Introductory Econometrics

Lecture 6: Multiple Regression, Model Selection

Alex Alekseev

November 21, 2022

University of Regensburg, Department of Economics

Previously on *Introductory Econometrics*...

Regression anatomy formula

$$\hat{\beta}_j = \frac{\widehat{Cov}(Y, U_j)}{\widehat{Var}(U_j)},$$

where U_j is the error term from the regression of X_j on all other predictors:

$$X_j = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \ldots + \gamma_k X_k + U_j.$$

Previously on *Introductory Econometrics*...

Omitted variables bias formula

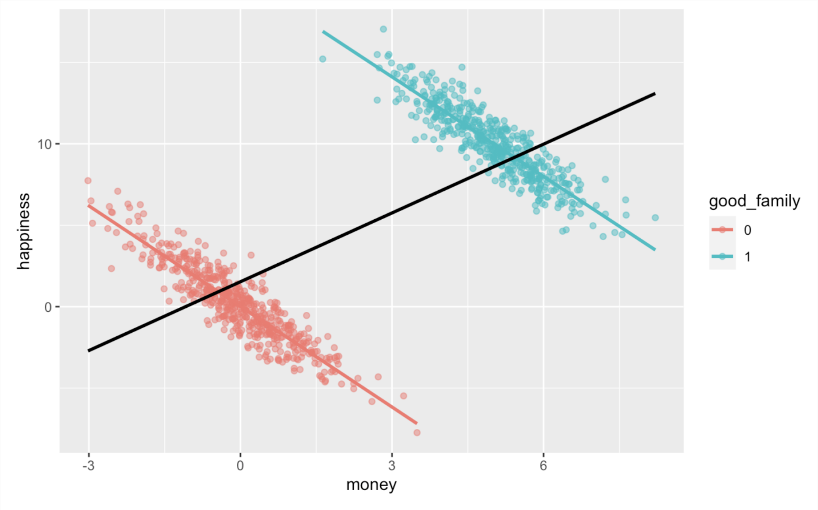
The expectation of the estimate of the slope coefficient from a simple regression of Y on Z is

$$\mathbb{E}\left[\frac{\widehat{Cov}(Y, Z)}{\widehat{Var}(Z)} \mid Z, \mathbf{x}\right] = \gamma + \beta_1\rho_1 + \ldots + \beta_k\rho_k,$$

where ρ_j are the slope coefficients from simple regressions of X_j on Z

OVB and the Simpson paradox

- Suppose we are studying the causal effect of money on happiness
- Let's assume that money has a truly negative effect on happiness
- Let's assume that we also have a third variable: whether a person is from a "good" family
- This variable will affect both the amount of money a person has (people from good families have more money than people from not so good families) and a person's happiness (people from good families are happier on average)
- Now we will illustrate graphically what happens when you estimate the "naive," unconditional effect of money on happiness and when you condition on the family background



happiness
10
0
-3
0
3
6
money
good_family
0
1

Variance of OLS

Assumptions

1. **Linear CEF** The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

2. **Random Sampling** Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^n$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed.
3. **No Perfect Collinearity** None of the predictors is a linear combination of other predictors.
4. **Homoskedasticity**

$$Var(u_i \mid x_{i1}, \ldots, x_{ik}) = \sigma^2, \quad i = 1, \ldots, n$$

Gauss-Markov Assumptions

Covariance of error terms

- Note that the **Random Sampling** assumptions implies that the error terms for each observation are uncorrelated

$$Cov(u_i, u_j \mid \mathbf{X}) = Cov(u_i, u_j) = 0, \quad i \neq j$$

Covariance matrix (aka variance-covariance matrix)

- If you have vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

  then its covariance matrix is

$$Var(\mathbf{x}) = \begin{pmatrix} Cov(x_1, x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(x_n, x_1) & Cov(x_n, x_2) & \dots & Cov(x_n, x_n) \end{pmatrix}$$

- Notice that its diagonal elements are

$$Var(x_1), \dots, Var(x_n)$$

Covariance matrix of the error term

- The covariance matrix of the error term is then

$$
\begin{aligned}
Var(\mathbf{u} \mid \mathbf{X}) &= \begin{pmatrix}
Var(u_1 \mid \mathbf{X}) & Cov(u_1, u_2 \mid \mathbf{X}) & \dots & Cov(u_1, u_n \mid \mathbf{X})) \\
Cov(u_2, u_1 \mid \mathbf{X}) & Var(u_2 \mid \mathbf{X}) & \dots & Cov(u_2, u_n \mid \mathbf{X})) \\
\vdots & \vdots & \vdots & \vdots \\
Cov(u_n, u_1 \mid \mathbf{X}) & Cov(u_n, u_2 \mid \mathbf{X})) & \dots & Var(u_n \mid \mathbf{X})
\end{pmatrix} \\
&= \begin{pmatrix}
\sigma^2 & 0 & \dots & 0 \\
0 & \sigma^2 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & \sigma^2
\end{pmatrix} \\
&= \sigma^2 \mathbf{I}
\end{aligned}
$$

- where $\mathbf{I}$ is the **identity** matrix

Variance of individual coefficients

- Under the **Gauss-Markov** assumptions

$$Var(\hat{\beta}_j \mid \mathbf{X}) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

where SST_j is the total sum of squares for the j-th predictor

$$SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

and R_j^2 is the R-squared from the auxilliary regression of X_j on all other predictors

$$X_j = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{j-1}X_{j-1} + \gamma_{j+1}X_{j+1} + \ldots + \gamma_k X_k + U_j.$$

Variance of individual coefficients explained

- The variance of coefficient j

$$Var(\hat{\beta}_j \mid \mathbf{X}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

- increases in σ^2 (not affected by the sample size)
- decreases in SST_j (i.e., we want to have as much variation in X_j as possible)
- increases in R_j^2 (if X_j is explained well by other predictors, then its variance will be high)

Variance inflation factors

- Write the variance of coefficient j as

$$Var(\hat{\beta}_j \mid \mathbf{X}) = \frac{\sigma^2}{SST_j} \frac{1}{1 - R_j^2}$$

- The term on the right

$$\frac{1}{1 - R_j^2}$$

  is called the **variance inflation factor** (VIF)

- It is used in practice to assess whether there are multicollinearity issues in the data
- The smallest possible value of VIF is 1
- Values that exceed 5 or 10 indicate the presence of multicollinearity

Special cases

- $R_j^2 = 0$: ideal case, X_j is uncorrelated with any other predictors
- $R_j^2 = 1$: perfect collinearity
- $0 < R_j^2 < 1$: typical case, multicollinearity

Estimation of the error variance

- The unbiased estimate of the error variance σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n - (k+1)}$$

- Then the estimate of standard deviation of β_j is

$$\hat{sd}(\hat{\beta}_j \mid X) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}$$

- This term is called the **standard error** of $\hat{\beta}_j$

Covariance matrix of the OLS

- The covariance between individual estimates is

$$Cov(\hat{\beta}_j, \hat{\beta}_l \mid \mathbf{X}) = \mathbb{E}[(\hat{\beta}_j - \beta_j)(\hat{\beta}_l - \beta_l) \mid \mathbf{X}]$$

- Stacking all the terms together, we get the **covariance matrix** of $\hat{\beta}$

$$Var(\hat{\beta} \mid \mathbf{X}) = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid \mathbf{X}]$$

The covariance matrix of the OLS estimator

$$Var(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Proof

- Start with the OLS formula and substitute for **y**

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

- Then

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Proof

- Plug this expression into the formula for the covariance matrix

$$
\begin{aligned}
\mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)' \mid \mathbf{X}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})' \mid \mathbf{X}] \\
&= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

Efficiency of OLS

Recall the assumptions

1. **Linear CEF** The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

2. **Random Sampling** Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^{n}$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed.
3. **No Perfect Collinearity** None of the predictors is a linear combination of other predictors.
4. **Homoskedasticity**

$$Var(u_i \mid x_{i1}, \ldots, x_{ik}) = \sigma^2, \quad i = 1, \ldots, n$$

Gauss-Markov Assumptions

OLS: unbiased and linear

- We have already established that OLS is **unbiased** under our assumptions
- It is also a **linear** estimator of the coefficients, in the sense that it can be written as a linear function of the outcome variable

OLS: unbiased and linear

- We have already established that OLS is **unbiased** under our assumptions
- It is also a **linear** estimator of the coefficients, in the sense that it can be written as a linear function of the outcome variable

Note

The linearity of the **estimator** is different from the linearity of the **model**

OLS: unbiased and linear

- We have already established that OLS is **unbiased** under our assumptions
- It is also a **linear** estimator of the coefficients, in the sense that it can be written as a linear function of the outcome variable
- Recall the regression anatomy formula

$$\hat{\beta}_j = \frac{\widehat{Cov}(Y, U_j)}{\widehat{Var}(U_j)} = \frac{\sum_{i=1}^{n} \hat{u}_{ij} y_i}{\sum_{i=1}^{n} \hat{u}_{ij}^2} = \sum_{i=1}^{n} w_{ij} y_i,$$

where

$$w_{ij} \equiv \frac{\hat{u}_{ij}}{\sum_{i=1}^{n} \hat{u}_{ij}^2}$$

- Each $\hat{\beta}_j$ is a linear function of y_i

Gauss-Markov theorem

Gauss-Markov theorem

Under the Gauss-Markov assumptions, the OLS estimator is the best linear unbiased estimator (BLUE)

- The **best** in the theorem means that OLS has the smallest variance among any other linear unbiased estimators $\tilde{\beta}_j$

$$Var(\hat{\beta}_j) \leqslant Var(\tilde{\beta}_j), \quad \text{for all } j$$

Model Selection Criteria

Goal of model selection

- In principle: find the population model
- In practice: find the "best" model for the purpose of the analysis
- More specific: Under the assumption that the population model is a multiple linear regression model find all regressors that are included in the regression and their appropriate transformations. Avoid omitting variables and including irrelevant variables

A brief theory of model selection

- There are two issues
  - variable (model) choice
  - estimation variance
- For the first issue, we choose an objective function to evaluate different models
- A popular objective function is the mean squared error (MSE) or root mean squared error (RMSE)

$$RMSE \equiv \sqrt{\frac{SSR}{n}}$$

- In-sample vs. out-of-sample

RMSE decomposition

- If the model includes all the relevant variables, the population model is a multiple linear regression, and MSE is minimized with respect to the parameters, then

$$MSE = Var(U) = \sigma^2$$

- If some relevant variables are missing, it can be shown that the *MSE* can be decomposed into variance and a squared bias
- For simplicity, suppose there are k relevant predictors but we use just one

$$Y = \beta_0 + \beta_1 X_1 + U$$

RMSE decomposition

- Then

$$MSE_1 = \sigma^2 + \mathbb{E}[(\mathbb{E}[Y \mid X_1, \ldots, X_k] - \mathbb{E}[Y \mid X_1])^2]$$

- The first term represents the deviation of the observed outcome from its conditional expectation in the true population model
- The second term captures the deviation of the conditional expectation of the **true** model from the conditional expectation of the **misspecified** model

MSE and estimation

- If parameters have to estimated, we have to add another term to MSE

$$
\begin{aligned}
MSE = &\ \text{variance of population model} \\
&+ \text{squared bias} \\
&+ \text{estimation variance}
\end{aligned}
$$

- The estimation variance in general increases with the number of variables
- It might be that minimizing MSE means choosing a model that omits some variables

Bias-variance trade-off

- When we start adding predictors to the model, a natural question arises of whether these extra predictors add value
- We run into an issue of how to compare different models
- Including more predictors typically improves a model's fit (reduces **variance**)
- However, the fit could improve even if predictors are truly irrelevant
- Hence more predictors could increase the **bias** of the model
- When the bias is high, taking our model to a new data set could lead to bad performance.

Model selection criteria

- One way to balance this trade-off is to use **model selection criteria**
- The two most popular model selection criteria are the Akaike Information Criterion (AIC) and the Bayesian (Schwartz) Information Criterion (BIC)
- The AIC formula is given by

$$\text{AIC} = 2k + n \ln \frac{SSR}{n} + n \ln(2\pi) + n$$

  where k is the number of estimated coefficients, n is the number of observations, and SSR is the residual sum of squares

Model selection criteria

- One way to balance this trade-off is to use **model selection criteria**
- The two most popular model selection criteria are the Akaike Information Criterion (AIC) and the Bayesian (Schwartz) Information Criterion (BIC)
- The AIC formula is given by

$$\text{AIC} = 2k + n \ln \frac{SSR}{n} + n \ln(2\pi) + n$$

  where k is the number of estimated coefficients, n is the number of observations, and SSR is the residual sum of squares

Note

Some formulas use the estimated number of coefficients + 1 for k when computing the AIC for the OLS. The function AIC in R does that.

Trade-off

$$\text{AIC} = 2k + n \ln \frac{SSR}{n} + n \ln(2\pi) + n$$

- When a new predictor is added to the model (k increases)
  - the first term $2k$ increases
  - the second term $n \ln \frac{SSR}{n}$ (weakly) decreases
  - the third term $n \ln(2\pi) + n$ remains constant
- This leads to a trade-off
- One should include an additional predictor if the corresponding decrease in the first term is larger than the increase in the second term
- When we compare different models, we pick the model with the **smallest** value of the AIC.

Note

- When computing the AIC for the OLS models, some prefer to omit the constant term $n \ln(2\pi) + n$
- This omission does not affect the comparison between the models but does change the absolute value of the AIC
- The function ‘extractAIC‘ in ‘R‘ does that. It also uses the number of coefficients for k, not the number of coefficients $+ 1$.

BIC

- The formula for the BIC slightly differs from the AIC in the first term:
$$\text{BIC} = k \ln n + n \ln \frac{SSR}{n} + n \ln(2\pi) + n$$
- The BIC penalizes extra predictors more than the AIC, hence the values of the BIC are typically higher than the AIC for a given model

Note

- In case you are wondering where the $n\ln(2\pi)+n$ term comes from, the formulas for the AIC and the BIC are actually derived for models estimated using the **Maximum Likelihood Estimation** method
- A model estimated using this method has a **likelihood**, L
- The original AIC formula is

$$\mathrm{AIC} = 2k - 2\ln(L).$$

- However, for the linear regression, one can show that the logarithm of the likelihood can be written in terms of SSR as follows:

$$\ln L = -\frac{n}{2}\ln\frac{SSR}{n} - \frac{n}{2}\ln(2\pi) - \frac{n}{2},$$

  which, after substitution, yields the formula in the beginning.

Using the information criteria

- When using the information criteria, there are a few of points worth keeping in mind
- First, the criteria are used to **compare** different models, they do not tell you about a model's fit
- You need a measure like R-squared for this purpose
- Second, you **should not** compare two models that use different transformations of the outcome variable
- For example, you should not compare a model where the outcome variable is not transformed with the model where the outcome variable is logged
- Third, it is a good idea to check both criteria, however, they **do not** always give the same results

Trade example

- The first model will be our very first simple linear regression

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + u_i.$$

- The second model will be the so-called gravity model that adds distance as a predictor

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i.$$

- The third model will add another predictor to the gravity model: the degree of a country's liberalization:

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + \beta_3 liberal + u_i.$$

- Finally, the fourth model will add a country's area to the previous model

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + \beta_3 liberal + \beta_4 \ln(area) + u_i.$$

Results

	simple	gravity	add liberal	add area
(Intercept)	−5.785	4.670	2.774	2.451
	(2.199)	(2.181)	(2.183)	(2.132)
log(gdp)	1.078	0.976	0.941	1.030
	(0.088)	(0.064)	(0.062)	(0.077)
log(distance)		−1.075	−0.973	−0.888
		(0.157)	(0.153)	(0.156)
liberal			0.497	0.333
			(0.193)	(0.207)
log(area)				−0.159
				(0.086)
Num.Obs.	48	48	48	48
AIC	166.333	134.103	129.337	127.634
BIC	171.947	141.588	138.693	138.861
R2	0.767	0.886	0.901	0.908
R2 Adj.	0.762	0.881	0.894	0.900

Next Time on *Introductory Econometrics*...

Variable transformations