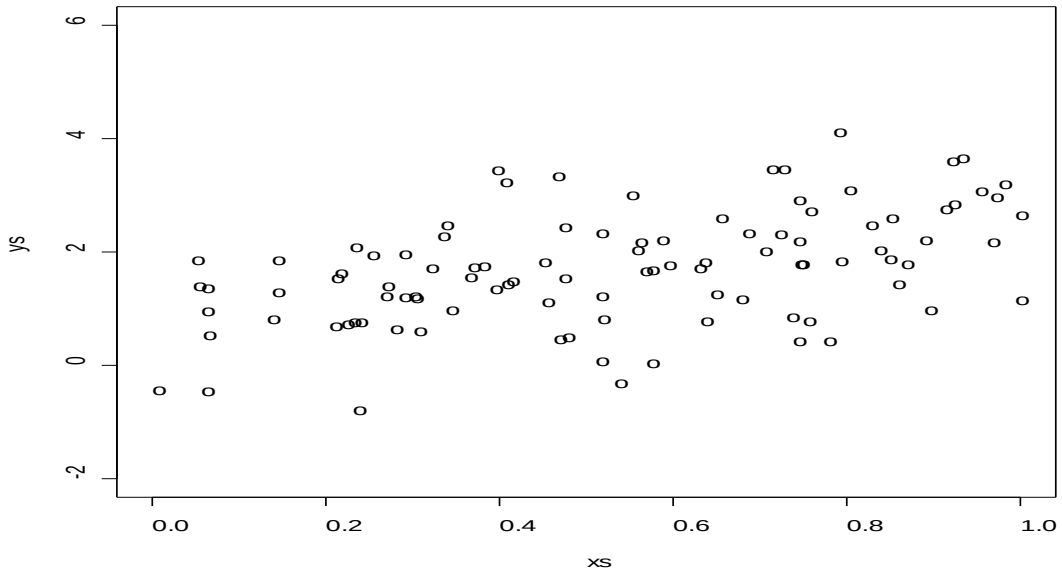
How would you estimate the population average Y at each value of X given only the sample of data?

- Divide the X-space (here axis) into bins; calculate the mean Y within each bin.
- Connect the means across X
- Maybe smooth out the line since we think the true population average Y changes as a smooth function of X.



Example 2: Quantifying model uncertainty

- How can you quantify how far the true regression curve might be from the one estimated from the data set?



Example 2: Model specification and interpretation

How would you answer the question: Does Y increase with X and if so, at what rate?

- Remember: model is a tool to address a scientific question
- Assume: $E(Y|X) = B0 + B1\ X$, average of Y is a straight line function of X
- NOTE: this model is wrong

- How do we interpret B1?

- What if we assume the model is true, what hypothesis can we test to address the question?

- Is the linear model useful in this case?

Two key uses / purposes for regression

1. Study the etiology of a process; how Y is caused by or associated with a set of Xs
   - Let X=(R,C); Study how risk factors R affect the outcome Y while controlling for potential confounders C
   - Let X=(R,C,E); Study how the effects of risk factors R are modified by variables E while controlling for confounders C

2. Predict Y using X

Regression fundamentals are the same for both!

Features of the regression model fit of most interest and strategy for model building can differ depending on your purpose.

Types of regression discussed in 140.653-654

- General: ave(Y|X)
- Linear model: ave(Y|X) = B_0 + sum(j=1,p) B_jX_j
- Additive models: ave(Y|X) = sum(j=1,p) $s_j(X_j)$
- Generalized linear models (GLMs): g(ave(Y|X)) = B_0 + sum(j=1,p) B_jX_j ; g- "link" function
  - Linear: g(u) = u
  - Logistic: g(u) = log(u/(1-u)) = "logit"(u)
  - Log-linear: g(u) = log(u)
  - Probit, tobit, complementary log-log,…
- Generalized additive models (GAMs): g(ave(Y|X)) = B_0 + sum(j=1,p) $s_j(X_j)$
- Classification and regression trees (CART): E(Y|X) is a "step function" in higher dimensional X-space
- Random forests: E(Y|X) is an average of a large number of "bootstrapped" trees

Key datasets

Nepali Children's Anthropometry (NCA) Data

- Cross-sectional nutrition survey of 4,000+ pre-school children
- Height, weight, arm-circumference and age on each
- Questions:

1. How does height vary with age. What is the average "growth rate" over the first 5 years of life?

2. How does shorter-term nutritional status vary by age; are younger children in better or worse status as measured by weight or arm-circumference controlled for height?

3. How well can you predict a child's weight given his height and age?

Key datasets

National Medical Expenditure Survey – Medical costs and smoking-caused diseases

- Now known as Medical Expenditure Panel Survey, conducted by AHRQ
- NEMS 1987 – national survey of 20,000 non-institutionalized adults, included supplemental survey on smoking behaviors
- Key variables: total medical expenditures, presence of smoking-caused disease (Lung cancer, COPD, CHD, Stroke,…), age, gender, SES, smoking status
- Questions:

1. How much more is spent per year on persons with smoking-caused diseases (SCDs) than on otherwise similar persons without SCDs?
2. Does this SCD-attributable expenditure differ by current smoking status or access to health care?
3. How does the risk of LC or COPD depend on the total pack-years of smoking and age?
4. How does the risk of CHD/Stroke change for former smokers as a function of the time since they quit?

Where to next?

- Basic tools for building regression models:

-- indicator variables

-- linear and cubic splines

-- interactions!

- Cover the basics in Lecture 2 and will do some hands-on work in Lecture 3