wooldRidge-vignette

Justin M Shea

2017-06-13

An excellent approach to learning is to find an example from your textbook and then recreate it. Below are examples from every chapter and the syntax provided here should get you through most of the book.

Load the `wooldRidge` package to access data in the manner specified in each example.

```
library(wooldRidge)
```

Chapter 2: The Simple Regression Model

`Example 2.10`: A Log Wage Equation

> "Using the wage1 data as in Example 2.4, but using log(wage) as the dependent variable, we obtain the following relationship:"

$$\widehat{log(wage)} = \beta_0 + \beta_1 educ$$

First, load the `wage1` data.

```
data(wage1)
```

Next, estimate a linear relationship between the log of *wage* and *education*.

```
log_wage_model <- lm(lwage ~ educ, data = wage1)
```

Finally, print the coefficients and R^2.

```
log_wage_model$coefficients
```

```
## (Intercept)        educ
##  0.58377267  0.08274437
```

```
summary(log_wage_model)$r.squared
```

```
## [1] 0.1858065
```

Chapter 3: Multiple Regression Analysis: Estimation

`Example 3.2`: Hourly Wage Equation

> "Using the 526 observations on workers in 'wage1', we include *educ*(years of education), *exper*(years of labor market experience), and *tenure*(years with the current employer) in an equation explain log(*wage*)."

$$\widehat{log(wage)} = \beta_0 + \beta_1 educ + \beta_3 exper + \beta_4 tenure$$

Estimate the model regressing *education*, *experience*, and *tenure* against *log(wage)*.

```
hourly_wage_model <- lm(lwage ~ educ + exper + tenure, data = wage1)
```

Again, print the estimated model coefficients:

```
hourly_wage_model$coefficients
```

```
## (Intercept)        educ       exper      tenure
## 0.284359541 0.092028988 0.004121109 0.022067218
```

Chapter 4: Multiple Regression Analysis: Inference

`Example 4.7` Effect of Job Training on Firm Scrap Rates

From the text:

> "The scrap rate for a manufacturing firm is the number of defective items - products that must be discarded - out of every 100 produced. Thus, for a given number of items produced, a decrease in the scrap rate reflects higher worker productivity."
>
> "We can use the scrap rate to measure the effect of worker training on productivity. Using the data in jtrain, but only for the year 1987 and for nonunionized firms, we obtain the following estimated equation:"

First, load the `jtrain` data set.

```
data("jtrain")
```

Next, create a logical index identifying which observations occur in 1987 and are non-union.

```
index <- jtrain$year == 1987 & jtrain$union == 0
```

Next, subset the jtrain data by the new index. This returns a data.frame of `jtrain` data of non-union firms for the year 1987.

```
jtrain_1987_nonunion <- jtrain[index,]
```

Now create the linear model regressing hrsemp(total hours training/total employees trained), the log of annual sales, and the log of the number of the employees, against the log of the scrape rate.

$$lscrap = \alpha + \beta_1 hrsemp + \beta_2 lsales + \beta_3 lemploy$$

```
linear_model <- lm(lscrap ~ hrsemp + lsales + lemploy, data = jtrain_1987_nonunion)
```

Finally, print the complete summary statistic diagnostics of the model.

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = lscrap ~ hrsemp + lsales + lemploy, data = jtrain_1987_nonunion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6301 -0.7523 -0.4016  0.8697  2.8273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.45837    5.68677   2.191   0.0380 *
## hrsemp      -0.02927    0.02280  -1.283   0.2111
## lsales      -0.96203    0.45252  -2.126   0.0436 *
## lemploy      0.76147    0.40743   1.869   0.0734 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 25 degrees of freedom
##   (97 observations deleted due to missingness)
## Multiple R-squared:  0.2624, Adjusted R-squared:  0.1739
## F-statistic: 2.965 on 3 and 25 DF,  p-value: 0.05134
```

Chapter 5: Multiple Regression Analysis: OLS Asymptotics

`Example 5.3`: Economic Model of Crime

We illustrate the **Lagrange multiplier (LM) statistics** test by using a slight extension of the crime model from example 3.5.

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + \mu$$

narr86: number of times arrested, 1986.

pcnv: proportion of prior arrests leading to convictions.

avgsen: average sentence served, length in months.

tottime: time in prison since reaching the age of 18, length in months.

ptime86: months in prison during 1986

qemp86: quarters employed, 1986

Load the `crime1` data set containing arrests during the year 1986 and other information on 2,725 men born in either 1960 or 1961 in California.

```r
data(crime1)
```

We use the *LM* statistic to test the null hypothesis that *avgsen* and *tottime* have no effect on *narr86* once other factors have been controlled for. First, estimate the restricted model by regressing *narr86* on *pcnv, ptime86,* and *qemp86*; the variables *avgsen* and *tottime* are excluded from this regression.

```r
restricted_model <- lm(narr86 ~ pcnv + ptime86 + qemp86, data = crime1)
```

We obtain the residuals $\tilde{\mu}$ from this regression, 2,725 of them.

```r
restricted_model_u <- restricted_model$residuals
```

Next, we run the regression of:

$$\tilde{\mu} = \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86$$

As always, the order in which we list the independent variables is irrelevant.

```r
LM_u_model <- lm(restricted_model_u ~ pcnv + ptime86 + qemp86 + avgsen + tottime,
    data = crime1)
```

This second regression produces R^2_μ, which turns out to be about 0.0015.

```r
summary(LM_u_model)$r.squared
```

```
## [1] 0.001493846
```

This may seem small, but we must multiple it by n to get the LM statistic:

$$LM = 2,725(0.0015)$$

```
LM_test <- nobs(LM_u_model) * 0.0015
LM_test
```

```
## [1] 4.0875
```

The 10% critical value in a chi-square distribution with two degrees of freedom is about 4.61 (rounded to two decimal places).

```
qchisq(1 - 0.10, 2)
```

```
## [1] 4.60517
```

Thus, we fail to reject the null hypothesis that $\beta_{avgsen} = 0$ and $\beta_{tottime} = 0$ at the 10% level.

The p-value is:

$$P(X_2^2 > 4.09) \approx 0.129$$

so we would reject the H_0 at the 15% level.

```
1-pchisq(LM_test, 2)
```

```
## [1] 0.129542
```

Chapter 6: Multiple Regression: Further Issues

Example 6.1:' Effects of Pollution on Housing Prices, standardized.

We use the data `hrprice2` to illustrate the use of beta coefficients. Recall that the key independent variable is *nox*, a measure of nitrogen oxide in the air over each community. One way to understand the size of the pollution effect-without getting into the science underling nitrogen oxide's effect on air quality-is to compute beta coefficients.

The population equation is the level-level model:

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + \mu$$

price: median housing price.

nox: Nitrous Oxide concentration; parts per million.

crime: number of reported crimes per capita.

rooms: average number of rooms in houses in the community.

dist: weighted distance of the community to 5 employment centers.

stratio: average student-teacher ratio of schools in the community.

The beta coefficients are reported in the following equation (so each variable has been converted to its z-score):

$$\widehat{zprice} = \beta_1 znox + \beta_2 zcrime + \beta_3 zrooms + \beta_4 zdist + \beta_5 zstratio$$

First, load the `hrpice2` data.

```
data(hrpice2)
```

Next, estimate the coefficient with the usual `lm` regression model, but this time standardized coefficients by wraping each variable with R's `scale` function:

```
housing_standard <- lm(scale(price) ~ 0 + scale(nox) + scale(crime) + scale(rooms) +
    scale(dist) + scale(stratio), data = hprice2)

housing_standard$coefficients
```

```
##     scale(nox)   scale(crime)   scale(rooms)    scale(dist) scale(stratio)
##     -0.3404460     -0.1432828      0.5138878     -0.2348385     -0.2702799
```

`Example 6.2:` Effects of Pollution on Housing Prices, Quadratic Interactive Term

We modify the housing model, adding a quadratic term in *rooms*:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + \mu$$

```
housing_interactive <- lm(lprice ~ lnox + log(dist) + rooms+I(rooms^2) + stratio, data = hprice2)
summary(housing_interactive)
```

```
##
## Call:
## lm(formula = lprice ~ lnox + log(dist) + rooms + I(rooms^2) +
##     stratio, data = hprice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04285 -0.12774  0.02038  0.12650  1.25272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.385478   0.566473  23.630  < 2e-16 ***
## lnox        -0.901682   0.114687  -7.862 2.34e-14 ***
## log(dist)   -0.086781   0.043281  -2.005  0.04549 *
## rooms       -0.545113   0.165454  -3.295  0.00106 **
## I(rooms^2)   0.062261   0.012805   4.862 1.56e-06 ***
## stratio     -0.047590   0.005854  -8.129 3.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 500 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.5988
## F-statistic: 151.8 on 5 and 500 DF,  p-value: < 2.2e-16
```

Chapter 7: Multiple Regression Analysis with Qualitative Information