Economics 326
Methods of Empirical Research in Economics

Lecture 7: Confidence intervals

Vadim Marmer
University of British Columbia

January 29, 2009

Point estimation

- Our model:
  1. $Y_i = \beta_0 + \beta_1 X_i + U_i, \qquad i = 1, \ldots, n.$
  2. $E\left(U_i | X_1, \ldots, X_n\right) = 0$ for all i's.
  3. $E\left(U_i^2 | X_1, \ldots, X_n\right) = \sigma^2$ for all i's.
  4. $E\left(U_i U_j | X_1, \ldots, X_n\right) = 0$ for all $i \neq j$.
  5. U's are jointly normally distributed conditional on X's.
- The OLS estimator $\hat{\beta}_1$ is a point estimator of β_1.
- For our model, conditional on X's:

$$\begin{aligned} \hat{\beta}_1 &\sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right), \\ Var\left(\hat{\beta}_1\right) &= \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}. \end{aligned}$$

- With probability one, we have that $\hat{\beta}_1 \neq \beta_1$.

Interval estimation problem

- We want to construct an interval estimator for β_1 :
  - The interval estimator is called a confidence interval (CI).
  - A CI contains the true value β_1 with some pre-specified probability $1-\alpha$, where α is a small probability of error.
  - For example, if $\alpha = 0.05$, then the random CI will contain β_1 with probability 0.95.
- $1-\alpha$ is called the coverage probability.
- Confidence interval: $CI_{1-\alpha} = [LB_{1-\alpha}, UB_{1-\alpha}]$. The lower bound (LB) and upper bound (UB) should depend on the coverage probability $1-\alpha$.
- The formal definition of CI: It is a random interval $CI_{1-\alpha}$ such that conditional on X's,

  $$P(\beta_1 \in CI_{1-\alpha}) = 1 - \alpha.$$

  Note that the random element is $CI_{1-\alpha}$.
- Sometimes, a CI is defined as $P(\beta_1 \in CI_{1-\alpha}) \geq 1 - \alpha$.

Symmetric CIs

- One approach to constructing CIs is to consider a symmetric interval around the estimator $\hat{\beta}_1$:

$$CI_{1-\alpha} = \left[\hat{\beta}_1 - c_{1-\alpha}, \hat{\beta}_1 + c_{1-\alpha}\right]$$

- The problem is choosing $c_{1-\alpha}$ such that $P\left(\beta_1 \in CI_{1-\alpha}\right) = 1 - \alpha$.
- In choosing $c_{1-\alpha}$ we will be relying on the fact that given our assumptions and conditionally on X's:

$$\hat{\beta}_1 \sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right) \text{ and } Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

- Note that conditionally on X's:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \sim N\left(0, 1\right).$$

Quantiles (percentiles) of the standard normal distribution.

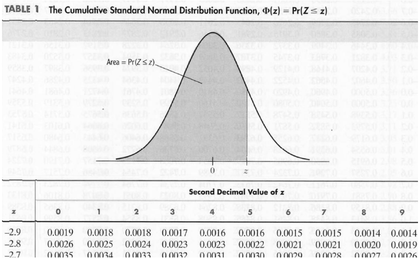


TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \le z)$

	Second Decimal Value of z									
z	0	1	2	3	4	5	6	7	8	9
–2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
–2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
–2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026

- Let $Z \sim N(0,1)$. The τ-th quantile (percentile) of the standard normal distribution is z_τ such that

$$P(Z \le z_\tau) = \tau.$$

- Median: $\tau = 0.5$ and $z_{0.5} = 0$. ($P(Z \le 0) = 0.5$).
- If $\tau = 0.975$ then $z_{0.975} = 1.96$. Due to symmetry, if $\tau = 0.025$ then $z_{0.025} = -1.96$.

σ^2 is known (infeasible CIs)

- Suppose (for a moment) that σ^2 is known, and we can compute exactly the variance of $\hat{\beta}_1$ as $Var\left(\hat{\beta}_1\right) = \sigma^2 / \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.
- Consider the following CI:

$$CI_{1-\alpha} = \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}\right].$$

- For example, if $1-\alpha = 0.95 \iff \alpha = 0.05 \iff z_{1-\alpha/2} = z_{0.975} = 1.96$, and

$$CI_{0.95} = \left[\hat{\beta}_1 - 1.96\sqrt{Var\left(\hat{\beta}_1\right)}, \hat{\beta}_1 + 1.96\sqrt{Var\left(\hat{\beta}_1\right)}\right].$$

Validity of the infeasible CIs (σ^2 is known)

- We need to show that $P\left(\beta_1 \in \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}\right]\right) = 1-\alpha.$
- Next,

$$
\begin{aligned}
& \hat{\beta}_1 - z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \leq \beta_1 \leq \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \\
\iff & -z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \leq \beta_1 - \hat{\beta}_1 \leq z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \\
\iff & -z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \leq \hat{\beta}_1 - \beta_1 \leq z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)} \\
\iff & -z_{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \leq z_{1-\alpha/2}
\end{aligned}
$$

Validity of the infeasible CIs (σ^2 is known)

- We have that

$$\beta_1 \in \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{Var\left(\hat{\beta}_1\right)}\right]$$

$$\iff \quad -z_{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \leq z_{1-\alpha/2}.$$

- Let $Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \sim N(0,1)$.

$$\begin{aligned} & P\left(-z_{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} \leq z_{1-\alpha/2}\right) \\ = \quad & P\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) \\ = \quad & P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) \\ = \quad & 1 - \alpha/2 - \alpha/2 = 1 - \alpha. \end{aligned}$$

Feasible confidence intervals (σ^2 is unknown)

- Since σ^2 is unknown, we must estimate it from the data:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{U}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2 .$$

- We can replace σ^2 by s^2, however, the result does not have a normal distribution any more:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}} \sim t_{n-2}, \text{ where } \widehat{Var}\left(\hat{\beta}_1\right) = \frac{s^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}.$$

  Here t_{n-2} denotes the t-distribution with $n-2$ degrees of freedom.
- The degrees of freedom depend on
  - the sample size (n),
  - and the number of parameters one have to estimate to compute s^2 (two in this case, β_0 and β_1).

Feasible confidence intervals (σ^2 is unknown)

- Let $t_{df,\tau}$ be the τ-th quantile of the t-distribution with the number of degrees of freedom df: If $T \sim t_{df}$ then
$$P\left(T \leq t_{df,\tau}\right) = \tau.$$
- Similarly to the normal distribution, the t-distribution is centered at zero and is symmetric around zero: $t_{n-2,1-\alpha/2} = -t_{n-2,\alpha/2}$.
- We can now construct a feasible confidence interval with $1-\alpha$ coverage as:

$$\begin{aligned} & CI_{1-\alpha} = \\ = & \left[\hat{\beta}_1 - t_{n-2,1-\alpha/2}\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}, \hat{\beta}_1 + t_{n-2,1-\alpha/2}\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}\right], \\ & \text{where } \widehat{Var}\left(\hat{\beta}_1\right) = \frac{s^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}. \end{aligned}$$

Example: Rent rates and average income

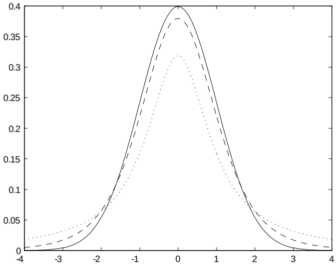
- Data (RENTAL.DTA): 64 cities in 1990, Rent = average rent, AvgInc = per capita income: $\text{Rent}_i = \beta_0 + \beta_1 \text{AvgInc}_i + U_i$.

```
. regress rent avginc

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  1,    62) =   78.34
       Model |  347069.249     1  347069.249           Prob > F      =  0.0000
    Residual |  274693.188    62  4430.53529           R-squared     =  0.5582
-------------+------------------------------           Adj R-squared =  0.5511
       Total |  621762.438    63  9869.24504           Root MSE      =  66.562

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |     .01158   .0013084     8.85   0.000     .0089646    .0141954
       _cons |   148.7764   32.09787     4.64   0.000      84.6137    212.9392
------------------------------------------------------------------------------
```

- $t_{62,0.975} \approx 2.00 \Longrightarrow$ The 95% confidence interval for β_1 is $[0.0115 - 2 \times 0.0013, 0.0115 + 2 \times 0.0013] = [0.0089, 0.0141]$.
- $t_{62,0.95} \approx 1.671 \Longrightarrow$ The 90% confidence interval for β_1 is $[0.0115 - 1.671 \times 0.0013, 0.0115 + 1.671 \times 0.0013] = [0.0093, 0.0137]$.

The effect of estimation of σ^2

- The t-distribution has heavier tails than normal.
  The graphs of normal (solid line), t_5 (dashed line), and t_1 (dotted line) PDFs:



- $t_{df,1-\alpha/2} > z_{1-\alpha/2}$, but as df increases $t_{df,1-\alpha/2} \to z_{1-\alpha/2}$.
- When the sample size n is large, $t_{n-2,1-\alpha/2}$ can be replaced with $z_{1-\alpha/2}$.

Interpretation of confidence intervals

- The confidence interval $CI_{1-\alpha}$ is a function of the sample $\{(Y_i, X_i) : i = 1, \ldots, n\}$, and therefore is random. This allows us to talk about probability of $CI_{1-\alpha}$ containing the true value of β_1.
- Once the confidence interval is computed given the data, we have its one realization. The realization of $CI_{1-\alpha}$ or (computed confidence interval) is not random, and it does not make sense anymore to talk about the probability that it includes the true β_1.
- Once the confidence interval is computed, it either contains the true value β_1 or it does not.