Week 13 Live Session

w203 Instructional Team

Aug 7, 2017

Announcements

Quiz 2 is available and due before the week 14 live session. You instructor has the password.

Interpreting Specifications

What is the interpretation of β_1 in each of the following specifications?

log-level: $\log y = \beta_0 + \beta_1 x + u$

level-log: $y = \beta_0 + \beta_1 \log x + u$

log-log: $\log y = \beta_0 + \beta_1 \log x + u$

added indicator: $y = \beta_0 + \beta_1 x + \beta_2 I(x = 0) + u$

no intercept: $y = \beta_1 x + u$

Issues with MLR: Using Logarithms

Using logarithms for the dependent or independent variables is one method used by statisticians to allow nonlinear relationships between the explained and explanatory variables.

Another potential benefit of using logs is that taking the log of a variable often narrows its range, which is useful when working with variables that are large monetary values.

Be careful not to use log transformation indiscriminantly - in some cases this can create extreme values. For example, when a variable y is between zero and one and takes on values close to zero, log(y) can be very large in magnitude whereas the original variable, y, is bounded between zero and one.

Hypothesis tests in MLR

Testing Hypotheses about a Single Population Parameter: The t Test

Hypothesis testing for a single coefficient is identical to the bivariate regression case with the t test statistic. The t statistic associated with any OLS coefficient can be used to test whether the corresponding unknown parameter in the population is equal to any given constant.

In most applications, our primary interest lies in testing the null hypothesis H_0: $\beta_j = 0$, where j corresponds to any of the k independent variables. The statistic we use to test (against any alternative) is called "the" t statistic or "the" t ratio of $\hat{\beta}_j$ and is defined as $t_{\hat{\beta}_j} = \hat{\beta}_j / se(\hat{\beta}_j)$

Since β_j measures the partial effect of x_j on (the expected value of) y, after controlling for all other independent variables, this hypothesis is saying that once $x_1, x_2, ., x_{j-1}, x_j, ., x_k$ have been accounted for, x_j has no effect on the expected value of y.

Note: We cannot state the null hypothesis as "x_j does have a partial effect on y" because this is true for any value of β_j other than zero.

Testing Hypotheses about a Single Linear Combination of the Parameters

In 4.4, Wooldridge shows how to test hypotheses about a single linear combination of the b_j by rearranging the equation and running a regression using transformed variables.

Remember to pay attention to the magnitude of the coefficient estimates in addition to the size of the t statistics.

Q. What is the difference between economical (or practical) and statistical significance?

The statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$. The economic or practical significance of a variable is related to the size (and sign) of $\hat{\beta}_j$.

Testing Multiple Linear Restrictions: The F Test

To test multiple hypotheses about the underlying parameters $\beta_0, \beta_1, ..., \beta_k$, we can use multiple restrictions to test whether a set of independent variables has no partial effect on a dependent variable.

It is often useful to test joint hypotheses together rather than use independent tests of the coefficients. For instance, the joint test that math and verbal SAT scores have no effect on W203 grades against the alternative that one or both scores has an effect.

Tests of joint hypotheses have test statistics that are distributed according to either the F or χ^2 distributions. These tests are often called Wald tests and may be quoted either as F or as χ^2 statistics.

When computing an F statistic, the numerator df is the number of restrictions being tested, while the denominator df is the degrees of freedom in the unrestricted model. If there is only one numerator degree of freedom, we are testing only a single hypothesis and the F-test becomes equivalent to the t-test. (Mathematically, if a random variable t follows the t_{N-K} distribution, then its square t^2 follows the $F_{(1,N-K)}$ distribution) The p-value you get from either test should be the same.

Exercises

Qualitative Data: Using Dummy Variables

Explain why the indicator variables have been included in the following models

$wage = \beta_0 + \beta_1 educ + \beta_2 I(educ = 12)$

$wage = \beta_0 + \beta_1 educ + \beta_2 female$

$wage = \beta_0 + \beta_1 educ + \beta_2 female + beta_3 educ * female$

$wage = \beta_0 + \beta_1 female + \beta_2 I(educ = 2) + \beta_3 I(educ = 3) + ... + \beta_{20} I(educ = 20)$

Handling more than two categories

Consider the following made-up data:

```
wage = c(12,52,35,64,65,76)
race = factor(c("black", "white", "white", "other", "black", "white"))
race_data = data.frame(wage, race)
```

Suppose you are interested in measuring the difference between the average wage of each race category. You build a linear model as follows:

```
race_model = lm(wage ~ race, data = race_data)
race_model$coefficients
```

```
## (Intercept)   raceother   racewhite
##    38.50000    25.50000    15.83333
```

Q1. Explain how R entered the factor variable into the linear model.

Q2. Based on the above output, how would you compute the average wage for black respondents? For white respondents?

Q3. Explain how R selected the base category above. Show how you would run the regression using white as the base category. (hint: a useful command is relevel)

```
?relevel
```

R Exercise

The file, engin.RData contains data from the Material Requirement Planning Survey carried out in Thailand during 1998. It was collected by Thada Chaisawangwong, a former graduate student at MSU. These data are for engineers in Thailand, and represents a more homogeneous group than data sets that consist of people across a variety of occupations.

```
library(car)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
load("engin.RData")
```

1. Use visualizations to investigate the bivariate relationship between wage and educ. Based on this analysis, what transformation, if any, would you apply to wage?
2. Create a linear model, model1, with just male and educ on the right hand side. Show how you would test the hypothesis that male has no effect on wage.
3. You are considering adding two variables representing experience to the model, exper and pexper. Show how you would test whether these variables are jointly significant.
4. You are considering adding a variable, swage, representing starting wage to the right hand side. Explain how this would affect your ability to understand the effects of gender.

5. Now show how you would alter your model to test whether each additional year of education has the same effect for men and for women.
6. As time allows, continue trying different model specifications, with the goal of understanding what effect gender has on wages.