Econometrics 1

Lecture 7: Linear Regression (2)
Linear regression with multiple regressors

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Website: https://huangjp.com

Review of
the linear regression with one regressor

The linear regression model

- The linear regression model with one regressor

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

coefficients

error term

dependent variable

independent variable / regressor

The OLS estimator, predicted values, and residuals

- The OLS estimators of the slope and the intercept are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- The OLS predicted value: $\hat{Y}_i = \boxed{\hat{\beta}_0 + \hat{\beta}_1 X_i}$

- The residuals: $\hat{u}_i = Y_i - \hat{Y}_i$

sample regression line/
sample regression function

The least squares assumptions

For the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n$$

it is assumed that:

1. The error term u_i has conditional mean zero given X_i:

   $$\mathrm{E}(u_i \mid X_i) = 0 \qquad (\Rightarrow \mathrm{corr}(X_i, u_i) = 0)$$

2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution; and

3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

Hypotheses concerning β_1

- Two-sided hypotheses

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_{1,0}$$

- The *t*-statistic

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

where

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \;\; \hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2]^2}$$

Heteroskedasticity-robust standard error (HC1)

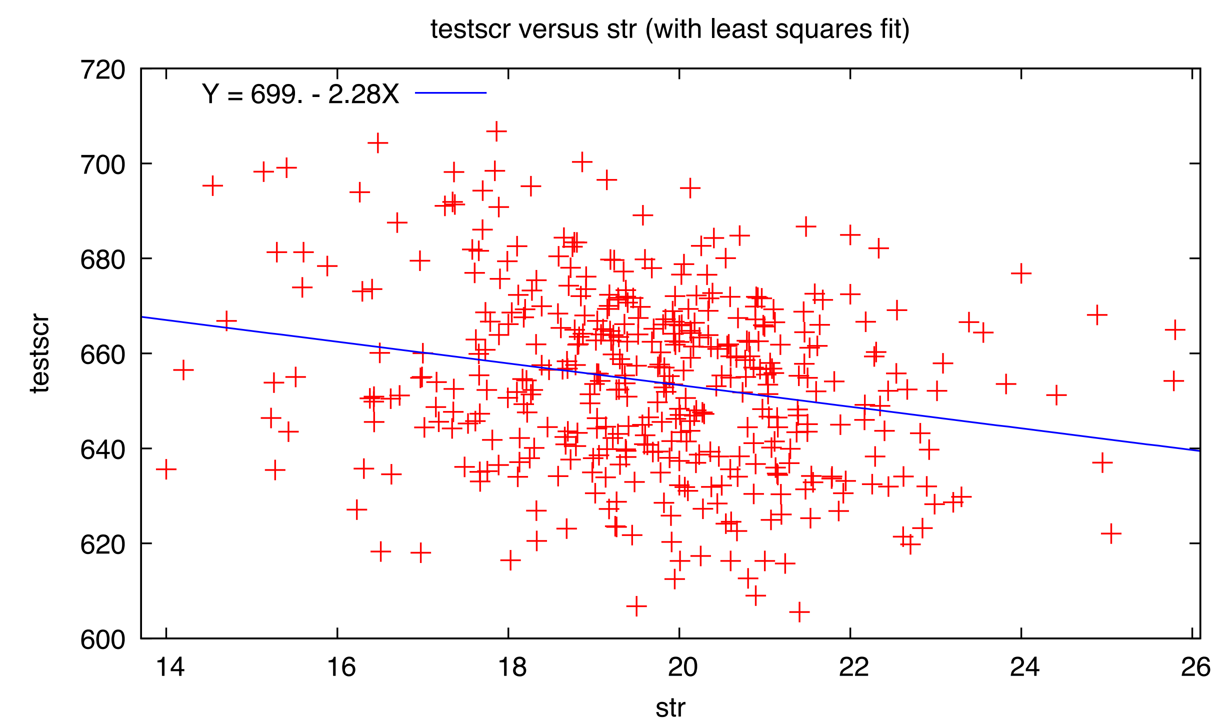
- The *p*-value

$$p\text{-value} = 2\Phi(-|t^{act}|)$$

Omitted variable bias

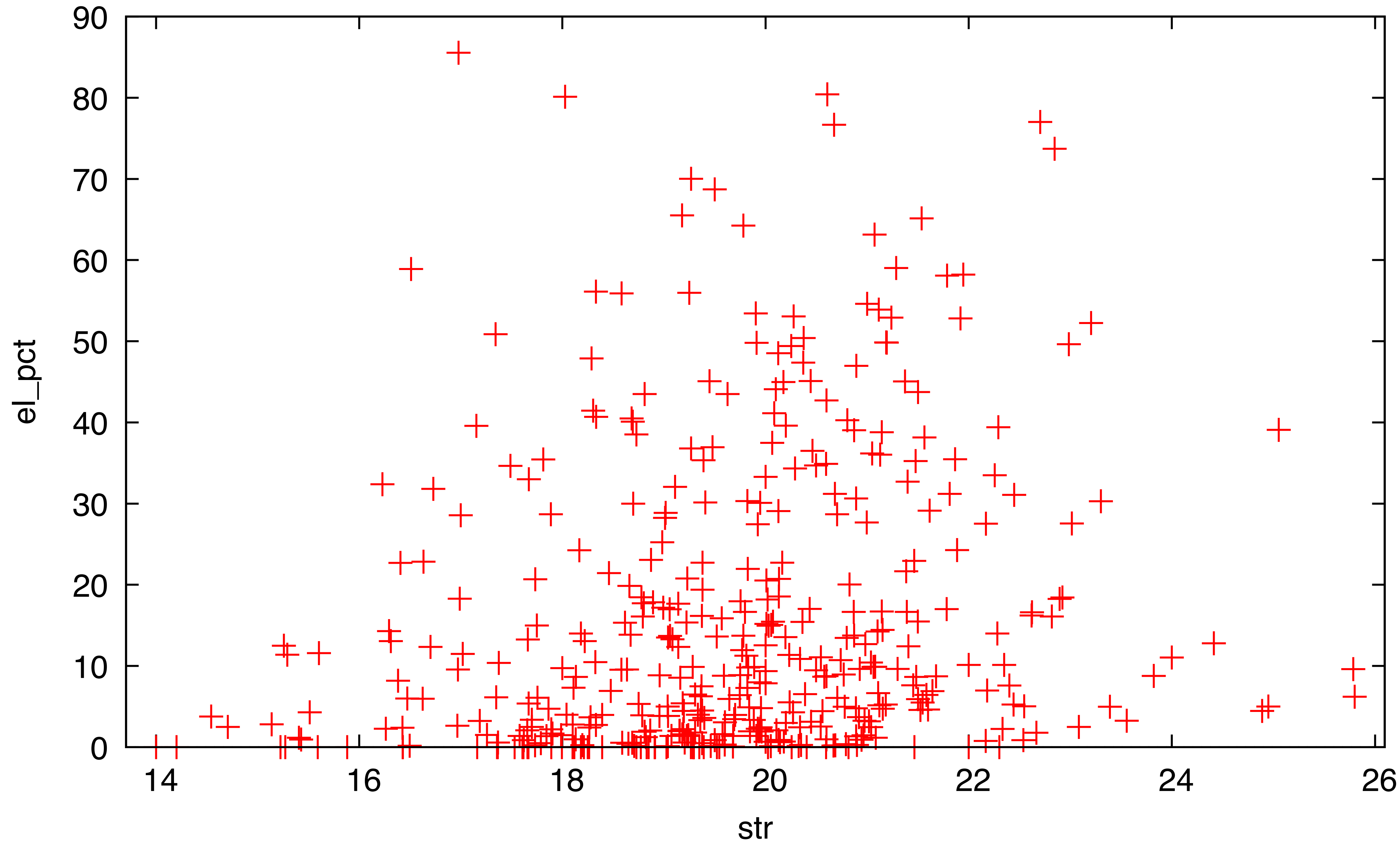
The STAR dataset

- Variables in the STAR dataset that may affect **test score**
  - total enrollment
  - number of teachers
  - number of computers
  - computers per student
  - expenditures per student
  - **student teacher ratio**
  - *percent of english learners*
  - percent qualifying for reduced-price lunch
  - percent qualifying for CalWORKs (California Work Opportunities and Responsibility to Kids program)
  - district average income



Student teacher ratio and percentage of English learners

The correlation between the two variable is 0.188



Omitted variable bias

- If the regressor is correlated with a variable that has been omitted from the analysis and that determines, in part, the dependent variable, then the OLS estimator will have **omitted variable bias**.

Omitted variable bias occurs when the following two conditions are true:

1. when the omitted variable is correlated with the included regressor, and
2. when the omitted variable is a determinant of the dependent variable.

Omitted variable bias

- The first least squares assumption

$$\mathrm{E}(u_i \mid X_i) = 0 \qquad (\Rightarrow \mathrm{corr}(X_i, u_i) = 0)$$

- If there is omitted variable bias, the error term is correlated with the independent variable, therefore this assumption is violated.

  The OLS estimator is then biased.

- Read the part “A Formula for Omitted Variable Bias” on page 224.

The multiple regression model

Multiple regression model

- Linear regression model with multiple regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + u_i$$

- The population regression line

$$\mathrm{E}(Y | X_{1i} = x_1, \ldots, X_{mi} = x_m) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- The intercept β_0 — the expected value of Y when all the X's equal 0.

- The coefficient β_k — the expected change in Y_i resulting from changing X_{ki} by one unit, holding constant the other X's.

The OLS estimator

- The OLS estimators $\hat{\beta}_0, \ldots, \hat{\beta}_m$ are the ones minimizing the sum of squares of prediction mistakes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_m X_{mi})^2$$

- The OLS estimators can be evaluated by local grid search (trial and error), or by explicit formulas (see Chapter 18, beyond the scope of this course). In practice, one can easily estimate them using statistical softwares.

Application to test scores

- The percentage of English learners, `el_pct`, can be another variable that explains test scores.
  - Dependent variable $Y = \texttt{testscr}$
  - Independent variables (regressors)
    $X_1 = \texttt{str}, \quad X_2 = \texttt{el\_pct}$

- The regression model is

$$\text{testscr}_i = \beta_0 + \beta_1 \, \text{str}_i + \beta_2 \, \text{el\_pct}_i + u_i$$

- In gretl:

```
ols testscr const str el_pct --robust
```

Regression results

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z      p-value
  -------------------------------------------------------------
  const      686.032       8.72822       78.60    0.0000    ***
  str         -1.10130     0.432847      -2.544   0.0109    **
  el_pct      -0.649777    0.0310318    -20.94    2.36e-97  ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

R^2 and adjusted R^2

- The R^2 measure

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

increases when the number of regressors increases, which does not depend on whether the fit of the model is improved.

- Adjusted R^2, or $\overline{R}^2$

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

k is the number of regressors

The least squares assumptions in the multiple regression model

For the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

it is assumed that:

1. u_i has conditional mean zero given $X_{1i}, X_{2i}, \ldots, X_{ki}$:

   $$E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$ (no omitted variables)

2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution; and

3. Large outliers are unlikely: $X_{1i}, X_{2i}, \ldots, X_{ki}$ and Y_i have nonzero finite fourth moments.

The least squares assumptions in the multiple regression model (cont.)

For the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

it is assumed that:

4. There is no *perfect multicollinearity*.

Perfect multicollinearity

The regressors are said to exhibit perfect multicollinearity if one of the regressors is a perfect linear function of the other regressors.

Multicollinearity

If the regressors have perfect multicollinearity:

- It is impossible to compute the OLS estimator.
  ⇒ E.g., *the dummy variable trap*

a problem

If the regressors have imperfect (nearly perfect) multicollinearity:

- The coefficients remain unbiased

- At least one of the coefficients will be imprecisely estimated (large sample variance).

a feature

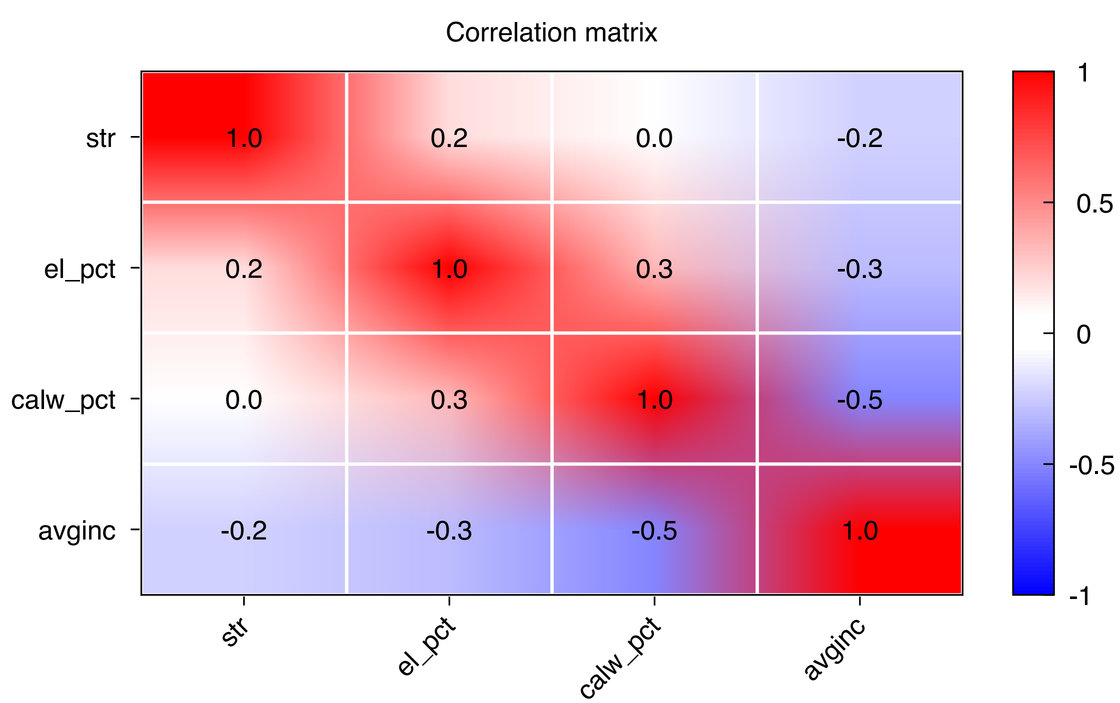
How to detect multicollinearity

- Check correlation matrix

```
corr str el_pct calw_pct avginc --plot=display
```

```
Correlation Coefficients, using the observations 1 - 420
5% critical value (two-tailed) = 0.0957 for n = 420

        str       el_pct      calw_pct       avginc
     1.0000       0.1876        0.0183      -0.2322  str
                  1.0000        0.3196      -0.3074  el_pct
                                1.0000      -0.5127  calw_pct
                                             1.0000  avginc
```



How to detect multicollinearity (cont.)

- Variance Inflation Factors (VIF)

  vif

  Execute this command after an `ols` command

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

         str    1.036
      el_pct    1.036

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation
coefficient between variable j and the other independent variables
```

How to deal with multicollinearity

- A simple solution to the problem of multicollinearity:

  Remove or replace the regressors that are perfectly (imperfectly) multicollinear with other regressors.

- Imperfect multicollinearity is not necessarily an error, but rather just a feature of OLS, your data, and the question you are trying to answer.

Hypothesis tests and model specification

Hypothesis tests for a single coefficient

- Hypotheses (two-sided):

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

- The *t*-statistic:

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

General *t*-statistic

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

- The *p*-value (large sample):

$$p\text{-value} = 2\Phi(-|t^{act}|)$$

Regression results

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z      p-value
  -------------------------------------------------------------
  const      686.032       8.72822        78.60    0.0000   ***
  str         -1.10130     0.432847       -2.544   0.0109   **
  el_pct      -0.649777    0.0310318     -20.94    2.36e-97 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

Summarize hypothesis testing results

- As an equation

$$\widehat{\text{testscr}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times \text{str} - \underset{(0.031)}{0.650} \times \text{el\_pct}$$

standard errors

Provide the most important information: **estimates** and **standard errors**. The *t*-statistics and *p*-values can be calculated.

- Use a table when you have several regression models

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

models

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.

Table 4

Individual Contribution to the Public Good

	Dep. var.: Individual contribution to the PGG			
	Model 1	Model 2	Model 3	Model 4
Northern Italy	1.213*	1.161**	1.066**	
	(0.580)	(0.432)	(0.429)	
Latitude				0.195***
				(0.057)
Individual choices over lotteries				
Strongly risk averse			0.806	0.806
			(0.572)	(0.569)
Risk neutral/risk loving			−0.921*	−0.895*
			(0.445)	(0.450)
Task comprehension (1 = low)		0.757	0.819	0.822
		(0.739)	(0.731)	(0.725)
Socio-demographic characteristics (control variables)	No	Yes	Yes	Yes
No. obs. (individuals)	372	372	372	372
R^2	0.015	0.085	0.101	0.106

Notes. OLS regression with standard errors robust for clustering at the session level (in parentheses). The dependent variable is the contribution of one participant averaged over all rounds of the PGG. The default category for risk preference is: moderately risk averse. Socio-demographic characteristics are listed in the main text. ***, **, and * indicate significance at the 1%, 5% and 10% level, respectively.

From Bigoni et al. (2016), *The Economic Journal*, 126:1318-1341

A guide for how to format tables and figures:
http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWtablefigs.html

Tests of Joint hypotheses

- The *overall* joint hypotheses of slope coefficients

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_m = 0$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, m\}$$

- Joint hypotheses with *q* restrictions

$$H_0 : \beta_{j_1} = \beta_{j_1,0},\ \beta_{j_2} = \beta_{j_2,0},\ \ldots,\ \beta_{j_q} = \beta_{j_q,0}$$
$$H_1 : \text{one or more of the } q \text{ restrictions under } H_0 \text{ does not hold}$$

- This test uses an *F*-statistic, which follows $F_{q,n-m-1}$ distribution ($F_{q,\infty}$ for large samples).

Regression results

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z      p-value
  --------------------------------------------------------
  const      686.032       8.72822        78.60    0.0000   ***
  str         -1.10130     0.432847       -2.544   0.0109   **
  el_pct      -0.649777    0.0310318     -20.94    2.36e-97 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

heteroskedasticity-robust

Testing single restrictions involving multiple coefficients

- Sometimes we need to test hypotheses like the following

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

- The null hypothesis has a single restriction, which can be tested using an *F-statistic* with an $F_{1,\infty}$ distribution in large sample.

Practice in gretl

1. OLS regression

```
ols testscr const str expn el_pct --robust
```

2. Test hypotheses (after an `ols` command)

```
restrict
    b[2] - b[3] = 0
end restrict
```

OLS results

```
Model 2: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z       p-value
  ---------------------------------------------------------
  const      649.578       15.4583       42.02     0.0000   ***
  str         -0.286399     0.482073     -0.5941   0.5524
  expn         3.86790      1.58072       2.447    0.0144   **
  el_pct      -0.656023     0.0317844   -20.64     1.21e-94 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    85699.71   S.E. of regression   14.35301
R-squared            0.436592   Adjusted R-squared   0.432529
F(3, 416)            147.2037   P-value(F)           5.20e-65
Log-likelihood      -1712.808   Akaike criterion     3433.615
Schwarz criterion    3449.776   Hannan-Quinn         3440.003

Excluding the constant, p-value was highest for variable 14 (str)
```

Test restricted model

```
Restriction:
 b[str] - b[expn] = 0

Test statistic: Robust F(1, 416) = 8.9403, with p-value =
0.00295511

Restricted estimates:

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const      685.822       11.3696       60.32    4.31e-208 ***
  str         -0.854052     0.459004     -1.861   0.0635    *
  expn        -0.854052     0.459004     -1.861   0.0635    *
  el_pct      -0.656690     0.0396393   -16.57    9.96e-48  ***

  Standard error of the regression = 14.5489
```

Model specification

- We often have to determine which variables to be included as regressors in a regression model.

- There is no single rule that applies in all situations.

- A base set of regressors should be chosen using a combination of *expert judgement*, *economic theory*, and *knowledge of how the data were collected*. The model with these regressors is referred to as a **base specification**.

- The base specification should contain *the variable of primary interest* and *control variables*.

Model specification (cont.)

- The nest step is to develop a list of candidate **alternative specification**, that is, alternative sets of regressors.
- If the estimates are numerically similar, it provides evidence that the base specification is reliable.
- If the estimates change substantially, it provides evidence that the base specification has bias.
- Do not rely solely on the R^2 or the adjusted R^2. See page 276-277.

Practice

- Read Section 7.6, and reproduce the analysis in it. The STAR data is given in “caschool.xlsx”.

- You can make tables similar to Table 7.1 in gretl, using the `modeltab` command. See command reference.

  For example:

```
modeltab free
ols testscr const str --robust --quiet
modeltab add
ols testscr const str el_pct --robust --quiet
modeltab add
modeltab show
```

References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.