Class 15 - Multiple Regression Model Inference (Part I)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Multiple Regression Analysis: Inference

❑ Statistical inference in the regression model

- Hypothesis tests about population parameters
- Construction of confidence intervals

❑ Sampling distributions of the OLS estimators

- The OLS estimators are random variables
- We already know their expected values and their variances
- However, for hypothesis testing we need to know their distribution
- In order to derive their distribution we need additional assumptions
- Assumption about distribution of errors: normal distribution

Assumption MLR.6 **Normality**

The population error u is *independent* of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0,\sigma^2)$.

Discussion of the normality assumption

- The normality of the error term is an empirical question
- In many cases, normality is questionable or impossible by definition
- In some cases, normality can be achieved through transformations of the dependent variable (e.g. use log(wage) instead of wage)
- Under normality, OLS is the best (even nonlinear) unbiased estimator
- Important: For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size

As we will see in Chapter 5:
Nonnormality of the errors is not a serious problem with large sample sizes

Terminology

$\underbrace{MLR.1 - MLR.5}$

"Gauss-Markov assumptions"

$\underbrace{MLR.1 - MLR.6}$

"Classical linear model (CLM) assumptions"

THEOREM 4.1

NORMAL SAMPLING DISTRIBUTIONS

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

$$\hat{\beta}_j \sim \text{Normal}(\beta_j, Var(\hat{\beta}_j))$$

The estimators are normally distributed around the true parameters with the variance that was derived earlier

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}$$

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

The standardized estimators follow a standard normal distribution

THEOREM 4.2 — ***t* DISTRIBUTION FOR THE STANDARDIZED ESTIMATORS**

Under the CLM assumptions MLR.1 through MLR.6,

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

If the standardization is done using the estimated standard deviation (= standard error), the normal distribution is replaced by a t-distribution

Note: The t-distribution is close to the standard normal distribution if **n-k-1** is **large**.

Hypothesis testing

What is hypothesis testing?

- ❑ Hypothesis testing and statistical inference allow us to answer questions about the **population** by looking at the **sample**.
- ❑ It's almost impossible to prove a theory is "correct" with hypothesis testing.
- ❑ All that can be done with hypothesis testing is to state that a particular sample conforms to a particular hypothesis.

The hypothesis test is broken into two hypotheses:

1. **Null** hypothesis, denoted "H_0:", typically is a statement of the values not expected.
2. **Alternative** hypothesis denoted "H_1:", typically is a statement of the values expected. (what you want to show!)

Pregnancy Diagnosis

H_0: *You are not pregnant (negative)*
H_1: *You are pregnant (positive)*

False Positive

False Negative

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

Hypothesis testing (cont'd)

Significance level (or simply the **level**) is defined as the probability of a Type I error:

Probability of rejecting a null given that it is true:

$$\alpha = \mathrm{P}(\text{Reject } \mathrm{H}_0 | \mathrm{H}_0)$$

- When we specify a value for α, we are essentially quantifying our tolerance for a Type I error.
- If α =0.05, then the researcher is willing to falsely reject the null 5% of the time, in order to detect deviations from null.

Example: Wage equation

Test whether, after controlling for **education** and **tenure**, higher work **experience** leads to higher hourly wages

$$\widehat{\log(wage)} = .284 + .092\,educ + .0041\,exper + .022\,tenure$$
$$(.104)\ \ (.007)\qquad (.0017)\qquad (.003)$$
$$n = 526,\ R^2 = .316,$$

Test $H_0 : \beta_{exper} = 0$ against $H_1 : \beta_{exper} > 0$

One would either expect a positive effect of experience on hourly wage or no effect at all.

What is your test statistic?

t-statistic (or t-ratio)

Null hypothesis :

$$H_0 : \beta_j = 0$$

The **population parameter** is equal to zero, i.e. **After controlling for the other independent variables**, there is no effect of x_j on y

---

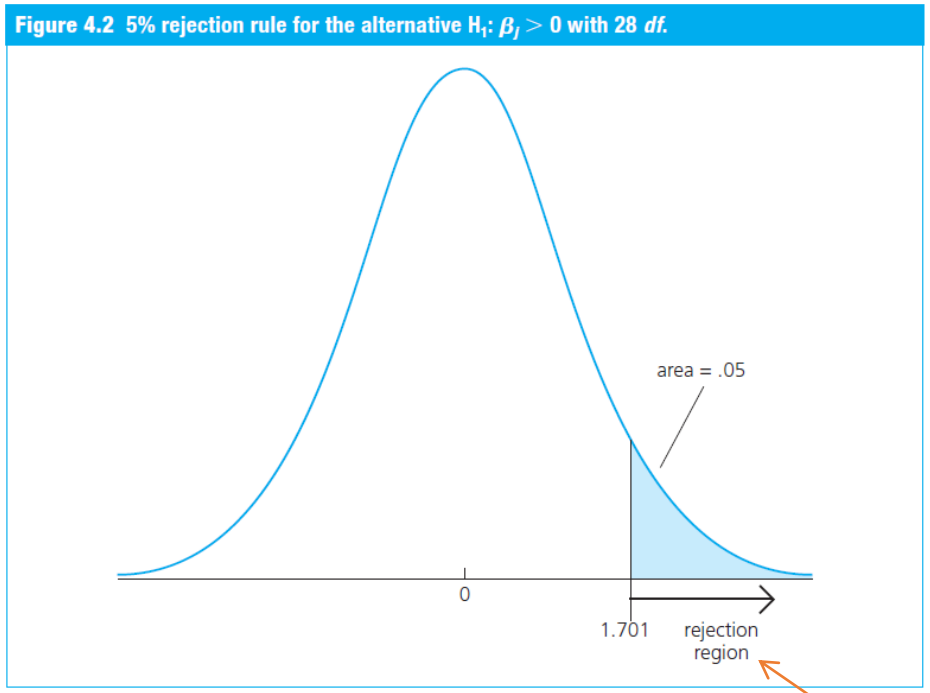
$$t_{\widehat{\beta_j}} = \frac{\widehat{\beta_j} - \beta_{j_{H_0}}}{se(\widehat{\beta_j})}$$

The t-statistic will be used to test the above null hypothesis. The farther the estimated coefficient is away from zero, the less likely it is that the null hypothesis holds true. But what does "far" away from zero mean?

This depends on the variability of the estimated coefficient, i.e. its standard deviation. The t-statistic measures how many **estimated standard deviations** the estimated coefficient is away from zero.

- Goal: Define a rejection rule so that, if it is true, H_0 is rejected

Testing against one-sided alternatives (greater than zero)



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j > 0$

- Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is "too large" (i.e. larger than a critical value).
- Significance level: the probability of rejecting the true null
- How do you calculate the critical value?
- What is the code in R?
- Rejection Rule: $t_{\hat{\beta}_j} > c$

Reject if t-statistic is greater than 1.701

Example: Wage equation (cont.)

$$t_{exper} = .0041/.0017 \approx 2.41$$

t-statistic

$$df = n - k - 1 = 526 - 3 - 1 = 522$$

Degrees of freedom; here the standard normal approximation applies **(why?)**

$$c_{0.05} = 1.645$$

$$c_{0.01} = 2.326$$

Critical values for the 5% and the 1% significance level (these are conventional significance levels).

The null hypothesis is rejected because the t-statistic exceeds the critical value.

"The effect of experience on hourly wage is statistically greater than zero at the 5% (and at the 1%) significance level."

Example: Student performance and school size

Test whether smaller **school size** leads to better **student performance,** after controlling for *staff* and *totcomp*.

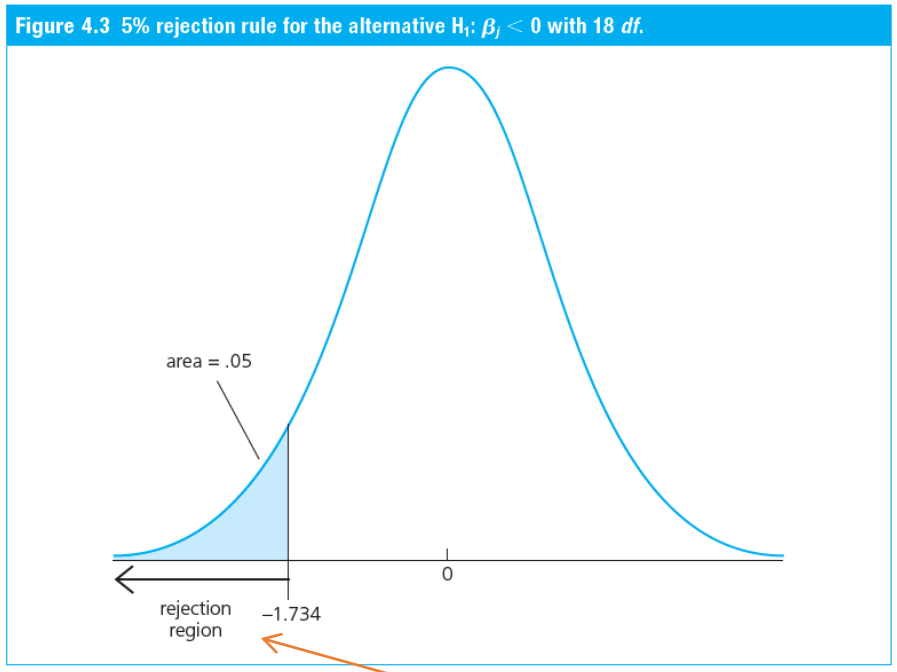
$$\widehat{math10} = 2.274 + .00046\, totcomp + .048\, staff - .00020\, enroll$$
$$(6.113)\ (.00010)\qquad (.040)\qquad (.00022)$$
$$n = 408,\ R^2 = .0541.$$

Test $H_0 : \beta_{enroll} = 0$ against $H_1 : \beta_{enroll} < 0$

Do larger schools hamper student performance or is there no such effect?

What is the t-stat?

Testing against one-sided alternatives (less than zero)



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j < 0$

- Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is "too small" (i.e. smaller than a critical value).
- Significance level: the probability of rejecting the true null
- How do you calculate the critical value?
- What is the code in R?
- Rejection Rule:

$$t_{\hat{\beta}_j} < -c$$

Reject if t-statistic is less than -1.734

Example: Student performance and school size (cont.)

$$t_{enroll} = -.00020/.00022 \approx -.91$$

t-statistic

$$df = n - k - 1 = 408 - 3 - 1 = 404$$

Degrees of freedom; here the standard normal approximation applies

$$c_{0.05} = -1.65$$

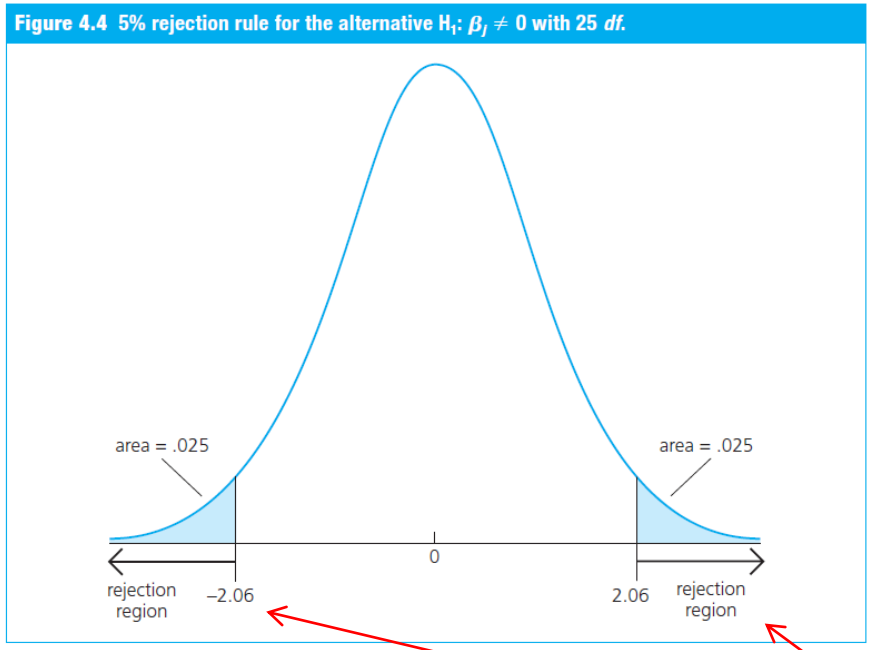
$$c_{0.15} = -1.04$$

Critical values for the 5% and the 15% significance level.

The null hypothesis is not rejected because the t-statistic is not smaller than the critical value.

We cannot reject the hypothesis that there is **no effect** of school size on student performance (not even for a large significance level of 15%).

Testing against two-sided alternatives



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$

- Reject the null hypothesis in favor of the alternative hypothesis if the absolute value of the estimated coefficient is too large.
- Significance level: the probability of rejecting the true null
- How do you calculate the critical value?
- What is the code in R?
- Rejection Rule:

$$|t_{\hat{\beta}_j}| > c$$

Reject if absolute value of t-statistic is less than -2.06 or greater than 2.06

Example: Determinants of college GPA

$$\widehat{colGPA} = 1.39 + .412\, hsGPA + .015\, ACT - .083\, skipped$$
$$(.33) \quad (.094) \qquad (.011) \qquad (.026)$$
$$n = 141, R^2 = .234.$$

```
Call:
lm(formula = colGPA ~ hsGPA + ACT + skipped, data = gpa1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85698 -0.23200 -0.03935  0.24816  0.81657

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38955    0.33155   4.191 4.95e-05 ***
hsGPA        0.41182    0.09367   4.396 2.19e-05 ***
ACT          0.01472    0.01056   1.393  0.16578
skipped     -0.08311    0.02600  -3.197  0.00173 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3295 on 137 degrees of freedom
Multiple R-squared:  0.2336,    Adjusted R-squared:  0.2168
F-statistic: 13.92 on 3 and 137 DF,  p-value: 5.653e-08
```

- ✓ The effects of hsGPA and skipped are significantly different from zero at the 1% significance level.
- ✓ The effect of ACT is not significantly different from zero, not even at the 10% significance level.

“Statistically significant” variables in a regression

- If a regression coefficient is different from zero in a **two-sided test**, the corresponding variable is said to be “statistically significant”

- If the number of degrees of freedom is large enough so that the normal approximation applies, the following rules of thumb apply:

$| t_{stat} | > 1.65$ → “statistically significant at 10% level”

$| t_{stat} | > 1.96$ → “statistically significant at 5% level”

$| t_{stat} | > 2.57$ → “statistically significant at 1% level”

Testing more general hypotheses about a regression coefficient

Null hypothesis $H_0 : \beta_j = a_j$ ← Hypothesized value of the coefficient

t-statistic $$t = \frac{(estimate - hypothesized\ value)}{standard\ error} = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)}$$

The test works **exactly as before**, except that the hypothesized value is substracted from the estimate when forming the statistic.

Example: Campus crime and enrollment

An interesting hypothesis is whether **crime increases by one percent** if enrollment is increased by one percent

$$\widehat{\log}(crime) = \underset{}{-} \underset{(1.03)}{6.63} + \underset{(0.11)}{1.27} \log(enroll)$$

$$n = 97, \ R^2 = .585$$

Estimate is different from one but is this difference statistically significant?

Build the Hypothesis test and find the t-stat:

$$H_0 : \beta_{\log(enroll)} = 1, \ H_1 : \beta_{\log(enroll)} \neq 1$$

$$t = (1.27 - 1)/.11 \approx 2.45 > 1.96 = c_{0.05}$$

The hypothesis is rejected at the 5% level

Important! This t-stat is **different** than what is reported by statistical packages!