Underfitting and Overfitting

Underfitting

Suppose the true model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\eta}$ is an unknown fixed vector and $\boldsymbol{\varepsilon}$ satisfies the GMM.

Suppose we incorrectly assume that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

This example of misspecifying the model is known as underfitting.

Note that $\boldsymbol{\eta}$ may equal $\boldsymbol{W\alpha}$ for some design matrix $\boldsymbol{W}$ whose columns could contain explanatory variables excluded from $\boldsymbol{X}$.

What are the implications of underfitting?

Find $E(\mathbf{c}'\hat{\boldsymbol{\beta}})$.

Find $E(\hat{\sigma}^2)$.

Example 1

Consider an experiment with two experimental units (mice in this case) for each of two treatments.

We might assume the GMM holds with

$$E(\boldsymbol{y}) = E\left(\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}\right) = \boldsymbol{X\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix}.$$

Example 1 (continued)

Suppose the person who conducted the experiment neglected to mention that, in each treatment group, one of the experimental units was male and the other was female.

Example 1 (continued)

Then the true model may require

$$E(\boldsymbol{y}) = \begin{bmatrix} \mu + \tau_1 + \alpha/2 \\ \mu + \tau_1 - \alpha/2 \\ \mu + \tau_2 + \alpha/2 \\ \mu + \tau_2 - \alpha/2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix} + \begin{bmatrix} \alpha/2 \\ -\alpha/2 \\ \alpha/2 \\ -\alpha/2 \end{bmatrix}$$

$$= \boldsymbol{X\beta} + \begin{bmatrix} 1/2 \\ -1/2 \\ 1/2 \\ -1/2 \end{bmatrix} \begin{bmatrix} \alpha \end{bmatrix} = \boldsymbol{X\beta} + \boldsymbol{W\alpha} = \boldsymbol{X\beta} + \boldsymbol{\eta}.$$

Example 1 (continued)

If we analyze the data assuming the GMM with $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, determine

1. $E(\widehat{\tau_1 - \tau_2})$, and
2. $E(\hat{\sigma}^2)$.

Example 2

Once again consider an experiment with two experimental units (mice) for each of two treatments.

Suppose we assume the GMM holds with

$$E(\boldsymbol{y}) = E\left(\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}\right) = \boldsymbol{X\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix}.$$

Example 2 (continued)

Suppose the person who conducted the experiment neglected to mention that both experimental units in treatment group 1 were female and that both experimental units in treatment group 2 were male.

Example 2 (continued)

Then the true model may require

$$
\begin{aligned}
E(\boldsymbol{y}) &= \begin{bmatrix} \mu+\tau_1+\alpha/2 \\ \mu+\tau_1+\alpha/2 \\ \mu+\tau_2-\alpha/2 \\ \mu+\tau_2-\alpha/2 \end{bmatrix} = \begin{bmatrix} \mu+\tau_1 \\ \mu+\tau_1 \\ \mu+\tau_2 \\ \mu+\tau_2 \end{bmatrix} + \begin{bmatrix} \alpha/2 \\ \alpha/2 \\ -\alpha/2 \\ -\alpha/2 \end{bmatrix} \\
&= \boldsymbol{X\beta} + \begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix} \begin{bmatrix} \alpha \end{bmatrix} = \boldsymbol{X\beta} + \boldsymbol{W\alpha} = \boldsymbol{X\beta} + \boldsymbol{\eta}.
\end{aligned}
$$

Example 2 (continued)

If we analyze the data assuming the GMM with $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, determine

1. $E(\widehat{\tau_1 - \tau_2})$, and
2. $E(\hat{\sigma}^2)$.

Overfitting

Now suppose we consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2] \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \quad \ni \quad \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2.$$

Furthermore, suppose that (unknown to us) $\boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{0}$.

In this case, we say that we are overfitting.

Note that we are fitting a model that is more complicated than it needs to be.

To examine the impact of the overfitting, consider the case where $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ is of full-column rank.

If we were to fit the simpler and correct model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, the LSE of $\boldsymbol{\beta}_1$ is $\tilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1\boldsymbol{y}$. Then

$$\begin{aligned} E(\tilde{\boldsymbol{\beta}}_1) &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{\beta}_1 \\ &= \boldsymbol{\beta}_1. \end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{\beta}}_1) &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathrm{Var}(\boldsymbol{y})\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} \\
&= \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} \\
&= \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}.
\end{aligned}
$$

If we were to fit the full model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

that is correct but more complicated than it needs to be, then the LSE of $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \left([\boldsymbol{X}_1, \boldsymbol{X}_2]'[\boldsymbol{X}_1, \boldsymbol{X}_2]\right)^{-1} [\boldsymbol{X}_1, \boldsymbol{X}_2]'\boldsymbol{y}$$

$$= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

If $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$, then

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}$$

$$= \begin{bmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1\boldsymbol{y} \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}_1 \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2\boldsymbol{y} \end{bmatrix}.$$

Now suppose $\boldsymbol{X}_1'\boldsymbol{X}_2 \neq \mathbf{0}$.

$$E\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y})$$
$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$
$$= \boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}.$$

Thus, $E(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1$.

$$
\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}\right) \\
&= \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1}.
\end{aligned}
$$

By Exercise A.72,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{E}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{E}^{-1} \\ -\boldsymbol{E}\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{E}^{-1} \end{bmatrix},$$

where $\boldsymbol{E} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$.

Thus, $\text{Var}(\hat{\beta}_1)$ is σ^2 times

$$(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$$
$$= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}.$$

Thus,

$$\text{Var}(\hat{\boldsymbol{\beta}}_1) - \text{Var}(\tilde{\boldsymbol{\beta}}_1) = \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}.$$

In a homework problem, you will show that

$$\text{Var}(\hat{\boldsymbol{\beta}}_1) - \text{Var}(\tilde{\boldsymbol{\beta}}_1) \quad \text{is NND.}$$

Thus, one cost of overfitting is increased variability of estimators of regression coefficients.

How is estimation of σ^2 affected?

Let $r_1 = rank(\boldsymbol{X}_1)$ and $r_2 = rank(\boldsymbol{X}_2)$.

If we fit a simpler model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, then

$$\tilde{\sigma}^2 = \frac{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{y}}{n - r_1} \quad \text{and}$$

$$E(\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{y}) = (n - r_1)\sigma^2$$

$$\Rightarrow E(\tilde{\sigma}^2) = \sigma^2.$$

If we overfit with the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, then

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}}{n - r} \qquad \text{and}$$

$$E(\hat{\sigma}^2) = \sigma^2.$$

Thus, overfitting does not lead to biased estimation of σ^2.

However, as we will see later in the course, overfitting leads to a loss of degrees of freedom $(n - r < n - r_1)$, which can lead to a loss of power for testing hypotheses about $\boldsymbol{\beta}$.