Inference

Devesh Raval

What is the point of inference? In general: two main goals:

Have some parameter of interest θ.

1. Hypothesis Testing: Can we use the data to determine whether θ is equal to a particular value? Since the data is random, the decision is random, and we may make mistakes sometimes.

2. Confidence Intervals: Can we use the data to construct an interval that contains θ? Since the data is random, the interval is also random, so it may not contain θ all of the time.

1 Sampling Distribution of Estimator

Let us define an estimator $T(Z_n)$- potentially an estimator of our parameter of interest. In general we want to know its properties- for example, what is its distribution under the null hypothesis? Alternative hypothesis? Then we can try to test hypotheses. Some examples of the estimator:

Estimator $T(Z_n)$

$$\begin{aligned} Z_n &= \{X_i, i = 1, ...N\} \\ T_N &= \frac{1}{n}\sum_i X_i \\ T_N &= \frac{\frac{1}{n}\sum_i (X_i - \bar{x})(Y_i - \bar{y})}{\frac{1}{n}\sum_i (X_i - \bar{x})^2} \end{aligned}$$

The first example is the sample mean, the second example is the bivariate regression slope coefficient.

The sampling distribution depends on four things:

1. The form of $T(Z_n)$
2. Sampling process (for example, iid sampling)
   (a) Other examples- balls out of an urn with replacement or without replacement, autocorrelation so not independent
3. Underlying Population Distribution P

(a) In regression context, P would be the distribution of the errors- normal (CNRM) or exponential or uniform, etc.

(b) In random sampling, if $X \sim P$, then $Z_n \sim F_N \sim P * P * .. * P = P^N$

4. Sample Size (N)

Example 1. Sample mean of normal data

The population distribution of X is $N(.75, .1875) = P$.

Draw a random sample, size N, from P.

Let $T_N = \frac{1}{N}\sum_i x_i = \bar{x}$ and $N = 3$.

Remember the Sample Mean Thm- $\bar{x} \sim N(\mu, \sigma^2/n)$. In this case, we then know that:

$$\bar{x} \sim N(.75, .1875/3)$$

We can then define the normalized variable z-

$$\begin{aligned} z &= \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \\ &= \frac{\bar{x}-.75}{\sqrt{.1875/3}} \sim N(0,1) \end{aligned}$$

Since z is distributed normally, we can look up the probability that z is in any interval in our computer!

$$Pr(|z| > 1.96) \quad = \quad .95$$

Draw the normal graph with both critical sides in tails (so .025 in each tail). If we use this probability statement, we have that

$$\begin{aligned} Pr(|z| > 1.96) &= .95 \\ Pr(|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}| > 1.96) &= .95 \\ Pr(|\bar{x}-\mu| > 1.96(\sigma/\sqrt{n})) &= .95 \\ Pr(\bar{x} - 1.96(\sigma/\sqrt{n}) \leq \bar{x} \leq \bar{x} + 1.96(\sigma/\sqrt{n})) &= .95 \end{aligned}$$

From this we get the Confidence Interval for the population mean:

$$\bar{x} \quad \pm \quad 1.96(\frac{\sigma}{\sqrt{n}})$$

In this case, because we know the true population mean and variance can say:

$$\begin{aligned} .75 \quad &\pm \quad 1.96 * .25 \\ [.26 \quad &, \quad 1.24] \end{aligned}$$

Otherwise would substitute sample mean and variance for population mean and variance.

What was the probability taken over? It is over the sampling distribution of the estimator. Thus, the CI does not say that the population mean μ lies in the CI with 95% probability- for any sample μ is either in the CI or not. This is instead called the Bayesian credibility set.

Instead, if 95% of times that we do experiment and create the confidence interval, μ would be in the interval that we created. Thus we are appealing to a long run sampling framework- in 95% of samples will be right. But if we only have one sample, does this make sense?

Also remember- Confidence Interval for the parameters- in this case μ- not for the estimator itself.

Example 2. Sample mean of Bernoulli RVs

Now we change the distribution of X to the Bernoulli distribution with $p = .75$. We still have the estimator $T_N = \bar{x}$ and $N = 3$. Now, the mean can only attain 4 different values:$0, 1/3, 2/3, 1$. Thus,

$$\bar{x} \quad \sim \quad Binomial(3, .75)$$

The distribution of $\bar{x}$ is then:

$$\begin{array}{rll} 0 & w/prob & .25^3 = .016 \\ 1/3 & & .75 * .25^2 * 3 = .141 \\ 2/3 & & .75^2 * .25^2 * 3 = .42 \\ 1 & & .75^3 = .42 \end{array}$$

What is the 95% CI? We can see that 42% prob is 1, 84% prob 1 or 2/3, 98% prob is 1, 2/3 or 1/3. Thus the 95% CI is:

$$[1/3 \quad , \quad 1]$$

even though the expectation and variance of the sample mean are the same (.75,.1875/3).

Thus, the sampling distribution here depended on the underlying distribution- as we changed the underlying distribution from the normal to binomial, CI changed. We can also see the CI will change with the sample size- in the normal case the variance of the normal density changes, in the binomial case the number of mass points will change.

What happens if we dont know the analytic solution to the sample distribution? If we know the true underlying population distribution and sample size, but have no closed form soln for the sample distribution- just simulate it. Draw a simulated sample from the population with a given sample size, compute the sample statistic- do this J times and get distribution of sample statistic. We call this simulation or "Monte Carlo".

1.1 Approximations

1. Asymptotic Approximation:

The basic idea here- replace finite sample sampling distribution with asymptotic distribution if N is large enough.

Consider $\bar{X}$ in random sampling. Then:

$$Z = \frac{\sqrt{N}(\bar{x}-\mu)}{\sigma} \rightarrow_D N(0,1)$$

under the Central Limit Thm.

If we assume that we can use the asymptotic approximation in place of the finite sample sampling distribution:

$$\bar{x} \sim_a N(\mu, \sigma^2/n)$$

where $\bar{x}$ is the estimator. However, in practice dont know μ, σ^2- so use their sample equivalents!

$$\bar{x} \sim_a N(\bar{x}, \sigma_N^2/n)$$

here the LHS $\bar{x}$ is the estimator (a random variable) and the RHS $\bar{x}$ the realization of that RV in the sample given.

Soon we will be using the OLS asymptotic distribution for inference.

2. Bootstrap Simulation:

Problem: we may not know the true underlying distribution P. Instead, approximate P(underlying distribution) with the sampling distribution P_N.

You already have the sample- simulate draws from P_N.

For example, have a sample of 100 observations. Calculate the mean.

Then make J random samples with replacement of 100 observations- calculate statistic (say mean) for these random samples- get distribution of statistic.

Bootstrap can be a better approximation than asymptotic approximation in finite samples.

1.2 Hypothesis Testing

Hypothesis testing begins with a specification of the hypothesis that is to be tested. This is known as the null hypothesis. What may be true if the null hypothesis is false is known as the alternative hypothesis.

Imagine we have a null hypothesis:

$$\begin{aligned} H_0: \quad \mu = &\quad \mu_0 \\ H_a: \quad \mu \neq &\quad \mu_0 \end{aligned}$$

How can we test this?

Two methods:

1. Does the null fall within the Confidence Interval? If not, reject the null (at the relevant (say 5%) significance level).

If so, accept (or do not reject) the null.

2. If

$$|\bar{Z}_0| = \frac{|\bar{X}-\mu_0|}{\sigma_N/\sqrt{N}} \geq 1.96$$

reject the null at the 5% significance level. Here we are forming a test stat Z_0- that depends on the null hypothesis (since we are imposing the null hypothesis, comparing with μ_0).

Can think of two types of error:

1. Type 1 error- probability that reject H_0 given that it is true.

2. Type 2 error- probability that do not reject H_0 given that it is false

In general, we call $\alpha = \Pr($ type 1 error) the size of the test or the significance level and set it to some value like .05. As we lower α we make it harder to reject the null- reducing the probability of Type 1 error.

However, we also have to consider $\beta=$ Pr(Type 2 error). We can also define Power$=1-\beta$. The power of the test is the probability that one correctly rejects the null hypothesis- this will depend on the alternative hypothesis!

Thus, we can not just consider the significance level of the test- need to consider the power properties. Take the 20 sided die test. Roll a 20 sided die- if it lands on 1, reject the null, otherwise accept the null.

This test will have an $\alpha = .05$- but its power is also always .05, regardless of the sample size or alternative hypothesis!

For example, lets test whether a certain regression coefficient is 0. If the true value is 10, the power will (for a good test anyway) be higher than if the true value is 2. But not for the 20 sided die test!

Some nice properties for a test include:

unbiasedness- $1-\beta \geq \alpha$ for all alternative hypotheses

consistency- $1-\beta \rightarrow_p 1$ for all alternative hypotheses.

Draw a graph of a test like this vs. 20 sided die test- y axis probability of rejecting.

In general, as we reduce α we also reduce the power of the test- so there is a tradeoff here.

Right now in economics/econometrics solve this by setting $\alpha = .05$- so as the sample size increases power increases for all alternative hypotheses but significance level remains the same. Why not have the significance level fall with the sample size too??

2 Univariate Tests and CIs

Let us start with the CI of one coefficient in the linear regression model.

Let

$$z' \quad = \quad \frac{b_{jN} - \beta_j}{\sigma_{Nj}}$$

here σ_{Nj} is an estimate of the variance of the LS estimator for variable j.

In random sampling, we have proved earlier that:

$$z' \quad \to_d \quad N(0,1)$$

Using this asymptotic approximation, we have that the estimator b_{Nj} (which is not the same as the estimate in the sample b_{jN}!!) is distributed roughly:

$$\begin{aligned} b_{Nj} \quad &\sim_a \quad N(\beta_j, \sigma_{Nj}^2) \\ &\sim_a \quad N(b_{jN}, \sigma_{Nj}^2) \end{aligned}$$

where the second line replaces the unobserved population parameter with the estimate in the sample- remember that the LHS is the estimator.

We then know that:

$$\begin{aligned} Pr(|z'| > c_\alpha) \quad &\approx \quad 1 - \alpha \\ Pr(b_{jN} - c_\alpha \sigma_{Nj}) \leq \beta_j \leq b_{jN} + c_\alpha \sigma_{Nj} \quad &\approx \quad 1 - \alpha \end{aligned}$$

So we can use $b_{jn} \pm c_\alpha \sigma_{Nj}$ as an approximate $1 - \alpha$ CI for β_j.

Hypothesis Testing:

$$\begin{aligned} H_0 : \quad \beta_j = \quad &\beta_{0j} \\ H_a : \quad \beta_j \neq \quad &\beta_{0j} \end{aligned}$$

We can then form the test statistic

$$z'_0 \quad = \quad \frac{b_{jN} - \beta_{j0}}{\sigma_{Nj}}$$

If $\beta_j = \beta_{0j}$ (that is under the null), then $z'_0 \to_d N(0,1)$.

And so $Pr(|z'_0| > c_\alpha) = 1 - \alpha$.

The most common such test is a significance test that sets $\beta_{0j} = 0$. Then if we accept the null, then cant reject the possibility that RV has no effect.

The test statistic also becomes:

$$z'_0 \quad = \quad \frac{b_{jN}}{\sigma_{Nj}}$$

However, this rejection or acceptance is not the final answer:

1. Could be that b_{jN} is close to zero- in our current sample little effect from the variable.

2. Or σ_{Nj} is high- so parameter is estimated imprecisely- would reject lots of nulls.

3. Statistical significance is not the same as economic significance-

Maybe can reject coefficient is not zero, but its not economically meaningful. What is the econ interpretation of the parameter? So- what does its magnitude mean?

Or- to quote Derek Neal- use the estimated coefficient in a sentence.

As sample size goes to infinity- will reject everything that is not truly zero- but coefficient of .00001 is not meaningful.

In a regression table- generally see t-stat, or p-value, or **,***- these are univariate tests!

3 Joint Tests

There are lot of joint tests- tests concerning more than one parameter. We can't just use a simple test (imagine $\beta_1 = 0$ and $\beta_2 = 0$ is our test- we cant just test both restrictions separately) because:

1. Coefficients could be correlated b/c variables are correlated (remember from the omitted variable bias results that unless variables are uncorrelated with each other, adding a variable will affect the other coefficients).

2. Joint significance is different from individual significance- even if all the variables are independent and uncorrelated.

Example: Lets take β_1 and β_2- can make a confidence interval for both separately- assume X_1 and X_2 independent. Then prob that in confidence interval of both is $.95^2 = .9025 < .95$.

Examples of joint tests:

1. log wage- dummies and interactions for race- test for race has no effect

2. include variable and quadratic, cubic terms- test that variable has no effect

3. CAPM and APT- test that APT not needed is test that all APT variables are insignificant

We begin the joint test case with the asymptotic normality of the LS estimator:

$$b_N \sim N(\beta, \Sigma_N)$$

where b_N is k by 1.

$$\begin{aligned} w^* &= (b_N - \beta)' \Sigma_N^{-1} (b_N - \beta) \\ &\to_d \chi^2(k) \end{aligned}$$

Here w^* is a scalar- why?

Proof. a) Since Σ_N is a positive definite matrix, it follows that $\Sigma_N^{-1} = p_N p_N', p_N = p_N'$.

b) Let $z^* = p_N(b_N - \beta) = \Sigma_N^{-1/2}(b_N - \beta) \rightarrow_d N(0, I)$

c) Thm. If $\{Z_1, Z_2, ..., Z_k\}$ are independent std normal rvs, then $\sum_i Z_i^2$ is distributed χ^2 with k degrees of freedom (as there are k sums of squared independent normal rvs.)

In our case, $\Sigma_N^{-1/2}(b_N - \beta)$ is composed of k independent std normal rvs, because in the multivariate normal distn case, uncorrelated implies independence.

Then $(b_N - \beta)'\Sigma_N^{-1/2'}\Sigma_N^{-1/2}(b_N - \beta)$ is really the sum of k independent std normal rvs. (look at the matrix form) so we can apply the thm and say:

$$\begin{aligned} w^* = z^{*'} z^* &= (b_N - \beta)'\Sigma_N^{-1/2'}\Sigma_N^{-1/2}(b_N - \beta) \\ &\rightarrow_d \chi^2(k) \end{aligned}$$

□

Test of Linear Functions of the RVs:
Examples- Cobb Douglas and returns to scale
Two coefficients are equal to each other (1 minus other =0)
What is the H matrix in these cases? In Zero null cases?
Linear Functions: Let $t_N = Hb_N, \theta = H\beta$.
where here H is a p by k nonrandom matrix with rank p.
It follows that

$$\begin{aligned} \sqrt{N}(t_N - \theta) &\rightarrow_d N(0, H\Sigma H') \\ w^* = (t_N - \theta)'(H\Sigma_N H')^{-1}(t_N - \theta) &\rightarrow_d \chi^2(p) \end{aligned}$$

Here p is the number of restrictions- go over some examples.
In the CRM, we have shown that:

$$\begin{aligned} \sqrt{N}(b_N - \beta) &\rightarrow_d N(0, \sigma^2 E(X'X)^{-1}) \\ b_N &\sim_a N(\beta, \sigma_N^2 (X'X)^{-1}) \\ \sigma_N^2 &= \frac{e'e}{N - k} \end{aligned}$$

Using this we have the Wald test stat becomes:

$$\begin{aligned} w^* &= (b_N - \beta)'(X'X)\frac{1}{\sigma_N^2}(b_N - \beta) \\ &\rightarrow_d \chi^2(k) \end{aligned}$$

where $X'X$ has been inverted twice.
With linear restrictions, we have:

$$w^* = (t_N - \theta)'(H(X'X)^{-1}\frac{1}{\sigma_N^2}H')^{-1}(t_N - \theta) \quad \rightarrow_d \quad \chi^2(p)$$

Confidence Intervals:
The confidence set is all θ that satisfy the following restriction:

$$w^* = (t_N - \theta)'(H(X'X)^{-1}\frac{1}{\sigma_N^2}H')^{-1}(t_N - \theta) \quad \leq \quad c_\alpha$$

where c_α is the appropriate critical value from a $\chi^2(p)$ distribution.

Example 3. Let p=2, and the confidence set be over both parameters, so H is the 2 by 2 identity matrix.

Then

$$\begin{aligned} H\Sigma_N H' &= \begin{matrix} 1 & r \\ r & 1 \end{matrix} \\ w^* &= (t_N - \theta)'[\begin{matrix} 1 & r \\ r & 1 \end{matrix}]\frac{1}{1-r^2}(t_N - \theta) \\ &= \frac{(t_{1N} - \theta_1)^2 + (t_{2N} - \theta_2)^2 - 2r(t_{1N} - \theta_1)(t_{2N} - \theta_2)}{(1-r^2)} \leq c_\alpha \end{aligned}$$

This is an ellipse.
If $r = 0$ (no correlation between the estimators) we get a circle:

$$w^* \quad = \quad (t_{1N} - \theta_1)^2 + (t_{2N} - \theta_2)^2 \leq c_\alpha$$

If did two independent t tests- get a square which is wrong- draw all of these (some points in square not in circle and v.v.)

Notice: means can not reject each univariate null hypo but reject joint.

Hypothesis Testing:

Draw graphs on confidence set to explain this:

$$\begin{aligned} H_0 : \quad \theta &= \theta_0 \\ H_a : \quad \theta &\neq \theta_0 \end{aligned}$$

$$w_0^* = (t_N - \theta_0)'(H\Sigma_N H')^{-1}(t_N - \theta_0) \quad \rightarrow_d \quad \chi^2(p)$$

under the null that $\theta = \theta_0$.

Test: If $w_0^* > c_\alpha$ then reject null at the α significance level. O/w accept.

Special Case: CRM with Zero Null Subvector Test

This is broader than it looks- if the null for $\beta_1 = 3$ the zero null becomes $\beta_1 - 3$.

The test here is, for β_2 that is k_2 by 1,

$$\begin{aligned} H_0 : \quad & \beta_2 && = 0 \\ H_a : \quad & \beta_2 && \neq 0 \\ H \quad & = && [0 \; I] \end{aligned}$$

where H is k_2 by k, 0 is k_2 by k_1, I is k_2 by k_2.

Under the null,

$$w^* = \frac{(b_2)'(H(X'X)^{-1}H')^{-1}(b_2)}{\sigma_N^2}$$

where $\sigma_N^2 = \frac{e'e}{N-k}$.

Remember from the omitted variable rule algebra that the short regression residuals from a regression of Y on X_1 alone were:

$$\begin{aligned} e^* &= M_1 Y \\ &= M_1(X_1 b_1 + X_2 b_2 + e) \\ &= M_1 X_2 b_2 + e \\ &= X_2^* b_2 + e \end{aligned}$$

Multiplying e^* by itself I get:

$$e^{*'} e^* = e'e + b_2' X_2^{*'} X_2^* b_2$$

Through some Partitioned Matrix Algebra results, $(H(X'X)^{-1}H')^{-1} = X_2^{*'} X_2^*$.

Combining these we have:

$$\begin{aligned} w^* &= \frac{b_2' X_2^{*'} X_2^* b_2}{\sigma_N^2} \\ &= \frac{(e^{*'} e^* - e'e)(N-k)}{e'e} \end{aligned}$$

Now divide the numerator and denominator by the total sum of squares, we will have:

$$\begin{aligned} w^* &= \frac{(e^{*'}e^* - e'e)(N-k)}{e'e} \\ &= \frac{(R^2 - R^{*2})(N-k)}{1-R^2} \end{aligned}$$

Here R^{2*} is the R^2 of a regression of Y on X_1- i.e. the short regression. Thus, the test statistic depends on the improvement in R^2 or goodness of fit after including the X_2 variables. The long regression does not impose the null hypothesis while the short regression does impose the null hypothesis.

Example:

F-test

Test the null that all slopes equal zero.

$$H_0 : \quad \beta_2 = \beta_3 = \beta_4 = ... = \beta_k \quad = 0$$

The alternative hypothesis is that any one of the slopes is not zero. The null is thus the constant regression whose R^2 is zero. We can then form the test stat:

$$\begin{aligned} w^* &= \frac{(R^2 - R^{*2})(N-k)}{1-R^2} \\ &= \frac{(R^2)(N-k)}{1-R^2} \rightarrow_d \chi^2(k-1) \end{aligned}$$

if the null is true.

Often the F-stat is reported:

$$F_0 = \frac{W_0^*}{p} = \frac{W_0^*}{k-1} = \frac{(R^2)(N-k)}{(1-R^2)(k-1)}$$

where p=k-1 is the number of restrictions.

Under the CNRM, $F_0 \sim F(k-1, n-k)$

However, if dont know CNRM, use asymptotic approx- sample size is large converges to chi-sq.

The F test- reported in Stata- says “Is this regression better than nothing i.e. the mean of y?”

A couple things not to do:

Pre-test bias:

Regression Strategy:

Run regression of Y on X_1 and X_2

If reject null of significance of X_2, report long regression.

If accept null, report short regression.

In this case, estimator on X_1 is $b_1(1-r)+rb_1^*$ where $r=1$ if b_2 not rejected. Thus, the coefficient in any sample is selected- cant apply previous results as distribution of OLS estimator no longer applies.

Why would people do this?

Regression fishing:

Run lots of regressions and report ones that come up significant.

But, multiple hypothesis testing problem- what is prob of at least 1 rejection if all null hypotheses true?

If test stats independent,

$$\begin{aligned} Pr(one-rej) &= 1-P(allacc) \\ &= 1-(1-\alpha)^k \end{aligned}$$

This is not the same as α!