STATISTICS 452/652: Statistical Learning and Prediction

August 19, 2020

Regression Review Part 1: Intro Stat

(Reading: ISLR Sections 3.1–3.4)

1 Goals of Lecture

I assume you've had two courses in statistics: an intro course that is pretty standard and a course featuring regression. In this lecture we will:

1. Review some of the most important ideas from your *first* course in statistics
   (a) What is a "model"
   (b) sampling distributions
   (c) standard errors
2. Review important features about linear regression, from your second course
   (a) What the model means
   (b) How model parameters are estimated
   (c) Sampling variability
   (d) Ways to extend the model to fit other shapes
3. Use these ideas to reconstruct what "regression modeling" is meant to achieve
   (a) Why are models wrong?
   (b) How are models wrong?
   (c) How do we use this information to guide our efforts?

2 Review of Introductory Statistics

The typical intro-stat course contains the following. You should recall and remember all of it, or review if there is anything you don't recognize.

- Basic probability
- Parameters: population quantities of interest
  - Means, standard deviations, probabilities of events
- Statistics: estimates of parameters
  - Sample mean, sample variance and standard deviation, sample proportions
- Distributions, including normal distribution
  - Empirical rule
- Measuring variability of statistics
  - Standard errors
  - Sampling distributions, central limit theorem
- Hypothesis tests
  - Null/alternative hypotheses
  - test statistics
  - Type 1 errors, α, p-value
  - Interpretations
- Confidence intervals
  - Coverage
  - Interpretations

2.1 Distributions as Models

In the introduction, I said that classical statistics starts with a MODEL. So what does that mean?

- Classical statistics starts with POPULATION of units and some particular measurement on those units
  - A population is just a word that means "the collection [or SET] of all possible units"
    - Equivalently it is the collection of all possible values that the particular measurement can take on

  - Ex: CGPA in first year for students entering Canadian universities from high school in 2010s
    * Population is students entering Uni from HS in these years
    * *OR* population is the CGPA after first year for these students

- A STATISTICAL MODEL is a probability distribution that we *assume* describes the measurements in the population
  - The probability distribution is just a **mathematical formula** that describes the shape of the histogram of measurements in the population
  - May be discrete (takes on only certain values) or continuous (takes on any value within an interval)
  - Usually chosen for combination of mathematical convenience and good fit
- **A MODEL IS JUST AN APPROXIMATION!**
  - There is no reason that real populations must adhere to man-made mathematical constructs
  - *No* population truly follows a normal distribution *exactly*
    * Implies that every measurement is taken to infinite number of decimals
    * Implies no lower or upper limits to size of measurement
  - **"All models are wrong. Some are useful." -George Box.**
  - *Many* populations are reasonably approximated by normal distributions
    * Central mound
    * Roughly symmetric tails on each side
- The model allows us to do probability calculations that would otherwise be impossible
  - Empirical rule
    * 68% ($\approx 2/3$) of population is within 1 standard deviation (SD) of the mean
    * 95% is within 2 SD of the mean
    * 99.7% is within 3 SD of the mean
  - Makes it possible to do tests with specific α and to compute p-values
  - Allows us to have XX% confidence that a confidence interval will cover a parameter.
  - ALL of classical statistical inference is based on models
- Calculations based on models give great answers when the approximation is good
  - Reliability of answers deteriorates when approximation becomes worse
  - Different statistics are more (or less) sensitive to vioplations from the model assumptions than others

- It's easy to check the quality of a model fit in simple problems
  - It can be impossible to check it in complex ones.
  - *Guess which type we will study?*

2.2 Effects of Sampling on Statistics

- When we "do statistics" we start by identifying one or more PARAMETERS we are interested in learning about
  - Parameter is just a quantity that could be computed from the population
  - Ex: Average CGPA or fraction of CGPAs below 2.0 among all students in the population.
- We draw a sample of "n" units from this population and measure them
- We use the sample to compute a statistic that estimates the parameter
  - Call this an "estimate"
  - Ex: average CGPA in sample or proportion of sample with CGPA<2.0
- **It is important to understand that the estimate we compute depends on the sample we draw!**
  - If we sampled a different set of n units, the estimate would be a little different
  - In fact, every different sample of n units has the potential to give a different estimate
- We could consider making a histogram of all the different possible values that a statistic can take on
  - The population histogram of all possible values that a statistic could take from a sample of size n is called THE SAMPLING DISTRIBUTION OF THE STATISTIC.
  - See the demonstration here: http://onlinestatbook.com/stat_sim/sampling_dist/
    - You will get to play with this in the homework.
- In some cases, mathematical analysis of a model tells us about the sampling distributions for certain statistics
  - The sample mean of a sample from a normal distribution also has a normal distribution
  - The CENTRAL LIMIT THEOREM (CLT) tells us that that sample means of samples from *ANY* distribution have an approximately normal sampling distribution.
    - The approximation is closer if the population distribution is closer to normal
    - The approximation is closer when the sample size is larger.

  - This is what we use for all of our inferences—tests and confidence intervals—in classical statistics.

- Even when we don't need to do formal inference, we still should have some measure of uncertainty of a statistic
  - A single "point estimate", like "100" is of limited use
  - Is it 100 ± 0.1 or 100 ± 50?
  - The difference is very meaningful for how we interpret the results and how we act upon them.

- We can get a measure of uncertainty of a statistic from its sampling distribution
  - If the sampling distribution of an estimate is very concentrated, then we have a good idea where the true parameter is.
  - If it is spread out, then we aren't so sure

- The standard deviation of a statistic's sampling distribution is called the STANDARD ERROR (SE) OF THE STATISTIC
  - Some simple statistics (means, proportions, estimated linear regression parameters) have simple formulas for standard errors
  - The SE for MANY more complicated statistics have complicated formulas or no formulas

- **In all cases, though, the larger n is, the smaller the SE.**
  - Estimates get "more precise (less variable) as the sample size increases
  - *This is the key thing I want you to remember about SEs, because this principle will carry through everything we do.*

3 Exercises (Due Friday of next week)

1. Visit http://onlinestatbook.com/stat_sim/sampling_dist/ and play with this app as described below. This is meant to help you better understand how sampling distributions work.

   (a) Start with the normal distribution that is represented on the initial screen as your "parent population". Select "Mean" in the third plot, "Variance" in the fourth plot, and "N=5" for both. Run 10,000 samples through this simulation. Take a screenshot of the results showing all four plots, including the statistics on the left and the settings on the right. Present it as your response. No comment needed yet.

   (b) Clear everything and repeat with "N=25" in both places.

   (c) Clear everything and repeat with "Skewed" in the first plot and "N=5" in the third and fourth.

   (d) Clear everything and repeat with "Skewed" in the first plot and "N=25" in the third and fourth.

   (e) Comment on the following, explaining what evidence these plots provide to support your answers. Please note that the scales on the X-axis sometimes change, so factor this into your explanations when necessary. You shouldn't need more than a sentence, or *maybe* two for each. If you say too much, you are missing the big points and will be penalized.

      i. Do different statistics computed on the same samples have to have the same sampling distribution?

      ii. What effects does increasing sample size have on the sampling distributions of statistics?

      iii. What effects does changing the parent distribution (the distribution from which data are sampled) from normal to skewed have on the sampling distributions of statistics?