Class 16 – Multiple Regression Model Inference (Part II)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Computing p-values for t-tests

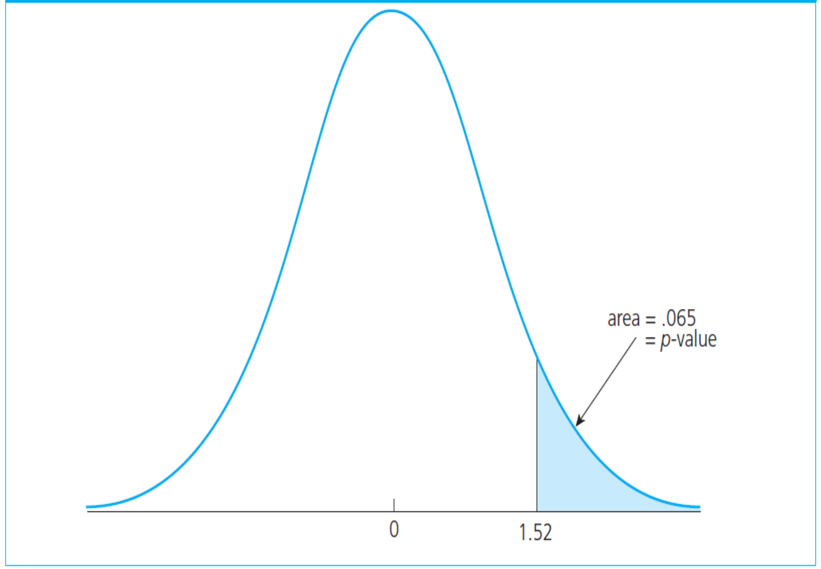
- Recall 1: The smallest significance level at which the null hypothesis is still rejected, is called the p-value of the hypothesis test
- Recall 2: **p-value** is the corresponding significance level of the test statistic.
- A small p-value is evidence against the null hypothesis **(a good thing!)** because one would reject the null hypothesis even at small significance levels
- A large p-value is evidence in favor of the null hypothesis **(a bad thing!)**
- P-values are more informative than tests at fixed significance levels
- The p-value is the significance level at which one is indifferent between rejecting and not rejecting the null hypothesis.

Computing and Using p-values (cont'd)

We said that **p-value** is the corresponding significance level of the test statistic.

P-values for one-tailed tests:

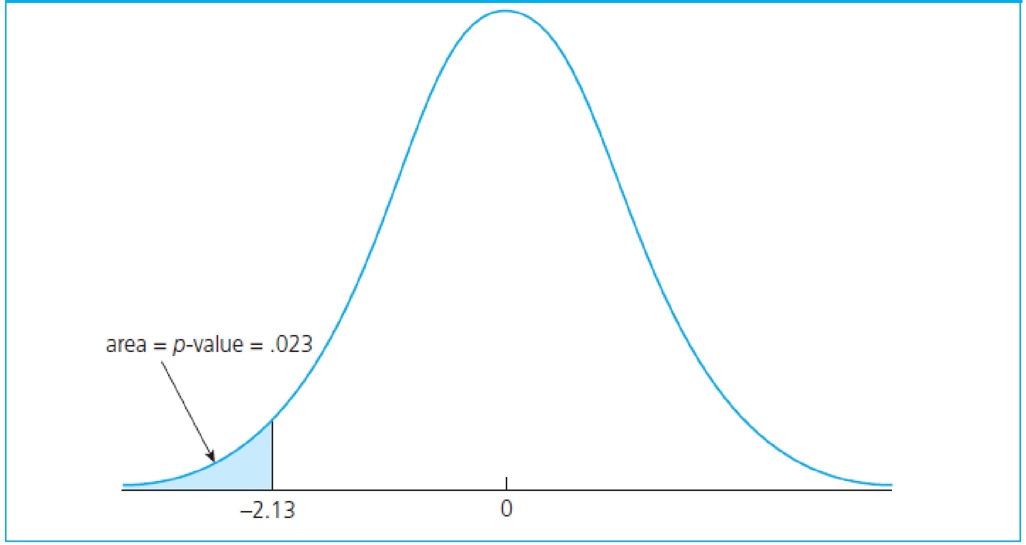
FIGURE C.7 The p-value when $t = 1.52$ for the one-sided alternative



$$p_{value} = P(T > 1.52) = 1 - CDF(1.52) = 0.065$$

n=200

FIGURE C.8 The p-value when $t = -2.13$ with 19 degrees of freedom for the one-sided alternative $\mu < 0$.



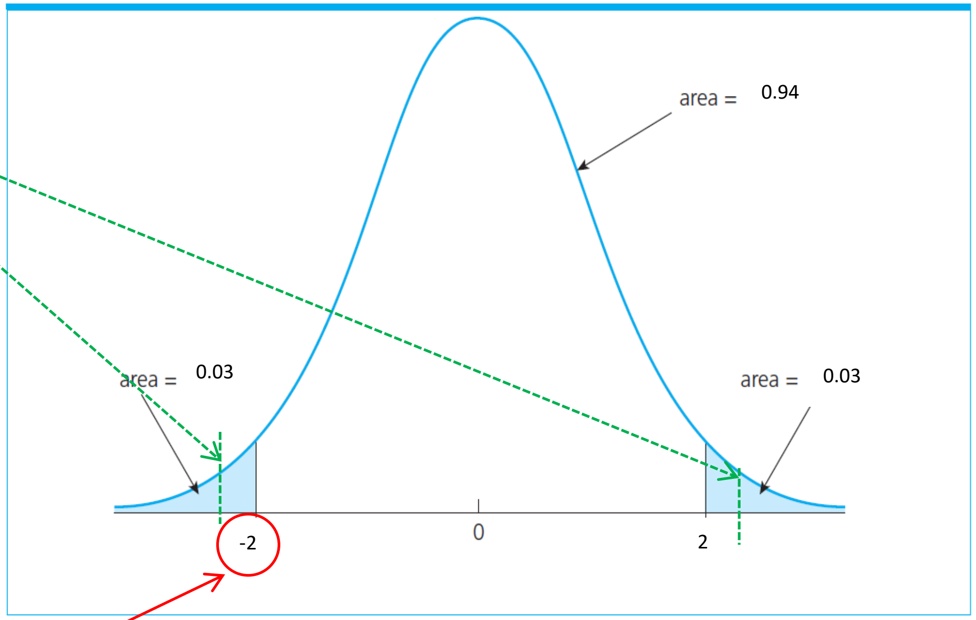
$$p_{value} = P(T < -2.13) = CDF(-2.13) = 0.023$$

n=20

Computing and Using p-values (cont'd)

P-values for Two-tailed tests:

These would be the critical values for a 5% significance level



Value of test statistic

$$p_{value} = P(|T| > |2|) =$$

$$2(1 - CDF(2)) = 0.06 \ ,$$

$$df = 20$$

A null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level.

Do you reject the null here?

$p_{value} = 6\%$, $\alpha = 5\%$

Economic / Practical significance VS. Statistical significance

- If a variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its economic or practical importance

- The fact that a coefficient is statistically significant does not necessarily mean it is economically or practically significant!

- If a variable is statistically and economically important but has the “wrong” sign, the regression model might be misspecified

- If a variable is NOT statistically significant at the usual levels (10%, 5%, or 1%), one may think of dropping it from the regression

- If the sample size is small, effects might be imprecisely estimated so that the case for dropping insignificant variables is less strong

Confidence intervals

Recall: CI is two-sided by nature $\hat{\beta}_j \pm C * se(\hat{\beta}_j)$

Critical value of **two-sided** test

$$P\Big(\underbrace{\hat{\beta}_j - c_{0.05} \cdot se(\hat{\beta}_j)} \leq \beta_j \leq \underbrace{\hat{\beta}_j + c_{0.05} \cdot se(\hat{\beta}_j)}\Big) = 0.95$$

Lower bound of the Confidence interval

Upper bound of the Confidence interval

Confidence level

Interpretation of the confidence interval:

- The bounds of the interval are random
- In repeated samples, the **interval** will **contain** the population regression coefficient (β) in $(1 - \alpha)\%$ of the cases

Confidence intervals for typical confidence levels

$$P\left(\hat{\beta}_j - c_{0.01} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.01} \cdot se(\hat{\beta}_j)\right) = 0.99$$

$$P\left(\hat{\beta}_j - c_{0.05} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.05} \cdot se(\hat{\beta}_j)\right) = 0.95$$

$$P\left(\hat{\beta}_j - c_{0.10} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.10} \cdot se(\hat{\beta}_j)\right) = 0.90$$

Use rules of thumb $c_{0.01} = 2.576, c_{0.05} = 1.96, c_{0.10} = 1.645$

Relationship between confidence intervals and hypotheses tests

$a_j \notin interval \;\Rightarrow\;$ reject $H_0 : \beta_j = a_j$ in favor of $H_1 : \beta_j \neq a_j$

Example: Model of firms‘ R&D expenditures

Spending on R&D

Annual sales

Profits as percentage of sales

$$\widehat{\log(rd)} = -4.38 + 1.084 \log(sales) + .0217\, profmarg$$
$$\quad (.47) \quad (.060) \qquad\qquad (.0128)$$
$$n = 32,\ R^2 = .918. \qquad df = 32-2-1 = 29 \Rightarrow c_{0.05} = 2.045$$

What are the CI for β_1 and β_2 ?

$$1.084 \pm 2.045(.060)$$
$$= (.961, 1.21)$$

The effect of sales on R&D is relatively **precisely** estimated as the interval is narrow. Moreover, the effect is significantly different from zero because zero is outside the interval.

$$.0217 \pm 2.045(.0218)$$
$$= (-.0045, .0479)$$

This effect is **imprecisely** estimated as the interval is very wide. It is not even statistically significant because zero lies in the interval.

R

```r
# chapter 4: MRM, Inference

library(wooldridge)
library(stargazer)


# Example 4-8
MRM <- lm(log(rd)~ log(sales)+ profmarg, rdchem)
summary(MRM)

# finding critical values
df    <- nobs(MRM) - 2-1
alpha <- 0.05
qt(1- alpha/2 , df)


# Look at t_stat
summary(MRM)$coefficients[,"t value"]


# Confidence Interval
confint(MRM, level =  1-alpha)
```

```
> summary(MRM)

Call:
lm(formula = log(rd) ~ log(sales) + profmarg, data = rdchem)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97681 -0.31502 -0.05828  0.39020  1.21783 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.37827    0.46802  -9.355 2.93e-10 ***
log(sales)   1.08422    0.06020  18.012  < 2e-16 ***
profmarg     0.02166    0.01278   1.694    0.101    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5136 on 29 degrees of freedom
Multiple R-squared:  0.918,	Adjusted R-squared:  0.9123 
F-statistic: 162.2 on 2 and 29 DF,  p-value: < 2.2e-16
```

```
> qt(1- alpha/2 , df)
[1] 2.04523
```

```
> # Look at t_stat
> summary(MRM)$coefficients[,"t value"]
(Intercept)  log(sales)    profmarg 
  -9.354916   18.011791    1.694150 
```

```
> # Confidence Interval
> confint(MRM, level =  1-alpha)
                   2.5 %     97.5 %
(Intercept) -5.335478450 -3.4210681
log(sales)   0.961107256  1.2073325
profmarg    -0.004487722  0.0477991
```

Testing hypotheses about a linear combination of the parameters

Example: Return to education at two-year vs. at four-year colleges

Number of years attending a **2-year** college

Number of years at a **4-year** college

Months in the workforce

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(.021)}{1.472} + \underset{(.0068)}{.0667}\, jc + \underset{(.0023)}{.0769}\, univ + \underset{(.0002)}{.0049}\, exper \\ n &= 6{,}763,\ R^2 = .222.\end{aligned}$$

- ✓ The **hypothesis** of interest is whether one year at a junior college is worth one year at a university
- ✓ the **alternative** of interest is one-sided: a year at a junior college is worth less than a year at a university

Testing hypotheses about a linear combination of the parameters

- ✓ The **hypothesis** of interest is whether one year at a junior college is worth one year at a university
- ✓ the **alternative** of interest is one-sided: a year at a junior college is worth less than a year at a university

---

Test $H_0 : \beta_1 - \beta_2 = 0$ against $H_1 : \beta_1 - \beta_2 < 0$

A possible test statistic would be:

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$

The difference between the estimates is normalized by the estimated standard deviation of the difference. The null hypothesis would have to be rejected if the statistic is "too negative" to believe that the true difference between the parameters is equal to zero.

Testing hypotheses about a linear combination of the parameters (cont'd)

Impossible to compute with standard regression output because

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

Usually not available in regression output

Alternative method:

Define $\theta_1 = \beta_1 - \beta_2$ and test $H_0 : \theta_1 = 0$ against $H_1 : \theta_1 < 0$

$$\log(wage) = \beta_0 + (\theta_1 + \beta_2) jc + \beta_2 univ + \beta_3 exper + u$$

$$= \beta_0 + \theta_1 jc + \beta_2 \boxed{(jc + univ)} + \beta_3 exper + u$$

Insert into original regression

a new regressor (= total years of college)

Testing hypotheses about a linear combination of the parameters (cont'd)

Total years of college

Estimation results

$$\widehat{\log(wage)} = 1.472 - .0102\ jc + .0769\ totcoll + .0049\ exper$$
$$(.021)\ \ (.0069)\qquad (.0023)\qquad\qquad (.0002)$$
$$n = 6,763,\ \ R^2 = .222$$

$$t = -.0102/.0069 = -1.48$$
$$p - value = P(t - ratio < -1.48) = .070$$

Hypothesis is rejected at 10% level but not at 5% level

Confidence Interval for $\theta_1 = \beta_1 - \beta_2$ $\longrightarrow$ $-.0102 \pm 1.96(.0069) = (-.0237, .0003)$

This method works always for single linear hypotheses