財金計量方法

2021/5/21 規劃

因應防疫期間，至學期末採遠距上課，這週安排課程與作業如下

1. 請同學閱讀 Quadratics python 例題解說影片。
   https://drive.google.com/file/d/1nI3fAY-MSkkgYH6MSNNTTevylTeeQ1xq/view?usp=sharing

2. 請同學閱讀逐步迴歸 python 解說影片。
   https://drive.google.com/file/d/1VfOF8040xlMN74fnVB4Ybj1H1UqRfmm1/view?usp=sharing

3. 在 Moodle 提供年報酬率換成日報酬率的簡單講義(針對之前 Fama 實作題目看到的問題)，請同學閱讀文件，並思考之前提交的作業是否需要修正的地方?
   - 有需要修正的同學，可以重新上傳到此作業區
   - Homework3--修正作業"以 python 試做 Fama 三因子模型"-上傳截止 5/25 5pm

4. 請各位同學實作之前 Wooldridge ch4 例題(example 4.2/4.5/4.8/4.9)，每一題須逐步完成以下方法。
   (1) 僅先用 R square 判斷方式，挑選應納入迴歸方程式的自變數，並且需要描述利用 R square 方式篩選的過程與結果，備註經由 R square 篩選方法，自己認哪一個為最合適的迴歸方程式，需解釋迴歸估計結果。
   (2) 利用 OVB、Multicollinearity、Quadratics 方式挑選應納入迴歸方程式的自變數，並描述利用此些方法篩選的過程與結果，並備註經由這些篩選方法，自己認為哪一個為最合適的迴歸方程式，需解釋迴歸估計結果。
   (3) 利用 Stepwise regression 方式，包含向前、向後、逐步挑選法，挑選應納入迴歸方程式的自變數，並描述利用此些方法篩選的過程與結果，備註經由逐步迴歸法，自己認哪一個為最合適的迴歸方程式，需解釋迴歸估計結果。
   - 請同學將此作業上傳至此作業區(需上傳 python code/樣本資料/若沒在 python code 內撰寫挑選過程與結果，可另傳一份 word 檔)
   - 檔名令為: [HW_4_學校_學號_系級_姓名]
   - Homework4-上傳截止 5/28 5pm

5. 請各位同學於 5/28 前完成此表單填寫。
   - https://forms.gle/m3nuo7SzpYYer4wb6

6. Wooldridge ch4 例題(example 4.2/4.5/4.8/4.9)相關資料整理如下:

- Wooldridge ch4 例題(example 4.2) –對應課本內容

EXAMPLE 4.2 **Student Performance and School Size**

There is much interest in the effect of school size on student performance. (See, for example, *The New York Times Magazine*, 5/28/95.) One claim is that, everything else being equal, students at smaller schools fare better than those at larger schools. This hypothesis is assumed to be true even after accounting for differences in class sizes across schools.

The file MEAP93 contains data on 408 high schools in Michigan for the year 1993. We can use these data to test the null hypothesis that school size has no effect on standardized test scores against the alternative that size has a negative effect. Performance is measured by the percentage of students receiving a passing score on the Michigan Educational Assessment Program (MEAP) standardized tenth-grade math test (*math*10). School size is measured by student enrollment (*enroll*). The null hypothesis is H_0: $\beta_{enroll} = 0$, and the alternative is H_1: $\beta_{enroll} < 0$. For now, we will control for two other factors, average annual teacher compensation (*totcomp*) and the number of staff per one thousand students (*staff*). Teacher compensation is a measure of teacher quality, and staff size is a rough measure of how much attention students receive.

The estimated equation, with standard errors in parentheses, is

$$\widehat{math10} = 2.274 + .00046\, totcomp + .048\, staff - .00020\, enroll$$
$$\qquad\qquad (6.113)\ (.00010)\qquad\quad (.040)\qquad (.00022)$$
$$n = 408,\ R^2 = .0541.$$

The coefficient on *enroll*, −.00020, is in accordance with the conjecture that larger schools hamper performance: higher enrollment leads to a lower percentage of students with a passing tenth-grade math score. (The coefficients on *totcomp* and *staff* also have the signs we expect.) The fact that *enroll* has an estimated coefficient different from zero could just be due to sampling error; to be convinced of an effect, we need to conduct a *t* test.

Since $n - k - 1 = 408 - 4 = 404$, we use the standard normal critical value. At the 5% level, the critical value is −1.65; the *t* statistic on *enroll* must be *less* than −1.65 to reject H_0 at the 5% level.

The *t* statistic on *enroll* is −.00020/.00022 ≈ −.91 which is larger than −1.65: we *fail* to reject H_0 in favor of H_1 at the 5% level. In fact, the 15% critical value is −1.04, and since −.91 > −1.04, we fail to reject H_0 even at the 15% level. We conclude that *enroll* is not statistically significant at the 15% level.

The variable *totcomp* is statistically significant even at the 1% significance level because its *t* statistic is 4.6. On the other hand, the *t* statistic for *staff* is 1.2, and so we cannot reject H_0: $\beta_{staff} = 0$ against H_1: $\beta_{staff} > 0$ even at the 10% significance level. (The critical value is $c = 1.28$ from the standard normal distribution.)

To illustrate how changing functional form can affect our conclusions, we also estimate the model with all independent variables in logarithmic form. This allows, for example, the school size effect to diminish as school size increases. The estimated equation is

$$\widehat{math10} = -207.66 + 21.16 \log(totcomp) + 3.98 \log(staff) - 1.29 \log(enroll)$$
$$\qquad\qquad (48.70)\ \ (4.06)\qquad\quad (4.19)\qquad\quad (0.69)$$
$$n = 408,\ R^2 = .0654.$$

The *t* statistic on log(*enroll*) is about −1.87; since this is below the 5% critical value −1.65, we reject H_0: $\beta_{\log(enroll)} = 0$ in favor of H_1: $\beta_{\log(enroll)} < 0$ at the 5% level.

In Chapter 2, we encountered a model where the dependent variable appeared in its original form (called *level* form), while the independent variable appeared in log form (called *level-log* model). The interpretation of the parameters is the same in the multiple regression context, except, of course, that we can give the parameters a ceteris paribus interpretation. Holding *totcomp* and *staff* fixed, we have $\Delta\widehat{math10} = -1.29[\Delta\log(enroll)]$, so that

$$\Delta\widehat{math10} \approx -(1.29/100)(\%\Delta enroll) \approx -.013(\%\Delta enroll).$$

Once again, we have used the fact that the change in log(*enroll*), when multiplied by 100, is approximately the percentage change in *enroll*. Thus, if enrollment is 10% higher at a school, $\widehat{math10}$ is predicted to be $.013(10) = 0.13$ percentage points lower (*math*10 is measured as a percentage).

Which model do we prefer: the one using the level of *enroll* or the one using log(*enroll*)? In the level-level model, enrollment does not have a statistically significant effect, but in the level-log model it does. This translates into a higher *R*-squared for the level-log model, which means we explain more of the variation in *math*10 by using *enroll* in logarithmic form (6.5% to 5.4%). The level-log model is preferred because it more closely captures the relationship between *math*10 and *enroll*. We will say more about using *R*-squared to choose functional form in Chapter 6.

- Wooldridge ch4 例題(example 4.2) –樣本變數

Example 4.2 - MEAP93

- A data.frame with 408 observations on 17 variables:
- **lnchprg:** perc of studs in sch lnch prog
- **enroll:** school enrollment
- **staff:** staff per 1000 students
- **expend:** expend. per stud, $
- **salary:** avg. teacher salary, $
- **benefits:** avg. teacher benefits, $
- **droprate:** school dropout rate, perc
- **gradrate:** school graduation rate, perc
- **math10:** perc studs passing MEAP math
- **sci11:** perc studs passing MEAP science
- **totcomp:** salary + benefits
- **ltotcomp:** log(totcomp)
- **lexpend:** log of expend
- **lenroll:** log(enroll)
- **lstaff:** log(staff)
- **bensal:** benefits/salary
- **lsalary:** log(salary)

- Wooldridge ch4 例題(example 4.5) –對應課本內容

EXAMPLE 4.5 **Housing Prices and Air Pollution**

For a sample of 506 communities in the Boston area, we estimate a model relating median housing price (*price*) in the community to various community characteristics: *nox* is the amount of nitrogen oxide in the air, in parts per million; *dist* is a weighted distance of the community from five employment centers, in miles; *rooms* is the average number of rooms in houses in the community; and *stratio* is the average student-teacher ratio of schools in the community. The population model is

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 stratio + u.$$

Thus, β_1 is the elasticity of *price* with respect to *nox*. We wish to test $H_0: \beta_1 = -1$ against the alternative $H_1: \beta_1 \neq -1$. The t statistic for doing this test is $t = (\hat{\beta}_1 + 1)/\text{se}(\hat{\beta}_1)$.

Using the data in HPRICE2, the estimated model is

$$\widehat{\log(price)} = \underset{(0.32)}{11.08} - \underset{(.117)}{.954}\log(nox) - \underset{(.043)}{.134}\log(dist) + \underset{(.019)}{.255}rooms - \underset{(.006)}{.052}stratio$$

$$n = 506, R^2 = .581.$$

The slope estimates all have the anticipated signs. Each coefficient is statistically different from zero at very small significance levels, including the coefficient on log(*nox*). But we do not want to test that $\beta_1 = 0$. The null hypothesis of interest is $H_0: \beta_1 = -1$, with corresponding t statistic $(-.954 + 1)/.117 = .393$. There is little need to look in the t table for a critical value when the t statistic is this small: the estimated elasticity is not statistically different from -1 even at very large significance levels. Controlling for the factors we have included, there is little evidence that the elasticity is different from -1.

- Wooldridge ch4 例題(example 4.5) –樣本變數

Example 4.5 - HPRICE2

- A data.frame with 506 observations on 12 variables:
- **price:** median housing price, $
- **crime:** crimes committed per capita
- **nox:** nit ox concen; parts per 100m
- **rooms:** avg number of rooms
- **dist:** wght dist to 5 employ centers
- **radial:** access. index to rad. hghwys
- **proptax:** property tax per $1000
- **stratio:** average student-teacher ratio
- **lowstat:** perc of people 'lower status'
- **lprice:** log(price)
- **lnox:** log(nox)
- **lproptax:** log(proptax)

- Wooldridge ch4 例題(example 4.8) –對應課本內容

EXAMPLE 4.8 **Model of R&D Expenditures**

Economists studying industrial organization are interested in the relationship between firm size—often measured by annual sales—and spending on research and development (R&D). Typically, a constant elasticity model is used. One might also be interested in the ceteris paribus effect of the profit margin—that is, profits as a percentage of sales—on R&D spending. Using the data in RDCHEM on 32 U.S. firms in the chemical industry, we estimate the following equation (with standard errors in parentheses below the coefficients):

$$\widehat{\log(rd)} = -4.38 + 1.084\log(sales) + .0217\,profmarg$$
$$\qquad\qquad (.47)\quad (.060)\qquad\qquad (.0128)$$
$$n = 32,\ R^2 = .918.$$

The estimated elasticity of R&D spending with respect to firm sales is 1.084, so that, holding profit margin fixed, a 1% increase in sales is associated with a 1.084% increase in R&D spending. (Incidentally, R&D and sales are both measured in millions of dollars, but their units of measurement have no effect on the elasticity estimate.) We can construct a 95% confidence interval for the sales elasticity once we note that the estimated model has $n - k - 1 = 32 - 2 - 1 = 29$ degrees of freedom. From Table G.2, we find the 97.5th percentile in a t_{29} distribution: $c = 2.045$. Thus, the 95% confidence interval for $\beta_{\log(sales)}$ is $1.084 \pm .060(2.045)$, or about (.961,1.21). That zero is well outside this interval is hardly surprising: we expect R&D spending to increase with firm size. More interesting is that unity is included in the 95% confidence interval for $\beta_{\log(sales)}$, which means that we cannot reject H_0: $\beta_{\log(sales)} = 1$ against H_1: $\beta_{\log(sales)} \neq 1$ at the 5% significance level. In other words, the estimated R&D-sales elasticity is not statistically different from 1 at the 5% level. (The estimate is not practically different from 1, either.)

The estimated coefficient on *profmarg* is also positive, and the 95% confidence interval for the population parameter, $\beta_{profmarg}$, is $.0217 \pm .0128(2.045)$, or about (−.0045,.0479). In this case, zero is included in the 95% confidence interval, so we fail to reject H_0: $\beta_{profmarg} = 0$ against H_1: $\beta_{profmarg} \neq 0$ at the 5% level. Nevertheless, the t statistic is about 1.70, which gives a two-sided p-value of about .10, and so we would conclude that *profmarg* is statistically significant at the 10% level against the two-sided alternative, or at the 5% level against the one-sided alternative H_1: $\beta_{profmarg} > 0$. Plus, the economic size of the profit margin coefficient is not trivial: holding *sales* fixed, a one percentage point increase in *profmarg* is estimated to increase R&D spending by $100(.0217) \approx 2.2\%$. A complete analysis of this example goes beyond simply stating whether a particular value, zero in this case, is or is not in the 95% confidence interval.

- Wooldridge ch4 例題(example 4.8) –樣本變數

Example 4.8 - RDCHEM

- A data.frame with 32 observations on 8 variables:
- **rd:** R&D spending, millions
- **sales:** firm sales, millions
- **profits:** profits, millions
- **rdintens:** rd as percent of sales
- **profmarg:** profits as percent of sales
- **salessq:** sales^2
- **lsales:** log(sales)
- **lrd:** log(rd)

- Wooldridge ch4 例題(example 4.9) –對應課本內容

EXAMPLE 4.9 **Parents' Education in a Birth Weight Equation**

As another example of computing an F statistic, consider the following model to explain child birth weight in terms of various factors:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u, \quad [4.42]$$

where

$bwght$ = birth weight, in pounds.
$cigs$ = average number of cigarettes the mother smoked per day during pregnancy.
$parity$ = the birth order of this child.
$faminc$ = annual family income.
$motheduc$ = years of schooling for the mother.
$fatheduc$ = years of schooling for the father.

Let us test the null hypothesis that, after controlling for *cigs*, *parity*, and *faminc*, parents' education has no effect on birth weight. This is stated as H_0: $\beta_4 = 0$, $\beta_5 = 0$, and so there are $q = 2$ exclusion restrictions to be tested. There are $k + 1 = 6$ parameters in the unrestricted model (4.42); so the *df* in the unrestricted model is $n - 6$, where n is the sample size.

We will test this hypothesis using the data in BWGHT. This data set contains information on 1,388 births, but we must be careful in counting the observations used in testing the null hypothesis. It turns out that information on at least one of the variables *motheduc* and *fatheduc* is missing for 197 births in the sample; these observations cannot be included when estimating the unrestricted model. Thus, we really have $n = 1{,}191$ observations, and so there are $1{,}191 - 6 = 1{,}185$ *df* in the unrestricted model. We must be sure to use these *same* 1,191 observations when estimating the restricted model (not the full 1,388 observations that are available). Generally, when estimating the restricted model to compute an F test, we must use the same observations to estimate the unrestricted model; otherwise, the test is not valid. When there are no missing data, this will not be an issue.

The numerator *df* is 2, and the denominator *df* is 1,185; from Table G.3, the 5% critical value is $c = 3.0$. Rather than report the complete results, for brevity, we present only the R-squareds. The R-squared for the full model turns out to be $R^2_{ur} = .0387$. When *motheduc* and *fatheduc* are dropped from the regression, the R-squared falls to $R^2_r = .0364$. Thus, the F statistic is $F = [(.0387 - .0364)/(1 - .0387)](1{,}185/2) = 1.42$; since this is well below the 5% critical value, we fail to reject H_0. In other words, *motheduc* and *fatheduc* are jointly insignificant in the birth weight equation. Most statistical packages these days have built-in commands for testing multiple hypotheses after OLS estimation, and so one need not worry about making the mistake of running the two regressions on different data sets. Typically, the commands are applied after estimation of the unrestricted model, which means the smaller subset of data is used whenever there are missing values on some variables. Formulas for computing the F statistic using matrix algebra—see Appendix E—do not require estimation of the retricted model.

- Wooldridge ch4 例題(example 4.9) –樣本變數

Example 4.9 - BWGHT

- A data.frame with 1388 observations on 14 variables:
- **faminc:** 1988 family income, $1000s
- **cigtax:** cig. tax in home state, 1988
- **cigprice:** cig. price in home state, 1988
- **bwght:** birth weight, ounces
- **fatheduc:** father's yrs of educ
- **motheduc:** mother's yrs of educ
- **parity:** birth order of child
- **male:** =1 if male child
- **white:** =1 if white
- **cigs:** cigs smked per day while preg
- **lbwght:** log of bwght
- **bwghtlbs:** birth weight, pounds
- **packs:** packs smked per day while preg
- **lfaminc:** log(faminc)