POLS 6481. Research Design and Quantitative Methods II

Lecture 7. Standard Errors of and Hypothesis Tests for Regression Coefficients

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

Outline:

1. Classical Linear Regression Model (CLRM) assumptions
2. Regression models and their sampling variances
3. Regression coefficients and their sampling distributions
4. The $t$ distribution
5. Confidence intervals for a regression coefficient
6. Null hypothesis retention region for a regression coefficient

POLS 6481. Research Design and Quantitative Methods II
Lecture 7. Standard Errors of and Hypothesis Tests for Regression Coefficients
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

## 0. Motivation: *Fun with Simulations*

Simulate $N$ = 1,000 observations from a known data process

$$y = \beta_0 + \beta_1 x + \beta_2 w + \beta_3 v + u$$

Assume $x$ and $w$ are modestly positively correlated

$v$ is Normally distributed and uncorrelated to $x$ or $w$ ; $v$ can be a disturbance term if we omit $v$

$u$ is uniform distributed and uncorrelated to $x$ or $w$ ; $u$ can be a disturbance term if we omit $u$

For concreteness, let $\beta_0 = 0$

$\beta_1 = 1$

$\beta_2 = 1$

$\beta_3 = 1$

Generate $x$, $w$, $v$, & $u$ by simulation, then generate $y$ using the population regression function

Next, draw samples, and regress $y$ on $x$ (and possibly on $w$ and/or $v$) to estimate parameters

Then, resample repeatedly and examine the distributions of parameter estimates

POLS 6481. Research Design and Quantitative Methods II
Lecture 7. Standard Errors of and Hypothesis Tests for Regression Coefficients
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

## 1. Classical Linear Regression Model assumptions

| | | |
|---|---|---|
| MLR.1 | Linear in Parameters | $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ |
| MLR.2 | Random Sampling | $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \; i = 1, \ldots n\}$<br>$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$ |
| MLR.3 | No Perfect Collinearity | $R_j^2 < 1 \qquad j = 0, 1, \ldots, k$ |
| MLR.4 | Zero Conditional Mean | $E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$ |
| MLR.5 | Homoskedasticity | $Var(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$ |

**MLR.1 – MLR.5 are the Gauss-Markov assumptions**

THM 3.4 = Properties of OLS

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \qquad j = 0, 1, \ldots, k$$

for all $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ for which $E(\tilde{\beta}_j) = \beta_j, j = 0, \ldots, k$

**THM 3.4 is the Gauss-Markov Theorem**; find "best" and "linear" and "unbiased"…

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

1. Classical Linear Regression Model assumptions

MLR.6 Normality of disturbances $u_i \sim N(0, \sigma^2)$ independently of $x_{i1}, x_{i2}, \ldots, x_{ik}$

Therefore it follows that: $y|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$

**MLR.1 – MLR.6 are the Classical Linear Regression Model (CLRM) Assumptions**

## 2. Regression models and their sampling variances

THM 3.3 Unbiased estimation of error variance*

$$MLR.1 - MLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2 \qquad\qquad \hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) / \,[n-k-1]$$

* THM 5.2 adds that under MLR.1 – MLR.5, $\widehat{\sigma^2}$ is a consistent estimator of $\sigma^2$ = *var*(*u*) (i.e., $plim\ \hat{\sigma}^2 = \sigma^2$ )

## 3. Regression coefficients and their sampling distributions

THM 3.1 Unbiased estimation of regression coefficients

$$MLR.1{-}MLR.4 \quad \Rightarrow \quad E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

THM 3.2 Sampling variance of regression slope coefficients

$$MLR.1 - MLR.5 \quad \Rightarrow \quad Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, \ldots, k$$

$$sd(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\sigma^2 / \left[SST_j(1 - R_j^2)\right]}$$

$$MLR.1 - MLR.5 \quad \Rightarrow \quad \widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}$$

$$se(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\hat{\sigma}^2 / \left[SST_j(1 - R_j^2)\right]}$$

3. Regression coefficients and their sampling distributions

THM 4.1 Normality of sampling distribution: Under MLR.1 – MLR.6 , OLS slope estimators are Normally distributed around the true parameters: $\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$

i.e., with mean: $E(\hat{\beta}_j) = \beta_j$ and variance: $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}, \quad j = 1, \ldots, k$

THM 4.1 Therefore, under MLR.1 – MLR.6 , the standardized slope estimators follow a standard Normal distribution: $\dfrac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$ for each *j*

THM 5.2 Under MLR.1 – MLR.5, OLS estimators are <u>asymptotically</u> Normally distributed, <u>and</u> the standardized slope estimators follow a standard Normal distribution:

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \stackrel{a}{\sim} N(0,1)$$

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

## 3. Regression coefficients and their sampling distributions

THM 4.1 Under MLR.1 – MLR.6 , the standardized slope estimators follow a standard Normal distribution: $\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$ for each $j$

Interpretation:

1. $P(-1.96 < z < 1.96) = .95$
2. $P(-1.96 < \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} < 1.96) = .95$
3. $P(-1.96 \cdot sd(\hat{\beta}_j) < \hat{\beta}_j - \beta_j < 1.96 \cdot sd(\hat{\beta}_j)) = .95$

4A. $P(\beta_j - 1.96 \cdot sd(\hat{\beta}_j) < \hat{\beta}_j < \beta_j + 1.96 \cdot sd(\hat{\beta}_j)) = .95$

The standard Normal distribution has the property that 95% of $z$ scores are between -1.96 and +1.96

There is a 95% probability that the sampling error will be less than 1.96 std. deviations (in either direction).

There is a 95% probability that $\hat{\beta}_j$ will be within 1.96 $sd(\hat{\beta}_j)$ of $\beta_j$

## 3. Regression coefficients and their sampling distributions

THM 4.1 Under MLR.1 – MLR.6 , the standardized slope estimators follow a standard Normal distribution: $\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$ for each $j$

Interpretation:

1. $P(-1.96 < z < 1.96) = .95$

2. $P(-1.96 < \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} < 1.96) = .95$

3. $P(-1.96 \cdot sd(\hat{\beta}_j) < \hat{\beta}_j - \beta_j < 1.96 \cdot sd(\hat{\beta}_j)) = .95$

There is a 95% probability that the sampling error will be less than 1.96 std. deviations (in either direction).

4B. $P(-\hat{\beta}_j - 1.96 \cdot sd(\hat{\beta}_j) < -\beta_j < -\hat{\beta}_j + 1.96 \cdot sd(\hat{\beta}_j)) = .95$

5. $P(\hat{\beta}_j + 1.96 \cdot sd(\hat{\beta}_j) > \beta_j > \hat{\beta}_j - 1.96 \cdot sd(\hat{\beta}_j)) = .95$

6. $P(\hat{\beta}_j - 1.96 \cdot sd(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + 1.96 \cdot sd(\hat{\beta}_j)) = .95$

95% confidence interval for $\beta_j$:
$\hat{\beta}_j \pm 1.96 \cdot sd(\hat{\beta}_j)$

## 4. The *t* distribution

THM 4.2 Under the CLRM assumptions (MLR.1 – MLR.6), $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$ [4.3]

where *n*–*k*–1 is the degrees of freedom (as *k*+1 is the number of unknown parameters in the population model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ )

Because the mean of the *t* distribution is zero, the numerator implies unbiasedness of the $\hat{\beta}_j$; the *t* distribution comes from the fact that the constant $\sigma$ in $sd(\hat{\beta}_j)$ has been replaced with the random variable $\hat{\sigma}$ [in $se(\hat{\beta}_j)$]. ... THM 4.2 allows us to test hypotheses involving the $\hat{\beta}_j$.

"The proof that [replacing $\sigma$ with $\hat{\sigma}$] leads to a *t* distribution with *n*–*k*–1 degrees of freedom is <u>difficult and not especially instructive</u>. Essentially, the proof shows that [4.3] can be written as the ratio of the standard normal random variable $(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j)$ over the square root of $(\widehat{\sigma^2}/\sigma^2)$. These random variables can be shown to be independent, and $(n\text{–}k\text{–}1)\cdot(\widehat{\sigma^2}/\sigma^2) \sim \chi^2_{n-k-1}$. The result follows from the definition of a *t* random variable."

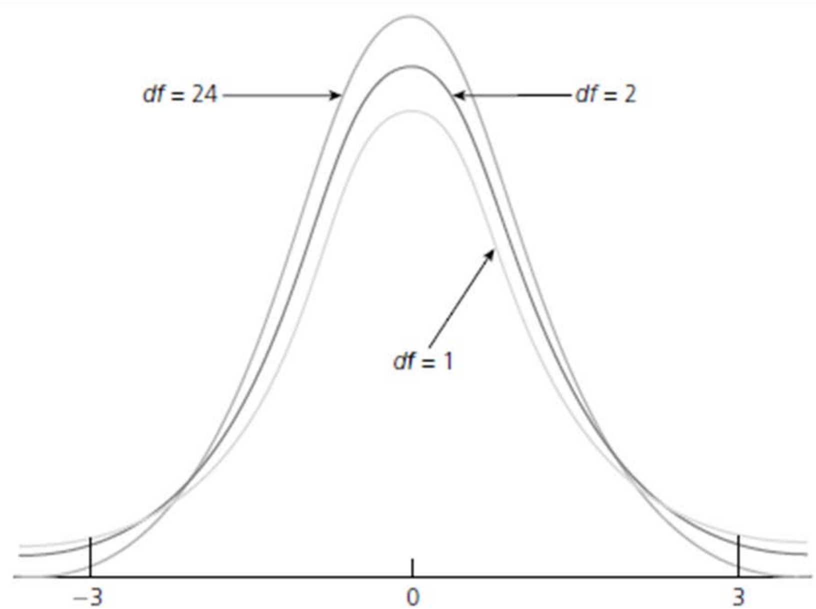
4. The *t* distribution

The pdf of the *t* distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails.

The expected value of a *t* distributed random variable is 0, and the variance is $^{n}/_{n-2}$ for $n > 2$.

As the degrees of freedom gets large, the *t* distribution approaches the standard normal distribution (which has expected value 0 and variance 1).

FIGURE B.10
The *t* distribution with various degrees of freedom.



POLS 6481. Research Design and Quantitative Methods II
Lecture 7. Standard Errors of and Hypothesis Tests for Regression Coefficients
Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

## 4. The *t* distribution

The pdf of the *t* distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails.

The expected value of a *t* distributed random variable is 0, and the variance is ${}^{n}/_{n-2}$ for $n > 2$.

As the degrees of freedom gets large, the *t* distribution approaches the standard normal distribution (which has expected value 0 and variance 1).

| | | Significance Level | | | | |
|---|---|---|---|---|---|---|
| 1-Tailed: | | .10 | .05 | .025 | .01 | .005 |
| 2-Tailed: | | .20 | .10 | .05 | .02 | .01 |
| Degrees of Freedom | 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| | 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| | 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| | 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| | 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| | 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| | 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| | 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| | 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| | 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| | 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| | 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| | 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| | 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| | 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| | 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| | 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| | 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| | 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| | 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| | 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| | 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| | 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| | 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| | 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| | 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| | 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| | 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| | 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| | 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| | 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| | 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| | 90 | 1.291 | 1.662 | 1.987 | 2.368 | 2.632 |
| | 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| | ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

## 5. Confidence intervals for a regression coefficient

Begin with a significance level (conventionally 5%)
The confidence level is 1 minus significance (i.e., 95% confidence)
The $t$ distribution gives a critical value for a given *confidence level* and *degrees of freedom*

A confidence interval is centered at $\widehat{\beta}_j$, and extends just far enough to cover 95% of the area under the $t$ distribution:

$$P\left(\widehat{\beta}_j - \underbrace{c_{0.05} \cdot se(\widehat{\beta}_j)}_{-\text{ margin of error}} \leq \beta_j \leq \widehat{\beta}_j + \underbrace{c_{0.05} \cdot se(\widehat{\beta}_j)}_{+\text{ margin of error}}\right) = 0.95$$

The margin of error equals the product of:

the $t$ critical value (denoted $c_{0.05}$) $\times$ the standard error of the coefficient (denoted $se(\widehat{\beta}_j)$)

One can carry out a hypothesis test using a confidence interval:

If the interval does not contain 0, reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j \neq 0$

Readings: Wooldridge, *Introductory Econometrics 5e*, 2.5 + 3.3 + 3.4 + 3.5 + 4.2 + 4.3

6. Null hypothesis retention region for a regression coefficient

Null hypothesis: $H_{0j} : \beta_j = 0$

Retention region is an interval centered at the null hypothesized value (0) , and extends just far enough to cover 95% of the area under the $t$ distribution

$$P(-c_{0.05} \cdot se(\widehat{\beta}_j) \leq 0 \leq c_{0.05} \cdot se(\widehat{\beta}_j)) = 0.95$$

Null hypothesis: $H_{0j} : \beta_j = a_j$

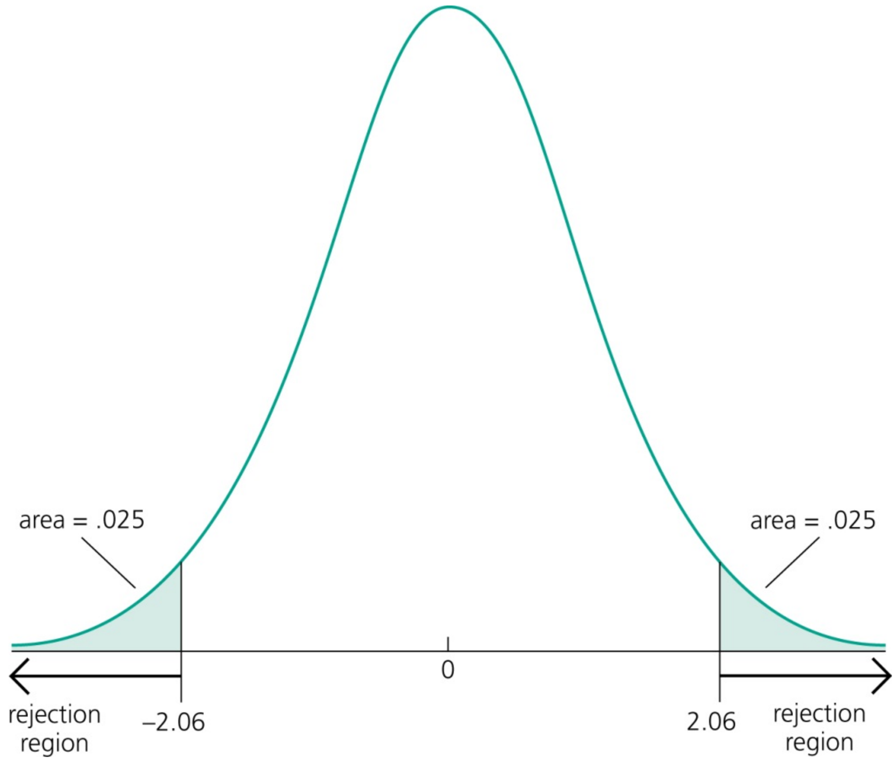
Retention region is an interval centered at the null hypothesized value ($a_j$) , and extends just far enough to cover 95% of the area under the $t$ distribution

$$P(a_j - c_{0.05} \cdot se(\widehat{\beta}_j) \leq a_j \leq a_j + c_{0.05} \cdot se(\widehat{\beta}_j)) = 0.95$$

The $t$ critical value for 95% confidence ($c_{0.05}$) is found using the $t$ table with $n-k-1$ *d.f.*

## 6. Null hypothesis retention region for a regression coefficient

For the following figure, note that the $t$ critical value with 25 *d.f.* equals ±2.06




retain $H_0 : \beta_j = 0$

Note that we never *accept* the alternative hypothesis ($H_{1j} : \beta_j \neq 0$); rather we *reject* the null if the $t$ statistic falls in the rejection region.