ECON4150 - Introductory Econometrics

Lecture 7: OLS with Multiple Regressors –Hypotheses tests–

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

Stock and Watson Chapter 7

Lecture outline

- Hypothesis test for single coefficient in multiple regression analysis
- Confidence interval for single coefficient in multiple regression
- Testing hypotheses on 2 or more coefficients
  - The F-statistic
  - The overall regression F-statistic
- Testing single restrictions involving multiple coefficients
- Measures of fit in multiple regression model
  - *SER*, R^2 and $\overline{R}^2$
  - Relation between (homoskedasticity-only) F-statistic and the R^2
  - Interpreting measures of fit
- Interpreting “stars” in a table with regression output

Hypothesis test for single coefficient in multiple regression analysis

```
. regress test_score class_size el_pct, robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     223.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4264
                                                Root MSE          =     14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

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).

Hypothesis test for single coefficient in multiple regression analysis

Under the 4 Least Squares assumptions of the multiple regression model:

Assumption 1: $E(u_i|X_{1i}, ..., X_{ki}) = 0$

Assumption 2: $(Y_i, X_{1i}, ..., X_{ki})$ for $i = 1, ..., n$ are $(i.i.d)$

Assumption 3: Large outliers are unlikely

Assumption 4: No perfect multicollinearity

The OLS estimators $\widehat{\beta}_j$ for $j = 1, .., k$ are approximately normally distributed in large samples

In addition

$$t = \frac{\widehat{\beta}_j - \beta_{j0}}{SE\left(\widehat{\beta}_j\right)} \sim N(0, 1)$$

We can thus perform, hypothesis tests in same way as in regression model with 1 regressor.

Hypothesis test for single coefficient in multiple regression analysis

$$H_0 : \beta_j = \beta_{j,0} \qquad H_1 : \beta_j \neq \beta_{j,0}$$

Step 1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_j X_{ji} + ... + \beta_1 X_{ki} + u_i$ by OLS to obtain $\widehat{\beta}_j$

Step 2: Compute the standard error of $\widehat{\beta}_j$ (requires matrix algebra)

Step 3: Compute the t-statistic

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

Step 4: Reject the null hypothesis if

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

Hypothesis test for single coefficient in multiple regression analysis

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)

$$H_0 : \beta_{ClassSize} = 0 \qquad H_1 : \beta_{ClassSize} \neq 0$$

Step 1: $\widehat{\beta}_{ClassSize} = -1.10$

Step 2: $SE\left(\widehat{\beta}_{ClassSize}\right) = 0.43$

Step 3: Compute the t-statistic

$$t^{act} = \frac{-1.10 - 0}{0.43} = -2.54$$

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

- $|-2.54| > 1.96$ and $p - value = 0.011 < 0.05$

Do we reject H_0 at a 1% significance level?

Confidence interval for single coefficient in multiple regression

```
. regress test_score class_size el_pct, robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     223.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4264
                                                Root MSE          =     14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

- 95% confidence interval for $\beta_{ClassSize}$ given in regression output
- 99% confidence interval for $\beta_{ClassSize}$:

$$\widehat{\beta}_{ClassSize} \pm 2.58 \cdot SE\left(\widehat{\beta}_{ClassSize}\right)$$

$$-1.10 \pm 2.58 \times 0.43$$

$$(-2.21 \quad ; \quad 0.01)$$

Hypothesis tests on 2 or more coefficients

We add two more variables, both measuring what % of students come from families with low income

- *meal pct*$_i$ measures % of students in district eligible for a free lunch
- *calw pct*$_i$ measures % of students in district eligible for social assistance under a program called “CALWORKS”

```
. 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

Testing 2 hypotheses on 2 or more coefficients

$$H_0 : \beta_{meal\ pct} = 0 \qquad H_1 : \beta_{meal\ pct} \neq 0$$

Step 1: $\widehat{\beta}_{meal\ pct} = -0.529$

Step 2: $SE\left(\widehat{\beta}_{meal\ pct}\right) = 0.038$

Step 3: $t_{meal\ pct} = \frac{-0.529-0}{0.038} = -13.87$

Step 4: Reject the null hypothesis at 5% significance level
$(|-13.87| > 1.96)$

$$H_0 : \beta_{calw\ pct} = 0 \qquad H_1 : \beta_{calw\ pct} \neq 0$$

Step 1: $\widehat{\beta}_{calw\ pct} = -0.048$

Step 2: $SE\left(\widehat{\beta}_{calw\ pct}\right) = 0.059$

Step 3: $t_{calw\ pct} = \frac{-0.048-0}{0.059} = -0.82$

Step 4: Don't reject the null hypothesis at 5% significance level
$(|-0.82| < 1.96)$

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?

$$H_0 : \beta_{meal\ pct} = 0 \ \& \ \beta_{calw\ pct} = 0 \qquad H_1 : \beta_{meal\ pct} \neq 0 \ and/or \ \beta_{calw\ pct} \neq 0$$

What if we reject H_0 if either $t_{meal\ pct}$ or $t_{calw\ pct}$ exceeds 1.96 (5% sign. level)?

- If $t_{meal\ pct}$ and $t_{calw\ pct}$ are uncorrelated:

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

- If $t_{meal\ pct}$ and $t_{calw\ pct}$ are correlated situation is even more complicated!

Testing 1 hypothesis on 2 or more coefficients

- 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 $q = 2$ restrictions: when testing the following hypothesis

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

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

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

- Equation illustrates that the F-statistic takes the potential correlation between the individual t-statistics into account
- In practice we use a software program to compute the F-statistic.

Testing 1 hypothesis on 2 or more coefficients

```
1 . 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

```
2 . test meal_pct calw_pct

 ( 1)  meal_pct = 0
 ( 2)  calw_pct = 0

       F(  2,   415) =   290.27
            Prob > F =    0.0000
```

Testing 1 hypothesis on 2 or more coefficients

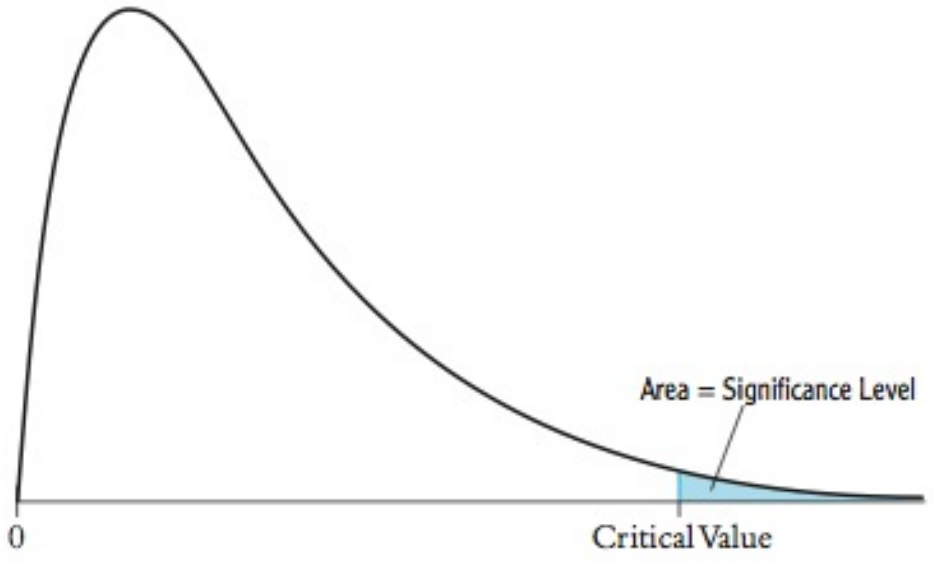
We want to test hypothesis that both the coefficient on % eligible for a free lunch and the coefficient on % eligible for calworks are zero?

$H_0 : \beta_{meal\ pct} = 0\ \&\ \beta_{calw\ pct} = 0 \qquad H_1 : \beta_{meal\ pct} \neq 0\ and/or\ \beta_{calw\ pct} \neq 0$

- 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*
- Previous slide shows $F = 290.27$
- Table 4 (S&W page 807) shows that the critical value at a 5% significance level equals 3.00 ($Pr_{H_0}(F_{2,\infty} > 3.00) = 0.05$)
- This implies that we reject H_0 at a 5% significance level because $290.27 > 3.00$.

Testing 1 hypothesis on 2 or more coefficients

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 : $\beta_j = \beta_{j,0}, ..., \beta_m = \beta_{m0}$ *for a total of q restrictions*

H_1 : *at least one of q restrictions under* H_0 *does not hold*

Step 1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_j X_{ji} + ... + \beta_1 X_{ki} + u_i$ by OLS

Step 2: Compute the F-statistic

Step 3: Reject the null hypothesis if F-statistic> $F_{q,\infty}^{critical}$ with $Pr\left(F_{q,\infty} > F_{q,\infty}^{critical}\right) = significance\ level$

Testing a joint hypothesis with q=3 restrictions

```
1 . 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

```
2 . test el_pct meal_pct calw_pct

 ( 1)  el_pct = 0
 ( 2)  meal_pct = 0
 ( 3)  calw_pct = 0

       F(  3,   415) =   481.06
            Prob > F =    0.0000
```

Testing a joint hypothesis with q=3 restrictions

$$H_0: \quad \beta_{el\ pct} = 0 \ \& \ \beta_{meal\ pct} = 0 \ \& \ \beta_{calw\ pct} = 0$$

$$H_1: \quad \beta_{el\ pct} \neq 0 \ and/or \ \beta_{meal\ pct} \neq 0 \ and/or \ \beta_{calw\ pct} \neq 0$$

Step 1: see previous slide

Step 2: F-statistic=481.06

Step 3: We reject the null hypothesis at a 5% significance level because F-statistic> $F^{critical}_{3,\infty} = 2.6$

Testing a joint hypothesis with q=1 restriction

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

- We have tested hypotheses that involve 1 restriction on 1 coefficient using the t-statistic
- We can also test these type of hypotheses using the F-statistic
- On slide 11 we saw that with $q = 2$ restrictions:

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

- With $q = 1$ restriction we have

$$F = t^2$$

Testing a joint hypothesis with q=1 restriction

```
. 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

```
.
. test class_size

 ( 1)  class_size = 0

       F(  1,   415) =    14.23
            Prob > F =     0.0002
```

The “overall” regression F-statistic

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

H_0 : $\beta_1 = 0, ..., \beta_k = 0$ *for a total of* $q = k$ *restrictions*

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

```
. 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

“Overall” regression *F*-statistic equals 361.68.

Testing single restrictions involving multiple coefficients

The % of students eligible for a free lunch and the % of students eligible for CALWORKS both measure whether students come from families with very low income

Suppose we want to test whether coefficient on *meal pct*$_i$ equals the coefficient on *calw pct*$_i$

$$H_0 : \beta_{\textit{meal pct}} = \beta_{\textit{calw pct}} \quad \textit{vs} \quad H_1 : \beta_{\textit{meal pct}} \neq \beta_{\textit{calw pct}}$$

There are two ways to perform this hypothesis test:

1. Transform regression model such that above H_0 is transformed to a null hypothesis that involves only 1 coefficient.
2. Test above restriction directly in software program (possible in Stata, not necessarily in all programs)

Testing single restrictions involving multiple coefficients

1.Transform regression model

.

$$H_0 : \beta_1 = \beta_2 \quad vs \quad H_1 : \beta_1 \neq \beta_2$$

Suppose we have regression model with only two regressors:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ & \textit{add \& subtract } \beta_2 X_{1i} \\ &= \beta_0 + \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} + u_i \\ & \textit{rewrite} \\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i \\ &= \beta_0 + \gamma X_{1i} + \beta_2 W_i + u_i \end{aligned}$$

Transformed hypothesis test:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0$$

Testing single restrictions involving multiple coefficients

1.Transform regression model

```
. gen W = meal_pct + calw_pct

.
. regress test_score class_size el_pct meal_pct W, 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	-.4807654	.0919803	-5.23	0.000	-.6615706	-.2999601
W	-.0478537	.0586541	-0.82	0.415	-.1631498	.0674424
_cons	700.3918	5.537418	126.48	0.000	689.507	711.2767

```
. test meal_pct

 ( 1)  meal_pct = 0

       F(  1,   415) =    27.32
            Prob > F =    0.0000
```

Testing single restrictions involving multiple coefficients

2. Test restriction directly in software program

```
. 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

```
.
. test meal_pct=calw_pct

 ( 1)  meal_pct - calw_pct = 0

       F(  1,   415) =    27.32
            Prob > F =     0.0000
```

.

Measures of fit in multiple regression model

There are 3 measures that measure how well the OLS estimate of the multiple regression line describes or "fits" the data

- The Standard Error of the Regression (SER)

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i}$$

- The R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- The "adjusted" $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

The SER in the multiple regression model

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i}$$

- The *SER* estimates the standard deviation of the error term u_i
- It measures spread of distribution of Y around regression line.
- $n - k - 1$ is degrees of freedom correction

```
. 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 *SER* = *Root MSE* = 9.0843.

The R^2

- The R^2 is the fraction of the sample variance of Y_i explained/predicted by $X_{1i}, ..., X_{ki}$
- The R^2 is the ratio of the sample variance of $\widehat{Y}_i$ and the sample variance of Y_i

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{SSR}{TSS}$$

- In the multiple regression model the R^2 always increases when a regressor is added
- Also when this additional regressor explains very little of Y!
- Reason: whenever estimated coefficient on additional regressor is not exactly zero the $SSR = \sum_{i=1}^{n} \widehat{u}_i$ is reduced.

The R^2

```
. regress test_score class_size el_pct meal_pct, robust

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     453.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7745
                                                Root MSE          =     9.0801
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.9983092	.2700799	-3.70	0.000	-1.529201	-.4674178
el_pct	-.1215733	.0328317	-3.70	0.000	-.18611	-.0570366
meal_pct	-.5473456	.0241072	-22.70	0.000	-.5947328	-.4999583
_cons	700.15	5.56845	125.74	0.000	689.2042	711.0958

```
.
. 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 R^2 and the homoskedasticity-only F-statistic

- There is a link between the F-statistic and the R^2
- This link can be easily displayed when we assume homoskedastic errors:

$$F^{homoskedasticity} = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}$$

  - with $SSR_{restricted}$ the sum of squared residuals assuming the q restriction under H_o hold
  - and $SSR_{unrestricted}$ the sum of squared residuals assuming the q restriction under H_o do not hold

- Since $R^2 = 1 - \frac{SSR}{TSS}$ we can also write

$$F^{homoskedasticity} = \frac{\left(R^2_{unrestricted} - R^2_{restricted}\right)/q}{\left(1 - R^2_{unrestricted}\right)/(n - k_{unrestricted} - 1)}$$

The R^2 and the homoskedasticity-only F-statistic

```
. 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 el_pct meal_pct calw_pct
```

Source	SS	df	MS	Number of obs =	420
				F(4, 415) =	357.05
Model	117862.131	4	29465.5327	Prob > F =	0.0000
Residual	34247.4629	415	82.524007	R-squared =	0.7749
				Adj R-squared =	0.7727
Total	152109.594	419	363.030056	Root MSE =	9.0843

test_score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.014353	.2397376	-4.23	0.000	-1.485605	-.5431019
el_pct	-.1298219	.0339973	-3.82	0.000	-.1966504	-.0629934
meal_pct	-.5286191	.0321901	-16.42	0.000	-.591895	-.4653432
calw_pct	-.0478537	.0609698	-0.78	0.433	-.1677019	.0719944
_cons	700.3918	4.697969	149.08	0.000	691.1571	709.6266

The R^2 and the homoskedasticity-only F-statistic

$$H_0: \quad \beta_{el\ pct} = 0 \ \& \ \beta_{meal\ pct} = 0 \ \& \ \beta_{calw\ pct} = 0$$

$$H_1: \quad \beta_{el\ pct} \neq 0 \ and/or \ \beta_{meal\ pct} \neq 0 \ and/or \ \beta_{calw\ pct} \neq 0$$

From previous slide:

- $R^2_{restricted} = 0.0512$
- $R^2_{unrestricted} = 0.7749$
- $n = 420$, $q = 3$ and $k_{unrestricted} = 4$.

$$F^{homoskedasticity} = \frac{(0.7749 - 0.0512)/3}{(1 - 0.7749)/(420 - 4 - 1)} = 444.74$$

The R^2 and the homoskedasticity-only F-statistic

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

Source	SS	df	MS	Number of obs =	420
				F(4, 415) =	357.05
Model	117862.131	4	29465.5327	Prob > F =	0.0000
Residual	34247.4629	415	82.524007	R-squared =	0.7749
				Adj R-squared =	0.7727
Total	152109.594	419	363.030056	Root MSE =	9.0843

test_score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.014353	.2397376	-4.23	0.000	-1.485605	-.5431019
el_pct	-.1298219	.0339973	-3.82	0.000	-.1966504	-.0629934
meal_pct	-.5286191	.0321901	-16.42	0.000	-.591895	-.4653432
calw_pct	-.0478537	.0609698	-0.78	0.433	-.1677019	.0719944
_cons	700.3918	4.697969	149.08	0.000	691.1571	709.6266

```
. test el_pct meal_pct calw_pct

 ( 1)  el_pct = 0
 ( 2)  meal_pct = 0
 ( 3)  calw_pct = 0

       F(  3,   415) =   444.59
            Prob > F =     0.0000
```

The "adjusted" $\overline{R}^2$

- The R^2 increases whenever a regressor is added.
- also when regressor explains very little of Y and does not improve the "fit" of the model
- We can correct for this by deflating the R^2 which gives the "adjusted" $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

- When we add a regressor:
  - $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$ decreases (whenever coefficient estimate is not exactly zero)
  - k increases, which increases $\frac{n-1}{n-k-1}$

The "adjusted" $\overline{R}^2$

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

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     453.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7745
                                                Root MSE          =     9.0801
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.9983092	.2700799	-3.70	0.000	-1.529201	-.4674178
el_pct	-.1215733	.0328317	-3.70	0.000	-.18611	-.0570366
meal_pct	-.5473456	.0241072	-22.70	0.000	-.5947328	-.4999583
_cons	700.15	5.56845	125.74	0.000	689.2042	711.0958

```
2 . display e(r2_a)
.77288978

3 . 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

```
4 . display e(r2_a)
.77267996
```

Interpreting measures of fit

The R^2 or "adjusted" $\overline{R}^2$ tell you whether the regressors are good at predicting or "explaining" the values of the dependent variable in the sample of data.

- An R^2($\overline{R}^2$) near 1 implies that X_{1i}, .., X_{ki} produce good predictions of Y_i
- An R^2($\overline{R}^2$) near 0 implies that X_{1i}, .., X_{ki} don't produce good predictions of Y_i
- It is important not rely too much on the R^2($\overline{R}^2$)
- Maximizing the R^2($\overline{R}^2$) rarely answers an economically meaningfull question!

Interpreting measures of fit

A high R^2 ($\overline{R}^2$) does **NOT** tell you whether:

An included variable is statistically significant.

- you need to perform an hypothesis test using a t- or F-statistic

The coefficients measure the true causal effect of the regressors

- The OLS estimators are only unbiased & consistent estimators if all Least Squares assumptions hold!

There is omitted variable bias

- Omitted variable bias implies violation of 1st Least Squares assumption, this cannot be assessed by R^2 ($\overline{R}^2$)

You have chosen the most appropriate set of regressors

- Appropriate depends on question of interest
- When estimating a causal effect, OLS assumptions should hold

Interpreting *stars* in a table with regression output

Dependent variable: district average test scores				
	1	2	3	4
Class size	-2.280***	-1.101**	-0.998***	-1.014***
	(0.519)	(0.433)	(0.270)	(0.269)
Percentage of English learners		-0.650***	-0.122***	-0.130***
		(0.031)	(0.033)	(0.036)
Percentage with free lunch			-0.547***	-0.529***
			(0.024)	(0.038)
Percentage with CALWORKS				-0.048
				(0.059)
N	420	420	420	420

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level