ECOM20001: Econometrics 1

Tutorial 9: Polynomial Regression

A. Getting Started

Please create a Tutorial9 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Tutorial9 folder:

- tute9.R
- tute9_cps.csv

The first file is the R code for tutorial 9, the second file is the .csv file that contains the dataset for the tutorial.[1] The dataset has the following 5 variables:

- year: year individual was randomly surveyed; either 1992 or 2012
- ahe: individual's average hourly earnings (in real terms, 2012=100)
- bachelor: equals 1 if individual has a bachelor degree, 0 otherwise
- female: equals 1 if individual is female, 0 otherwise
- age: age of the individual at time of survey

In total, the dataset contains this information for 15,052 individuals in the U.S.

B. Go to the Code

With the R file downloaded into your Tutorial9 folder, you are ready to proceed with the tutorial. Please go to the tute9.R file to continue with the tutorial.

[1] The reference for these data is the Current Population Survey (CPS) which is collected by the U.S. Department of Labor Statistics and provides individual-level data on the population, employment, and earnings. It is constructed from randomly sampling the U.S. population. For details, see https://www.census.gov/programs-surveys/cps.html

C. Questions

Having worked through the tute9.R code and graphs, please answer the following:

1. Using either the ggplot()[2] command in RStudio, create a scatter plot with age on the horizontal axis and ahe on the vertical axis. Note that the ggplot() code in the tute9.R file overlays predicted values from quadratic or cubic regressions with ahe as the dependent variable and age as the independent variable, which helps with interpreting the results.
   - Does there appear to be a nonlinear relationship between age and ahe?
   - Provide an economic explanation for the relationship you find
2. Using sequential hypothesis testing (see slide 31 of lecture note 8) estimate 4 polynomial regressions, where ahe is the dependent variable in each regression, and where the sets of regressors in each respective regression are:
   - age, age^2, age^3, age^4, bachelor, female
   - age, age^2, age^3, bachelor, female
   - age, age^2, bachelor, female
   - age, bachelor, female

Report heteroskedasticity-robust standard errors for each regression above, as well as for every regression used for the remainder of the tutorial. Please answer the following questions:

- Using the sequential hypothesis testing method, which is the preferred polynomial regression model?
- Do any of the polynomial regressions appear to suffer from imperfect multicollinearity?
- Report the overall regression F-statistic for your preferred regression model and interpret the joint test result.
- Which test result confirms, statistically, that there is indeed a nonlinear relationship between ahe and age, holding fixed education and gender?

[2] R is known for making very nice graphs using ggplot(). So we are taking the opportunity in this tutorial to learn this great graphs-building package. The comments in the tute9.R code describe step-by-step how to install ggplot2, the package in R for using the ggplot() command. It is the exact same steps as you took to install the AER package for using the coeftest() command except you install the ggplot2 package.

3. Run another nonlinear regression with the following set of regressors, without any other control variables:

   - age, age^2

Compare your regression results to the results that you found in question 2 where you estimate a quadratic regression model where bachelor and females are included as controls. Is it important to include bachelor and female as control variables in estimating the nonlinear relationship between ahe and age.

4. Using an estimated quadratic regression model with ahe as the dependent variable and age, age^2, bachelor, and female as regressors, interpret two separate partial effects of increasing age from 25 to 28, and from 28 to 31, using only the estimated regression coefficient on age, and ignoring the regression coefficient on age^2. Are the two (incorrect) partial effects the same?

5. Using the quadratic regression model you estimated in question 4, compute the nonlinear partial effects on ahe from:

   - increasing age from 25 to 28
   - increasing age from 28 to 31
   - increasing age from 31 to 35

Use the general approach described in slides 16 to 18 of lecture note 8 for your calculations.

   - Contrast your partial effects from increasing age from 25 to 28 and from 28 to 31 to the (incorrect) partial effects you computed in question 4. Why do your results differ in questions 4. and 5.?
   - Interpret and contrast each of the 3 (correct) partial effects you have computed. Are they larger or smaller in magnitude with higher levels of age?

6. Compute the standard errors and 95% confidence intervals (CIs) for each of the three partial effects you computed in question 5. Use the general approach described in slides 19-21 of lecture note 8 for your calculations.

   - Report the standard errors and briefly interpret each of the 95% CIs
   - Are the 95% CIs the same width around each of the three partial effects that you calculated?

See the next page for the derivation of the joint hypothesis tests that you are required to run for computing the standard errors and 95% CIs for each of the three partial effects computed in question 5.

Supplemental: Derivations of Joint Hypothesis Tests for Computing Standard Errors of Nonlinear Partial Effects in Question 6.

For details on where these derivations fit into the procedure for computing standard errors for nonlinear partial effects, see lecture note 8, slides 16-24.

Regression model used in questions 5 and 6:

$$ahe_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 bachelor_i + \beta_4 female_i + u_i$$

Nonlinear partial effect on ahe from increasing age from **25 to 28**, and joint null to test as part of computing the standard error for this partial effect:

$$\begin{aligned}
\Delta\widehat{ahe}_i &= \left(\hat{\beta}_0 + \hat{\beta}_1 28 + \hat{\beta}_2 28^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 25 + \hat{\beta}_2 25^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) \\
&= \hat{\beta}_1 28 + \hat{\beta}_2 28^2 - \hat{\beta}_1 25 - \hat{\beta}_2 25^2 \\
&= \hat{\beta}_1 28 + \hat{\beta}_2 784 - \hat{\beta}_1 25 - \hat{\beta}_2 625 \\
&= 3\hat{\beta}_1 + 159\hat{\beta}_2 \\
&\Rightarrow \text{Test joint null that } 3\hat{\beta}_1 + 159\hat{\beta}_2 = 0, \text{obtain corresponding } F\text{-statistic from the test, } F \\
&\Rightarrow \text{compute } SE(\Delta\widehat{ahe}_i) = \frac{|\Delta\widehat{ahe}_i|}{\sqrt{F}}
\end{aligned}$$

Nonlinear partial effect on ahe from increasing age from **28 to 31,** and joint null to test as part of computing the standard error for this partial effect:

$$\begin{aligned}
\Delta\widehat{ahe}_i &= \left(\hat{\beta}_0 + \hat{\beta}_1 31 + \hat{\beta}_2 31^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 28 + \hat{\beta}_2 28^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) \\
&= \hat{\beta}_1 31 + \hat{\beta}_2 31^2 - \hat{\beta}_1 28 - \hat{\beta}_2 28^2 \\
&= \hat{\beta}_1 31 + \hat{\beta}_2 961 - \hat{\beta}_1 28 - \hat{\beta}_2 784 \\
&= 3\hat{\beta}_1 + 177\hat{\beta}_2 \\
&\Rightarrow \text{Test joint null that } 3\hat{\beta}_1 + 177\hat{\beta}_2 = 0, \text{obtain corresponding } F\text{-statistic from the test, } F \\
&\Rightarrow \text{compute } SE(\Delta\widehat{ahe}_i) = \frac{|\Delta\widehat{ahe}_i|}{\sqrt{F}}
\end{aligned}$$

Nonlinear partial effect on ahe from increasing age from **31 to 35**, and joint null to test as part of computing the standard error for this partial effect:

$$\begin{aligned}
\Delta\widehat{ahe}_i &= \left(\hat{\beta}_0 + \hat{\beta}_1 35 + \hat{\beta}_2 35^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 31 + \hat{\beta}_2 31^2 + \hat{\beta}_3 bachelor_i + \hat{\beta}_4 female_i\right) \\
&= \hat{\beta}_1 35 + \hat{\beta}_2 35^2 - \hat{\beta}_1 31 - \hat{\beta}_2 31^2 \\
&= \hat{\beta}_1 35 + \hat{\beta}_2 1225 - \hat{\beta}_1 961 - \hat{\beta}_2 784 \\
&= 4\hat{\beta}_1 + 264\hat{\beta}_2 \\
&\Rightarrow \text{Test joint null that } 4\hat{\beta}_1 + 264\hat{\beta}_2 = 0, \text{obtain corresponding } F\text{-statistic from the test, } F \\
&\Rightarrow \text{compute } SE(\Delta\widehat{ahe}_i) = \frac{|\Delta\widehat{ahe}_i|}{\sqrt{F}}
\end{aligned}$$