Economics 361
Two "Case Studies" for Hypothesis Testing

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

Overview

For all of these "case studies," assume that you are given a random sample of some population (X, Y) which satisfies the Gauss-Markov assumptions:

1. Linearity: $E[Y|X] = X\beta$
2. Spherical Errors: $V(Y|X) = \sigma^2 I$
3. Full Rank: $(X'X)^{-1}$ exists

In addition, let us add a fourth assumption

4. Normality: $Y|X$ is distributed multivariate Normal

You should be able to explain why the four assumptions, together, imply that $Y|X \sim N(X\beta, \sigma^2\ I)$

In all cases, you do not observe β. You will need to estimate β using OLS. In some cases, you also do not observe σ^2.

Later, we will revisit how these hypothesis tests change when we are unable or unwilling to make the fourth assumption – in which case, we are forced to use *asymptotic* distributions ...

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

1 Case Study: Vegas Line and Efficient Market Hypothesis

Consider the two earlier readings on using sports gambling data to test the Efficient Market Hypothesis. For this case study, your size N random sample of (X, Y) is

$$Y = \begin{pmatrix} PS_1 - VL_1 \\ \vdots \\ PS_N - VL_N \end{pmatrix} \qquad X = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_N \end{pmatrix}$$

For simplicity, assume there is only one "public information" item, Z. Y is $(N \times 1)$ and X is $(N \times 2)$

The Efficient Market Hypothesis (EMH) essentially argues that

$$BP_{MSE}(PS_i|Z_i) = E[PS_i|Z_i] = VL_i$$

i.e. Vegas Line (VL_i) captures all relevant predictive information about PS_i contained in Z_i.

We know from the Overview that the random sample satisfies the G-M assumptions. Therefore

$$E[Y|X] = X\beta = \beta_0 + \beta_1 Z$$

But also note that

$$\begin{aligned} E[Y|X] &= E[PS - VL\ |X] = E[PS - VL\ |Z] \quad \text{as the only random variable in } X \text{ is } Z \\ &= E[PS|Z] - E[VL|Z] = E[PS|Z] - E[\ \underbrace{E[PS|Z]}_{=VL \text{ under EMH}}\ |Z] \\ &= 0 \quad \text{as } E[\ E[PS|Z]\ |Z] = E[PS|Z] \end{aligned}$$

Therefore, EMH along with the G-M assumptions implies that $\beta_0 = 0$ and $\beta_1 = 0$.

$\beta_0 = 0$ portion of the implication mainly argues that the Vegas Line should not be systematically under or over predicting the Point Spread. Otherwise, gamblers can improve their performance by taking the Vegas Line and adjusting it by the systematic bias. The spirit of the EMH lies with the second portion, $\beta_1 = 0$, which indicates that factoring the actual realized value of Z does not improve a gambler's performance once the Vegas Line is considered.

So, the relevant hypothesis test of the EMH in this setting can be considered

- $H_o : \beta_1 = 0$ (EMH holds)
- $H_a : \beta_1 \neq 0$ (EMH does not hold)

We can test the above hypothesis using one of two test statistic: the Z-statistic or t-statistic. Which statistic we use depends on whether we know (or assume to know) the value of σ^2

1.1 Z-statistic: σ^2 Known

For notational simplicity, we will use Y_i in place of $PS_i - VL_i$.

We know from the four assumptions given in the Overview that

$$\begin{pmatrix} b_0^{ols} \\ b_1^{ols} \end{pmatrix} \mid X \;\sim\; N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\; \sigma^2 (X'X)^{-1} \right)$$

Note: $(X'X)$ is a (2×2) matrix and, therefore, so is $(X'X)^{-1}$. Without loss of generality, denote elements of $(X'X)$ as

$$(X'X) \;=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Remember: the values of (a, b, c, d) are known (can be calculated from the data). You should be able to explain why for this case

$$a \;=\; \sum_{i=1}^{N} (1 \times 1) \;=\; N \qquad b \;=\; \sum_{i=1}^{N} (1 \times Z_i) \;=\; c \qquad d \;=\; \sum_{i=1}^{N} (Z_i)^2$$

Inverting a (2×2) matrix is not difficult

$$(X'X)^{-1} \;=\; \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Therefore

$$\begin{pmatrix} b_0^{ols} \\ b_1^{ols} \end{pmatrix} \mid X \;\sim\; N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\; \frac{\sigma^2}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \right)$$

Recall that random variables jointly distributed (multivariate) Normal have marginal distributions that are also Normal. Therefore

$$b_1^{ols} \mid X \;\sim\; N(\, \beta_1 \,,\; \sigma^2 \frac{a}{ad - bc} \,)$$

Using change of variables

$$\underbrace{\frac{b_1^{ols} - r}{\sqrt{\sigma^2 \frac{a}{ad-bc}}}}_{\text{Z-stat}} \mid X \;\sim\; N(\, \frac{\beta_1 - r}{\sqrt{\sigma^2 \frac{a}{ad-bc}}} \,,\; 1) \qquad \text{for any known real valued } r$$

r is defined by the null hypothesis being tested with the Z-statistic:

- $H_o : \beta_1 - r = 0$

For this case study testing the EMH, $r = 0$ and, therefore, the particular Z-statistic

$$\underbrace{\frac{b_1^{ols}}{\sqrt{\sigma^2 \frac{a}{ad-bc}}}}_{\text{Z-stat}} \mid X \quad \sim \quad N(\frac{\beta_1}{\sqrt{\sigma^2 \frac{a}{ad-bc}}} , 1)$$

When H_o is true ($\beta_1 = 0$), the above distribution simplifies to the Standard Normal, $N(0, 1)$. The Z-statistic is a test statistic whose sampling distribution, when H_o is true, is Standard Normal.

Note that if we know (or willing to assume) σ^2, the above Z-statistic can be calculated just from the data (sample). But if σ^2 is unknown, we cannot calculate the above Z-statistic; we will have to use a different test statistic!

So, we now have the hypotheses (null and alternate) and the test statistic. We just need to specify the critical region – the values of the test statistic for which we reject H_o. The critical region depends critically (pun intended) on two things

- Alternate Hypothesis (H_a)
- Chosen Significance Level – i.e. Prob(Type I Error) when H_o is true

For this case study, $H_a : \beta_0 \neq 0$. This means that we want to reject when our data suggests a β_1 value that is much larger than zero or much smaller than zero. Alternate hypotheses of this "$\neq$" type lead to **"two-sided"** tests, tests whose critical regions draw from both extremes.

Additionally, as discussed in lecture, we want to draw from both extremes symmetrically, suggesting that the critical region should be of the following type

Reject if Z-statistic $> C$ or Z-statistic $< -C$ where C is some positive real value

This C is often dubbed the "Critical Value" (i.e. the value that defines the critical region).

The value of C depends on the chosen significance level (α). C should be chosen such that

$$\text{Prob}(\{\text{Z-stat} > C\} \cup \{\text{Z-stat} < -C\}) = \alpha \qquad \text{when } H_0 \text{ true}$$

Recall from above that when H_o is true, Z-stat is distributed N(0,1), which is symmetric around 0. Therefore

$$\text{Prob}(\{\text{Z-stat} > C\}) \; = \; \text{Prob}(\{\text{Z-stat} < -C\}) \; = \; \frac{\alpha}{2}$$

The appropriate value for C can be solved using the Distribution Table for a Standard Normal.

Example: for $\alpha = 0.1$, $C = 1.64$. Reject H_o if the Z-stat is greater than 1.64 or less than -1.64.

1.2 t-statistic: σ^2 Unknown

When σ^2 is unknown, the Z-statistic cannot be calculated. The denominator, a function of σ^2, is unknown! A different statistic must be used. This different statistic is credited to William Gosset. The sampling distribution that characterizes this statistic is named after the pseudonym he used to publish his pioneering work.

Gosset's work suggests that the following may be used as an estimate for the unknown σ^2

$$s^2 \quad = \quad \frac{1}{N-k} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{k-1} b_j^{ols} X_{ji} \right)^2$$

where X_{ji} is the value of the X_j random variable for the i^{th} observation, N the number of observations in the sample and k the number of variables in X.

Note that the above is just the sum of squared residuals (SSR) divided by $N-k$: $\frac{1}{N-k}\sum_i (e_i)^2$

Heuristically, the residual acts like the sample analog to the regression error ($\epsilon_i \equiv Y_i - \sum_j X_{ji}\beta_j$). As the variance of ϵ_i conditional on X is σ^2, we are using the sample variance of the residual as a (method of moments) estimate of σ^2. Note that $\sum_i e_i = 0$. We divide by $N-k$ rather than N to account for the fact that calculating the residual requires estimates for the k elements in β.

Using Gosset's work, we can show that

$$\underbrace{\frac{b_1^{ols} - r}{\sqrt{s^2 \frac{a}{ad-bc}}}}_{\text{t-stat}} \mid X \quad \sim \quad t_{N-k} \qquad \text{when } \beta_1 = r$$

(The change of variables involved is within your grasp except for two steps, which requires further knowledge of Linear Algebra and sampling theory; see Goldberger for full details)

The t-statistic is simply the Z-statistic with σ^2 replaced by s^2. This statistic, unlike the Z-statistic, can be calculated even if σ^2 is unknown.

For our case study, $r = 0$ and the particular t-statistic

$$\underbrace{\frac{b_1^{ols}}{\sqrt{s^2 \frac{a}{ad-bc}}}}_{\text{t-stat}} \mid X \quad \sim \quad t_{N-k} \qquad \text{when } \beta_1 = 0$$

The derivation of the appropriate critical region is analogous to that for the Z-statistic.

C should be chosen such that

$$\text{Prob}(\ \{\text{t-stat} > C\} \cup \{\text{t-stat} < -C\}\ ) = \alpha \qquad \text{when } H_0 \text{ true}$$

The t-distribution is also symmetric around 0. Therefore

$$\text{Prob}(\ \{\text{t-stat} > C\}\ )\ =\ \text{Prob}(\ \{\text{t-stat} < -C\}\ )\ =\ \frac{\alpha}{2}$$

The appropriate value for C can be solved using the Distribution Table for a t-distribution with $N-k$ degrees of freedom. For this case study, note that $k=2$.

Example: for $\alpha = 0.1$ and $N = 22$ (and, therefore, $N-k=20$), $C = 1.725$. Reject H_o if the t-stat is greater than 1.725 or less than -1.725.

2 Case Study: Cobb-Douglas and Returns to Scale

Consider the simple Cobb-Douglas production function discussed in the problem sets, where output is produced using two inputs: capital (K) and labor (L). We showed that the Cobb-Douglas production function is log-linear. Therefore, for this case study

$$Y\ =\ \begin{pmatrix} ln(Y_1) \\ \vdots \\ ln(Y_N) \end{pmatrix} \qquad X\ =\ \begin{pmatrix} 1 & ln(K_1) & ln(L_1) \\ \vdots & \vdots & \vdots \\ 1 & ln(K_N) & ln(L_N) \end{pmatrix}$$

From the G-M assumptions

$$E[lnY_i|X]\ =\ \beta_0 + \beta_1 ln(K_i) + \beta_2 ln(L_i)$$

A key concept in production theory is that of "returns to scale." When all inputs are doubled, by how much does output increase? When the increase in output is also double, the production function is said to exhibit *constant* returns to scale. When the increase in output is more than double, *increasing* returns to scale. When less than double, *decreasing* returns to scale. Additionally, when the production function exhibits increasing returns to scale, we say that there are "scale economies."

Whether "real world" production functions exhibited constant, increasing, or decreasing returns to scale used to be one of the major empirical issues in economics. Microeconomic theory shows that the "perfectly competitive" market is generally efficient only if production functions exhibit non-increasing returns to scale. Additionally, for some markets, increasing returns to scale could prohibit a market from achieving perfect competition: a firm could use its scale economies to drive out rivals.[1] Some researchers have used these results to conclude that if a market is (or nearly) perfectly competitive, production function must exhibit non-increasing returns to scale.[2]

Thus, hypothesis tests on the returns to scale for production in competitive markets are usually of the following type

- H_o : Constant Returns to Scale
- H_a : Decreasing Returns to Scale

[1] The issue of scale economies has died down since the advent of imperfect competition models – e.g. "oligopoly" models – and the realization that most real world markets are characterized by imperfect competition

[2] Yes, not the greatest application of logic ...

For Cobb-Douglas production function, returns to scale corresponds to the value of the sum of the coefficients before the factors of production.[3] In this case study, that sum is namely $\beta_1 + \beta_2$ and the above hypothesis test

- $H_o : \beta_1 + \beta_2 - 1 = 0$ (constant returns to scale)
- $H_a : \beta_1 + \beta_2 - 1 < 0$ (decreasing returns to sale)

Again, we can test the above hypothesis using either the Z-statistic or the t-statistic, depending on whether σ^2 is known

2.1 Z-statistic: σ^2 Known

We know from the four assumptions given in the Overview that

$$\begin{pmatrix} b_0^{ols} \\ b_1^{ols} \\ b_2^{ols} \end{pmatrix} \mid X \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 (X'X)^{-1} \right)$$

Note that $k = 3$ and, therefore, $(X'X)^{-1}$ is a (3×3) matrix. For notational simplicity, let

$$(X'X)^{-1} = \Omega = \begin{pmatrix} \Omega_{00} & \Omega_{01} & \Omega_{02} \\ \Omega_{10} & \Omega_{11} & \Omega_{12} \\ \Omega_{20} & \Omega_{21} & \Omega_{22} \end{pmatrix}$$

Therefore

- $V(b_1^{ols})|X) = \sigma^2 \Omega_{11}$
- $\text{Cov}(b_1^{ols}, b_2^{ols}|X) = \sigma^2 \Omega_{12} = \sigma^2 \Omega_{21}$

As $(b_0^{ols}, b_1^{ols}, b_2^{ols})$ are jointly distributed multivariate normal, linear functions of $(b_1^{ols}, b_2^{ols}, b_3^{ols})$ are distributed normal as well. Therefore

$$\underbrace{\frac{b_1^{ols} + b_2^{ols} - r}{\sqrt{V(b_1^{ols} + b_2^{ols} - r|X)}}}_{\text{Z-stat}} \mid X \sim N\left(\frac{\beta_1 + \beta_2 - r}{\sqrt{V(b_1^{ols} + b_2^{ols} - r|X)}}, 1 \right)$$

Note that when $\beta_1 + \beta_2 = r$ the above distribution simplifies to the Standard Normal. For our particular case study, $r = 1$ and therefore

$$\underbrace{\frac{b_1^{ols} + b_2^{ols} - 1}{\sqrt{V(b_1^{ols} + b_2^{ols} - 1|X)}}}_{\text{Z-stat}} \mid X \sim N\left(\frac{\beta_1 + \beta_2 - 1}{\sqrt{V(b_1^{ols} + b_2^{ols} - 1|X)}}, 1 \right)$$

[3] You should be able to explain why

The remaining task is to solve for $V(b_1^{ols} + b_2^{ols} - 1|X)$ using the above joint distribution.

$$\begin{aligned} V(b_1^{ols} + b_2^{ols} - 1|X) &= V(b_1^{ols}|X) + V(b_2^{ols}|X) + 2\ \mathrm{Cov}(b_1^{ols}, b_2^{ols}|X) \\ &= \sigma^2\Omega_{11} + \sigma^2\Omega_{22} + 2\sigma^2\Omega_{12} \end{aligned}$$

Therefore

$$\underbrace{\frac{b_1^{ols} + b_2^{ols} - 1}{\sqrt{\sigma^2(\Omega_{11} + \Omega_{22} + 2\Omega_{12})}}}_{\text{Z-stat}} \mid X \sim N\left(\frac{\beta_1 + \beta_2 - 1}{\sqrt{\sigma^2(\Omega_{11} + \Omega_{22} + 2\Omega_{12})}},\ 1 \right)$$

Again, the above Z-stat is distributed Standard Normal, N(0,1), when H_o is true: $\beta_1 + \beta_2 - 1 = 0$. If σ^2 is known, the Z-stat can be calculated as b^{ols} and Ω are simply functions of the sample.

The critical region for this Z-statistic differs from that of the earlier case study as the alternate hypothesis is no longer"two-sided." We want to reject mainly when the sample implies a value of $\beta_1 + \beta_2$ that is much smaller than 1 (consistent with decreasing returns to scale). We call such tests, where we favor rejection based on one of the extremes, **"one-sided"** tests. Therefore, we reject if the Z-statistic is less than some critical value C.

C should be chosen such that

$$\text{Prob}(\ \{\text{Z-stat} < C\}\ ) = \alpha \qquad \text{when } H_0 \text{ true}$$

α is, again, the chosen significance level.

Example: For $\alpha = 0.05$, C = -1.64. Reject H_o if Z-stat is less than -1.64.

2.2 t-statistic: σ^2 Unknown

When σ^2 is unknown, we cannot calculate the Z-statistic. But for reasons analogous to those in the first case study, we can show that

$$\underbrace{\frac{b_1^{ols} + b_2^{ols} - 1}{\sqrt{s^2(\Omega_{11} + \Omega_{22} + 2\Omega_{12})}}}_{\text{t-stat}} \mid X \sim t_{N-k} \qquad \text{when } H_o \text{ is true}$$

Here, $k = 3$ as there are three variables in X. The critical value is chosen in an analogous manner as the Z-stat, except the Distribution Tables for the t is used instead of the Standard Normal. Note: again, the t-stat is the Z-stat with s^2 instead of σ^2

Example: for $\alpha = 0.05$ and $N = 22$ (and, therefore, $N - k = 20$), $C = -1.725$. Reject H_o if the t-stat is less than -1.725.

3 Asymptotic Hypothesis Testing

Suppose we are unwilling to make the fourth assumption in the Overview: $Y|X \sim$ Multivariate Normal. All of the hypothesis tests discussed in the two earlier case studies require this fourth assumption.

However, if we have a random sample and are willing to make the following assumptions about the property of our sample as $N \to \infty$

$$\lim_{N\to\infty} \frac{(X'X)}{N} = Q_{XX} \qquad \lim_{N\to\infty} \frac{(X'Y)}{N} = Q_{XY}$$

where Q_{XX} and Q_{XY} are some well defined matrices with finite valued elements

we can show that

$$\sqrt{N}\left(b^{ols} - \beta\right) \quad \xrightarrow{d} \quad N(0, \sigma^2 Q_{XX}^{-1})$$

using the Central Limit Theorem (CLT).

The above essentially states that as $N \to \infty$ (sample size approaches infinite), the distribution of $\sqrt{N}\left(b^{ols} - \beta\right)$ calculated from the sample converges to that of $N(0, \sigma^2 Q_{XX}^{-1})$

By practice (not by proper mathematics/statistics), change of variables is "applied" to the above to write the following statement

$$b^{ols} \quad \overset{a}{\sim} \quad N(\beta \ , \ \frac{\sigma^2}{N} Q_{XX}^{-1})$$

which is "read" as saying that b^{ols} has an **asymptotic distribution** of $N(\beta, \frac{\sigma^2}{N} Q_{XX}^{-1})$

Note that $Q_{XX}^{-1} = (Q_{XX})^{-1} = (\ \lim_{N\to\infty} \frac{(X'X)}{N}\ )^{-1}$. Consequently, again by practice (not proper mathematics/statistics), we "argue" that

$$(X'X)^{-1} \quad \approx \quad \frac{1}{N}\, Q_{XX}^{-1} \qquad \text{for "large" values of } N$$

Note also that by the Law of Large Numbers (LLN), s^2 converges to σ^2 as $N \to \infty$. Therefore, by practice, if we have a random sample and are willing to make the above assumptions about Q_{XX} and Q_{XY}, we "argue" that

$$\underbrace{\frac{b_j^{ols} - r}{\sqrt{s^2 (X'X)_{jj}^{-1}}}}_{\text{Asymp. Z-stat}} \quad \overset{a}{\sim} \quad N(\frac{\beta_j - r}{\sqrt{s^2 (X'X)_{jj}^{-1}}} \ , \ 1)$$

The rest of the hypothesis testing process is the same as that of a regular Z-statistic. Essentially, we calculate a t-statistic and treat it like a Z-statistic. Asymptotic hypothesis testing is something of a desperation move. The mathematics/statistics underlying such tests is, at best, *maybe* reasonable for tests involving "large" N random samples. But these asymptotic hypothesis tests are, currently, among the most popular undertaken, not only in economics but also other empirical disciplines.