Data 5600 4/20/2022

Chapter 4 - Inference in the MLP

knowing that

$$E(\hat{\beta}_j) = \beta_j$$

and that

$$s.e.(\hat{\beta}_j) = \frac{\sigma}{[SST_j(1-R_j^2)]^{1/2}}$$

isn't quite enough to do inference.

We need to make an assumption about the distribution.

We will make one additional assumption:

<u>Assumption MLR.6 - Normality</u>

The population error u is independent of the explanatory variables $x_1, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2.

$$u \sim N(\mu = 0, \sigma^2)$$

With Assumptions MLR.1 through MLR.6 we have the Classical Linear Regression Model (CLR).

A succinct way to summarize the population assumptions of the CLM is

$$y \mid \mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \sigma^2)$$

Where $\mathbf{x}$ (bold) is shorthand for $(x_1, x_2, \ldots, x_k)$

Normal Sampling Distributions

Under the CLM assumptions MLR.1 through MLR.6, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim N\left(\beta_j, \mathrm{Var}(\hat{\beta}_j)\right)$$

Note:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j\,(1 - R_j^2)}$$

Where is given in chapter 3. Therefore

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

Testing Hypotheses about a Single Population Parameter

Our population model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

and we assume that it satisfies the CLR assumptions. Then

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Where $k+1$ is the number of unknowns in the population model. (k slope parameters and the intercept)

Primarily, we will be interested in testing the null hypothesis

$$H_0 : \beta_j = 0$$

where j corresponds to any of the k explanatory variables. In simple language, this means that after the variables $x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_k$ have been accounted for, x_j has <u>no effect</u> on the expected value of y.

Ex: p. 121

Consider the wage regression

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

The null hypothesis $H_0 : \beta_2 = 0$ means that, once tenure and education have been accounted for, the number of years in the work force (exper) has no effect on hourly wage. This is economically interesting. If true, it implies that a person's work history does not affect wage. If $\beta_2 > 0$, then prior work experience contributes to productivity, and hence wage.

The statistic we use to conduct the null hypothesis is the t statistic of $\hat{\beta}_j$

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

The Appropriate Rejection Regions:

- When $H_1: \beta_j > 0$ the rejection region is $t_{\hat{\beta}_j} > c$
- When $H_1: \beta_j < 0$ the rejection region is $t_{\hat{\beta}_j} < -c$
- When $H_1: \beta_j \neq 0$ the rejection region is $|t_{\hat{\beta}_j}| > c$

Note:
c is the critical value from a t dist with $df = n-k-1$ for some given α level say $\alpha = .1$ or $\alpha = .05$

Ex: pp. 126-127

Our sample contains $n = 408$ high schools in michigan in 1993.

We can use these data to test the null hypothesis that school size has no effect on standardized test scores against the alternative that size has a negative effect.

Performance is measured by the percentage of students receiving a passing score on the Michigan Educational Assessment Program (MEAP) standardized tenth-grade math test (math10). School size is measured by student enrollment (enroll).

The null hypothesis is

$$H_0 : \beta_{enroll} = 0$$

The alternative hypothesis is

$$H_1 : \beta_{enroll} < 0$$

We control for average annual teacher compensation (a proxy for quality) (totcomp) and the number of staff per 1000 students (staff).

The estimated equation, with standard errors is

$$\widehat{math10} = \underset{(6.113)}{2.274} + \underset{(.0001)}{.00046}\, totcomp + \underset{(.04)}{.048}\, staff - \underset{(.00022)}{.0002}\, enroll$$

$n = 408$

$R^2 = .0541$

The coefficient on enroll is $-.0002$ is in agreement that larger schools hamper performance.

Since $\underset{=408}{n} - \underset{=3}{k} - 1 = 404$, we can use the standard normal distribution.

For $\alpha = .05$, the critical value is -1.65

Our t statistic is

$$t = \frac{-0.0002}{0.00022} \approx -0.91$$

$$t > \underset{-1.96}{\text{critical value}} \implies \text{we fail to reject } H_0.$$

We conclude that enroll is not statistically significant at the $\alpha = .05$ level.

Ex: R code for example 4.3 on pp. 128–129

Wooldridge's estimates and standard errors

$$\widehat{colGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{0.412}\, hsGPA + \underset{(.011)}{0.015}\, ACT - \underset{(.026)}{0.083}\, skipped$$

$n = 141$, $R^2 = .234$

$$t = \left| \frac{\hat{\beta}_{skipped}}{se(\hat{\beta}_{skipped})} \right| = \left| -\frac{0.083}{0.026} \right| = |-3.19| = 3.19$$

α	z_α
$\alpha = .05$	1.96
$\alpha = .01$	2.58

so skipped is statistically significant at the 5% and 1% levels!

Testing Other Hypotheses About β_j

$H_0: \beta_j = 0$ is the most common hypothesis, but sometimes we want to test whether β_j is equal to some other given constant.

Generally,

$$H_0: \quad \beta_j = a_j$$

where a_j is the hypothesized value of β_j.

the appropriate t statistic is

$$t = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)}$$

The general way to remember this is

$$t = \frac{(\text{estimate} - \text{hypothesized value})}{\text{standard error}}$$

Computing P-Values

The p-value for testing the null hypothesis

$$H_0: \beta_j = 0$$

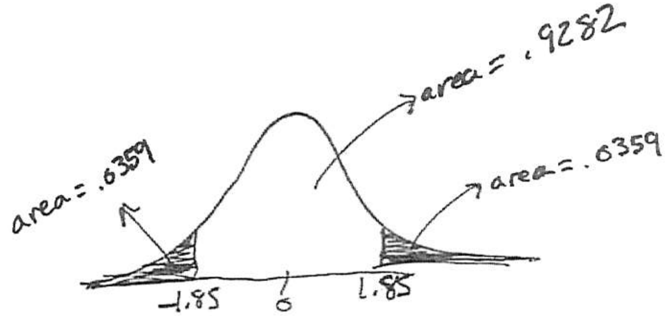
against the two-sided alternative is

$$P(|T| > |t|)$$

in which T is a t distributed random variable with $n-k-1$ degrees of freedom and t is the numerical value of our t statistic.

<u>See Figure 4.6</u>

$t = 1.85$ and $df = 40$



$$\text{p-value} = P(|T| > 1.85) = 2 * P(T > 1.85)$$
$$= 2 * (.0359)$$
$$= .0718$$

oops

NB: in R code:

pval = 2*(1 - pt(1.85, df=40))

= 0.0717 1068

$\approx$ 0.0718

R FTW!

<u>Economic vs. Statistical Significance</u>

Hypothesis testing focuses on the <u>statistical significance</u> of X_j. We also need to pay attention to the magnitude of β_j in addition to the size of the t statistic.

Statistical significance of X_j is entirely determined by the size of $t_{\hat{\beta}_j}$, whereas <u>economic significance</u> or <u>practical significance</u> is related to the size (and sign) of $\hat{\beta}_j$.

<u>Note:</u> $t_{\hat{\beta}_j}$ can be statistically significant either because $\hat{\beta}_j$ is "large" or because "$se(\hat{\beta}_j)$" is small.

A variable can seem important even if its effect is very small in practical terms!

Confidence Intervals

Using the fact that $\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim t_{n-k-1}$

leads to a simple rule for confidence intervals for the unknown population β_j. A 95% CI is

$$\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$$

in which c is the 97.5^{th} percentile in a t_{n-k-1} distribution

Testing Multiple Linear Restrictions – The F-Test

We know how to test whether a particular variable has no partial effect on the dependent variable: the t test!

We may want to test whether a group of variables has no effect on the dependent variable. More precisely, the null hypothesis is that a group of variables has no effect on y, once another set of variables has been controlled for.

<u>Ex: example on p. 143</u>

Consider the model that explains major league baseball player's salaries:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

where

salary = total 1993 salary

years = years in the MLB

gamesyr = average games played per year

bavg = career batting average (e.g. bavg = 250)

hrunsyr = home runs per year

rbisyr = runs batted in per year

NB: for the curious sabermetrics is the study of baseball statistics

Suppose we want to test that once years in the league has been and gamesyr controlled for, statistics measuring performance (bavg, hrunsyr, rbisyr) have no effect on salary. The null hypothesis is

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1: \text{Not } H_0$$

The null has 3 exclusion restrictions. If H_0 true then bavg, hrunsyr, and rbisyr have no effect on log(salary) after years and gamesyr have been controlled for.

We call this a joint hypothesis test.

The model without these three variables is

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u$$

In the context of hypothesis testing, we call this the restricted model and the original model the unrestricted model.

The restricted model always has fewer parameters than the restricted model.

Now we need a test statistic. This is the

F statistic defined by

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

in which

SSR_r = the residual sum of squares of the restricted model

SSR_{ur} = the residual sum of squares of the unrestricted model

$q = df_r - df_{ur}$ is the numerator degrees of freedom

Recall that $df =$ number of observations − number of parameters estimated

Note: $df_r > df_{ur}$ b/c $n =$ the same for both

The SSR in the denominator of F is divided by the degrees of freedom in the unrestricted model

$n-k-1 =$ denominator degrees of freedom $= df_{ur}$

Ex: in the baseball example if $n = 353$

$$\left.\begin{aligned} df_{ur} &= 353 - 6 = 347 \\ df_r &= 353 - 9 = 344 \end{aligned}\right\} \Rightarrow q = 3$$

Assuming that the CLM assumptions hold under the null

$$F \sim F_{q,\, n-k-1}$$

Once a critical value is selected the rejection region is

$$F > c$$

If H_0 is rejected we say that the set of explanatory variables excluded from the restricted model are jointly statistically significant.

Ex: for baseball (see p. 147) we have

$$F = \frac{(198.311 - 183.186)}{183.186} \cdot \frac{.347}{3} \approx 9.55$$

For $\alpha = .05$ $c = 2.76$, for $\alpha = .01$ $c = 4.13$

NB: ALSO $F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \frac{(n-k-1)}{q}$

$F = 9.55$ is well above the 1% critical value

so we soundly reject the null that barg, hrunsyr

and rbisyr have no effect on log(salary).

$\Rightarrow$ They are jointly statistically significant!

P-Values for F Tests

$$\text{p-value} = P(\mathcal{F} > F)$$

in which $\mathcal{F}$ is an F random variable with $(q, n-k-1)$ degrees of freedom and F is the actual value of our F statistic given our sample of data.

F Statistic for Overall Significance

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1:$ Not H_0

the restricted model is

$$y = \beta_0 + u$$

NB: just the constant

Note: this is the F statistic that R spits out from lm with a corresponding p-value