Univariate Regression: Hypothesis Tests and Confidence Intervals (SW Ch. 5)

Part 1

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Mar 21, 2022

Outline

1. **The standard error of $\hat{\beta}_1$**
2. **Hypothesis tests concerning β_1**
3. **Confidence intervals for β_1**
4. Regression when X is binary
5. Heteroskedasticity and homoskedasticity
6. Efficiency of OLS and the Student t distribution

Review (Lecture 5):
Sample Mean Hypothesis Test Example

The average adult male height in a certain country is 170 cm. We suspect that the men in a certain city in that country might have a different average height due to some environmental factors. We pick a random sample of size 9 from the adult males in the city and obtain the following values for their heights (in cm):

$$176.2 \quad 157.9 \quad 160.1 \quad 180.9 \quad 165.1 \quad 167.2 \quad 162.9 \quad 155.7 \quad 166.2$$

Assume that the height distribution in this population is normally distributed. Here, we need to decide between

$$H_0 : \mu = 170$$
$$H_1 : \mu \neq 170$$

Based on the observed data, is there enough evidence to reject H_0 at significance level $\alpha = 0.05$?

Review:
Sample Mean Hypothesis Test Example (cont'd)

Solution:

Let's first compute the sample mean and the sample standard deviation. The sample mean is

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9}{9}$$
$$= 165.8$$

The sample variance is given by

$$S^2 = \frac{1}{9-1} \sum_{k=1}^{9} (X_k - \bar{X})^2 = 68.01$$

The sample standard deviation is given by $S = \sqrt{S^2} = 8.25$

Review:
Sample Mean Hypothesis Test Example (cont'd)

Now, our test statistic is

$$W(X_1, X_2, \cdots, X_9) = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{165.8 - 170}{\frac{8.25}{3}} = -1.52$$

Thus, $|W| = 1.52$. Also, we have

$$t_{\frac{\alpha}{2}, n-1} = t_{0.025, 8} \approx 2.31$$

Thus, we conclude

$$|W| \leq t_{\frac{\alpha}{2}, n-1}$$

Therefore, we cannot reject H_0. In other words, we do not have enough evidence to conclude that the average height in the city is different from the average height in the country.

A big picture review of where we are going…

We want to learn about the slope of the population regression line. We have data from a sample, so there is sampling uncertainty. There are five steps towards this goal:

1. State the population object of interest
2. Provide an estimator of this population object
3. Derive the sampling distribution of the estimator (this requires certain assumptions). In large samples this sampling distribution will be normal by the CLT.
4. The square root of the estimated variance of the sampling distribution is the standard error (SE) of the estimator
5. Use the *SE* to construct *t*-statistics (for hypothesis tests) and confidence intervals.

Object of interest: $\boldsymbol{\beta_1}$ (1 of 2)

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots, n$$

$$\beta_1 = \text{slope of population regression line}$$

Estimator: the OLS estimator $\hat{\beta}_1$.

The Sampling Distribution of $\hat{\beta}_1$:

Because the population regression line is $E(Y|X) = \beta_0 + \beta_1 X$, $E(u_i \mid X_i) = 0$.

To derive the large-sample distribution of $\hat{\beta}_1$ assume :

- (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.
- Large outliers in X and/or Y are rare (X and Y have four moments)

These are the second and third least squares assumptions.

Object of interest: β_1 (2 of 2)

The Sampling Distribution of $\hat{\beta}_1$:

For n large, $\hat{\beta}_1$ is approximately distributed,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ where } v_i = (X_i - \mu_X)u_i$$

Hypothesis Testing and the Standard Error of $\hat{\beta}_1$ (Section 5.1)

The objective is to test a hypothesis, like $\beta_1 = 0$, using data – to reach a tentative conclusion whether the (null) hypothesis is correct or incorrect.

General setup

Null hypothesis and **two-sided** alternative:

$$H_0\colon\ \beta_1 = \beta_{1,0} \text{ vs. } H_1\colon\ \beta_1 \neq \beta_{1,0}$$

where $\beta_{1,0}$ is the hypothesized value under the null.

Null hypothesis and **one-sided** alternative:

$$H_0\colon\ \beta_1 = \beta_{1,0} \text{ vs. } H_1\colon\ \beta_1 < \beta_{1,0}$$

General approach: construct *t*-statistic, and compute *p*-value (or compare to the *N*(0,1) critical value)

- ***In general***: $$\boxed{t = \frac{\text{estimator - hypothesized value}}{\text{standard error of the estimator}}}$$

  where the *SE* of the estimator is the square root of an estimator of the variance of the estimator.

- ***For testing the mean of Y***: $$t = \frac{\bar{Y} - \mu_{Y,0}}{S_Y / \sqrt{n}}$$

- ***For testing β_1***, $$\boxed{t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)},}$$

  where $SE(\hat{\beta}_1)$ = the square root of an estimator of the variance of the sampling distribution of $\hat{\beta}_1$

THE t TEST OF SIGNIFICANCE: DECISION RULES

Type of hypothesis	H_0 : the null hypothesis	H_1 : the alternative hypothesis	Decision rule: reject H_0 if
Two-tail	$\beta_1 = \beta_{1,0}$	$\beta_1 \neq \beta_{1,0}$	$\lvert t \rvert > t_{\alpha/2,\ \mathrm{df}}$
Right-tail	$\beta_1 \leq \beta_{1,0}$	$\beta_1 > \beta_{1,0}$	$t > t_{\alpha,\ \mathrm{df}}$
Left-tail	$\beta_1 \geq \beta_{1,0}$	$\beta_1 < \beta_{1,0}$	$t < -t_{\alpha,\ \mathrm{df}}$

Notes:

- $\beta_{1,0}$ is the hypothesized numerical value of β_1.
- $|t|$ means the absolute value of t.
- $t_{\alpha,\ \mathrm{df}}$ or $t_{\alpha/2,\ \mathrm{df}}$ means the critical t value at the α or $\alpha/2$ level of significance.
- df: degrees of freedom, (n-2) for the two-parameter model, (n-3) for the three-parameter model, and so on.

Formula for $SE(\hat{\beta}_1)$ (1 of 2)

Recall the expression for the variance of (large *n*):

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n(\sigma_X^2)^2}, \text{ where } v_i = (X_i - \mu_X)u_i.$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_v^2 and σ_X^2 by estimators constructed from the data:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\text{estimator of } \sigma_v^2}{(\text{estimator of } \sigma_X^2)^2} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

where $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$.

Formula for $SE(\hat{\beta}_1)$ (2 of 2)

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}, \text{ where } \hat{v}_i = (X_i - \bar{X})\hat{u}_i.$$

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \text{the standard error of } \hat{\beta}_1$$

This is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates [var(X)]^2.
- Why the degrees-of-freedom adjustment n – 2? Because two coefficients have been estimated (β_0 and β_1).
- $SE(\hat{\beta}_1)$ is computed by regression software
- Your regression software has memorized this formula so you don’t need to.

Summary:
To test H_0: $\beta_1 = \beta_{1,0}$ v. H_1: $\beta_1 \neq \beta_{1,0}$,

- Construct the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Reject at 5% significance level if $|t| > 1.96$
- The p-value is $p = \Pr[|t| > |t^{act}|]$ = probability in tails of normal outside $|t^{act}|$; you reject at the 5% significance level if the p-value is < 5%.
- This procedure relies on the large-n approximation that $\hat{\beta}_1$ is normally distributed; typically $n = 50$ is large enough for the approximation to be excellent.

Example: *Test Scores* and *STR*, California data (1 of 2)

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

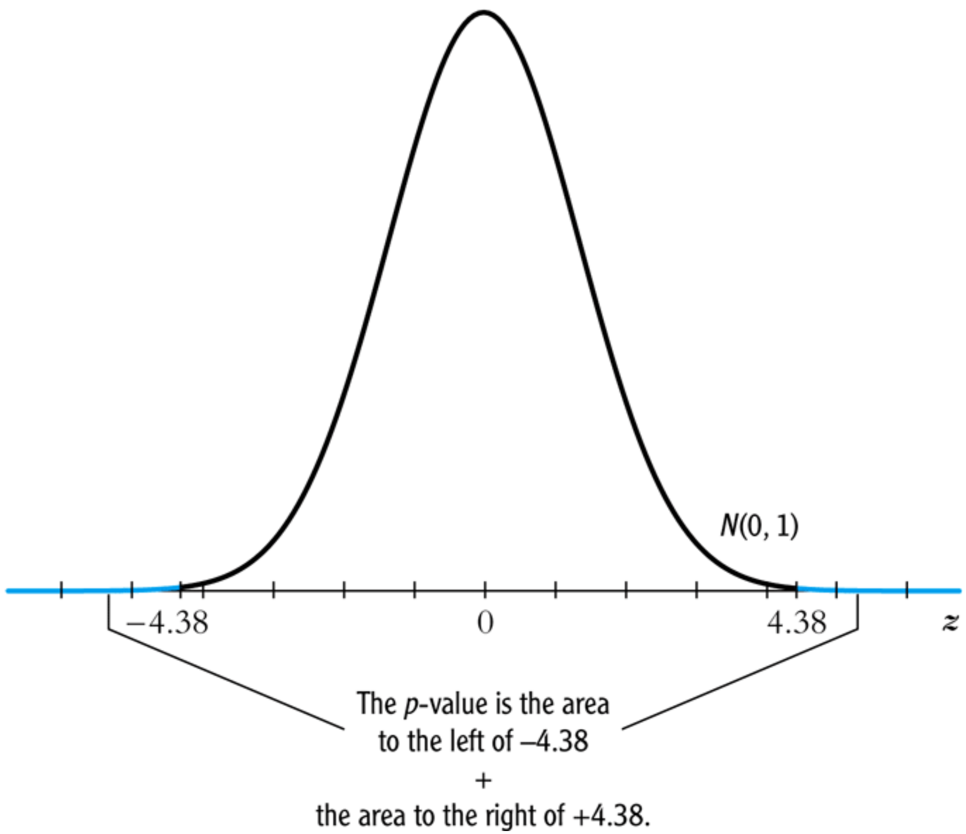
Regression software reports the standard errors:

$$SE(\hat{\beta}_0) = 10.4 \qquad SE(\hat{\beta}_1) = 0.52$$

$$t\text{-statistic testing } \beta_{1,0} = 0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.52} = -4.38$$

- The 1% 2-sided significance level is 2.58, so we reject the null at the 1% significance level.
- Alternatively, we can compute the *p*-value…

Example: *Test Scores* and *STR*, California data (2 of 2)



The *p*-value based on the large-*n* standard normal approximation to the *t*-statistic is 0.00001 (10^{-5})

Confidence Intervals for β_1 (Section 5.2)

Recall that a 95% confidence is, equivalently:

- The set of points that cannot be rejected at the 5% significance level;
- A set-valued function of the data (an interval that is a function of the data) that contains the true parameter value 95% of the time in repeated samples.

Because the *t*-statistic for β_1 is N(0,1) in large samples, construction of a 95% confidence for β_1 is just like the case of the sample mean:

$$\text{95\% confidence interval for } \beta_1 = \{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$$

Confidence interval example: Test Scores and STR
Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$$SE(\hat{\beta}_0) = 10.4 \qquad\qquad SE(\hat{\beta}_1) = 0.52$$

95% confidence interval for $\hat{\beta}_1$:

$$\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\} = \{-2.28 \pm 1.96 \times 0.52\}$$

$$= (-3.30,\ -1.26)$$

The following two statements are equivalent *(why?)*

- The 95% confidence interval does not include zero;
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level

A concise (and conventional) way to report regressions: Put standard errors in parentheses below the estimated coefficients to which they apply.

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR,\ R^2 = .05,\ SER = 18.6$$

<u>This expression gives a lot of information</u>

- The estimated regression line is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

  - The standard error of $\hat{\beta}_0$ is 10.4
  - The standard error of $\hat{\beta}_1$ is 0.52

- The R^2 is .05; the standard error of the regression is 18.6

OLS regression: reading STATA output

```
regress testscr str, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.38    0.000    -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44    0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

so:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \;,\; R^2 = .05,\; SER = 18.6$$

$t(\beta_1 = 0) = -4.38$, $p-$value $= 0.000$ (2−sided)

95% 2−sided conf. interval for β_1 is $(-3.30,\ -1.26)$

Summary of statistical inference about β_0 and β_1

Estimation:

- OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have approximately normal sampling distributions in large samples

Testing:

- H_0: $\beta_1 = \beta_{1,0}$ v. $\beta_1 \neq \beta_{1,0}$ ($\beta_{1,0}$ is the value of β_1 under H_0)
- $t = (\hat{\beta}_1 - \hat{\beta}_{1,0})/SE(\hat{\beta}_1)$
- p-value = area under standard normal outside t^{act} (large n)

Confidence Intervals:

- 95% confidence interval for β_1 is $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$
- This is the set of β_1 that is not rejected at the 5% level
- The 95% CI contains the true β_1 in 95% of all samples.