POLS 503: Advanced Quantitative Political Methodology: The Notes

Jeffrey B. Arnold

2016-04-04

Contents

Chapter 1

Introduction

hello, world!

Chapter 2

Review of Statistics

Before getting started it is worth recalling a few concepts and terms from statistics.

2.1 Terms

2.1.1 Statistic

A **statistic** is a function of sample data. Let $x = (x_1, \ldots, x_N)$ be our sample data. Examples of statistics include:

mean $\text{mean}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$
variance $\text{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - mean(x))^2$

Statistics do not need to be single numbers. The sample data x is, itself statistic, albeit output of the use.

2.1.2 Parameter

A **parameter** is a function of the population. While the value of a statistic depends on a particular sample drawn from the population, the value of a parameter is fixed. Perhaps the most commonly used parameter is the population mean, $\text{E}(X)$. Another example is the population variance, $\text{Var}(X) = \text{E}((X - E(X))^)$.

A single population can have many (infinite) parameters that we could consider, although in practice, inference usually is concerned with only a few of these, such as the mean or variance, or parameters by which a distribution is commonly parameterized. A parameter usually represents one particular aspect of the population, not all the features of the population; like a a statistic, it throws away information in doing so. And like a statistic, by throwing away this information, we will often be able to proceed

While statistics are random variables, parameters are a fixed feature of the population. In statistical inference we use statistics of the sample to make inferences about parameters of the population.

2.2 Estimators and Estimates

An **estimator** is a statistic (function of the sample data) used to estimate a population parameter. An **estimate** is the value (number) of an estimator for a specific sample.

Many statistics can be used as an estimator for a given parameter, but not all of them will be good estimators. For example, the sample mean, sample median, sample variance, and always guessing 0 are all estimators of the population mean. However, we will prefer using the sample mean as the estimator of the population parameter, for reasons other than it having the sample name To understand why we prefer some estimators over others, we need to consider the sampling distribution of the statistic.

2.2.1 Sampling distribution

Since a statistic is a function of the sample, it varies sample to sample. The distribution of the sample statistic over repeated samples from the sample population is called the **sampling distribution**.

2.2.2 Standard Error

The standard deviation of a sampling distribution is called the **standard error**. The larger the standard error, the variable the statistic is acrros samples. All else equal, it is preferrable to have a statistic with a lower standard error, since it is less liable to vary wildly acrros samples.

Since the standard error is a property of the population distribution from which the samples were drawn, this means that the standard error of a statistic is a *parameter*. But, then, what are the standard errors that are returned by statistical software? The standard errors that are calculated from a sample are an estimate of the population standard error. This means that in addition to the properties of the statistic of interest, we can, and often have to evaluate, different methods of calculating the standard error of the stampling distribution of that statistic.

2.2.3 Methods for evaluating statistics

- Bias
- Variance
- Mean squared error
- Consistency
- Robustness / resilience

2.3 Example

TODO

Chapter 3

Simple Linear Regression

- Interpret the

Simple linear regression, also called *bivariate linear* regression is linear regression with one outcome variable and one covariate,

$$Y = \beta_0 + \beta_1 X$$

3.1 Conditional Expectation Function

When we want to describe an outcome or estimate the effect of a covariate on an outcome, one way to model this is through a conditional expectation—what is the expected value (mean) of the outcome for a given value of the covariate(s).

The **conditional expectation function (CEF)**, also called the **regression function** of Y given X is the function that gives the expected value (mean) of Y at various values of x. It is denoted $\mathrm{E}(Y|X = x)$.

Note that the CEF is a function of the population. In the same way that the expected value of a distribution is a parameter, the conditional expected values of the distribution given values of x is also a parameter.

Like other parameters, the CEF throws away a lot of data. It does not describe all the ways that the distribution of Y changes as a function of x; it only describes how the *mean* of Y changes as a function of x.

3.2 Ordinary Least Squares

Ordinary least squares is an estimator of the slope an intercept of the regression line. It finds the slope and intercept that minimizes the sum of the squared residuals. The estimates

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1) &= \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2 \\ &= \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ &= \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{\epsilon}_i^2\end{aligned}$$

where, $\hat{y}_i$ are the fitted values,

$$\hat{y}_i = b_0 + b_1 x_i$$

and $\hat{\epsilon}_i$ are the esimated residuals,

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

INSERT PLOT

A nice feature of OLS is that that the solution to the minimization problem can be found analytically.

The solutions to OLS are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var} x}$$

In other words the regression slop is the covariance between the outcome and predictor variable, scaled by the variance of the predictor variable. Since $\operatorname{Cor}(x, y) = \operatorname{Cov}(x, y)/(\operatorname{sd} x \operatorname{sd} y)$, the regression slope can be expressed as a function of the correlation between x and y,

$$\hat{y} = \operatorname{Cor}(x, y) \cdot \frac{\operatorname{sd} x}{\operatorname{sd} y}.$$

Questions:

- How does the slope of the regression vary with the sample correlation? covariance?
- What happens when x doesn't vary? In other words $x_1 = x_2 = \cdots = x_n$. Answer this using the equation, and by sketching what this data and the regression line would look like?
- What happens when y doesn't vary?

In more complicated problems, the soluations have to be found through an iterative optimization problem.

3.2.1 Mechanical properties of the OLS statistic

These properties are related to the algorithm used the estimate OLS, and are not directly related to how well OLS works as an estimator. These properties follow directly from the first-order conditions used to find the OLS estimates.

1. Residuals are zero on average. And by implications, the residuals sum to zero.

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\epsilon}_i = 0$$

2. The residuals are uncorrelated with the covariate,

$$\operatorname{Cov}(x_i, \hat{\epsilon}_i) = 0$$

3. The residuals are uncorrelated with the fitted values

$$\operatorname{Cov}(\hat{y}_i, \hat{\epsilon}_i) = 0$$

4. The regression line goes through $(\bar{y}, \bar{x})$.

Note that 2 and 3 are properites of the estimated residuals, not the population residuals. If the population residuals are correlated with the covariate or outcome, then OLS estimator will have problems.

3.2.2 OLS slope is a weighted sum of the outcomes

The OLS estimator for the slope, $\hat{\beta}_1$, can be written as a weighted sum of the outcomes,

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i$$

where

$$w_i = \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This is important for several reasons. Even though the regression line includes all observations in its calculation, not all observations are equally important in determining the regression line. Those observations far from the average value of x are given more weight.[^ch2-outlier]

See the paper by XXXX on panel data and regression weighting.

This also implies that $\hat{\beta}$ is a random variable, since it is the sum of random variables, Y_i.

For the derivation, see http://www.mattblackwell.org/files/teaching/s07-simple-regression.pdf

3.3 Properties of the OLS Estimator

Remember, OLS is a statistic—a mechanical function of the data. We plug in data and get a result. It may be more complicated than sample means, variances, medians, or other summary stats seen earlier, but it is analgous.

Since it is a function of the sample data, an the sample is a random variable, the OLS statistic has a sampling distribution. This sampling distribution will determine how well the OLS statistic works as an estimator of the population linear regression line.

INSERT FAKE DATA EXAMPLE

3.4 Goodness of fit

How well does the regression line fit the data? There are two commonly used statistics to judge this.

- prediction error, or the standard error of the regression
- R^2 (R-squared) and adjusted R^2

3.4.1 Prediction error

3.5 Properties of the Estimator

- Unbiased
- Consistent
- Asymptotically normal
- Efficiency. An efficient estimator is the estimator with the lowest *standard error*. OLS is BLUE—the Best Linear Unbiased Estimator. This is proven by the *Gauss-Markov Theorem*. For the set of the class of linear, unbiased, estimators (with a few regularity conditions), OLS is the efficient estimator.

Figure 3.1: OLS Is BLUE (Best Linear Unbiased Estimator)

3.6 Assumptions for unbiasedness and consistency of OLS

1. Linearity
2. Random (i.i.d.) sample
3. Variation in x_i
4. Zero conditional mean of the errors

These assumptions are about using the OLS statistic as an estimator of the population linear regression line. The only assumption that is necessary to calculate a sample regression line is variation in x. If the other assumptions are not met, then the sample OLS statistic can be calculated, but it may be a poor estimator of the population regression line.

3.7 References

- Much of this is taken from Matthew Blackwell lecture notes for Govt 2002 http://www.mattblackwell.org/files/teaching/s07-simple-regression.pdf
- Tobias Funke meme http://memegenerator.net/instance2/5082792. Generated by Jeffrey Arnold. Hattip to Brenton Kenkel for the refence. http://bkenkel.com/psci8357/notes/02-reintroduction.pdf.

Bibliography