Confidence Intervals and Bootstrapping

Charlie Gibbons
ARE 212

Spring 2011

Notes

Outline

Notes

Confidence interval definition

Suppose that we are interested in a univariate regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

We'd like to find a range of values b_1^L to b_1^U such that the expected probability of β_1 lying in that range is equal to $1 - \alpha$. This is known as a $100(1-\alpha)\%$ *confidence interval*.

In math, we want

$$\Pr\left(b_1^L < \beta_1 < b_1^U\right) = 1 - \alpha.$$

Notes

Solving

Let's subtract our value of $\hat{\beta}_1$ from all sides:

$$\Pr\left(b_1^L - \hat{\beta}_1 < \beta_1 - \hat{\beta}_1 < b_1^U - \hat{\beta}_1\right) = 1 - \alpha.$$

Now, divide all sides by the standard deviation of $\hat{\beta}_1$

$$\Pr\left(\frac{b_1^L - \hat{\beta}_1}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} < \frac{\beta_1 - \hat{\beta}_1}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} < \frac{b_1^U - \hat{\beta}_1}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}}\right) = 1 - \alpha.$$

Notes

Solving, con't

Lastly, multiply everything by -1—this changes the direction of the inequalities!

$$\Pr\left(\frac{\hat{\beta}_1 - b_1^L}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} > \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} > \frac{\hat{\beta}_1 - b_1^U}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}}\right) = 1 - \alpha.$$

Notes

Separation

Now we have

$$\Pr\left(\frac{\hat{\beta}_1 - b_1^L}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} > Z > \frac{\hat{\beta}_1 - b_1^U}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}}\right) = 1 - \alpha.$$

This is the probability that our estimate is between two values. This probability is equal to 1 minus the probability of being above the high value and the probability of being below the low value:

$$1 - \Pr\left(\frac{\hat{\beta}_1 - b_1^L}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} < Z\right) - \Pr\left(Z < \frac{\hat{\beta}_1 - b_1^U}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}}\right) = 1 - \alpha.$$

Notes

Equal tail probabilities

Let the probability of being in each tail be the same. Then we have:

$$1 - 2 \times \Pr\left(Z < \frac{\hat{\beta}_1 - b_1^U}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} \right) = 1 - \alpha$$

$$\Pr\left(Z < \frac{\hat{\beta}_1 - b_1^U}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} \right) = \frac{\alpha}{2}.$$

We solve for b_1^U using the lower critical value

$$b_1^U = \hat{\beta}_1 + |z_{\alpha/2}| \sqrt{\text{Var}\left(\hat{\beta}_1\right)},$$

with b_1^L found analogously.

Notes

Why equal tail probabilities?

Why do we set the probability of being in each tail of this confidence interval to be equal?

This makes the shortest possible interval. Additionally, it chooses the region of highest density for the estimator.

Notes

Summary

We can write the general expression for confidence intervals as:

$$\left[\hat{\beta}_1 - z_{\alpha/2}, \sqrt{\text{Var}\left(\hat{\beta}_1\right)},\ \hat{\beta}_1 + z_{\alpha/2}, \sqrt{\text{Var}\left(\hat{\beta}_1\right)} \right].$$

This means that, if we repeated our estimation on many samples, then the true parameters would lie in these regions 95% of the time; we say that this confidence interval has 95% *coverage*.

This *does not* mean that there is a 95% chance that β_1 is really in this range in this particular case.

This is because β_1 *is not random* so it is either in this interval or not—the probability is either exactly 0 or exactly 1.

Notes

Confidence v. prediction intervals

We can create a confidence interval for the expected value of y conditional upon x:

$$\Pr(\hat{y}_L < \mathbb{E}[y|X] < \hat{y}_U|X) = \Pr(\hat{y}_L < \beta_0 + \beta_1 x < \hat{y}_U|X) = 1 - \alpha.$$

A *prediction interval* predicts y itself, not its expected value:

$$\Pr(\hat{y}_L < y < \hat{y}_U|X) = \Pr(\hat{y}_L < \beta_0 + \beta_1 x + \epsilon < \hat{y}_U|X) = 1 - \alpha.$$

Though both intervals have the same midpoint, the prediction interval has a higher variance because it takes into account the variability of our estimates as well as the variability of our error term.

Both are constructed as described for confidence intervals; just the estimate of the standard error is different.

Notes

The prediction interval

The prediction of y for some x' is $x'\hat{\beta}$. The prediction variance is

$$\begin{aligned}\text{Var}\left(y \mid X, x\right) &= \text{Var}\left(x'\hat{\beta} + \epsilon \mid X, x\right) \\ &= x'\text{Var}\left(\hat{\beta} \mid X, x\right) x + \text{Var}\left(\epsilon \mid X, x\right) \\ &= \sigma^2\left(1 + x'\left(X'X\right)^{-1} x\right)\end{aligned}$$

in the case of homoskedasticity and no serial correlation.

Notes

Reformulating the prediction variance

If the regression includes a constant term, this can be written

$$\sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{i=1}^{K-1}\left(x_j - \bar{x}_j\right)\left(x_i - \bar{x}_i\right)\left(X'_{-1}M_1X_{-1}\right)^{-1}_{ij}\right],$$

where X_{-1} is the matrix that contains all but the constant term and M_1 demeans the predictors (*i.e.*, by getting the residuals from their regression on a constant term).

We see that our most precise predictions will be for x near $\bar{x}$ and thus for y near $\bar{y}$.

Notes

Intuitive notions

Confidence intervals and hypothesis tests are very similar.

A confidence interval asks, given a tail probability α and the assumption that $\beta_1 = \hat{\beta}_1$ (*i.e.*, that our unbiased estimator gives us an estimate that is the true mean), what critical values produce this tail probability?

A hypothesis test asks, given critical values and the assumption that $\beta_1 = b$ (*i.e.*, that our null hypothesis is true), what is the probability of being in the tails?

Notes

Relating these concepts

A confidence interval contains all the values for null hypotheses that cannot be rejected at the α level.

A hypothesis that is rejected at the α level is outside of the $100(1-\alpha)\%$ confidence interval and a hypothesis that cannot be rejected at that level is contained in that confidence interval.

Thus, it is said that a hypothesis test is an *inverted* confidence interval and vice-versa.

Notes

Variance of our estimator

The bootstrap is used primarily to investigate the variance of an estimator.

Why is our estimator random? Because our data are random. For example, our estimate of the average height of a population will vary because the average height in a sample will vary and will not precisely equal the population average.

Without bootstrapping or parametric results, how could we consider the variance of our estimator? Get a sample, calculate the estimator, draw a new sample, calculate the estimator again, and repeat many times to get the distribution.

The bootstrap approximates this process.

Notes

Non-parametric bootstrap

The non-parametric bootstrap is the most intuitive expression of this idea.

There is a population joint distribution $F(y, x)$ and our estimator $\hat{\beta}$ has a population distribution $G(y, x)$. To form our sample, we draw from $F(\cdot)$ independently, giving a sample distribution $\hat{F}(y, x)$. Our estimator has distribution $\hat{G}(y, x)$.

Our first assumption is that $\hat{F}(\cdot) \xrightarrow{p} F(\cdot)$ and $\hat{G}(\cdot) \xrightarrow{p} G(\cdot)$.

We can't draw new samples from F, but we can resample from $\hat{F}$. We create B bootstrap samples by drawing samples of size N with replacement from $\{(y_i, x_i)\}_{i=1}^{N}$.

Notes

Replacement

Why with replacement?

What if we drew a sample of size N without replacement from our full sample? We get the exact same sample back.

Additionally, we assume that our observations are draw independently from F, so we maintain that assumption when drawing from $\hat{F}$; drawing with replacement introduces dependence.

Notes

Bootstrapped estimates

For each bootstrap sample b, we calculate our estimator *for that sample*. Then, we use the distribution of our estimators, $\{\hat{\beta}^b\}_{b=1}^{B}$, to get $\widehat{\hat{G}}(\cdot)$.

Hence, our next assumption is that $\widehat{\hat{G}}(\cdot) \xrightarrow{p} \hat{G}(\cdot)$.

Consistency here depends upon the number of bootstraps B, *which we can control.*

Notes

Summary

In totality, we have to assume that

$$\widehat{\hat{G}}(\cdot) \xrightarrow{p} \hat{G}(\cdot) \xrightarrow{p} G(\cdot);$$

our bootstapped distribution of estimators has to approximate the distribution of our sample estimator, which has to approximate the population distribution of our estimator.

The first linkage depends upon the properties of the bootstrap and the number of bootstrap samples B, while the latter depends upon consistency of our estimator and our sample size N.

Notes

Applications

Bootstrapping is most commonly used to calculate

1. standard errors, especially for estimators with properties that are difficult to calculate analytically (*e.g.*, medians),
2. p-values for hypothesis testing, and
3. confidence intervals.

Notes

Standard error

To calculate the standard error, simply calculate

$$\widehat{\text{Var}}\left(\hat{\beta}\right) = \frac{\sum_{i=1}^{B}\left(\hat{\beta}^b - \hat{\beta}^*\right)^2}{B-1},$$

where $\hat{\beta}^*$ is the average of the bootstrapped values (this should be close to the full sample $\hat{\beta}$) and take the square root. This is robust to heteroskedasticity (for the non-parametric bootstrap).

Notes

Hypothesis testing

You can use the estimate of the standard error of $\hat{\beta}$ to construct normal-based hypothesis tests. A better alternative would be to use the bootstrapped results directly.

The p-value for the null hypothesis of b (where b is a scalar) is

$$\frac{\sum_{b=1}^{B} \mathbb{I}\left\{|\hat{\beta}^b - \hat{\beta}| > |b - \hat{\beta}|\right\}}{B};$$

what proportion of bootstrapped values are further from the observed $\hat{\beta}$ than the null hypothesis?

Note: this is formulated differently than standard hypothesis testing. Draw a picture to ensure that you understand the distinction.

Notes

Confidence intervals

Again, you can use the estimated standard error to calculate a confidence interval, but you can use the bootstrapped samples directly. Specifically, find the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of your bootstrapped results and use these as your interval's end points.

This permits the confidence interval to be asymmetric, which is the most important improvement that bootstrapping provides.

Additionally, the CI is *range-respecting*; since all bootstrapped values actually stem from the application of the estimator to data, the CI contains only values that are possible for the estimator. For example, a bootstrapped CI will never give you a correlation larger than 1 in absolute value.

Notes

Parametric bootstrap

The parametric procedure assumes homoskedasticity and linearity; the non-parametric approach gives reasonable answers if these assumptions fail.

Changing notation, let the errors be draw from the distribution $F(\epsilon \mid X) = F(\epsilon)$ and that $\epsilon = y - x'\beta$.

Using our full data set, we calculate $\hat{\beta}$ and the sets $\hat{y}$ and $\hat{\epsilon}$. Now, our bootstrapped sets consist of the *fixed* X matrix and the vector $y^b = \hat{y} + \hat{\epsilon}^b$; the predictors and their relationship to y is fixed, but the errors are shuffled and added to the fixed predicted values. We use these data to calculate $\hat{\beta}^b$ and conduct inference as discussed in the non-parametric case.

We may inflate the residuals by a factor $((1 - h_i)^{-1/2}$, for example) to ensure that the variance of the residuals is the same as the error distribution (and homoskedastic).

Notes

Comparison

This is parametric because we assume that the linear model holds. Shuffling only the error terms also breaks any linkage between the errors and X; hence, there cannot be heteroskedasticity.

The benefit is that the standard errors here will tend to be smaller. This difference gets smaller as n gets larger.

As DM illustrate, there are actually many ways to bootstrap in regression, each one different based upon the precise DGP being assumed.

Notes

Notes

Notes