Ch04. Multiple Regression Analysis: Inference

Ping Yu

HKU Business School
The University of Hong Kong

Sampling Distributions of the OLS Estimators

Statistical Inference in the Regression Model

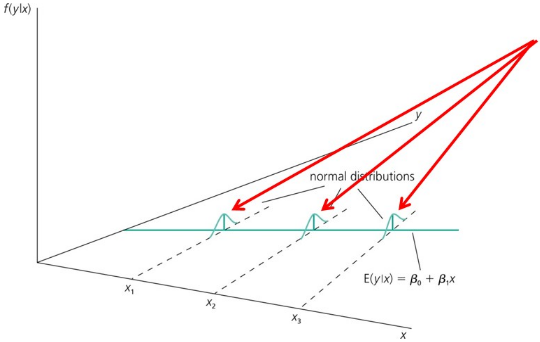
- Hypothesis tests about population parameters.
- Construction of confidence intervals.
  - These two tasks are closely related.

We need to study sampling distributions of the OLS estimators for statistical inference!

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

Standard Assumptions for the MLR Model (continue)

- Assumption MLR.6 (Normality): u_i is independent of $(x_{i1}, \cdots, x_{ik})$, and $u_i \sim N\left(0, \sigma^2\right)$.
  - It is stronger than MLR.4 (zero conditional mean) and MLR.5 (homoskedasticity).



It is assumed that the unobserved factors are normally distributed around the population regression function.

The form and the variance of the distribution does not depend on any of the explanatory variables.

It follows that:

$$y|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$

Discussion of the Normality Assumption

- The error term is the sum of "many" different unobserved factors.
- Sums of many independent and similarly distributed factors are normally distributed (central limit theorem or CLT).
- Problems:
  - How many different factors? Number large enough?
  - Possibly very heterogeneous distributions of individual factors.
  - How independent are the different factors?
  - Are they additive?
- The normality of the error term is an empirical question.
- In some cases, the error distribution should be "close" to normal, e.g., test scores.
- In many other cases, normality is questionable or impossible by definition.

continue

- Examples where normality cannot hold:
  - Wages (nonnegative; also: minimum wage). [figure here]
  - Number of arrests (takes on a small number of integer values). [figure here]
  - Unemployment (indicator variable, takes on only 1 or 0).
- 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, i.e., it is the minimum variance unbiased estimator.
- Important: For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size (Chapter 5, will not be covered).
- Terminology:
  - MLR.1-MLR.5: Gauss-Markov assumptions
  - MLR.1-MLR.6: classical linear model (CLM) assumptions

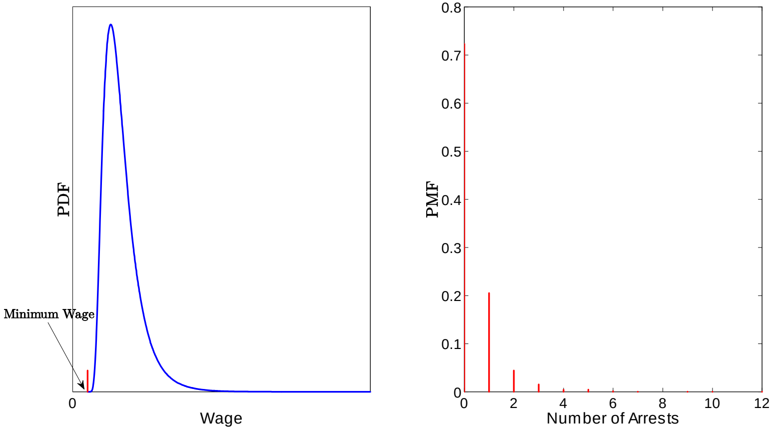


Figure: PDF of Wage and PMF of Number of Arrests: the minimum wage in HK is HK$37.5/hr, and in Beijing RMB24/hr at present.

Normal Sampling Distributions

- Theorem 4.1: Under assumptions MLR.1-MLR.6,

$$\widehat{\beta}_j \sim N\left(\beta_j, Var\left(\widehat{\beta}_j\right)\right).$$

Therefore,

$$\frac{\widehat{\beta}_j - \beta_j}{sd\left(\widehat{\beta}_j\right)} \sim N(0,1).$$

- The estimators are normally distributed around the true parameters with the variance that was derived earlier.
  - Note that as before, we are conditioning on $\{\mathbf{x}_i, i = 1, \cdots, n\}$.
- The standardized estimators follow a standard normal distribution.
  - Recall that if $X \sim N\left(\mu, \sigma^2\right)$, then

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$

Testing Hypotheses about a Single Population Parameter: The t Test

t-Distribution for Standardized Estimators

- Theorem 4.2: Under assumptions MLR.1-MLR.6,

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} \sim t_{n-k-1} = t_{df}.$$

- If the standardization is done using the estimated standard deviation (= standard error), the normal distribution is replaced by a t-distribution. [why? see the slide after introducting the t-distribution.]
- The t-distribution is close to the standard normal distribution if $n-k-1$ is large.
- The t-distribution is named after Gosset (1908)[photo here], "The probable error of a mean". At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the student's t rather than Gosset's t! The name "t" was popularized by R.A. Fisher (we will discuss him later).

History of the t-Distribution

William S. Gosset (1876-1937)

[Review] *t*-Distribution

- If Z is a standardized normal variable,

$$Z \sim N(0,1),$$

and variable X has a χ^2 (chi-square) distribution with ν degrees of freedom,

$$X \sim \chi^2_\nu,$$

independent of Z, then

$$\frac{Z}{\sqrt{X/\nu}} = \frac{\text{standard normal variable}}{\sqrt{\text{independent chi-square variable}/df}} \sim t_\nu,$$

a t-distribution with ν degrees of freedom.

continue

- If $Z_1, \cdots, Z_\nu$ are independently distributed random variables such that $Z_i \sim N(0,1)$, $i = 1, \cdots, \nu$, then
$$X = \sum_{i=1}^{\nu} Z_i^2$$
follows the χ^2 distribution with ν degrees of freedom.
- Note that
$$\sum_{i=1}^{\nu} Z_i^2 \Big/ \nu \to E\left[Z_i^2\right] = 1 \text{ as } \nu \to \infty,^1$$
so
$$t_\nu \to N(0,1) \text{ as } \nu \to \infty.$$
- Recall that $Var(Z_i) = E\left[Z_i^2\right] - E[Z_i]^2$, so $E\left[Z_i^2\right] = Var(Z_i) + E[Z_i]^2 = 1 + 0^2 = 1$.

[1] $E\left[\sum_{i=1}^{\nu} Z_i^2 / \nu\right] = E\left[Z_i^2\right]$ and $Var\left(\sum_{i=1}^{\nu} Z_i^2 / \nu\right) = Var\left(Z_i^2\right)/\nu \to 0$.

continue

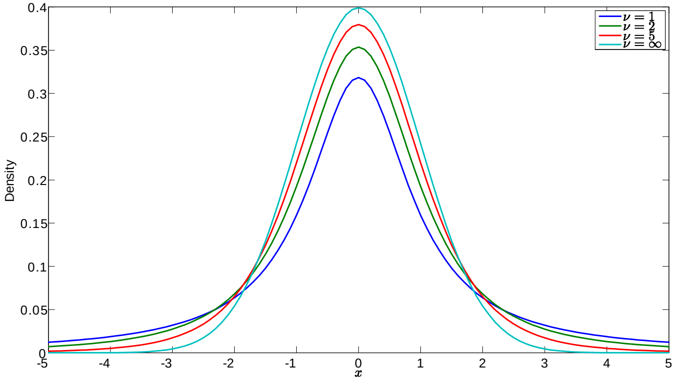


Figure: Density of the t Distribution for $\nu = 1, 2, 5, \infty$

(*) Why the Standardized $\widehat{\beta}_j$ Using Its SE Follows the t-Distribution?

- We provide only a rough idea here.
- Note that

$$
\begin{aligned}
\frac{\widehat{\beta}_j-\beta_j}{se\left(\widehat{\beta}_j\right)} &= \frac{\left(\widehat{\beta}_j-\beta_j\right)/sd\left(\widehat{\beta}_j\right)}{\sqrt{\frac{\sum_{i=1}^n \widehat{u}_i^2/(n-k-1)}{SST_j\left(1-R_j^2\right)}\Big/ Var\left(\widehat{\beta}_j\right)}} \\
&= \frac{\left(\widehat{\beta}_j-\beta_j\right)\Big/\sqrt{\frac{\sigma^2}{SST_j\left(1-R_j^2\right)}}}{\sqrt{\frac{\sum_{i=1}^n \widehat{u}_i^2/(n-k-1)}{SST_j\left(1-R_j^2\right)}\Big/\frac{\sigma^2}{SST_j\left(1-R_j^2\right)}}} \\
&= \frac{\left(\widehat{\beta}_j-\beta_j\right)/sd\left(\widehat{\beta}_j\right)}{\sqrt{\sum_{i=1}^n\left(\frac{\widehat{u}_i}{\sigma}\right)^2\Big/(n-k-1)}} \\
&\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-k-1}\Big/(n-k-1)}} = t_{n-k-1}
\end{aligned}
$$

t Statistic or t Ratio

- Suppose the null hypothesis (for more general hypotheses, see below) is

$$H_0 : \beta_j = 0.$$

- 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.
- The t statistic or t ratio of $\widehat{\beta}_j$ for **this** H_0 is

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se\left(\widehat{\beta}_j\right)}.$$

- 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. [figure here]
- If the null hypothesis is true,

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se\left(\widehat{\beta}_j\right)} = \frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} \sim t_{n-k-1}.$$

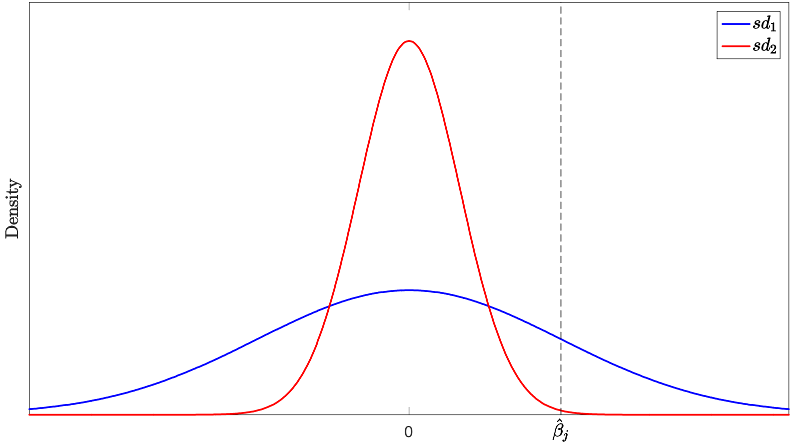


Figure: Same $\widehat{\beta}_j$ Values BUT Different Standard Deviations

a: Testing against One-Sided Alternatives (greater than zero)

- Goal: Define a rejection rule so that, if H_0 is true, it is rejected only with a small probability (= significance level or level for short, e.g., 5%).
- To determine the rejection rule, we need to decide the relevant alternative hypothesis.
- First consider a one-sided alternative of the form

$$H_1 : \beta_j > 0.$$

- Reject the null hypothesis in favor of the alternative hypothesis if the estimated coefficient is "too large" (i.e., larger than a critical value). (why?)
- Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases. [figure here]
- In the figure, this is the point of the t-distribution with 28 degrees of freedom that is exceeded in 5% of the cases.
- So the rejection rule is to reject if the t statistic is greater than 1.701.
- Analog: evidences cannot happen if you are not a criminal in testing

$$H_0 : \text{you are innocent vs. } H_1 : \text{you are guilty}$$

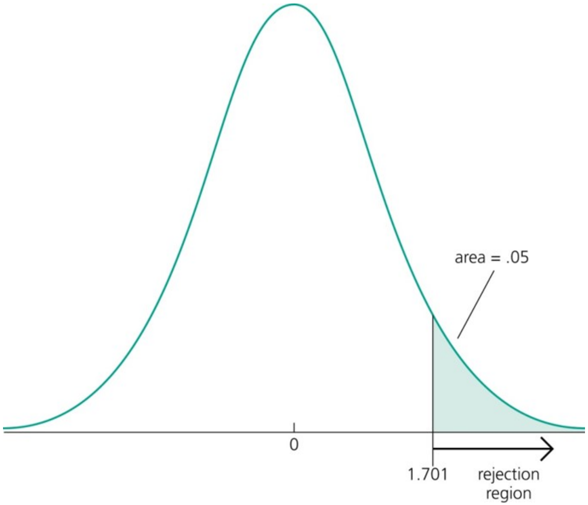


Figure: 5% Rejection Rule for the Alternative $H_1 : \beta_j > 0$ with 28 df

- Rejection region is the set of values of t statistic at which we will reject the null.

Example: Hourly Wage Equation

- Test whether, after controlling for education and tenure, higher work experience leads to higher hourly wages.
- The fitted regression line is

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

  where standard errors appear in parentheses below the estimated coefficients.
- Test

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

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

continue

- The t statistic for $\widehat{\beta}_{exper}$ is

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

- $df = n - k - 1 = 526 - 3 - 1 = 522$, quite large, so the standard normal approximation applies.
- The 5% critical value is $c_{0.05} = 1.645$, and the 1% critical value is $c_{0.01} = 2.326$.
  - 5% and 1% are conventional significance levels.
- The null hypothesis is rejected because the t statistic exceeds the critical value.
- The conclusion is that the effect of experience on hourly wage is statistically greater than zero at the 5% (and even at the 1%) significance level.

Testing against One-Sided Alternatives (less than zero)

- We want to test
$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j < 0$$
- Reject the null hypothesis in favor of the alternative hypothesis if the estimated coefficient is "too small" (i.e., smaller than a critical value). (why?)
- Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases. [figure here]
- In the figure, this is the point of the t-distribution with 18 degrees of freedom so that 5% of the cases are below the point.
- So the rejection rule is to reject if the t statistic is less than -1.734.

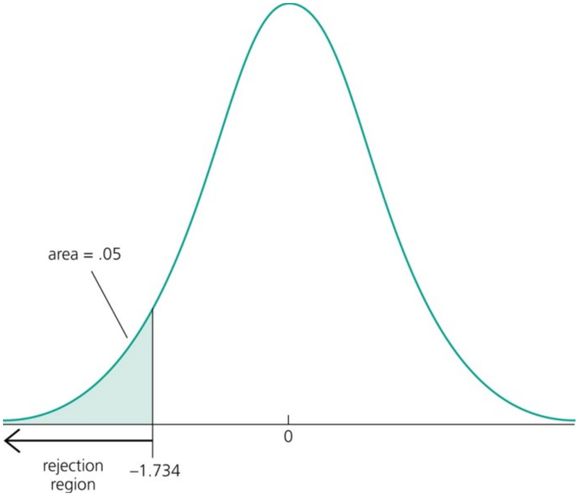


Figure: 5% Rejection Rule for the Alternative $H_1 : \beta_j < 0$ with 18 df

Example: Student Performance and School Size

- Test whether smaller school size leads to better student performance.
- The fitted regression line is

$$\begin{aligned} \widehat{math10} &= \underset{(6.113)}{2.274} + \underset{(.00010)}{.00046} totcomp + \underset{(.040)}{.048} staff - \underset{(.00022)}{.00020} enroll \\ n &= 408, R^2 = .0541 \text{ (quite small)} \end{aligned}$$

where
$math10$ = percentage of students passing 10th-grade math test
$totcomp$ = average annual teacher compensation
$staff$ = staff per one thousand students
$enroll$ = school enrollment (= school size)

continue

- Test

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

  - Do larger schools hamper student performance or is there no such effect?
- The t statistic for $\widehat{\beta}_{enroll}$ is

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

- $df = n - k - 1 = 408 - 3 - 1 = 404$, quite large, so the standard normal approximation applies.
- The 5% critical value is $c_{0.05} = -1.645$, and the 15% critical value is $c_{0.15} = -1.04$.
- The null hypothesis is not rejected because the t statistic is not smaller than the critical value.
- The conclusion is that one cannot reject the hypothesis that there is no effect of school size on student performance (not even for a lax significance level of 15%).

continue

- Using an alternative specification of functional form, we have

$$\begin{aligned}\widehat{math10} &= \underset{(48.70)}{-207.66} + \underset{(4.06)}{21.16}\log(totcomp) + \underset{(4.19)}{3.98}\log(staff) - \underset{(0.69)}{1.29}\log(enroll) \\ n &= 408, R^2 = .0654 \text{ (slightly higher, but still quite small)}\end{aligned}$$

- Test

$$H_0 : \beta_{\log(enroll)} = 0 \text{ against } H_1 : \beta_{\log(enroll)} < 0.$$

- The t statistic for $\widehat{\beta}_{\log(enroll)}$ is

$$t_{\log(enroll)} = \frac{-1.29}{0.69} \approx -1.87.$$

- $c_{0.05} = -1.645$, and $t_{\log(enroll)} < c_{0.05}$, so the hypothesis that there is no effect of school size on student performance can be rejected in favor of the hypothesis that the effect is negative.
- How large is the effect? quite small:

$$-1.29 = \frac{\partial math10}{\partial \log(enroll)} = \frac{\partial math10}{\partial enroll/enroll} = \frac{-1.29/100}{1/100} = \frac{-0.0129}{+1\%}.$$

b: Testing against Two-Sided Alternatives

- We want to test
$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0$$
- Reject the null hypothesis in favor of the alternative hypothesis if the **absolute value** of the estimated coefficient is too large. (why?)
- Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases. [figure here]
- In the figure, these are the points of the t-distribution so that 5% of the cases lie in the two tails.
- So the rejection rule is to reject if the t statistic is greater than 2.06 or less than -2.06.

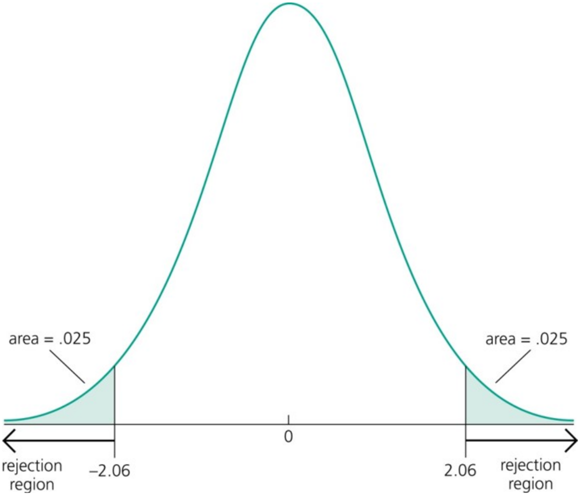


Figure: 5% Rejection Rule for the Alternative $H_1 : \beta_j \neq 0$ with 25 df

Example: Determinants of College GPA

- The fitted regression line is

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

where

$skipped$ = average number of lectures missed per week

- $df = n - k - 1 = 141 - 3 - 1 = 137$, quite large, so the standard normal approximation applies.
- $t_{hsGPA} = 4.38 > c_{0.01} = 2.576$
- $t_{ACT} = 1.36 < c_{0.10} = 1.645$
- $|t_{skipped}| = |-3.19| > c_{0.01} = 2.576$
- 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:

$$\begin{aligned} |t \text{ ratio}| &> 1.645 \Longrightarrow \text{statistically significant at 10\% level} \\ |t \text{ ratio}| &> 1.96 \Longrightarrow \text{statistically significant at 5\% level} \\ |t \text{ ratio}| &> 2.576 \Longrightarrow \text{statistically significant at 1\% level} \end{aligned}$$

- $1.96 (\approx 2)$ is a magic number in practice.

Guidelines for Discussing Economic and Statistical Significance

- If a variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its economic or practical significance.
  - "Statistically significance" means $\left|\widehat{\beta}_j / se\left(\widehat{\beta}_j\right)\right|$ is large, but this may be due to a small $se\left(\widehat{\beta}_j\right)$. "Economic significance" means a large $\left|\widehat{\beta}_j\right|$ in the ecomomic sense.
- The fact that a coefficient is statistically significant does not necessarily mean it is economically or practically significant!
  - E.g, $\log(enroll)$ in the above example is statistically significant, but economically insignificant.
- If a variable is statistically and economically important but has the "wrong" sign, the regression model might be misspecified.
- If a variable is statistically insignificant at the usual levels (10%, 5%, 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.
  - $se\left(\widehat{\beta}_j\right) = \sqrt{\frac{\widehat{\sigma}^2}{SST_j\left(1-R_j^2\right)}}$ tends to be large when n is small because SST_j tends to be small and R_j^2 tends to be large. As a result, $\left|t_{\widehat{\beta}_j}\right|$ tends to be small and the conclusion of "statistical insignificance" is not reliable.

c: Testing More General Hypotheses About a Regression Coefficient

- The general null is stated as

$$H_0 : \beta_j = a_j,$$

where a_j is the hypothesized value of the coefficient.

- The t statistic is

$$t = \frac{(\text{estimate} - \text{hypothesized value})}{\text{standard error}} = \frac{\widehat{\beta}_j - a_j}{se\left(\widehat{\beta}_j\right)}.$$

- 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.
- The fitted regression line is

$$\begin{aligned} \log\widehat{(crime)} &= \underset{(1.03)}{-6.63} + \underset{(.11)}{1.27}\log(enroll) \\ n &= 97, R^2 = .585 \end{aligned}$$

  where
  $crime =$ annual number of crimes on college campuses

- Although $\widehat{\beta}_{\log(enroll)}$ is different from one but is this difference statistically significant? We want to test

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

- The t statistic is

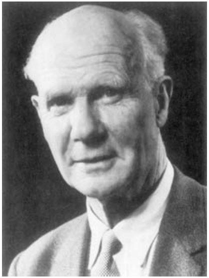
$$t = \frac{1.27 - 1}{.11} \approx 2.45 > 1.96 = c_{0.05},$$

  so the null is rejected at the 5% level.

Different Traditions of Hypothesis Testing

- Rejection/Acceptance Dichotomy:

Jerzy Neyman (1894-1981), Berkeley



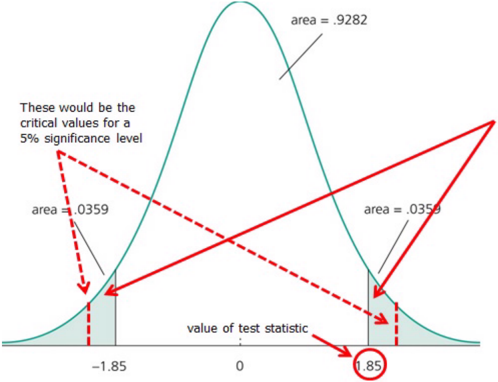
Egon Pearson (1895-1980)[2], UCL

- *p*-Value: R.A. Fisher (we will discuss more about him later).

[2] He is the son of Karl Pearson.

d: Computing p-Values for t Tests

- If the significance level is made smaller and smaller, there will be a point where the null hypothesis cannot be rejected anymore.
- The reason is that, by lowering the significance level, one wants to avoid more and more to make the error of rejecting a correct H_0.
- The smallest significance level at which the null hypothesis is still rejected, is called the p-value of the hypothesis test.
  - The p-value is the significance level at which one is **indifferent** between rejecting and not rejecting the null hypothesis. [figure here]
  - Alternatively, the p-value is the probability of observing a t statistic as extreme as we did if the null is true. [$p = P(|T| > |t|)$]
  - A null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level. [$\alpha = P(|T| > c_\alpha)$, so $|t| > c_\alpha \Longleftrightarrow p < \alpha$]
- A small p-value is evidence against the null hypothesis because one would reject the null hypothesis even at small significance levels.
- A large p-value is evidence in favor of the null hypothesis.



In the two-sided case, the p-value is thus the probability that the t-distributed variable takes on a larger absolute value than the realized value of the test statistic, e.g.:

$$P(|T| > |t|) = P(|T| > 1.85) = 2 \times .0359 = .0718,$$

where T denotes a t distributed random variable with n-k-1 df and t denotes the numerical value of the test statistic.

Figure: Obtaining the P-Value Against a Two-Sided Alternative, When $t = 1.85$ and $df = 40$

- P-values are more informative than tests at fixed significance levels because you can choose your own significance level.

Confidence Intervals

Confidence Intervals

- Simple manipulation of the result in Theorem 4.2 implies that

$$P\left(\underbrace{\widehat{\beta}_j - c_{0.05}\cdot se\left(\widehat{\beta}_j\right)}_{\text{lower bound of the CI}} \le \beta_j \le \underbrace{\widehat{\beta}_j + c_{0.05}\cdot se\left(\widehat{\beta}_j\right)}_{\text{upper bound of the CI}}\right) \quad (*)$$

$$= P\left(\left|\frac{\widehat{\beta}_j-\beta_j}{se\left(\widehat{\beta}_j\right)}\right| \le c_{0.05}\right) = 0.95,$$

  where $c_{0.05}$ is the 5% critical value of two-sided test, and 0.95 is called the confidence level.

- Interpretation of the confidence interval:

  - The bounds of the interval are random. (the length = $2c_{0.05}\cdot se\left(\widehat{\beta}_j\right)$ is also random!)

  - In repeated samples, the interval that is constructed in the above way will cover the population regression coefficient in 95% of the cases.

- Analog: catch a butterfly using a net.

How to Construct CIs? Inverting the Test Statistic

- Collecting all b_j's that are not rejected in the t test of $H_0 : \beta_j = b_j$ vs. $H_1 : \beta_j \neq b_j$. [why? see (*)]
- Relationship between confidence intervals and hypotheses tests: If $b_j \notin$CI, then we will reject $H_0 : \beta_j = b_j$ in favor of $H_1 : \beta_j \neq b_j$ at the level of $($1$-$confidence level$)$.
- Confidence intervals for typical confidence levels:

$$\begin{aligned} P\left(\widehat{\beta}_j - c_{0.01} \cdot se\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.01} \cdot se\left(\widehat{\beta}_j\right)\right) &= 0.99, \\ P\left(\widehat{\beta}_j - c_{0.05} \cdot se\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.05} \cdot se\left(\widehat{\beta}_j\right)\right) &= 0.95, \\ P\left(\widehat{\beta}_j - c_{0.10} \cdot se\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.10} \cdot se\left(\widehat{\beta}_j\right)\right) &= 0.90. \end{aligned}$$

- Use rules of thumb: $c_{0.01} = 2.576, c_{0.05} = 1.96$ and $c_{0.10} = 1.645$.
  - to catch a butterfly with a higher probability, we must use a larger net.

Example: Model of R&D Expenditures

- The fitted regression line is

$$\begin{aligned} \widehat{\log(rd)} &= -4.38 + 1.084 \log(sales) + .0217 profmarg \\ &\quad\;\, (.47) \quad\; (.060) \qquad\qquad\qquad (.0128) \\ n &= 32, R^2 = .918 \text{ (very high)} \end{aligned}$$

  where
  rd = firm's spending on R&D
  $sales$ = annual sales
  $profmarg$ = profits as percentage of sales

- $df = 32 - 2 - 1 = 29$, so $c_{0.05} = 2.045$.
- The 95% CI for $\beta_{\log(sales)}$ is $1.084 \pm 2.045(.060) = (.961, 1.21)$. The effect of $\log(sales)$ on $\log(rd)$ is relatively precisely estimated as the interval is narrow. Moreover, the effect is significantly different from zero because zero is outside the interval.
- The 95% CI for $\beta_{profmarg}$ is $.0217 \pm 2.045(.0128) = (-.0045, .0479)$. The effect of *profmarg* on $\log(rd)$ is imprecisely estimated as the interval is very wide. It is not even statistically significant because zero lies in the interval.

Testing Hypotheses about a Single Linear Combination of the Parameters

Example: Return to Education at 2-Year vs. at 4-Year Colleges

- Suppose the model is

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

where
jc = years of education at 2-year colleges
$univ$ = years of education at 4-year colleges

- Suppose we want to test

$$H_0 : \beta_1 - \beta_2 = 0 \text{ vs. } H_1 : \beta_1 - \beta_2 < 0$$

- A possible test statistic would be

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

- 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.

continue

- It is impossible to compute such a t statistic with standard regression output because

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

  where $\widehat{Cov}\left(\widehat{\beta}_1,\widehat{\beta}_2\right)$ is usually not available in regression output.
- Alternative method: Define $\theta_1=\beta_1-\beta_2$ and test $H_0:\theta_1=0$ vs. $H_1:\theta_1<0$.
- Now, $\beta_1=\theta_1+\beta_2$. Inserting it into the original regression, we have

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

  where $jc+univ$ is a new regressor, representing total years of college.

continue

- The fitted regression line is

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

- Now,

$$t = \frac{-.0102}{.0069} = -1.48 \in (c_{0.05}, c_{0.10}) = (-1.645, -1.282),$$

so the null is rejected at 10% level but not at 5% level.

- Alternatively, the p-value is

$$P(T < -1.48) = .070 \in (.05, .10),$$

and the 95% CI for θ_1 is

$$-.0102 \pm 1.96(.0069) = (-.0237, .0003)$$

covering 0.

- This method works always for **single linear** hypotheses.

Testing Multiple Linear Restrictions: The F Test

a: Testing Exclusion Restrictions

- Consider the following model that explains major league baseball players' salaries:

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

  where
  $salary$ = the 1993 total salary
  $years$ = years in the league
  $gamesyr$ = average games played per year
  $bavg$ = career batting average
  $hrunsyr$ = home runs per year
  $rbisyr$ = runs batted in per year

- The hypotheses are

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

  where H_0 is not true means "at least one of β_3, β_4 and β_5 is not zero".

- This is to test whether performance measures have no effects or can be excluded from regression.

Estimation of the Unrestricted Model

- The estimated unrestricted model is

$$\begin{aligned}
\widehat{\log(salary)} &= \underset{(.29)}{11.19} + \underset{(.0121)}{.0689} years + \underset{(.0026)}{.0126} gamesyr \\
&\quad + \underset{(.00110)}{.00098} bavg + \underset{(.0161)}{.0144} hrunsyr + \underset{(.0072)}{.0108} rbisyr \\
n &= 353, SSR_{ur} = 181.186, R^2 = .6278
\end{aligned}$$

  where the subscript *ur* in *SSR* indicates the SSR in the unrestricted model.
- None of these three variables is statistically significant when tested individually.
- Idea: How would the model fit (measured in *SSR*) be if these variables were dropped from the regression?

Estimation of the Restricted Model

- The estimated restricted model is

$$\begin{aligned}\widehat{\log(salary)} &= 11.22 + .0713 years + .202 gamesyr \\ &\quad\ (.11) \quad (.0125) \qquad\quad (.0013) \\ n &= 353, SSR_r = 198.311, R^2 = .5971\end{aligned}$$

  where the subscript r in SSR indicates the SSR in the restricted model.
- The sum of squared residuals (SSR) necessarily increases in the restricted model [why? recall from Chapter 3 for the case with $H_0 : \beta_{k+1} = 0$ and $q = 1$], but is the increase statistically significant?
- The rigorous test statistic is

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

  where, $q = df_r - df_{ur} = (n-(k-q)-1) - (n-k-1)$ is the number of restrictions, and $n-k-1 = df_{ur}$.
- The relative increase of the sum of squared residuals when going from H_1 to H_0 follows a F-distribution (if the null hypothesis H_0 is correct).

History of the F-Distribution

Ronald A. Fisher (1890-1962), UCL

- Ronald A. Fisher (1890-1962) is one iconic founder of modern statistical theory. The name of F-distribution was coined by G.W. Snedecor, in honor of R.A. Fisher. The p-value is also credited to him.

[Review] F-Distribution

- If X_1 follows a χ^2 distribution with d_1 degrees of freedom,
$$X_1 \sim \chi^2_{d_1},$$
and X_2 follows a χ^2 distribution with d_2 degrees of freedom,
$$X_2 \sim \chi^2_{d_2},$$
independent of X_1, then
$$\frac{X_1/d_1}{X_2/d_2} = \frac{\text{chi-square variable}/df}{\text{independent chi-square variable}/df} \sim F_{d_1,d_2},$$
an F-distribution with degrees of freedom d_1 and d_2.
- As in the t-distribution, $X_2/d_2 \to 1$ as $d_2 \to \infty$. So
$$F_{d_1,d_2} \to \chi^2_{d_1}/d_1$$
as $d_2 \to \infty$.

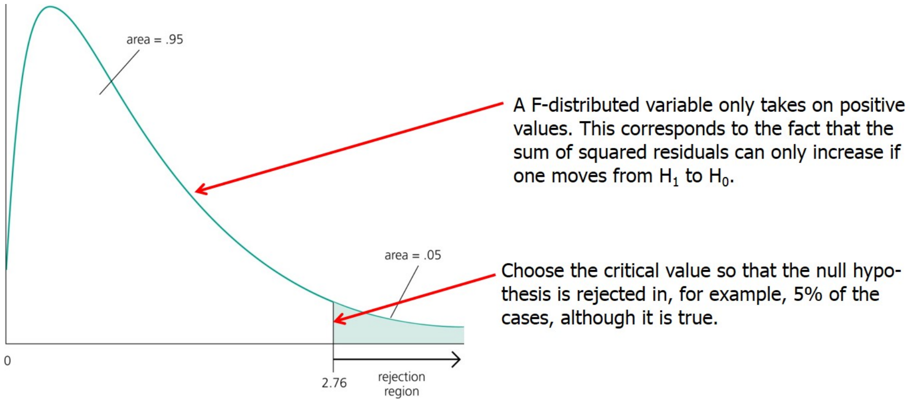


Figure: The 5% Critical Value and Rejection Region in an $F_{3,60}$ Distribution

- We need only to show that $(SSR_r - SSR_{ur})/\sigma^2 \sim \chi^2_q$ and is independent of SSR_{ur} to show $F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$, since from the t statistic, we know that $SSR_{ur}/\sigma^2 \sim \chi^2_{n-k-1}$.

Test Decision in Example

- The F statistic is
$$F = \frac{(198.311 - 181.186)/3}{181.186/(353-5-1)} \approx 9.55,$$
where $q = 3, n = 353$ and $k = 5$.
- Since $F \sim F_{3,347} \approx \chi_3^2/3$, $c_{0.01} = 3.78$, thus the null is rejected.
- Alternatively, p-value$= P(F_{3,347} > 9.55) = 0.0000$, so the null hypothesis is overwhelmingly rejected (even at very small significance levels).
- Discussion:
  - If H_0 is rejected, we say that the three variables are "jointly significant".
  - They were not significant when tested individually.
  - The likely reason is multicollinearity between them.

(*) b: Relationship between F and t Statistics

- When there is only one restriction, we can use both the t test and F test.
- It turns out that they are equivalent (in testing against two-sided alternatives) in the sense that $F = t^2$.
  - Recall that

$$F_{1,n-k-1} = \frac{\chi_1^2}{\chi_{n-k-1}^2 \big/ (n-k-1)} = \left[\frac{N(0,1)}{\sqrt{\chi_{n-k-1}^2 \big/ (n-k-1)}} \right]^2 = t_{n-k-1}^2$$

- The t statistic is more flexible for testing a single hypothesis since it can be used to test against one-sided alternatives.
- Since t statistics are easier to obtain than F statistic, there is no reason to use an F statistic to test hypotheses about a single parameter.
- The F statistic is intended to detect whether a **set** of coefficients is different from zero, but it is never the best test for determining whether a **single** coefficient is different from zero. The t test is best suited for testing a single hypothesis.
  - It is possible that β_1 and/or β_2 is significant based on the t test, but (β_1, β_2) are jointly insignificant based on the F test.

c: The R-Squared Form of the F Statistic

- Recall that

$$R^2 = 1 - \frac{SSR}{SST} \Longrightarrow SSR = SST\left(1 - R^2\right).$$

- As a result,

$$\begin{aligned} F &= \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/(n-k-1)} \\ &= \frac{\left[SST\left(1-R_r^2\right) - SST\left(1-R_{ur}^2\right)\right]/q}{SST\left(1-R_{ur}^2\right)/(n-k-1)} \\ &= \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1-R_{ur}^2\right)/(n-k-1)}. \end{aligned}$$

- In the example,

$$F = \frac{(.6278 - .5971)/3}{(1 - .6278)/347} \approx 9.54,$$

very close to the result based on SSR (difference due to rounding error).

e: The F Statistic for Overall Significance of a Regression

- In the regression
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$
we want to check whether $(x_1, \cdots, x_k)$ do not help to explain y.
- Rigorously,
$$H_0 : \beta_1 = \cdots = \beta_k = 0.$$
- The restricted model is
$$y = \beta_0 + u,$$
which is a regression on constant, and $\widehat{\beta}_0 = \overline{y}$ from Assignment I.
- The F statistic is
$$F = \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/(n-k-1)} = \frac{R_{ur}^2/k}{\left(1 - R_{ur}^2\right)/(n-k-1)} \sim F_{k,n-k-1},$$
where $q = k$ and $R_r^2 = 0$ from Assignment I.
- The test of overall significance is reported in most regression packages; the null hypothesis is usually overwhelmingly rejected.

f: Testing General Linear Restrictions

- Example: Test whether house price assessments are rational, where the model is

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u,$$

where
$price$ = house price
$assess$ = the assessed housing value (before the house was sold)
$lotsize$ = size of the lot, in feet
$sqrft$ = square footage
$bdrms$ = number of bedrooms

- The null is

$$H_0 : \beta_1 = 1, \beta_2 = \beta_3 = \beta_4 = 0,$$

- $\beta_1 = 1$ means that if house price assessments are rational, a 1% change in the assessment should be associated with a 1% change in price.
- $\beta_2 = \beta_3 = \beta_4 = 0$ means that in addition, other known factors should not influence the price once the assessed value has been controlled for.

Example continue

- The unrestricted regression is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u.$$

- The restricted regression is

$$y = \beta_0 + x_1 + u \Longrightarrow y - x_1 = \beta_0 + u$$

  - The restricted model is actually a regression of $y - x_1$ on a constant, and the resulting $\widehat{\beta}_0$ is the sample mean of $y - x_1$.

- The test statistic is

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(1.880 - 1.822)/4}{1.822/(88-4-1)} \approx .661,$$

  where SSR_{ur} is obtained from the next slide.

- $F \sim F_{4,83} \Longrightarrow c_{0.05} = 2.50 \Longrightarrow H_0$ cannot be rejected.

Regression Output for the Unrestricted Regression

- The fitted unrestricted regression line is

$$\begin{aligned}\widehat{\log(price)} &= \underset{(.570)}{.264} + \underset{(.151)}{1.043}\log(assess) + \underset{(.0386)}{.0074}\log(lotsize) \\ &\quad \underset{(.1384)}{-.1032}\log(sqrft) + \underset{(.0221)}{.0338}bdrms \\ n &= 88, SSR = 1.822, R^2 = .773\end{aligned}$$

- The F test works for general multiple **linear** hypotheses.
- For all tests and confidence intervals, validity of assumptions MLR.1-MLR.6 (esp. homoskedasticity) has been assumed. Tests may be invalid otherwise.
- The p-value for the F test is similarly defined as in the t test, see slide 51.
- (*) Like the CI, the confidence region of (β_1, β_2) can be constructed by collecting all values (b_1, b_2) that cannot be rejected in testing $H_0 : \beta_1 = b_1, \beta_2 = b_2$ vs. $H_1 : \beta_1 \neq b_1$ and/or $\beta_2 \neq b_2$.