Econometrics 1 *Applied Econometrics with R*

Lecture 8: Linear Regression (2)

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼1726

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Office hour: Mon./Tue. 13:00-14:00

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

Linear Regression with Multiple Regressors

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 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

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

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

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

Using the `lm` command

- The function for fitting linear models in R is `lm`
- Basic usage of `lm`

```
> lm(y ~ x)
> lm(y ~ x1 + x2)
```

- It returns an object that contains informations about the linear regression

```
> z <- lm(y ~ x1 + x2)
> summary(z)
```

Practice

- Import data from `caschool.xlsx`
- Discover the results of applying the `lm` command to the regression model

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

- Can you find the OLS estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$?

```
> f2 <- lm(testscr ~ str + el_pct)
> summary(f2)

Call:
lm(formula = testscr ~ str + el_pct)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 ** 
el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

```
> f <- lm(testscr ~ str)
> summary(f)

Call:
lm(formula = testscr ~ str)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.727 -14.251   0.483  12.822  48.540 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
str          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,Adjusted R-squared:  0.04897 
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

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 $X_{1i}, X_{2i}, \ldots, X_{ki}$:

$$\mathrm{E}(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$

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

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.

If the regressors have imperfect (nearly perfect) multicollinearity:

- The coefficients remain unbiased

- At least one of the coefficients will be imprecisely estimated.

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.

- Further reading:

  Multicollinearity in R
  https://datascienceplus.com/multicollinearity-in-r/

Hypothesis Tests

Why hypothesis tests

- The OLS estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ have different values when different samples are used.
- What can we say about the relation of dependent and independent variables in the population?
- Hypothesis tests for regression coefficients.

Hypothesis tests for a single coefficient

- Hypotheses:

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

- The *t*-statistic:

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

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

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

```
> f2 <- lm(testscr ~ str + el_pct)
> summary(f2)

Call:
lm(formula = testscr ~ str + el_pct)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 ** 
el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Related commands

- OLS estimates (fitted coefficients)

```
> coef(f2)
```
or
```
> f2$coefficients
```

- Standard errors

```
> coef(summary(f2))[,2]
```

- Confidence intervals

```
> confint(f2)
```

Summarize hypothesis testing results

- As an equation

$$\widehat{\text{testscr}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10} \times \text{str} - \underset{(0.04)}{0.65} \times \text{el\_pct}$$

  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 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	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

Tests of Joint hypotheses

- The overall joint hypotheses of slope coefficients

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

- This test use an *F*-statistic, which follows $F_{m,n-m-1}$ distribution.

- The `lm` command returns the *F*-statistic with the corresponding *p*-value.

```
> f2 <- lm(testscr ~ str + el_pct)
> summary(f2)

Call:
lm(formula = testscr ~ str + el_pct)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 ** 
el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264,	Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

Some useful commands

Table 3.1. Generic functions for fitted (linear) model objects.

Function	Description
`print()`	simple printed display
`summary()`	standard regression output
`coef()`	(or `coefficients()`) extracting the regression coefficients
`residuals()`	(or `resid()`) extracting residuals
`fitted()`	(or `fitted.values()`) extracting fitted values
`anova()`	comparison of nested models
`predict()`	predictions for new data
`plot()`	diagnostic plots
`confint()`	confidence intervals for the regression coefficients
`deviance()`	residual sum of squares
`vcov()`	(estimated) variance-covariance matrix
`logLik()`	log-likelihood (assuming normally distributed errors)
`AIC()`	information criteria including AIC, BIC/SBC (assuming normally distributed errors)

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*. Such regressors are referred to as a **base specification**.

- The nest step is to develop a list of candidate **alternative specification**, that is, alternative sets of regressors.

Model specification

- Comparing the estimates of coefficients between the base specification and the alternative specifications.

- If the estimates are numerically similar, it provides evidence that the base specification are reliable.

- If the estimates change substantially, it provides evidence that the base specification has bias.

- Interpreting the R^2 and the adjusted R^2 carefully.

Practice

- Add the percentage of students who are eligible for receiving a reduced priced lunch at school (`meal_pct`) to your regressors.

- Write down the regression model of `testscr` on `str`, `el_pct`, and `meal_pct`.

- Perform the OLS estimation and test hypotheses of coefficients.

Assignment 2

References

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

2. Kleiber, C. and Zeileis, A., *Applied Econometrics with R*, Springer, 2008.