Lecture 4: Hypothesis Test and Confidence Intervals

Big Data Analytics,Spring 2019

Zhaopeng Qu

Nanjing University

4/20/2019

Simple OLS: Hypothesis Test

Introduction: Class size and Test Score

- Our Regression Result

$$\widehat{TestScore} = 698.9 - 22.8 \times STR,\ R^2 = 0.051, SER = 18.6$$

- How can you be sure about the result?
- Don't Forget. We only get the result from **the sample**.
- Eg.can you reject the claim that cutting the class size will not help boost test scores?

Review: Hypothesis testing:

- Hypothesis testing is one of a fundamental problems in statistics.
- A hypothesis is (usually) an *assertion* about the unknown **population parameters** such as β_1 in a simple OLS

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- such as whether β_1 equals to zero or not

$$\beta_1 = 0$$

- Using the data, we want to determine whether an assertion is true or false.

Review: Testing a hypothesis concerning a population mean

- **the null hypothesis**: $H_0 : E(Y) = \mu_{Y,0}$, **the alternative hypothesis**: $H_1 : E(Y) \neq \mu_{Y,0}$
- Step 1 Compute the *sample mean* $\bar{Y}$
- Step 2 Compute the *standard error* of $\bar{Y}$

$$SE(\bar{Y}) = \frac{s_Y}{\sqrt{n}}$$

- Step 3 Compute the *t-statistic* actually computed

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}$$

- Step 4 See if we can **Reject** the null hypothesis at a certain significance levle α like 5%.

$$|t^{act}| > critical\ value$$

$$p - value < significance\ level$$

Two-Sided Hypotheses Concerning β_1

- **the null hypothesis**: $H_0 : \beta_1 = \beta$ and **the alternative hypothesis**: $H_1 : \beta_1 \neq \beta$
- Step1: Estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ by OLS to obtain $\hat{\beta}_1$
- Step2: Compute the standard error of $\hat{\beta}_1$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)}$$

- Step4: Reject the null hypothesis if $\mid t^{act} \mid > critical\ value$ or if $p - value < significance\ level$

General Form of the t-Statistics

$$t = \frac{estimator - hypothesized\ value}{standard\ error\ of\ the\ estimator}$$

The Standard Error of $\hat{\beta}_1$(1)

- The standard error of $\hat{\beta}_1$ is an **estimator** of the standard deviation of the sampling distribution $\sigma_{\hat{\beta}_1}$, thus

$$SE(\hat{\beta}_1) = \sqrt{\sigma^2_{\hat{\beta}_1}}$$

The Standard Error of $\hat{\beta}_1$

- Recall

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{1}{n}\frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}}$$

- Use the sample variance of $(X_i - \mu_X)u_i$, thus $\frac{1}{n-2}\sum(X_i - \bar{X})^2\hat{u}_i^2$ to estimate population covariance $Var[(X_i - \mu_X)u_i]$
- Use the sample variance of X_i, thus $\frac{1}{n}\sum(X_i - \bar{X})^2$ to replace population covariance of X_i, thus $Var(X_i)$
- Then it can be shown that

$$SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}}$$

Application to Test Score and Class Size

- The regression equation:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + u_i$$

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

OLS regression results

- the OLS regression line

$$\begin{aligned}\widehat{TestScore} =& 698.9 - 22.8 \times STR,\ R^2 = 0.051, SER = 18.6\\ & (10.4)\ \ (0.52)\end{aligned}$$

Testing a two-sided hypothesis concerning β_1

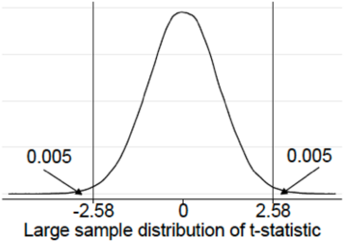
- **the null hypothesis** $H_0 : \beta_1 = 0$, and **the alternative hypothesis** $H_1 : \beta_1 \neq 0$
- Step1: Estimate $\hat{\beta}_1 = -2.28$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.52$
- Step3: Compute the t-statistic

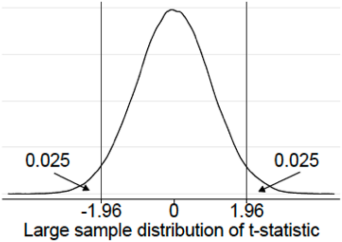
$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - 0}{0.52} = -4.39$$

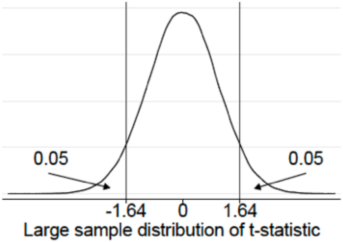
- Step4: Reject the null hypothesis if
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 1.96$
  - $p - value < significance\ level = 0.05$

Critical value of the t-statistic

The critical value of t-statistic depends on significance level α







1% and 10% significant levels

- Step4: Reject the null hypothesis at a **10%** significance level
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 1.64$
  - $p - value = 0.00 < significance\ level = 0.1$
- Step4: Reject the null hypothesis at a **1%** significance level
  - $\mid t^{act} \mid = \mid -4.39 \mid > critical\ value = 2.58$
  - $p - value = 0.00 < significance\ level = 0.01$

Two-Sided Hypotheses Concerning β_1 in a certain value

- Let $\beta_{1,0} = -2$, then Null $H_0 : \beta_1 = -2$, Alternative $H_1 : \beta_1 \neq -2$
- Step1: Estimate $\hat{\beta}_1 = -2.28$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.52$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

- Step4: We can't reject the null hypothesis at 5% significant level because
- $\mid t^{act} \mid = \mid -4.54 \mid > critical\ value = 1.96$
- $p - value < significance\ level = 0.05$

One-sided Hypotheses Concerning β_1

- Let $\beta_{1,0} = -2$, then Null $H_0 : \beta_1 = -2$, Alternative $H_1 : \beta_1 < -2$
- Step1: Estimate $\hat{\beta}_1 = -2.28$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.52$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

- Step4: we can't reject the null hypothesis at 5% significant level because $t^{act} = -0.54 > critical\ value. - 1.96$

Wrap up

Condidence Intervals

Confidence interval for a regression coefficient β_1

- Method for constructing a confidence interval for a population mean can be easily extended to constructing a confidence interval for a regression coefficient.
- Using a two-sided test, a hypothesized value for β_1 will be rejected at 5% significance level if $\mid t^{act} \mid > critical\ value = 1.96$.
- So $\hat{\beta}_1$ will be in the confidence set if $\mid t^{act} \mid \leq critical\ value = 1.96$
- Thus the 95% confidence interval for β_1 are within ± 1.96 standard errors of $\hat{\beta}_1$

$$\hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right)$$

Confidence interval for $\beta_{ClassSize}$

```
. regress test_score class_size

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(1, 418)     =   22.58
       Model |  7794.11004     1  7794.11004           Prob > F      =  0.0000
    Residual |  144315.484   418  345.252353           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0490
       Total |  152109.594   419  363.030056           Root MSE      =  18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

- Thus the 95% confidence interval for β_1 are within ± 1.96 standard errors of $\hat{\beta}_1$

$$\hat{\beta}_1 \pm 1.96 \cdot SE\left(\hat{\beta}_1\right) = -2.28 \pm (1.96 \times 0.48) = [-3.3, -1.34]$$

Confidence interval for predicted effets of changing X

- Consider changing X by a given amount, ΔX. The predicted change in Y associated with this change in X is $\beta_1 \Delta$.
- the 95% confidence interval for $\beta_1 \Delta X$ is

$$\hat{\beta}_1 \Delta X \pm 1.96 \cdot SE\left(\hat{\beta}_1\right) \times \Delta X$$

- eg reducing the student–teacher ratio by 2. then the 95% confidence interval is

$$[-3.3 \times 2, -1.34 \times 2] = [-6.6, -2.68]$$

Regression When X is a Binary Variable

$$
\begin{aligned}
\widehat{TestScore} = & 650 + 7.4 \times D, \ R^2 = 0.037, SER = 18.7 \\
& (1.3) \ \ (1.8)
\end{aligned}
$$

Gauss-Markov theorem and Heteroskedasticity

Introduction

- Recall we discussed the properties of $\bar{Y}$ in Chapter 2.
  - an unbiased estimator of μ_Y
  - a consistent estimator of μ_Y
  - has an approximate normal sampling distribution for large n
  - the **Best Linear Unbiased Estimator(BLUE)**: it is the most efficient estimator of μ_Y among all unbiased estimators.

the fourth OLS assumption

- Three Basic OLS Regression Assumptions
  - Assumption 1
  - Assumption 2
  - Assumption 3
- Assumption 4: The error terms are **homoskedastic**

$$Var(u_i \mid X_i) = \sigma_u^2$$

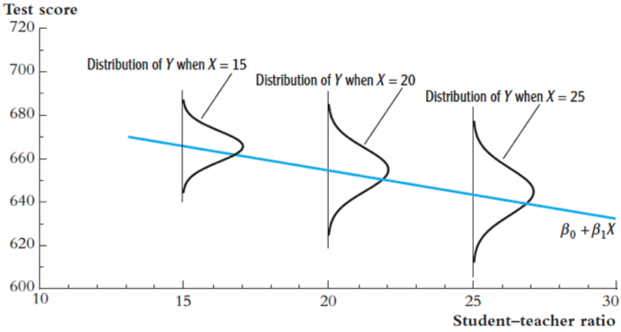
- Then $\hat{\beta}^{OLS}$ is the **Best Linear Unbiased Estimator(BLUE)**: it is the most efficient estimator of β_1 among all conditional unbiased estimators that are a linear function of $Y_1, Y_2, ..., Y_n$.

Heteroskedasticity & homoskedasticity

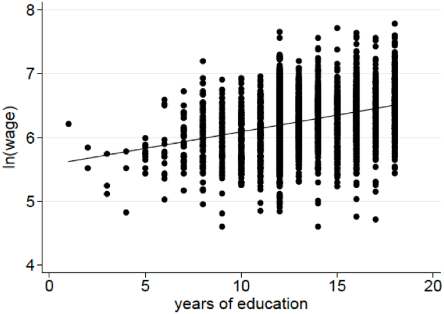
- The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, ...n$, in particular does not depend on X_i. Otherwise, the error term is **heteroskedastic**.

FIGURE 5.2 An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of u given X, $\mathrm{var}(u|X)$, depends on X, u is heteroskedastic.



An Example: the returns to schooling



- The spread of the dots around the line is clearly increasing with years of education X_i.
- Variation in (log) wages is higher at higher levels of education.
- This implies that

$$Var(u_i \mid X_i) \neq \sigma_u^2$$

Heteroskedasticity & homoskedasticity

- If the error terms are heteroskedastic we should use the following heteroskedasticity robust standard errors

$$SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}}$$

- If we assume that the error terms are homoskedastic the standard errors of the OLS estimators simplify to

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{s^2_{\hat{u}}}{\sum(X_i - \bar{X})^2}}$$

- In many applications homoskedasticity is not a plausible assumption.If the error terms are heteroskedastic, then you use the homoskedastic assumption to compute the S.E. of $\hat{\beta}_1$
  - The standard errors are wrong (often too small)
  - The t-statistic does NOT have a $N(0,1)$ distribution (also not in

Heteroskedasticity & homoskedasticity

- Since homoskedasticity is a special case of heteroskedasticity, these heteroskedasticity robust formulas are also valid if the error terms are homoskedastic.
- Hypothesis tests and confidence intervals based on above SE's are valid both in case of homoskedasticity and heteroskedasticity.
- In reality, since in many applications homoskedasticity is not a plausible assumption It is best to use heteroskedasticity robust standard errors. (we lose nothing)
- In **Stata**, the default option of regression is to assume homoskedasticity, to obtain heteroskedasticity robust standard errors use the option "robust":

$$regress\ y\ x\ ,\ robust$$

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------

. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------

. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Heteroskedasticity

- If the error terms are heteroskedastic
  - The fourth OLS assumption is violated
  - The Gauss-Markov conditions do not hold
  - The OLS estimator is not BLUE (not efficient)
- But (given that the other OLS assumptions hold)
  - The OLS estimators are unbiased
  - The OLS estimators are consistent
  - The OLS estimators are normally distributed in large samples

OLS with Multiple Regressors: Hypotheses tests

Assumptions of the Multiple OLS

- Fourth Basic Assumption
  - Assumption 1 : $E[u_i \mid X_{1i}, X_{2i}..., X_{ki}] = 0$
  - Assumption 2 : i.i.d sample
  - Assumption 3 : Large outliers are unlikely.
  - Assumption 4 : No perfect multicollinearity.
- the OLS estimators $\hat{\beta}_j$ for $j = 1, ..., k$ are approximately normally distributed in large samples.
- In addition

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE\left(\hat{\beta}_j\right)} \sim N(0,1)$$

Hypothesis test for single coefficient

- $H_0 : \beta_j = \beta_{j,0}$ $H_1 : \beta_1 \neq \beta_{j,0}$
- Step1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i$ by OLS to obtain $\hat{\beta}_j$
- Step2: Compute the standard error of $\hat{\beta}_j$ (requires matrix algebra)
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_j - \beta_{j,0}}{SE\left(\hat{\beta}_j\right)}$$

- Step4: Reject the null hypothesis if

  ∗ $| t^{act} | > critical\ value$

  ∗ or if $p - value < significance\ level$

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

Case: Class Size and test scores

- Does changing class size, while holding the percentage of English learners constant, have a statistically significant effect on test scores? (using a 5% significance level)
- $H_0 : \beta_{ClassSize} = 0 \; H_1 : \beta_{ClassSize} \neq 0$
- Step1: Estimate $\hat{\beta}_1 = -1.10$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.43$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} = \frac{-1.10 - 0}{0.43} = -2.54$$

- Step4: Reject the null hypothesis if
  - $\mid t^{act} \mid = \mid -2.54 \mid > critical\ value. 1.96$
  - or $p - value = 0.011 < significance\ level = 0.05$

Testing 1 hypothesis on 2 or more coefficients

- Suppose we want to test hypothesis that both the coefficient on % eligible for a free lunch and the coefficient on % eligible for calworks are zero?
- $H_0 : \beta_{meal\,pct} = 0 \ \&\beta_{calw\,pct} = 0$,
  $H_1 : \beta_{meal\,pct} \neq 0 \ and/or \ \beta_{calw\,pct} \neq 0$
- If either $t_{meal\,pct}$ or $t_{calw\,pct}$ exceeds 1.96, we should reject?

Testing 1 hypothesis on 2 or more coefficients

- We assume that $t_{meal\,pct}$ and $t_{calw\,pct}$ are *uncorrelated*:

$$
\begin{aligned}
&Pr(t_{meal\,pct} > 1.96\, and/or\, t_{calw\,pct} > 1.96)\\
&= 1 - Pr(t_{meal\,pct} > 1.96\, and\, t_{calw\,pct} > 1.96)\\
&= 1 - Pr(t_{meal\,pct} > 1.96) * Pr(t_{calw\,pct} > 1.96)
\end{aligned}
$$

$$
\begin{aligned}
=&1 - 0.95 \times 0.95\\
=&0.0975 > 0.05
\end{aligned}
$$

- if $t_{meal\,pct}$ and $t_{calw\,pct}$ are correlated, then *it is more complicated*.

Heteroskedasticity & homoskedasticity

- If we want to test joint hypotheses that involves multiple coefficients we need to use an **F-test** based on the **F-statistic**
- F-Statistic with $q = 2:$ when testing the following hypothesis

$$H_0 : \beta_1 = 0 \;\&\; \beta_2 = 0 \quad H_1 : \beta_1 \neq 0 \, and/or \, \beta_2 \neq 0$$

- the F-statistic combines the two t-statistics as follows

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1 t_2} t_1 t_2}{1 - \hat{\rho}_{t_1 t_2}^2}\right)$$

where $\hat{\rho}_{t_1 t_2}$ is an estimator of the correlation between the two t-statistics.

Hypothesis test for single coefficient

- $H_0 : \beta_j = \beta_{j,0}$ $H_1 : \beta_1 \neq \beta_{j,0}$
- Step1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i$ by OLS to obtain $\hat{\beta}_j$

– Step2: Compute the standard error of $\hat{\beta}_j$ (requires matrix algebra)

– Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_j - \beta_{j,0}}{SE\left(\hat{\beta}_j\right)}$$

- Step4: Reject the null hypothesis if

  $*$ $\mid t^{act} \mid > critical\ value$ $*$ or if $p - value < significance\ level$

Test Scores and Class Size

```
. regress test_score class_size

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
  test_score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  class_size |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

Case: Class Size and test scores

- Does changing class size, while holding the percentage of English learners constant, have a statistically significant effect on test scores? (using a 5% significance level)
- $H_0 : \beta_{ClassSize} = 0 \; H_1 : \beta_{ClassSize} \neq 0$
- Step1: Estimate $\hat{\beta}_1 = -1.10$
- Step2: Compute the standard error: $SE(\hat{\beta}_1) = 0.43$
- Step3: Compute the t-statistic

$$t^{act} = \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} = \frac{-1.10 - 0}{0.43} = -2.54$$

- Step4: Reject the null hypothesis if
  - $\mid t^{act} \mid = \mid -2.54 \mid > critical\, value.1.96$
  - $p - value = 0.011 < significance\; level = 0.05$

Testing 1 hypothesis on 2 or more coefficients

- Suppose we want to test hypothesis that both the coefficient on % eligible for a free lunch and the coefficient on % eligible for calworks are zero?
- $H_0 : \beta_{meal\,pct} = 0\ \&\beta_{calw\,pct} = 0$,
  $H_1 : \beta_{meal\,pct} \neq 0\ and/or\ \beta_{calw\,pct} \neq 0$
- If either $t_{meal\,pct}$ or $t_{calw\,pct}$ exceeds 1.96, we should reject?

Testing 1 hypothesis on 2 or more coefficients

- We assume that $t_{meal\,pct}$ and $t_{calw\,pct}$ are *uncorrelated*:

$$
\begin{aligned}
&Pr(t_{meal\,pct} > 1.96\,and/or\,t_{calw\,pct} > 1.96) \\
&= 1 - Pr(t_{meal\,pct} > 1.96\,and\,t_{calw\,pct} > 1.96) \\
&= 1 - Pr(t_{meal\,pct} > 1.96) * Pr(t_{calw\,pct} > 1.96) \\
=&1 - 0.95 \times 0.95 \\
=&0.0975 > 0.05
\end{aligned}
$$

- if $t_{meal\,pct}$ and $t_{calw\,pct}$ are correlated, then *it is more complicated*.

Heteroskedasticity & homoskedasticity

- If we want to test joint hypotheses that involves multiple coefficients we need to use an **F-test** based on the **F-statistic**
- F-Statistic with $q = 2$: when testing the following hypothesis

$$H_0 : \beta_1 = 0 \: \& \: \beta_2 = 0 \quad H_1 : \beta_1 \neq 0 \: and/or \: \beta_2 \neq 0$$

- the F-statistic combines the two t-statistics as follows

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1 t_2} t_1 t_2}{1 - \hat{\rho}_{t_1 t_2}^2}\right)$$

where $\hat{\rho}_{t_1 t_2}$ is an estimator of the correlation between the two t-statistics.

Case: Class Size and test scores

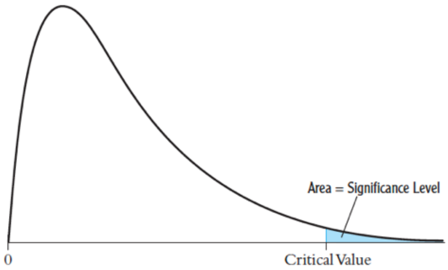
- We want to test hypothesis that both the coefficient on % eligible for a free lunch and the coefficient on % eligible for calworks are zero?
  - $H_0 : \beta_{meal\, pct} = 0 \,\&\beta_{calw\, pct} = 0$
  - $H_1 : \beta_{meal\, pct} \neq 0 \; and/or \; \beta_{calw\, pct} \neq 0$
- The null hypothesis consists of two restrictions $q = 2$
- It can be shown that the F-statistic with two restrictions has an approximate $F_{2,\infty}$ distribution in large samples

$$F = 290.27$$

- Table 4 (S&W page 795) shows that the critical value at a 5% significance level equals 3.
- This implies that we reject H_0 at a 5% significance level because $290.27 > 3$

F-Distribution

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



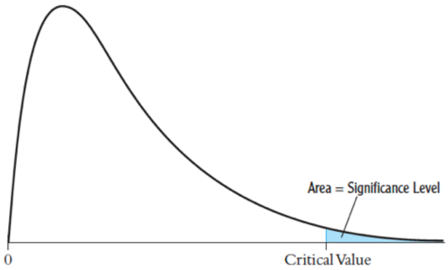
	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78

General procedure for testing joint hypothesis with q restrictions

- $H_0 : \beta_j = \beta_{j,0}, ..., \beta_m = \beta_{m,0}$ for a total of q restrictions.
- H_1:at least one of q restrictions under H_0 does not hold.
- Step1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i$ by OLS
- Step2: Compute the **F-statistic**
- Step3 : Reject the null hypothesis if $F - Statistic > F^{act}_{q,\infty}$ or $p - value = Pr[F_{q,\infty} > F^{act}]$

Case: Class Size and test scores: q=3 restrictions

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78

Case: Class Size and test scores: q=3 restrictions

- $H_0 : \beta_{el\,pct} = \beta_{meal\,pct} = \beta_{calw\,pct} = 0$
- H_1: at least one of q restrictions under H_0 does not hold

1. Step1: Estimate by Multiple OLS
2. Step2: $F - Statistic = 481.06$
3. Step3: We reject the null hypothesis at a 5% significance level because
$$F - Statistic > F_{3,\infty} = 2.6$$

The "overall" regression F-statistic

- The "overall" F-statistic test the joint hypothesis that all the k slope coefficients are zero

– $H_0 : \beta_j = \beta_{j,0}, ..., \beta_m = \beta_{m,0}$ for a total of $q = k$ restrictions.

– H_1: at least one of $q = k$ restrictions under H_0 does not hold.

The "overall" regression F-statistic

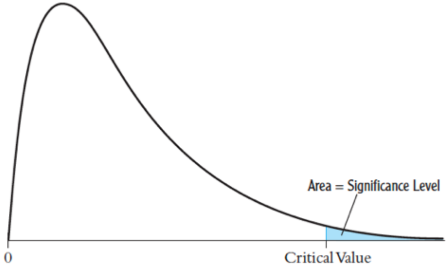
```
. regress test_score class_size el_pct meal_pct calw_pct, robust

Linear regression                               Number of obs     =        420
                                                  F(4, 415)         =     361.68
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.7749
                                                  Root MSE          =     9.0843
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.014353	.2688613	-3.77	0.000	-1.542853	-.4858534
el_pct	-.1298219	.0362579	-3.58	0.000	-.201094	-.0585498
meal_pct	-.5286191	.0381167	-13.87	0.000	-.6035449	-.4536932
calw_pct	-.0478537	.0586541	-0.82	0.415	-.1631498	.0674424
_cons	700.3918	5.537418	126.48	0.000	689.507	711.2767

- The overall $F-Statistics = 361.68$

The "overall" regression F-statistic

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



Degrees of Freedom	Significance Level 10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78

The "Star War" and Regression Table

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31* (0.34)	−1.01* (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547* (0.024)		−0.529* (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix (4.1). Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.