Multiple Regression: Multicollinearity

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

Problems and Applications

Consider the regression of average hourly earnings AHE (in dollars) on Age (in years) and several binary variables for characteristics such as sex, education, and region of employment:

$$\widehat{AHE} = 0.33 + 10.42\,College - 4.57\,Female + 0.61\,Age$$
$$+ 0.74\,Northeast - 1.54\,Midwest - 0.44\,South$$
$$R^2 = 0.185, \quad SER = 12.01, \quad n = 7178$$

a. Do there appear to be important regional differences?

b. Why is the regressor $West$ omitted from the regression? What would happen if it were included?

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Problems and Applications

Consider the regression of average hourly earnings AHE (in dollars) on Age (in years) and several binary variables for characteristics such as sex, education, and region of employment:

$$\begin{aligned}\widehat{AHE} &= 0.33 + 10.42\,College - 4.57\,Female + 0.61\,Age \\ &\quad + 0.74\,Northeast - 1.54\,Midwest - 0.44\,South \\ R^2 &= 0.185, \quad SER = 12.01, \quad n = 7178\end{aligned}$$

a. Do there appear to be important regional differences?

b. Why is the regressor $West$ omitted from the regression? What would happen if it were included?

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Problems and Applications

Consider the regression of average hourly earnings AHE (in dollars) on Age (in years) and several binary variables for characteristics such as sex, education, and region of employment:

$$\begin{aligned}\widehat{AHE} &= 0.33 + 10.42\,College - 4.57\,Female + 0.61\,Age \\ &\quad + 0.74\,Northeast - 1.54\,Midwest - 0.44\,South \\ R^2 &= 0.185, \quad SER = 12.01, \quad n = 7178\end{aligned}$$

a. Do there appear to be important regional differences?

b. Why is the regressor $West$ omitted from the regression? What would happen if it were included?

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Problems and Applications

a. Do there appear to be important regional differences?

Since the variables for $West$ is omitted from the regression, it is the reference group to which the other regional variables can be compared to. On average, and controlling for other variables in the regression, workers in the $Northeast$ earn $\$0.74$ more per hour than workers in the $West$; while workers in the $Midwest$ earn $\$1.54$ less than workers in the West; and workers in the $South$ earn $\$0.44$ less than workers in the $West$.

Problems and Applications

a. Do there appear to be important regional differences?

Since the variables for $West$ is omitted from the regression, it is the reference group to which the other regional variables can be compared to. On average, and controlling for other variables in the regression, workers in the $Northeast$ earn $\$0.74$ more per hour than workers in the $West$; while workers in the $Midwest$ earn $\$1.54$ less than workers in the West; and workers in the $South$ earn $\$0.44$ less than workers in the $West$.

Problems and Applications

b. Why is the regressor $West$ omitted from the regression? What would happen if it were included?

The regressor $West$ is omitted to avoid perfect multicollinearity. Perfect multicollinearity would arise because the data is divided into exactly 4 groups: $West$, $Midwest$, $Northeast$, and $South$. Since the 4 categories are exchaustive and mutually exclusive, by construction, they add up to 1 for every observation in the dataset. This is known as the "dummy variable trap". Perfect multicollinearity among regressors is usually easy to detect. Some software will produce an error, others will drop one of the perfectly multicollinear regressors and issue a warning. Imperfect multicollinearity is another issue, much less easy to deal with.

b. Why is the regressor $West$ omitted from the regression? What would happen if it were included?

The regressor $West$ is omitted to avoid perfect multicollinearity. Perfect multicollinearity would arise because the data is divided into exactly 4 groups: $West$, $Midwest$, $Northeast$, and $South$. Since the 4 categories are exchaustive and mutually exclusive, by construction, they add up to 1 for every observation in the dataset. This is known as the "dummy variable trap". Perfect multicollinearity among regressors is usually easy to detect. Some software will produce an error, others will drop one of the perfectly multicollinear regressors and issue a warning. Imperfect multicollinearity is another issue, much less easy to deal with.

Problems and Applications

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

The expected difference in earnings between Juanita and Jennifer is:

$$\begin{aligned} & AHE_{\text{Juanita}} - AHE_{\text{Jennifer}} \\ & = (AHE|College = 1, Female = 1, Age = 28, \\ & \qquad Northeast = 0, Midwest = 0, South = 1) \\ & - (AHE|College = 1, Female = 1, Age = 28, \\ & \qquad Northeast = 0, Midwest = 1, South = 0) \\ & = (-0.44) - (-1.54) \\ & = +1.10 \end{aligned}$$

The expected difference in earnings between Juanita and Jennifer, based on the information used in the regression, is $\$1.10$ per hour.

c. Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

The expected difference in earnings between Juanita and Jennifer is:

$$\begin{aligned} & AHE_{\text{Juanita}} - AHE_{\text{Jennifer}} \\ & \quad = (AHE | College = 1, Female = 1, Age = 28, \\ & \qquad\qquad Northeast = 0, Midwest = 0, South = 1) \\ & \quad - (AHE | College = 1, Female = 1, Age = 28, \\ & \qquad\qquad Northeast = 0, Midwest = 1, South = 0) \\ & \quad = (-0.44) - (-1.54) \\ & \quad = +1.10 \end{aligned}$$

The expected difference in earnings between Juanita and Jennifer, based on the information used in the regression, is $\$1.10$ per hour.