13 Effects of Departures from Assumptions

The standard assumption for the hypothesis tests and confidence intervals discussed was $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$. We will consider the effects on the inference for various departures from the above assumption.

13.1 Effects of Underfitting

Suppose that the true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, but we fit the model $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ instead. We assume that the columns of $\mathbf{Z}$ are linearly independent of the columns of $\mathbf{X}$, and that $\mathbf{X}$ has full rank, i. e. $\text{rank}(\mathbf{X}_{n\times p}) = p$.

13.1 Theorem: If we assume the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we get $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\eta}$, and therefore the bias for the parameter estimates is equal to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\eta}$. The fitted values are biased as well.

13.2 Example: Fit $E[Y] = \beta_0 + \beta_1 x$, when the true model is $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. Then

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

and

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

Therefore the bias in $\hat{\boldsymbol{\beta}}$ is

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

13.3 Example: Fit $E[Y_{ij}] = \mu_i$, when the true model is $Y_{ij} = \mu_i + \eta z_{ij} + \varepsilon_{ij}$, with $i = 1, 2$, $j = 1, \dots, n_i$. In other words, we are comparing two groups, but ignore the covariate z. In matrix form the true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, or

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

Then the bias in $(\hat{\mu}_1, \hat{\mu}_2)'$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\eta = (\bar{z}_1, \bar{z}_2)'\,\eta$, and so the group comparison $\hat{\mu}_1 - \hat{\mu}_2$ is unbiased if $\bar{z}_1 = \bar{z}_2$.

13.4 Note: Example 13.3 illustrates the effect of randomization. Suppose we randomly assign experimental units (for example patients) to the two groups. Then $\bar{z}_1 \approx \bar{z}_2$ for any covariate z, as long as groups are fairly large. Thus, randomization controls for bias due to unfitted covariates.

13.5 Theorem: If we assume the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we still have $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. However, the estimate of the error variance σ^2 is biased, since

$$E[S^2] = \sigma^2 + \frac{\boldsymbol{\eta}'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}\boldsymbol{\eta}}{n-p} > \sigma^2.$$

13.6 Note: The lesson in Theorem 13.5 is that underfitting leads to overestimation of the error variance.

13.2 Effects of Overfitting

Suppose the true model is $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, but we fit the model $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$. In other words, we are fitting the unnecessary terms in $\mathbf{X}_2$.

13.7 Theorem: The parameter estimates and the fitted values in the above scenario are still unbiased.

13.8 Note: For $\text{cov}(\hat{\boldsymbol{\beta}}_1)$ we have

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

where

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

and

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

Therefore,

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

compared with $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ which would result from fitting the true model $E[\mathbf{Y}] = \mathbf{X}_1\boldsymbol{\beta}_1$.

13.9 Theorem: In the above, $\mathbf{F}\mathbf{E}^{-1}\mathbf{F}'$ is positive definite unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$.

13.10 Note: The lesson in Theorem 13.9 is that the variance of individual components of $\hat{\boldsymbol{\beta}}_1$ will be inflated by overfitting unless the unnecessary terms fitted are orthogonal to the other terms in the model.

13.11 Theorem: S^2 remains unbiased, i. e. $E[S^2] = \sigma^2$.

13.12 Note: The lesson in this subsection is that overfitting does not introduce bias into regression coefficient estimates, but it does inflate their variances. In comparison to underfitting, we have the following:

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

13.3 Effects of a Mis-Specified Covariance Matrix

Assume that we have specified $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ correctly, but suppose that $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, when we assume that $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

13.13 Theorem: In the full rank case the parameter estimates are still unbiased, but

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Also, in most cases S^2 is biased, since

$$E[S^2] = \frac{\sigma^2}{n-p}\text{tr}[\mathbf{V}(\mathbf{I}-\mathbf{P})].$$

13.14 Example: The effect of non-constant variance in the two-sample t-test:

Assume the model $Y_{ij} = \mu_i + \varepsilon_{ij}$, $\text{var}(\varepsilon_{ij}) = \sigma_i^2$, $i = 1, 2$, $j = 1, \ldots, n_i$. The usual t-statistic for forming a confidence interval for $\mu_1 - \mu_2$ is

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S(n_1^{-1} + n_2^{-1})^{1/2}},$$

where

$$S^2 = \frac{1}{n-2}\sum_i\sum_j (Y_{ij} - \bar{Y}_i)^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}.$$

Here, $n = n_1 + n_2$, and s_i^2 is the sample variance in the ith group. Now, if $\sigma_1^2 = \sigma_2^2$ and ε_{ij} is normally distributed, then $T \sim t_{n-2} \approx N(0,1)$ for large n. However, assume that $\sigma_1^2 \neq \sigma_2^2$. Then, heuristically, $S^2 \approx \frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)$ for large n, and T is approximately normally distributed with mean 0 and

$$\text{var}(T) \approx \frac{\text{var}(\bar{Y}_1 - \bar{Y}_2)}{\frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)(n_1^{-1} + n_2^{-1})} = \frac{n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}{\frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)(n_1^{-1} + n_2^{-1})} = \frac{\frac{\sigma_1^2}{\sigma_2^2} + \frac{n_1}{n_2}}{\frac{n_1}{n_2}\frac{\sigma_1^2}{\sigma_2^2} + 1}.$$

In other words, $\text{var}(T) \approx 1$ and therefore $T \approx N(0,1)$ for large n if either $\sigma_1^2 = \sigma_2^2$ (i.e. the equal variance assumption holds), or if $n_1 = n_2$ (i.e. the sample sizes are equal, regardless of equality of variances).

13.15 Example: (cont.) Recall the 95% confidence interval for $\mu_1 - \mu_2$:

$$\text{CI} = [\ \bar{Y}_1 - \bar{Y}_2 - t^{.025}_{n-2} S(n_1^{-1} + n_2^{-1})^{1/2}\ ,\ \bar{Y}_1 - \bar{Y}_2 + t^{.025}_{n-2} S(n_1^{-1} + n_2^{-1})^{1/2}\ ].$$

The error rate of this confidence interval is

$$P(\mu_1 - \mu_2 \notin \text{CI}) = P(|T| > t^{.025}_{n-2}) \approx P(|N(0, v)| > t^{.025}_{n-2}),$$

where $v = (\frac{\sigma_1^2}{\sigma_2^2} + \frac{n_1}{n_2})/(\frac{n_1}{n_2}\frac{\sigma_1^2}{\sigma_2^2} + 1)$.

Some values of the error rate based on the above normal approximation are given in the table below. The error rate does not deviate too far from the nominal value of 0.05 unless both the sample sizes and the variances differ substantially between groups.

$\downarrow$	σ_1^2/σ_2^2						
n_1/n_2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8
$\frac{1}{2}$	0.011	0.016	0.028	0.050	0.080	0.110	0.133
1	0.050	0.050	0.050	0.050	0.050	0.050	0.050
2	0.133	0.110	0.080	0.050	0.028	0.016	0.011
4	0.237	0.179	0.110	0.050	0.016	0.004	0.001
8	0.331	0.237	0.133	0.050	0.011	0.001	0.000

13.4 Effects of Non-normality

Suppose we have correctly specified the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ E[\boldsymbol{\varepsilon}] = \mathbf{0},\ \text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I},$ but suppose that $\boldsymbol{\varepsilon}$ is not necessarily multivariate normal.. We have seen previously that in the full rank case $\hat{\boldsymbol{\beta}}$ is unbiased, and $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, without any distributional assumptions. Under some regularity conditions, the usual distributional properties of $\hat{\boldsymbol{\beta}}$ and the F test statistic still hold approximately for large n. In particular,

$$\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

and for a testable hypothesis $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$,

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}})}{qS^2} \approx \chi^2_q/q,$$

where $\text{rank}(\mathbf{A}_{q\times p}) = q$. Note that the usual distribution of F when $\boldsymbol{\varepsilon} \sim$ MVN, $F_{q,n-p}$, is also approximately χ^2_q/q for large n. Therefore, inferences based on either the F or χ^2 distribution will be approximately correct.

13.16 Note: The effect of non-normality on the type I error rate of F-tests depends more critically on the kurtosis of the distribution (heaviness of the tails) rather than the skewness. But beware of one-sided t-tests with skewed data!

13.17 Note: The effect of non-normality tends to be less severe in balanced ANOVA designs.