The Simple Analytics of Regression Inference in a Finite Population

Richard Startz

Douglas Steigerwald[[1]](#footnote-1)\*

October 2020

Revised November 2020

Abstract

We provide asymptotic variances of the regression estimator in a sample taken from a finite population considering several different sources of uncertainty. The general model we present here nests several popular specifications of the regression model, encompassing both the classical regression and the potential outcomes approaches. Derivation of the required formulas is fairly trivial. We emphasize that some parameters required to compute the asymptotic variance are not identified by the data, but may be pinned down approximately by economic theory. We also emphasize that the correct choice of variance estimator depends in part on the statistical question being investigated. In particular, there may be a difference in the estimator applicable for internal versus external validity.

It is useful to note that this paper is now complete except for a few minor details, to wit:

1. The results are for a 1-RHS variable regression. I hope it is relatively easy to convert to a *k-*RHS variable regression. Not sure how to do this. There is an issue of some columns being attributes and others treatment.
2. The simple clustered result is done, but ones allowing for more things to be random look like the algebra might be messy.
3. In some cases, it may be possible to re-arrange formulas to give bounds. Not sure.
4. Tying all this into AAIW is important…no idea how to do it. We need to make clear the marginal contribution while also keeping AAIW happy.
5. I believe extending this allow for heteroskedasticity is trivial. I think it just involves unsimplifying some formulas.
6. I am not sure how to handle non-constant treatment effects. I’m not sure how to think of ATE when the RHS-variable is not just 0/1. More importantly, if is not constant we end up with being random which adds a cross-term everywhere.
7. It is customary to include an empirical application. Seems stupid in this case, but maybe we need to. Note that AAIW just do simulations, which we could do as well.
8. The paper is in Word. I imagine at some point when you take over control you would prefer it in TeX.

If you can figure out 1-6, I can take care of 7 and 8.

# Introduction

With rare exception, economists model their statistical data as a draw from an infinite population. As a practical matter, populations may be large or small but are almost always finite. The real issue for frequentist inference is how the sources of random variation enter the problem. What analysis of finite populations adds is that if sample sizes are close to the size of the population, then (possible hypothetical) samples will consist of mostly the same observations, so there will be little variance across estimators drawn from different samples. In contrast, if the sample size is small compared to the population, then the usual inferential rules apply.

In this introduction we lay out a fairly general regression model. We then discuss how various assumptions about the data generating process allow nesting of the classical regression model and the classic potential outcomes model. We discuss the how inference may be different in what is sometimes termed internal versus external validity. In the next section, we derive asymptotic variances for the general case and then illustrate for interesting special cases. We emphasize that once set up properly, *derivation of asymptotic standard errors is fairly trivial.* In the section which follows we provide a similar analysis for the case of clustering.

As a first step, we wish to acknowledge our intellectual debt to a series of papers by Abadie, Athey, Imbens, and Wooldridge (2014, 2017, 2020) (henceforth AAIW). While we extend their results, we have endeavored to keep our notation as close to AAIW as possible.

We begin by defining the data generating model and the sampling model. The finite population is generated as follows

where are independent for all , are independent, and are independent.

Note that there are two explicit sources of variation, which we might call heterogeneity and randomness. Data differs across observations (heterogeneity) and data differs across (possibly hypothetical) generations of the population (randomness). In another words, heterogeneity means that a particular variable may be different for observation than for observation . Randomness means that if we were to return and resample observation , we might get a different value. As a first step, we show how this DGP nests various common ways of thinking about a regression model.

Consider first a potential outcomes framework a la Neyman/Rubin/Imbens. In the potential outcomes model there is a different value of for each value of often just two possible values so we are interested in . For this to be an interesting statistical problem, different observations need to have different potential outcomes. But potential outcomes should be the same if the observation is resampled. In our framework, this means that is permanently assigned to observation , which we can write as and . In other words, is heterogenous but not random. Of course, in many situations modelled in the potential outcomes framework, is random as well as heterogenous. In many RCTs if we were to revisit and retreat person with the same value of , we would expect to change because of unaccounted for random factors (hopefully) orthogonal to the treatment.

Consider next the classic regression model. In the classic regression model is a random variable, typically with expected value zero. Thus we would specify and . In other words, is random, but not heterogenous.

The remaining issue is the behavior of . For identification, must be either heterogenous or random or both. AAIW label as an “attribute” if is heterogenous but not random. The traditional—if somewhat inaccurate—example would be gender, where would vary and . AAIW label as a “treatment” if it is random, The classic example is random assignment of treatment in an RCT, where usually is constant across , although this last is not necessary.

Analysis of finite populations introduces a new source of randomness, drawing (possibly hypothetical) samples from the finite population. Let be a vector of independent Bernoulli random variables with We assume to be independent of both and and to be independent across repeated draws. If , observation is included in the sample and if the observation is not included.

Thus we are potentially faced with two sources of heterogeneity, and , and three sources of randomness, , , and . In what follows we call variability in “treatment variability,” variability in “outcome variability,” and variability in “sampling variability.”

In the original work on finite populations, Neyman (1923), Neyman dealt only with sampling variability and outcome heterogeneity. He held and—since he studied the sample mean—he had . Neyman also specified a fixed sample size. We instead specify a fixed probability of an observation being selected. This enormously simplifies the analysis, indeed makes the details fairly trivial. The cost is that we offer asymptotic results only.

The Neyman result makes it easy to see the intuition of why sampling variability matters in finite populations. Suppose the sample is actually a census, that is the sample is the entire population. One calculates a sample mean. One then takes a second sample of the entire population. The observations are the same, so the calculated mean is the same. The variance of the calculated mean is zero. In contrast, if the sample size is small compared to the population size few observations will be repeated and the variance will be close the standard result.

But consider a different example. Individuals in a tour group enter an establishment with equal amounts of money. On departure, they are surveyed to find out how much money they depart with. The statistician computes the sample mean. How should inference be conducted? The answer to this critical question cannot be given simply by looking at the data.

Suppose that the tourists have been buying tickets to attractions inside the establishment. The Neyman-style finite population correction gives the right answer if the question is that is the mean spending among this group. There is internal validity. Now suppose we wanted to predict spending of the same group of tourists on another day, with the necessary assumption that the behavioral characteristics remained unchanged. The Neyman correction also gives us correct inference for this group, so there is external validity. In contrast, suppose that rather than buying attraction tickets the tourists had been exploring the one-armed bandit problem. In this case variability is entirely due to outcome variability. If one wishes to calculate winnings for the group of tourists on this particular visit, what AAIW call a “descriptive estimand,” then the finite population is correct—there is internal validity. However, if one were interested in outcomes of the group on the second visit, then only the standard estimate, , would maintain external validity.

# The Regression Model

We have claimed that derivation of the asymptotic variance of the least squares estimator is fairly trivial. Write the estimator as

|  |  |
| --- | --- |
|  |  |

If we define

then uncertainty about arises only from the second term in equation (1), the ratio . We can find the asymptotic variance in two steps. First, find the asymptotic variance of , which is easy since is linear in independent variables. Second, apply Slutsky’s theorem to find . It is useful to define notation for the population regression

We can write the population regression coefficient as . It follows immediately that .

We now make the usual assumption that so that is a consistent estimator. Note also that since , we have and . It follows immediately that .

To find the asymptotic variance of we twice deploy the standard result on the variance of the product of independent random variables, . First, the variance of .

Next, note that

Since and since is Bernoulli with variance , we have

Simplifying the algebra by collecting the last two terms and noting that variance plus square of the expectation is the expectation of the square

Since the observations in the population are independent, the variance of the sum in is the sum of the variances,

Using Slutsky’s theorem and simplifying some we have

|  |  |
| --- | --- |
|  |  |

where is the expected number of observations.

Not all the terms in equation (2) can be estimated from the data. For example, while the sum of is observable for observations in the sample and the number of observations gives an unbiased estimate of , neither can be inferred. But presumably the investigator has an opinion from other sources about . In contrast, finding and in the completely general case is more difficult. However, the required values can be inferred from the data in special cases, and some bounds are available. We turn now to these special cases.

##### Sampling randomness only, :

We have , so we get the traditional finite population correction to the Eicker-Huber-White variance. Note that in this case that .

|  |  |
| --- | --- |
|  |  |

##### Outcome randomness only, :

We have and .

Note that so in a large sample this is the classic regression variance formula

##### Treatment randomness only, :

We have , and .

|  |  |
| --- | --- |
|  |  |

Note that when . In a large sample this is the classic regression variance formula plus an adjustment term reflecting the relative contributions of treatment heterogeneity and treatment randomness.

Consideration of treatment randomness provides an example in which the investigator may have information not revealed in the data but which is known from the experimental design. Suppose that , which is to say the treatment is applied randomly with probability . In this case and . The formula for the asymptotic variance simplifies to

##### Sampling randomness and treatment randomness, :

We have .

|  |  |
| --- | --- |
|  |  |

Note that this is very close to the sum of the variances in equations (3) and (4). Note also that in the case of a randomly assigned binary treatment variable, this simplifies to

##### Sampling randomness and outcome randomness, :

We need

We have . We have

|  |  |
| --- | --- |
|  |  |

This simplifies to

##### Outcome randomness and treatment randomness, :

We have and .

|  |  |
| --- | --- |
|  |  |

Again, this simplifies in the case of a randomly assigned binary treatment,

# Clustering

The analysis of the clustered case is similar; the two changes are that we need to consider randomness across clusters as well as randomness across observations and that the asymptotics are taken with respect to the number of clusters rather than the number of observations. Draws within a cluster may be correlated.

Indexing clusters by we can write the population regression coefficient as

Cluster selection depends on , where and that is independent of the other random variables. Observation in cluster is selected iff . We have , , and .

We can write the sample regression coefficient as

Note that and .

We begin with the simple case in which the only random element is sample selection. We have

|  |  |
| --- | --- |
|  |  |

Because clusters are independent, if follows immediately that

|  |  |
| --- | --- |
|  |  |

Applying Slutsky’s theorem we have

|  |  |
| --- | --- |
|  |  |

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014): “Finite Population Causal Standard Errors,” Technical Report, National Bureau of Economic Research.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): “When Should You Adjust Standard Errors for Clustering?” Technical Report, National Bureau of Economic Research.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2020): “Sampling‐Based versus Design‐Based Uncertainty in Regression Analysis,” *Econometrica,* Vol. 88, No. 1, 265-296.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition, Wiley, New York.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press, Baltimore and London.

Manski, C. F. and J. V. Pepper (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics* 100(2), 232–244.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. section 9. Statistical Science 5(4), 465–472.

Appendix with Boring Algebra

The basic rule for the variance of independent random variables is that

The basic rule for the asymptotic variance of a ratio, , where the numerator, , converges in distribution to an asymptotic distribution with variance and the denominator, , converges in probability to , then Slutsky’s theorem tells us

We are interested in a regression in a population of the form

The population regression coefficient is

It is useful to substitute in for as in

|  |  |
| --- | --- |
|  |  |

In order to conduct frequentist inference, we need to specify distributions across (hypothetical) draws of the population. We will use the notation

Going forward we assume are mutually independent.

It is useful to define notation for the numerator and denominator in the last term in equation (10).

Going forward, we make the usual least squares assumption that

In fact, we are going to make the slightly stronger assumption that.

Because and are independent, we can write the expectation of the numerator in the first term in equation (10) as

If there is a homogenous treatment effect, , then . If there is a heterogenous treatment effect *and* treatments equal 0 or 1, then is the average treatment effect. More generally, is a weighted average of individual treatment effects. In general, the target of a regression on data sampled from the population is .

Now turn to sampling from the population.

Let be a Bernoulli random variable with mean . We assume the are mutually independent and that is independent from . As a useful reminder, .

The sample regression coefficient is

|  |  |
| --- | --- |
|  |  |

It is useful to define notation for the numerator and denominator in the last term in equation (11)(10).

Note

Note also

It follows then that

From this point on, we assume a non-random treatment effect, .

We first consider the case of a homogenous treatment effect, so . This implies and also that only the second term on the right of equation (11) is random.

We need to find the variance of . Re-write this as . Note first that

From the formula at the beginning of the appendix we have

Note now that

And applying the variance of a product formula again

Simplifying the algebra by collecting the last two terms and noting that variance plus square of the expectation is the expectation of the square

Note that the last term is a statement about repeated sampling of the observation rather than an expectation taken across the population.

Since the observations in the population are independent, the variance of the sum is the sum of the variances

Now we turn to the use of Slutsky’s theorem.

|  |  |
| --- | --- |
|  |  |

Define the expected number of observations as . We can re-write equation (12) as

|  |  |
| --- | --- |
|  |  |

Consider now several special cases:

##### Sampling variation only, :

We have , so we get the traditional finite population correction formula. Note that in this case that .

|  |  |
| --- | --- |
|  |  |

##### Outcome variation only, :

We have and .

|  |  |
| --- | --- |
|  |  |

Note that so in a large sample this is the classic regression variance formula

##### Treatment variation only, :

We have , and .

|  |  |
| --- | --- |
|  |  |

Note that when . In a large sample this is the classic regression variance formula plus an adjustment term.

##### Sampling variation and outcome variation, :

We have , since .

|  |  |
| --- | --- |
|  |  |

As , this simplifies to

Note that the second term is the standard robust variance formula with a finite population adjustment. Observe also from a single sample it is not possible to identify from the data.

##### Sampling variation and treatment variation, :

We have .

|  |  |
| --- | --- |
|  |  |

Note that this is very close to the sum of the variances in equations (14) and (16).

##### Outcome variation and treatment variation, :

We have and .

|  |  |
| --- | --- |
|  |  |

1. \* Department of Economics, University of California Santa Barbara. Thanks for Matthew Grant for bringing these issues to our attention. [↑](#footnote-ref-1)