Econometric Methods
Prediction Intervals

Prediction Intervals: Using Sample Data to Predict

1. Setting the Stage

a. We have an iid random sample $\{Y_1, Y_2, \ldots Y_n\}$ from the distribution of Y. Y has unknown mean μ, and unknown variance, σ^2.

b. Previously, we have focused on estimating μ, the unknown mean of the distribution. And for that purpose, we gravitated towards the sample mean, $\bar{Y} = \frac{1}{n}\sum Y_i$, because it is a Best Linear Unbiased Estimator (***BLUE***) of the population mean (which is to say that it has minimum variance in the class of linear unbiased estimators).

2. Predicting Y_{n+1} (the next sampled value)

a. As before we will focus on linear estimators, and search for the BLUE estimator, which has minimum variance in the class of linear unbiased estimators: $W = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \ldots + \beta_n Y_n$

b. This estimator will be unbiased if the expected residual is 0, or $E(W - Y_{n+1}) = 0$. Since the expected residual is $E(W - Y_{n+1}) = \beta_0 + \mu \sum \beta_i - \mu = \beta_0 + \mu\left(\sum \beta_i - 1\right)$, for W to always be unbiased we must have: $\beta_0 = 0 \text{ and } \sum_{i=1}^{n} \beta_i = 1$.

c. So if we only consider the set (or class) of linear unbiased estimators, we are considering only estimators of the form:

$W = \beta_1 Y_1 + \beta_2 Y_2 + \ldots + \beta_n Y_n$, where $\sum_{i=1}^{n} \beta_i = 1$.

(This should sound familiar.)

d. Since Y_i's $(i = 1, \ldots, n+1)$ are pairwise independent, the $\beta_i Y_i$'s are pairwise independent and the variance of the sum is the sum of the variances. And so, $Var(W - Y_{n+1}) = \beta_1^2 Var(Y_1) + \beta_2^2 Var(Y_2) + \ldots + \beta_n^2 Var(Y_n) + Var(Y_{n+1})$. This is $\sigma^2 \sum \beta_i^2 + \sigma^2 = \sigma^2\left(\sum \beta_i^2 + 1\right)$, since $Var(Y_i) = \sigma^2$ for each i.

e. This sets up the optimization problem:

min $Var(W - Y_{n+1}) = \sigma^2\left(\sum \beta_i^2 + 1\right)$ subject to $\sum_{i=1}^{n} \beta_i = 1$.

f. But we know from before that the solution is $\beta_i^* = \dfrac{1}{n}$ for all i… and so W is just the Sample Mean: $W = \bar{Y} = \dfrac{1}{n}\sum Y_i$.

g. Accordingly, the sample mean is not just a BLUE estimator of the population mean. It is as well a BLUE estimator of the next sampled value from the distribution.

3. Inference: Prediction Intervals

a. As usual, we'll assume that Y is Normally distributed: $Y \sim N(\mu, \sigma^2)$. Since the Y_i's $(i = 1, ..., n+1)$ are independent, $\bar{Y} \sim N(\mu, \dfrac{\sigma^2}{n})$ and $Y_{n+1} \sim N(\mu, \sigma^2)$. And so $\bar{Y} - Y_{n+1}$ is normally distributed: $\bar{Y} - Y_{n+1} \sim N(0, \dfrac{\sigma^2}{n} + \sigma^2)$.

b. Put differently: $\dfrac{\bar{Y} - Y_{n+1}}{\sigma\sqrt{(1+1/n)}} \sim N(0,1)$.

c. Since we don't know the variance σ^2, we'll estimate it using the Sample Variance: $S_{YY} = \dfrac{\sum (Y_i - \bar{Y})^2}{n-1}$.

d. As usual, we can use $S_Y = \sqrt{S_{YY}}$ to estimate σ, and given the assumptions above, we have a t distribution with n-1 degrees of freedom: $\dfrac{\bar{Y} - Y_{n+1}}{S_Y\sqrt{(1+1/n)}} \sim t_{n-1}$.

e. Consider a critical value c defined by $\text{Prob}(-c < t_{n-1} < c) = .95$. Then we have $\text{Prob}(-c < \dfrac{\bar{Y} - Y_{n+1}}{S_Y\sqrt{(1+1/n)}} < c) = .95$.

f. Or put differently, $\text{Prob}(\bar{Y} - cS_Y\sqrt{(1+1/n)} < Y_{n+1} < \bar{Y} + cS_Y\sqrt{(1+1/n)}) = .95$.

g. And so the 95% ***Prediction Interval*** $\left[\bar{Y} \pm cS_Y\sqrt{(1+1/n)}\right]$ has the property that 95% of the time, intervals formed in this fashion will contain the (to be) sampled value of Y_{n+1}.

4. Prediction Intervals v. Confidence Intervals

a. There is a close similarity between:

i. *Confidence Intervals* for the unknown mean μ: $\left[\bar{Y} \pm cS_Y \sqrt{1/n}\right]$

ii. *Prediction Intervals* for the unknown value of Y_{n+1}: $\left[\bar{Y} \pm cS_Y \sqrt{(1+1/n)}\right]$

b. Both are centered around the sample mean $\bar{Y}$, but the prediction interval has a larger standard error. Both use the t_{n-1} distribution.

i. The standard error for the prediction interval reflects the uncertainty in estimating the mean, captured by the $\sqrt{1/n}$ term, as well as the variance in Y itself (which is estimated using the sample variance S_{YY}).