Lecture 1 Handout

Elizabeth Colantuoni

1/24/2021

Objectives:

Upon completion of this session, you will be able to do the following:

- Describe the scientific method
- Appreciate the role of statistics in applying the scientific method
- Understand the joint probability distribution of random variables X and Y
- Define the conditional distribution of scalar random variable Y given random vector X
- Define and calculate the regression of Y on X
- Understand that a model is a potentially useful approximation to reality
- Follow principles for useful model construction when addressing a scientific question using data

Definitions

1. What is science?

- Search for parsimonious hypotheses (laws) that explain observations of nature
- Search for "truth" (and "beauty")
- John Keats Poem : "Beauty is truth, truth beauty,—that is all Ye know on earth, and all ye need to know."

2. What is the scientific method?

Methodology for generating and interpreting data as evidence to support some competing hypotheses about nature more than others

3. What is statistics?

Principles and quantitative methods for implementing the scientific method

4. What is a 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 a person = population mean + person's deviation from mean: $BP_i = \mu + \epsilon_i$
- The "Model": is an approximation to reality; simplification of reality; cartoon that captures key characteristics of a problem. REMEMBER: Models are (almost) never true; they are more or less useful for a given problem

5. What makes a statistical model "useful"?

- 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 quantitation 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
- Is there a "right model" or "best model" or "correct model" Visit http://www.biostat.jhsph.edu/~cfrangak/cominte/How_to_choose_wrong_model.pdf

6. What is regression?

Answer 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.

Answer 2. Average Y at each value of X; E(Y|X) = fct(X)

Why the name "regression"?

Galton's study of the heights of fathers and sons:

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

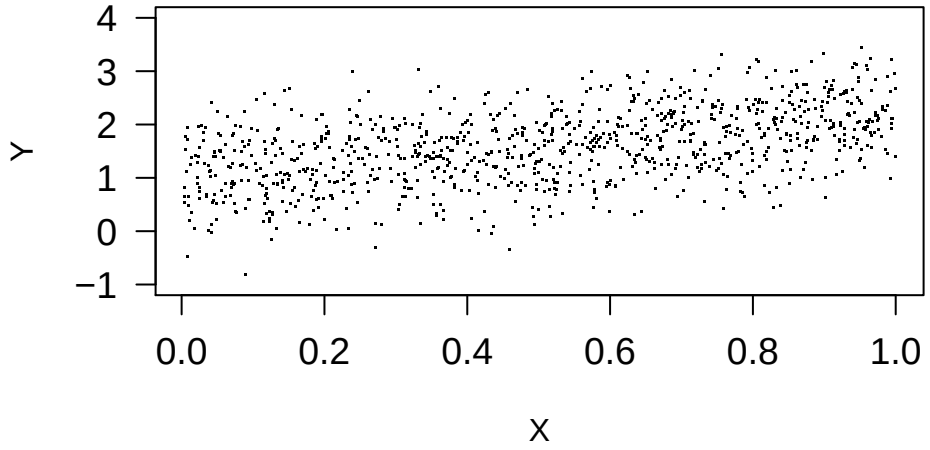
Example 1:

Calculate the regression of Y on X for the population below; average Y at each value of X.

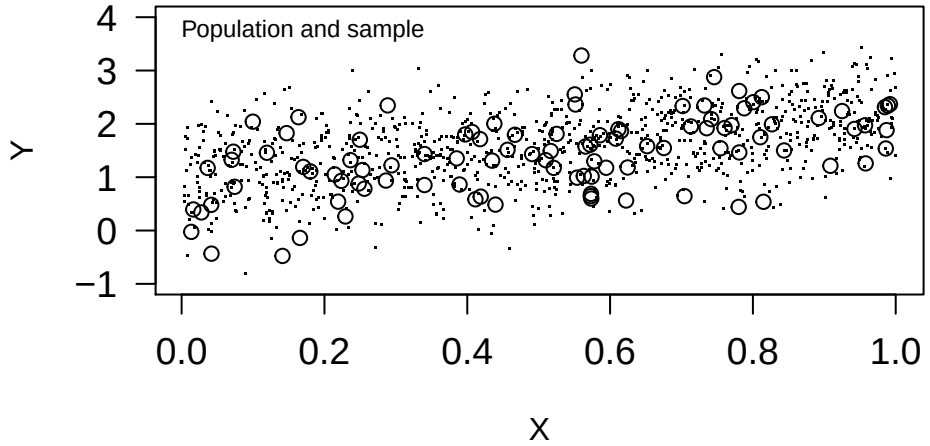
	Y=0	Y=1	Total
X=0	100	100	200
X=1	50	150	200
Total	150	250	400

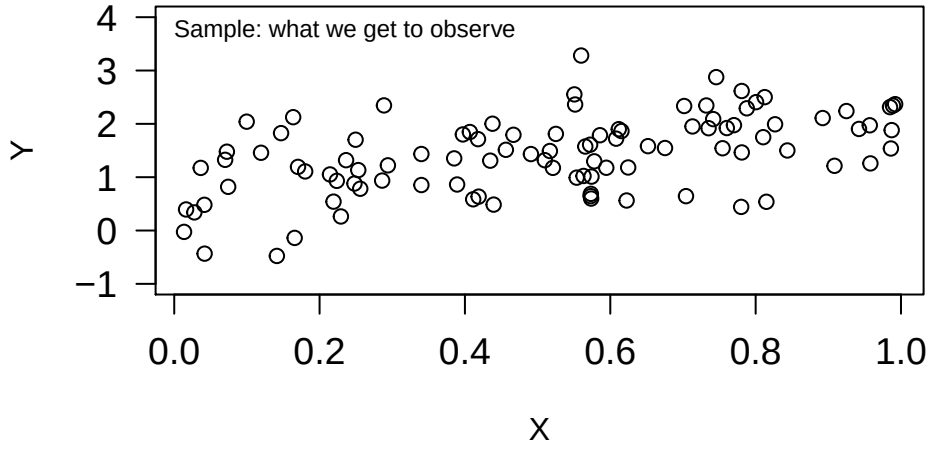
Example 2.

Consider the figure below. The figure represents a population of (X, Y).



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. The sample we observe is displayed in the figure below by the open circles.





1. 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.

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

- Draw a 95% confidence interval about each bin mean and imagine all the smooth curves that cut through the intervals

3. 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: ave(Y|X) = B0 + B1 X, i.e. average Y changes as a straight line function of X (It really isn't)
- But, in this model, B1 is the difference in average Y comparing persons with X = x+1 to persons with X = x. Alternatively, B1 is the change in average Y per unit change in X.

4. If we tentatively assume this model to be true, the scientific question becomes: Is B1 > 0?

- Does the average Y increase as X increases?
- Is Y linearly related (in the positive direction) with X?
- The scientific question has been translated into a statistical one of estimation. We can estimate the true B1 and quantify how far our estimate is from the true value.

5. Is the linear model useful in this case?

- To estimate the average rate of change in ave(Y|X) for X in (0,1)
- Some evidence the slope is smaller near 0 and greater near 1.

7. What are the purposes of regression?

There are two main purpores of regression:

- 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
- Predict Y using X

8. What are the types of regression of Y on $X = (X_1, X_2, \ldots, X_p)$ discussed in this course?

- General: ave(Y|X)
- Additive models: ave(Y|X) = sum(j=1,p) sj(Xj)
- Linear model: ave(Y|X) = B0 + sum(j=1,p) BjXj
- Generalized linear models (GLMs): g(ave(Y|X)) = B0 + sum(j=1,p) BjXj; 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)) = B0 + sum(j=1,p) sj(Xj)
- 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

Two Examples for 653

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:
  - How does height vary with age. What is the average "growth rate" over the first 5 years of life?
  - 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?
  - How well can you predict a child's weight given his height and age?

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