ECON20003 Quantitative Methods 2

Tutorial 7 (Week 4 - Tuesday)

Zheng Fan

The University of Melbourne

Introduction

Zheng Fan

- Ph.D candidate in Economics at Unimelb
- know more about me: zhengfan.site

If you need help,

- Consultation & Ed discussion board (your first priority)
- Email Dr. Xuan Vu for all subject matters
- Consult Stop 1 for special consideration
- Email: fan.z@unimelb.edu.au (last resort!)

Before posting any questions, make sure you have reviewed the materials on Canvas and questions on Ed discussion board!

Learning Objectives

By the end of this tutorial, you should be able to

- Understand the structure of a multiple linear regression model
- Interpret regression coefficients correctly
- Assess model fit using R squared and adjusted R squared
- Conduct hypothesis testing using F tests and t tests
- Evaluate regression assumptions using residual diagnostics

Multiple Linear Regression Model

Multiple linear regression extends simple regression by allowing more than one explanatory variable.

The population model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

The sample regression model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

In this subject, estimation is always done using R.

Example: Household Food Consumption

We study the relationship between

- Household food consumption
- Household income
- Household size

The model is

$$foodcon_i = \beta_0 + \beta_1 income_i + \beta_2 size_i + \varepsilon_i$$

Before estimating the model, we should think about the expected signs of coefficients.

(a) Expected Signs of Coefficients

Economic intuition suggests

- Higher income should increase food consumption, holding size constant
- Larger households should consume more food, holding income constant

Therefore, both slope coefficients are expected to be positive.

(b) Exploratory Data Analysis - visual plot

Before running a regression, always inspect the data visually.

We plot

- Food consumption versus income
- Food consumption versus household size

Scatter plots help assess

- Direction of relationships
- Linearity
- Potential outliers

(c) Estimating the Model in R

We estimate the model using the `lm()` function in R.

R Code

```
m <- lm(foodcon ~income + size, data = t7e1)
# or
m <- lm(t7e1$foodcon ~t7e1$income + t7e1$size)

summary(m)
```

The summary output provides coefficient estimates, standard errors, test statistics, and goodness of fit measures.

(c) Regression Output

```
Call:
lm(formula = foodcon ~ income + size, data = t7e1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9748 -0.3340 -0.1127  0.1496  2.7894

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7943798  0.4363349   6.404 1.55e-06 ***
income      -0.0001639  0.0065644  -0.025     0.98
size         0.3834845  0.0718867   5.335 2.04e-05 ***

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.7188 on 23 degrees of freedom
Multiple R-squared: 0.558, Adjusted R-squared: 0.5196
F-statistic: 14.52 on 2 and 23 DF, p-value: 8.363e-05
```

(d) Interpreting Coefficients

Each slope coefficient measures the effect of one variable **holding all other variables constant**.

- The income coefficient estimate (-0.00016) means that, keeping household size constant, an extra \$1,000 household income is likely to be accompanied by a \$0.16 decrease of household food consumption.
- The size coefficient estimate (0.383) is positive as expected. It means that, keeping household income constant, with every additional household member household food consumption is expected to increase by \$383.

This conditional interpretation is crucial in multiple regression.

(e) Goodness of Fit

The unadjusted coefficients of determination (R^2) is 0.558.

- It suggests that about 56% of the total variation in household food consumption can be explained by the variations in household income and size.

(f) Goodness of Fit

Adjusted R squared accounts for

- Sample size
- Number of explanatory variables

Adjusted R squared is especially useful when comparing models with different numbers of regressors.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right) = 1 - \frac{25}{23}(1 - 0.558) = 0.520$$

(g) Overall Model Significance

We test whether the regression model is useful using an F test.

$$H_0 : \beta_1 = \beta_2 = 0,$$
$$H_A : \beta_1 \neq 0 \text{ or/and } \beta_2 \neq 0$$

A small p value leads us to conclude that the model explains variation in the dependent variable.

(h) Confidence Intervals

A 95 percent confidence interval gives a range of plausible values for the true coefficient.

- If zero is inside the interval, the coefficient is not statistically significant
- If zero is outside the interval, the coefficient is statistically significant

In R, confidence intervals are obtained using

R Code

```
confint(m, level = 0.95)
```

Regression Output

From R output:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7943798  0.4363349   6.404 1.55e-06 ***
income      -0.0001639  0.0065644  -0.025     0.98
size         0.3834845  0.0718867   5.335 2.04e-05 ***
```

Note: t value is calculated based on null = 0

p value is calculated based on two-tail test: null = 0

You need to calculate for any other hypothesis.

(i) Individual Significance Tests

Each coefficient is tested using a t test.

- $H_0 : \beta = 0$
- $H_0 : \beta > 0$

(j) Regression Assumptions

For valid inference in small samples, we assume

- Linearity
- Independence
- Homoskedasticity
- Normality of errors

Normality is assessed using regression residuals.

Residual Diagnostics

Residuals are obtained using

R Code

```
m_res <- residuals(m)
```

We examine

- Histograms with normal curves
- Normal Q Q plots
- Formal normality tests

Implications of Non Normality

If residuals are not normally distributed

- OLS estimates remain unbiased
- Confidence intervals and hypothesis tests may be unreliable in small samples

This is especially important when the sample size is limited.