ECON4150 - Introductory Econometrics

Lecture 5: OLS with One Regressor: Hypothesis Tests

Monique de Haan
(moniqued@econ.uio.no)

Stock and Watson Chapter 5

Lecture outline

- Testing Hypotheses about one of the regression coefficients
  - Repetition: Testing a hypothesis concerning a population mean
  - Testing a 2-sided hypothesis concerning β_1
  - Testing a 1-sided hypothesis concerning β_1
- Confidence interval for a regression coefficient
- Efficiency of the OLS estimator
  - Best Linear Unbiased Estimator (BLUE)
  - Gauss-Markov Theorem
  - Heteroskedasticity & homoskedasticity
- Regression when X_i is a binary variable
  - Interpretation of β_0 and β_1
  - Hypothesis tests concerning β_1

Repetition: Testing a hypothesis concerning a population mean

$$H_0 : E(Y) = \mu_{Y,0} \qquad H_1 : E(Y) \neq \mu_{Y,0}$$

Step 1: Compute the sample average $\overline{Y}$

Step 2: Compute the standard error of $\overline{Y}$

$$SE\left(\overline{Y}\right) = \frac{s_Y}{\sqrt{n}}$$

Step 3: Compute the t-statistic

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

Step 4: Reject the null hypothesis at a 5% significance level if

- $|t^{act}| > 1.96$
- or if $p - value < 0.05$

Repetition: Testing a hypothesis concerning a population mean

Example: California test score data; mean test scores

Suppose we would like to test

$$H_0 : E\left(TestScore\right) = 650 \qquad H_1 : E\left(TestScore\right) \neq 650$$

using the sample of 420 California districts

Step 1: $\overline{TestScore} = 654.16$

Step 2: $SE\left(\overline{TestScore}\right) = 0.93$

Step 3: $t^{act} = \frac{\overline{TestScore} - 650}{SE\left(\overline{TestScore}\right)} = \frac{654.16 - 650}{0.93} = 4.47$

Step 4: If we use a 5% significance level, we reject H_0 because $|t^{act}| = 4.47 > 1.96$

Repetition: Testing a hypothesis concerning a population mean

Example: California test score data; mean test scores

```
. ttest test_score=650

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
test_s~e |     420    654.1565    .9297082    19.05335    652.3291     655.984
------------------------------------------------------------------------------
    mean = mean( test_score)                                      t =   4.4708
Ho: mean = 650                                   degrees of freedom =      419

   Ha: mean < 650               Ha: mean != 650               Ha: mean > 650
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Testing a 2-sided hypothesis concerning β_1

- Testing procedure for the population mean is justified by the Central Limit theorem.
- Central Limit theorem states that the t-statistic (standardized sample average) has an approximate $N(0,1)$ distribution in large samples
- Central Limit Theorem also states that
  - $\widehat{\beta}_0$ & $\widehat{\beta}_1$ have an approximate normal distribution in large samples
  - and the standardized regression coefficients have approximate $N(0,1)$ distribution in large samples
- We can therefore use same general approach to test hypotheses about β_0 and β_1.
- We assume that the Least Squares assumptions hold!

Testing a 2-sided hypothesis concerning β_1

$$H_0 : \beta_1 = \beta_{1,0} \qquad H_1 : \beta_1 \neq \beta_{1,0}$$

Step 1: Estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ by OLS to obtain $\widehat{\beta}_1$

Step 2: Compute the standard error of $\widehat{\beta}_1$

Step 3: Compute the t-statistic

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

Step 4: Reject the null hypothesis if

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

Testing a 2-sided hypothesis concerning β_1

The standard error of $\widehat{\beta}_1$

The standard error of $\widehat{\beta}_1$ is an estimate of the standard deviation of the sampling distribution $\sigma_{\widehat{\beta}_1}$

Recall from previous lecture:

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

It can be shown that

$$SE\left(\widehat{\beta}_1\right) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \widehat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^2}}$$

Testing a 2-sided hypothesis concerning β_1

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

```
. 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
------------------------------------------------------------------------------
```

Suppose we would like to test the hypothesis that class size does not affect test scores ($\beta_1 = 0$)

Testing a 2-sided hypothesis concerning β_1
5% significance level

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Step 1: $\widehat{\beta}_1 = -2.28$

Step 2: $SE(\widehat{\beta}_1) = 0.52$

Step 3: Compute the t-statistic

$$t^{act} = \frac{-2.28 - 0}{0.52} = -4.39$$

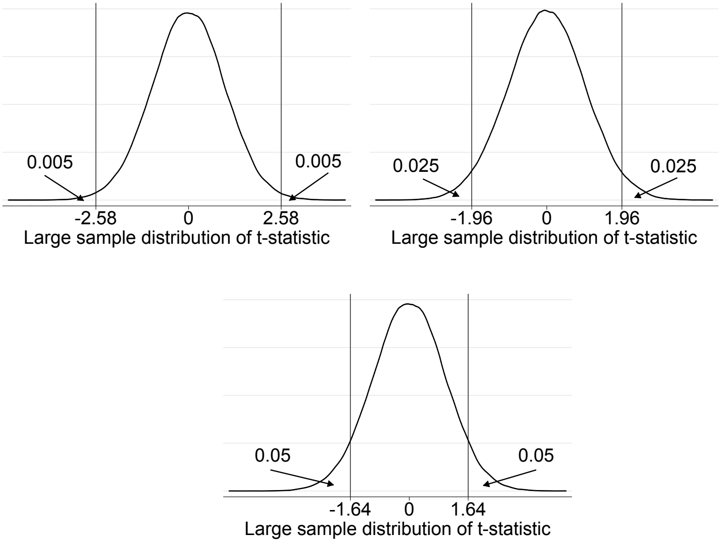
Step 4: We reject the null hypothesis at a 5% significance level because

- $|-4.39| > 1.96$
- $p - value = 0.000 < 0.05$

Testing a 2-sided hypothesis concerning β_1

Critical value of the *t*-statistic

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



Testing a 2-sided hypothesis concerning β_1

1% and 10% significance levels

Step 1: $\widehat{\beta}_1 = -2.28$

Step 2: $SE(\widehat{\beta}_1) = 0.52$

Step 3: Compute the t-statistic

$$t^{act} = \frac{-2.28 - 0}{0.52} = -4.39$$

Step 4: We reject the null hypothesis at a 10% significance level because

- $|-4.39| > 1.64$
- $p - value = 0.000 < 0.1$

Step 4: We reject the null hypothesis at a 1% significance level because

- $|-4.39| > 2.58$
- $p - value = 0.000 < 0.01$

Testing a 2-sided hypothesis concerning β_1
5% significance level

$$H_0 : \beta_1 = -2 \qquad H_1 : \beta_1 \neq -2$$

Step 1: $\widehat{\beta}_1 = -2.28$

Step 2: $SE(\widehat{\beta}_1) = 0.52$

Step 3: Compute the t-statistic

$$t^{act} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

Step 4: We don't reject the null hypothesis at a 5% significance level because

- $|-0.54| < 1.96$

Testing a 2-sided hypothesis concerning β_1
5% significance level

```
. 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
```

test_score	Coef.	Robust 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

$$H_0 : \beta_1 = -2 \quad \longrightarrow \quad H_0 : \beta_1 - (-2) = 0$$

```
. lincom class_size-(-2)

 ( 1)   class_size = -2
```

test_score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-.2798083	.5194892	-0.54	0.590	-1.300945	.7413286

Testing a 1-sided hypothesis concerning β_1
5% significance level

$$H_0 : \beta_1 = -2 \qquad H_1 : \beta_1 < -2$$

Step 1: $\widehat{\beta}_1 = -2.28$

Step 2: $SE(\widehat{\beta}_1) = 0.52$

Step 3: Compute the t-statistic

$$t^{act} = \frac{-2.28 - (-2)}{0.52} = -0.54$$

Step 4: We don't reject the null hypothesis at a 5% significance level because

- $-0.54 > -1.64$

Confidence interval for a regression coefficient

- 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 $|t| > 1.96$
- and will be in the confidence set if $|t| \leq 1.96$
- Thus the 95% confidence interval for β_1 are the values of $\beta_{1,0}$ within ± 1.96 standard errors of $\widehat{\beta}_1$

95% confidence interval for β_1

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

Confidence interval for $\beta_{ClassSize}$

```
. 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
```

test_score	Coef.	Robust 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

- 95% confidence interval for β_1 (shown in output)

$$(-3.30 \quad , \quad -1.26)$$

- 90% confidence interval for β_1 (not shown in output)

$$\widehat{\beta}_1 \pm 1.64 \cdot SE\left(\widehat{\beta}_1\right)$$

$$-2.27 \pm 1.64 \cdot 0.52$$

$$(-3.12 \quad , \quad -1.42)$$

Properties of the OLS estimator of β_1

Recall the 3 least squares assumptions:

Assumption 1: $E\left(u_i|X_i\right) = 0$

Assumption 2: $\left(Y_i, X_i\right)$ for $i = 1, ..., n$ are $i.i.d$

Assumption 3: Large outliers are unlikely

If the 3 least squares assumptions hold the OLS estimator $\widehat{\beta}_1$

- Is an unbiased estimator of β_1
- Is a consistent estimator β_1
- Has an approximate normal sampling distribution for large n

Properties of $\overline{Y}$ as estimator of μ_Y

In lecture 2 we discussed that:

- $\overline{Y}$ is an unbiased estimator of μ_Y
- $\overline{Y}$ a consistent estimator of μ_Y
- $\overline{Y}$ has an approximate normal sampling distribution for large n

AND

$\overline{Y}$ is the Best Linear Unbiased Estimator (BLUE): it is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, ..., Y_n$

Let $\hat{\mu}_Y$ be an unbiased estimator of μ_Y

$$\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i \qquad \textit{with } a_1, ...a_n \textit{ nonrandom constants}$$

then $\overline{Y}$ is more efficient than $\hat{\mu}_Y$, that is

$$Var\left(\overline{Y}\right) < Var\left(\hat{\mu}_Y\right)$$

Best Linear Unbiased Estimator (BLUE)

If we add a fourth OLS assumption:

Assumption 4: The error terms are homoskedastic

$$Var\left(u_i|X_i\right) = \sigma_u^2$$

$\widehat{\beta}_1^{OLS}$ 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 $Y_1, ..., Y_n$

Let $\widetilde{\beta}_1$ be an unbiased estimator of β_1

$$\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$$

where $a_1, ..., a_n$ can depend on $X_1, ..., X_n$ (but not on $Y_1, ..., Y_n$)

then $\widehat{\beta}_1^{OLS}$ is more efficient than $\widetilde{\beta}_1$, that is

$$Var\left(\widehat{\beta}_1^{OLS}|X_1, ..., X_n\right) < Var\left(\widetilde{\beta}_1|X_1, ..., X_n\right)$$

Gauss-Markov theorem for $\widehat{\beta}_1$

The Gauss-Markov theorem states that if the following 3 Gauss-Markov conditions hold

1. $E\left(u_i|X_1,...,X_n\right) = 0$
2. $Var\left(u_i|X_1,...,X_n\right) = \sigma_u^2\,, \quad 0 < \sigma_u^2 < \infty$
3. $E\left(u_i u_j|X_1,...,X_n\right) = 0\,, \quad i \neq j$

The OLS estimator of β_1 is BLUE

It is shown in S&W appendix 5.2 that the following 4 Least Squares assumptions imply the Gauss-Markov conditions

Assumption 1: $E\left(u_i|X_i\right) = 0$

Assumption 2: (Y_i, X_i) for $i = 1, ..., n$ are *i.i.d*

Assumption 3: Large outliers are unlikely

Assumption 4: The error terms are homoskedastic: $Var\left(u_i|X_i\right) = \sigma_u^2$

Heteroskedasticity & homoskedasticity

The fourth least Squares assumption

$$Var\left(u_i|X_i\right) = \sigma_u^2$$

states that the conditional variance of the error term does not depend on the regressor X

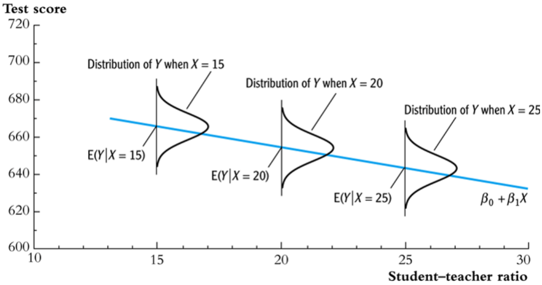
Under this assumption the variance of the OLS estimators simplify to

$$\sigma^2_{\widehat{\beta}_0} = \frac{E\left(X_i^2\right)\sigma_u^2}{n\sigma_X^2}$$

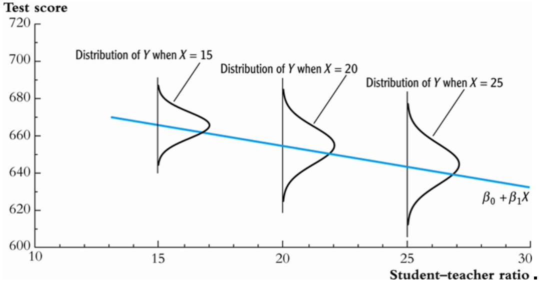
$$\sigma^2_{\widehat{\beta}_1} = \frac{\sigma_u^2}{n\sigma_X^2}$$

Is homoskedasticity a plausible assumption?

Example of **homoskedasticity** $Var(u_i|X_i) = \sigma_u^2$:

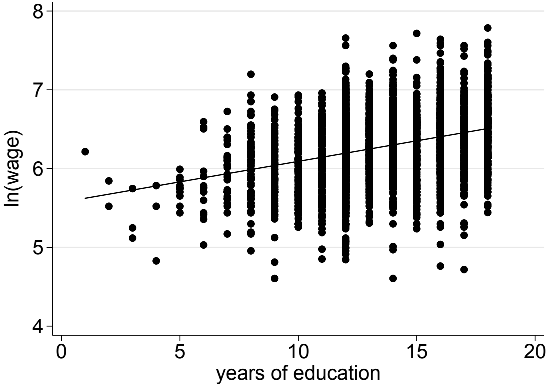


Example of **heteroskedasticity** $Var(u_i|X_i) \neq \sigma_u^2$



Heteroskedasticity & homoskedasticity

Example: The returns to education



- 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 $Var\left(u_i|X_i\right) \neq \sigma_u^2$.

Heteroskedasticity & homoskedasticity

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

$$SE\left(\widehat{\beta}_1\right) = \frac{s_{\widehat{u}}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$SE\left(\widehat{\beta}_0\right) = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)s_{\widehat{u}}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

- In many applications homoskedasticity is not a plausible assumption
- If the error terms are heteroskedastic, that is $Var\left(u_i|X_i\right) \neq \sigma_u^2$ and the above formulas are used to compute the standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$
  - The standard errors are wrong (often too small)
  - The t-statistic does not have a $N\left(0,1\right)$ distribution (also not in large samples)
  - The probability that a 95% confidence interval contains true value is not 95% (also not in large samples)

Heteroskedasticity & homoskedasticity

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

$$SE\left(\widehat{\beta}_1\right) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \widehat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}}$$

$$SE\left(\widehat{\beta}_0\right) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{H}_i^2 \widehat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{H}_i^2\right]^2}}$$

$$\textit{with} \qquad \widehat{H}_i = 1 - \left(\overline{X} / \frac{1}{n}\sum_{i=1}^{n} X_i^2\right) X_i$$

- 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.

Heteroskedasticity & homoskedasticity

- In Stata the default option is to assume homoskedasticity
- Since in many applications homoskedasticity is not a plausible assumption
- It is best to use heteroskedasticity robust standard errors
- To obtain heteroskedasticity robust standard errors use the option “robust”:

```
Regress y x , robust
```

Heteroskedasticity & homoskedasticity

```
. regress test_score class_size
```

Source	SS	df	MS
Model	**7794.11004**	**1**	**7794.11004**
Residual	**144315.484**	**418**	**345.252353**
Total	**152109.594**	**419**	**363.030056**

Number of obs	=	**420**
F(1, 418)	=	**22.58**
Prob > F	=	**0.0000**
R-squared	=	**0.0512**
Adj R-squared	=	**0.0490**
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**

test_score	Coef.	Robust 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 & homoskedasticity

If the error terms are heteroskedastic

- The fourth OLS assumption: $Var(u_i|X_i) = \sigma_u^2$ 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

Regression when X_i is a binary variable

Sometimes a regressor is binary:

- $X = 1$ if small class size, $= 0$ if not
- $X = 1$ if female, $= 0$ if male
- $X = 1$if treated (experimental drug), $= 0$ if not

Binary regressors are sometimes called “dummy” variables.

So far, β_1 has been called a “slope,” but that doesn’t make sense if X is binary.

How do we interpret regression with a binary regressor?

Regression when X_i is a binary variable

Interpreting regressions with a binary regressor

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

- When $X_i = 0$,

$$\begin{aligned} E\left(Y_i|X_i=0\right) &= E\left(\beta_0 + \beta_1 \cdot 0 + u_i|X_i=0\right) \\ &= \beta_0 + E\left(u_i|X_i=0\right) \\ &= \beta_0 \end{aligned}$$

- When $X_i = 1$,

$$\begin{aligned} E\left(Y_i|X_i=1\right) &= E\left(\beta_0 + \beta_1 \cdot 1 + u_i|X_i=1\right) \\ &= \beta_0 + \beta_1 + E\left(u_i|X_i=0\right) \\ &= \beta_0 + \beta_1 \end{aligned}$$

- This implies that $\beta_1 = E(Y_i|X_i = 1) – E(Y_i|X_i = 0)$ is the population difference in group means

Regression when X_i is a binary variable

Example: The effect of being in a small class on test scores

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

Let $SmallClass_i$ be a binary variable:

$$SmallClass_i \begin{cases} = 1 \text{ if } Class\ size < 20 \\ \\ = 0 \text{ if } Class\ size \geq 20 \end{cases}$$

Interpretation of β_0: population mean test scores in districts where class size is large (not small)

$$\beta_0 = E\left(TestScore_i | SmallClass_i = 0\right)$$

Interpretation of β_1: the difference in population mean test scores between districts with small and districts with larger classes (not small).

$$\beta_1 = E\left(TestScore_i | SmallClass_i = 1\right) - E\left(TestScore_i | SmallClass_i = 0\right)$$

Regression when X_i is a binary variable

Example: The effect of being in a small class on test scores

```
. tab small_class
```

small_class	Freq.	Percent	Cum.
0	**182**	**43.33**	**43.33**
1	**238**	**56.67**	**100.00**
Total	**420**	**100.00**	

```
. bys small_class: sum class_size
```

-> small_class = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
class_size	**182**	**21.28359**	**1.155685**	**20**	**25.8**

-> small_class = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
class_size	**238**	**18.38389**	**1.283886**	**14**	**19.96154**

Regression when X_i is a binary variable

Example: The effect of being in a small class on test scores

```
. regress test_score small_class, robust

Linear regression                               Number of obs     =        420
                                                   F(1, 418)         =      16.34
                                                   Prob > F          =     0.0001
                                                   R-squared         =     0.0369
                                                   Root MSE          =     18.721
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
small_class	7.37241	1.823578	4.04	0.000	3.787884	10.95694
_cons	649.9788	1.322892	491.33	0.000	647.3785	652.5792

- $\widehat{\beta}_0 = 649.98$ is the sample average of test scores in districts with an average class size ≥ 20.
- $\widehat{\beta}_1 = 7.37$ is the difference in the sample average of test scores in districts with class size < 20 and districts with average class size≥ 20

Regression when X_i is a binary variable

Example: The effect of being in a small class on test scores

```
. ttest test_score, by(small_class) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	182	649.9788	1.323379	17.85336	647.3676	652.5901
1	238	657.3513	1.254794	19.35801	654.8793	659.8232
combined	420	654.1565	.9297082	19.05335	652.3291	655.984
diff		-7.37241	1.823689		-10.95752	-3.787296

```
    diff = mean( 0) - mean( 1)                                   t =   -4.0426
Ho: diff = 0                    Satterthwaite's degrees of freedom =   403.607

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) =  0.0000         Pr(|T| > |t|) =  0.0001          Pr(T > t) =  1.0000
```

Regression when X_i is a binary variable

Testing a 2-sided hypothesis concerning β_1, 1% significance level

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Step 1: $\widehat{\beta}_1 = 7.37$

Step 2: $SE(\widehat{\beta}_1) = 1.82$

Step 3: Compute the t-statistic

$$t^{act} = \frac{7.37 - 0}{1.82} = 4.04$$

Step 4: We reject the null hypothesis at a 1% significance level because

- $|4.04| > 2.58$
- $p - value = 0.000 < 0.01$

Regression when X_i is a binary variable

Example: The effect of high per student expenditure on test scores

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

Let $HighExpenditure_i$ be a binary variable:

$$HighExpenditure_i \begin{cases} = 1 \textit{ if per student expenditure} > \$6000 \\ = 0 \textit{ if per student expenditure} \leq \$6000 \end{cases}$$

Interpretation of β_0: population mean test scores in districts with low per student expenditure

$$\beta_0 = E\left(TestScore_i | HighExpenditure_i = 0\right)$$

Interpretation of β_1: the difference in population mean test scores between districts with high and districts with low per student expenditures.

$$\beta_1 = E\left(TestScore_i | HighExpenditure_i = 1\right) - E\left(TestScore_i | HighExpenditure_i = 0\right)$$

Regression when X_i is a binary variable

Example: The effect of high per student expenditure on test scores

```
. regress test_score high_expenditure, robust

Linear regression                               Number of obs     =        420
                                                   F(1, 418)         =       8.02
                                                   Prob > F          =     0.0048
                                                   R-squared         =     0.0295
                                                   Root MSE          =     18.792
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
high_expenditure	10.01216	3.535408	2.83	0.005	3.062764	16.96155
_cons	652.9408	.9311991	701.18	0.000	651.1104	654.7712

- $\widehat{\beta}_0 = 652.94$ is the sample average of test scores in districts with low per student expenditures.
- $\widehat{\beta}_1 = 10.01$ is the difference in the sample average of test scores in districts with high and districts with low per student expenditures.

Regression when X_i is a binary variable

Testing a 2-sided hypothesis concerning β_1, 10% significance level

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Step 1: $\widehat{\beta}_1 = 10.01$

Step 2: $SE(\widehat{\beta}_1) = 3.54$

Step 3: Compute the t-statistic

$$t^{act} = \frac{10.01 - 0}{3.54} = 2.83$$

Step 4: We reject the null hypothesis at a 10% significance level because

- $|2.83| > 1.64$
- $p - value = 0.005 < 0.10$