Advanced Methods in Biostatistics II

Lecture 5

November 7, 2017

Linear model

- Consider the linear model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- Today we will revisit the problem where either irrelevant explanatory variables are included, or relevant variables are omitted.

- In addition, we will address the effects of other types of model misspecification.

Model misspecification

- In linear models, we can characterize different forms of model misspecification.
- To illustrate, let us consider the following models:

$$\text{Model 1: } \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon$$

$$\text{Model 2: } \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon$$

where the ε are assumed iid normals with variance σ^2.

Model misspecification

- Let us further differentiate between the assumed and the true model.
- For example, if we assume Model 1 but Model 2 is true, we have underfit the model (i.e., omitted variables that were necessary).
- In contrast, if we assume Model 2 but Model 1 is true, we have overfit the model (i.e., included variables that were unnecessary).

Impact of underfitting

- Let us begin by considering underfitting, i.e., assume Model 2 is true, but we instead use the model:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon.$$

- In this setting the least-squares estimator is given by

$$\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}.$$

Impact of underfitting

- Computing the expectation, we see that

$$\begin{aligned} E(\hat{\beta}_1) &= E((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' E(\mathbf{y}) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) \\ &= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 \end{aligned}$$

- Thus, the estimate of β_1 is biased.
- Note that the bias disappears if either $\beta_2 = 0$ or $\mathbf{X}_1'\mathbf{X}_2 = 0$.

Impact of underfitting

- Consider the case where both design matrices are mean-centered.
- Now the term
$$\frac{1}{n-1}\mathbf{X}_1'\mathbf{X}_2$$
represents the empirical variance-covariance matrix between $\mathbf{X}_1$ and $\mathbf{X}_2$.
- Thus, if the omitted variables are uncorrelated with the included variables, then no bias exists.

Example

- Suppose we fit
$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \varepsilon,$$
when the true model is
$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + \varepsilon.$$
- In this situation
$$\mathbf{X}_1' = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}$$
and
$$\mathbf{X}_2' = \begin{pmatrix} x_1^2 & \dots & x_n^2 \end{pmatrix}.$$

Example

- Thus, we can write:

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

and

$$\mathbf{X}_1'\mathbf{X}_2 = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix} = \begin{pmatrix} \sum x_i^2 \\ \sum x_i^3 \end{pmatrix}.$$

Example

- Therefore we can express the bias in $\hat{\beta}$ as follows:

$$
\begin{aligned}
bias &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 \\
&= \frac{\beta_2}{\sum(x_i - \bar{x})^2} \begin{pmatrix} (\sum x_i^2)^2/n - \bar{x}\sum x_i^3 \\ -\bar{x}\sum x_i^2 + \sum x_i^3 \end{pmatrix}.
\end{aligned}
$$

Example

- Suppose we fit
$$y_{ij} = \mu_i + \varepsilon_{ij},$$
when the true model is
$$y_{ij} = \mu_i + \eta z_{ij} + \varepsilon_{ij},$$
with $i = 1, 2,\ j = 1, \ldots, n_i$.
- In other words, we are comparing two groups, but ignore the covariate z.

Example

- In matrix form the true model is $\mathbf{y} = \mathbf{X}_1\beta + \mathbf{X}_2\boldsymbol{\eta} + \varepsilon$, or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} z_{11} \\ \dots \\ z_{1n_1} \\ z_{21} \\ \dots \\ z_{2n_2} \end{pmatrix} \eta + \begin{pmatrix} \varepsilon_{11} \\ \dots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2n_2} \end{pmatrix}.$$

Example

- Then the bias in $(\hat{\mu}_1, \hat{\mu}_2)'$ is given by

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\eta = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} \eta.$$

- Hence, the group comparison given by

$$\hat{\mu}_1 - \hat{\mu}_2$$

is unbiased if $\bar{z}_1 = \bar{z}_2$.

Example

- This example illustrates the effect of randomization.
- Suppose we randomly assign experimental units to the two groups.
- Then we will have $\bar{z}_1 \approx \bar{z}_2$ for any covariate z, as long as groups are fairly large.
- Thus, randomization helps controls for bias due to unfitted covariates.

Impact of underfitting

- The theoretical standard errors for $\hat{\beta}_1$ is still correct in that

$$\mathrm{var}(\hat{\beta}_1) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\sigma^2.$$

- However, we still need to estimate σ^2.

Impact of underfitting

- The estimate of σ^2 will be biased, with

$$E(s^2) = \sigma^2 + \frac{1}{n-p}\boldsymbol{\beta}_2'\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2\boldsymbol{\beta}_2.$$

- This can be seen by noting that:

$$\begin{aligned} E(\mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{y}) &= (\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2)'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &\quad + \mathrm{tr}[(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\sigma^2\mathbf{I})] \\ &= (\mathbf{X}_2\boldsymbol{\beta}_2)'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})(\mathbf{X}_2\boldsymbol{\beta}_2) + (n-p)\sigma^2 \end{aligned}$$

Impact of underfitting

- Because the term $\mathbf{I} - \mathbf{H}_{\mathbf{X}_1}$ is positive definite, the term s^2 is biased upward.
- In this setting, variation due to unmodeled systematic variation is incorrectly attributed to the error.

Overfitting

- Now, let us consider the case of overfitting.
- Assume the correctly specified model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \epsilon.$$

- However, suppose we instead use the model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \\ &= \mathbf{X}\boldsymbol{\beta} + \epsilon \end{aligned}$$

where $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \;\; \boldsymbol{\beta}_2]'$.

Impact of overfitting

- In this setting, our estimate of β_1 will be unbiased.
- This holds because the true model is a special case of the fitted model with $\beta_2 = \mathbf{0}$.

Block matrix inversion

Theorem

If $\mathbf{A}$ and $\mathbf{D}$ are symmetric and all inverses exist,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{G} & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{G} & \mathbf{E}^{-1} \end{pmatrix},$$

where $\mathbf{E} = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})$, $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$, and $\mathbf{G} = \mathbf{C}\mathbf{A}^{-1}$.

Impact of overfitting

- Using this result and the fact that $\mathbf{G} = \mathbf{F}'$, we can write:

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \\
&= \sigma^2 \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix},
\end{aligned}
$$

where

$$\mathbf{F} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2,$$

and

$$\mathbf{E} = \mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2.$$

Impact of overfitting

- Therefore,

$$\text{var}(\hat{\beta}_1) = \sigma^2[(\mathbf{X}_1'\mathbf{X}_1)^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}'],$$

- Compare this to $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ which would result from fitting the true model where $E[\mathbf{Y}] = \mathbf{X}_1\beta_1$.
- In the above, $\mathbf{F}\mathbf{E}^{-1}\mathbf{F}'$ is positive definite unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$.

Impact of overfitting

- Therefore, the variance assuming Model 2 will always be greater than the variance assuming Model 1.
- Note at no point did we actually utilize which model was actually true.
- This illustrates the key point that adding more regressors into a linear model necessarily increases the standard error of the ones already included.

Impact of overfitting

- This is called "variation inflation".
- Note that the estimated variances need not go up, since σ^2 will decrease as we include additional variables.

Impact of overfitting

- If we fit Model 2 but Model 1 is correct, then our variance estimate will be unbiased.
- Again, this holds because we fit the correct model, and simply allowed for the possibility that β_2 was non-zero when it is in fact exactly zero.
- Therefore s^2 is an unbiased estimate for σ^2.

Impact of overfitting

- However, recall that

$$\frac{(n-p_1-p_2)s_2^2}{\sigma^2} \sim \chi^2_{n-p_1-p_2},$$

  where s_2^2 is the variance assuming Model 2.

- Similarly,

$$\frac{(n-p_1)s_1^2}{\sigma^2} \sim \chi^2_{n-p_1},$$

  where s_1^2 is the variance assuming Model 1.

Impact of overfitting

- Using the fact that the variance of a χ^2-distributed random variable is twice the degrees of freedom, we get that

$$\frac{Var(s_2^2)}{Var(s_1^2)} = \frac{(n - p_1)}{(n - p_1 - p_2)}.$$

- Thus, despite both estimates being unbiased, the variance of the estimated variance under Model 2 is higher.

Summary

	Effect of Underfitting	Effect of Overfitting
$\hat{\beta}$	biased	unbiased
$\hat{\mathbf{y}}$	biased	unbiased
s^2	biased upward	unbiased
$\text{var}(\hat{\beta})$	still $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$>$ than necessary

Mis-specified variance-covariance

- Next, let us assume that we have specified $E[\mathbf{Y}] = \mathbf{X}\beta$ correctly, but the variance-covariance matrix incorrectly.
- To illustrate, suppose that $\mathrm{var}(\varepsilon) = \sigma^2\mathbf{V}$, but we assume that $\mathrm{var}(\varepsilon) = \sigma^2\mathbf{I}$.
- In the full rank case the parameter estimates $\hat{\beta}$ are still unbiased.

Mis-specified variance-covariance

- However,

$$\mathrm{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

- Also, in most cases s^2 will be biased, since

$$E[s^2] = \frac{\sigma^2}{n-p}\mathrm{tr}[\mathbf{V}(\mathbf{I}-\mathbf{H})].$$

Effects of non-normality

- Finally, let us suppose we have correctly specified the model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad E[\varepsilon] = \mathbf{0}, \quad \mathrm{cov}(\varepsilon) = \sigma^2\mathbf{I},$ but suppose that ε is not necessarily multivariate normal.

- We have seen previously that $\hat{\beta}$ is unbiased, and $\mathrm{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, without requiring any distributional assumptions.

Effects of non-normality

- Thus, normality is not required to fit a linear model.
- However, normality of the coefficient estimates $\hat{\beta}$ is needed to compute confidence intervals and perform tests.
- As $\hat{\beta}$ is a weighted sum of $\mathbf{y}$, the Central Limit Theorem guarantees that it will be normally distributed if the sample size is large enough.
- Thus, tests and confidence intervals can be based on the associated t-statistic in these settings.

Bootstrap

- However, in many settings, bootstrap procedures may be more appropriate.
- There are several alternative ways of performing the bootstrap on linear models.
- The most straightforward approach is to link the response and explanatory variables for each observation and resample observations.
- However, this treats the explanatory variables as random rather than fixed.

Bootstrap

- To circumvent this, an alternative strategy is to select bootstrap samples of the residuals, and use these to create new observations, i.e.

$$y_i^* = \hat{y}_i + e_i^*.$$

- One can now link the bootstrapped y values with the fixed x values to obtain bootstrap model coefficients.