Ch13. Analysis-of-Covariance

In general, an analysis of covariance model can be writen as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Z}$ contains 0s and 1s, $\boldsymbol{\alpha}$ contains μ and parameters such as α_i, β_j, and γ_{ij} representing factors and interactions (or other effects); $\boldsymbol{X}$ contains the covariate values; and $\boldsymbol{\beta}$ contains coefficients of the covariates.

1 Estimation and Testing

1.1 Estimation

We assume that $\boldsymbol{Z}$ is less than full rank as in overparameterized ANOVA models amd that $\boldsymbol{X}$ is full-rank as in regression models. We also assume that

$$E(\boldsymbol{\epsilon}) = 0, \quad \text{and} \quad Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}.$$

The model can be expressed as

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&= (\boldsymbol{Z}, \boldsymbol{X}) \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\epsilon} \qquad (1) \\
&:= \boldsymbol{U}\boldsymbol{\theta} + \boldsymbol{\epsilon}.
\end{aligned}$$

We can express the normal equation as two sets of equations:

$$\begin{aligned}
\boldsymbol{Z}'\boldsymbol{Z}\hat{\boldsymbol{\alpha}} + \boldsymbol{Z}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{Z}'\boldsymbol{y}, \\
\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{\alpha}} + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y}.
\end{aligned}$$

Using a generalized inverse of $\boldsymbol{Z}'\boldsymbol{Z}$, we can solve for $\boldsymbol{\alpha}$ as

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{y} - (\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&:= \hat{\boldsymbol{\alpha}}_0 - (\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{X}\hat{\boldsymbol{\beta}},
\end{aligned}$$

where $\hat{\boldsymbol{\alpha}}_0 = (\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{y}$ is a solution for the normal equations for the model

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

without the covariates.

Let $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. We see that

$$\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

Finally, we can get

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} := \boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy},$$

where $\boldsymbol{E}_{xx} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}$ and $\boldsymbol{e}_{xy} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$.

For the analysis-of-covariance model, we define SSE as $\mathrm{SSE}_{y\cdot x}$ and it can expressed as

$$\begin{aligned} SSE_{y\cdot x} &= \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\theta}}'\boldsymbol{U}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{y} - (\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')\begin{pmatrix}\boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{x}'\boldsymbol{y}\end{pmatrix} \\ &= \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\alpha}}_0'\boldsymbol{Z}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\ &= SSE_y - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\ &= SSE_y - \boldsymbol{e}_{xy}'\boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy}, \end{aligned}$$

where SSE_y is the same as the SSE for the ANOVA model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ without covariates. Since $\boldsymbol{E}_{xx}$ is nonsingular, we always have

$$SSE_{y\cdot x} < SSE_y.$$

1.2 Testing Hypotheses

The hypotheses $H_0 : \boldsymbol{C}\boldsymbol{\theta} = 0$ can be performed using the general linear hypothesis tests. In particular, for testing $H_0 : \boldsymbol{\beta} = 0$, we have

$$SSH = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}\hat{\boldsymbol{\beta}},$$

by noting that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$. By the above derivation, we also have

$$SSH = \boldsymbol{e}_{xy}'\boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy}.$$

In addition, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2[\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}]^{-1}$.

2 One-Way Model with One Covariate

The model is given by

$$y_{ij} = \mu + \alpha_i + \beta x_{ij}, \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, n.$$

In the matrix form, the model is given by

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{x}\beta + \boldsymbol{\epsilon}.$$

2.1 Estimation

By the results of previous chapters, one estimator of $\boldsymbol{\alpha}$ is obtained as

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \hat{\boldsymbol{\alpha}}_0 - (\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{x}\hat{\beta} \\ &= \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{k\cdot} \end{pmatrix} - \begin{pmatrix} 0 \\ \hat{\beta}\bar{x}_{1\cdot} \\ \hat{\beta}\bar{x}_{2\cdot} \\ \vdots \\ \hat{\beta}\bar{x}_{k\cdot} \end{pmatrix}.\end{aligned}$$

In this case

$$\begin{aligned}\boldsymbol{E}_{xx} &= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2, \\ \boldsymbol{e}_{xy} &= \sum_{ij}(x_{ij} - \bar{x}_{i\cdot})(y_{ij} - \bar{y}_{i\cdot}), \\ \boldsymbol{e}_{yy} &= \sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2.\end{aligned}$$

Then the estimator of β is given by

$$\hat{\beta} = \frac{\boldsymbol{e}_{xy}}{\boldsymbol{e}_{xx}} = \frac{\sum_{ij}(x_{ij} - \bar{x}_{i\cdot})(y_{ij} - \bar{y}_{i\cdot})}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2}.$$

The $\mathrm{SSE}_{y\cdot x}$ is given by

$$SSE_{y\cdot x} = \boldsymbol{e}_{yy} - \boldsymbol{e}_{xy}'\boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy} = \sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2 - \frac{[\sum_{ij}(x_{ij} - \bar{x}_{i\cdot})(y_{ij} - \bar{y}_{i\cdot})]^2}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2}, \tag{2}$$

which has $k(n-1) - 1$ degrees of freedom.

2.2 Testing Hypotheses

For testing hypotheses, we assume that the ϵ_{ij}'s are independently distributed as $N(0, \sigma^2)$.

Treatments To test

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k,$$

adjusted for the covariate, we can use a full-reduced-model approach. The reduced model is given by

$$y_{ij} = \mu^* + \beta x_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n.$$

In particular,

$$SSE_{rd} = \sum_{ij}(y_{ij} - \bar{y}_{..})^2 - \frac{[\sum_{ij}(x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})]^2}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})^2},$$

which has $kn - 2$ degrees of freedom.

The test statistic for the hypothesis is given by

$$F = \frac{SS(\boldsymbol{\alpha}|\mu, \beta)/(k-1)}{SSE_{y\cdot x}/[k(n-1)-1]} = \frac{[SSE_{rd} - SSE_{y\cdot x}]/(k-1)}{SSE_{y\cdot x}/[k(n-1)-1]},$$

which is distributed as $F(k-1, k(n-1)-1)$ under H_0.

The ANOVA table is given in Table 1.

Table 1: Analysis of Covariance for testing $H_0 : \alpha_1 = \cdots = \alpha_k$ in the One-way Model with one covariate, where $\mathrm{SST}_{y\cdots x} = SSE_{rd}$.

hline Source	SS Adjusted for Covariate	Adjusted df
Treatment	$\mathrm{SST}_{y\cdot x} - SSE_{y\cdot x}$	$k-1$
Error	$\mathrm{SSE}_{y\cdot x}$	$k(n-1)-1$
Total	$\mathrm{SST}_{y\cdot x}$	$kn-2$

Slope We consider a test for $H_0 : \beta = 0$. The full-reduced-model approach leads to

$$SSH = \boldsymbol{e}'_{xy}\boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy} = \frac{e_{xy}^2}{e_{xx}}.$$

The resulting test statistic is given by

$$F = \frac{e_{xy}^2/e_{xx}}{SSE_{y\cdot x}/[k(n-1)-1]},$$

which is distributed as $F(1, k(n-1)-1)$ when H_0 is true.

Homogeneity of Slopes The tests of $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$ and $H_0 : \beta = 0$ assume a common slope for all k groups. To check this assumption, we can test the hypothesis of equal slopes in the groups:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k,$$

where β_i is the slope in the ith group.

The full model is given by

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n,$$

which can be expressed in the matrix form:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_k \end{pmatrix},$$

and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$.

Applying the formulas derived before, we have

$$\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-}\boldsymbol{Z}' = diag\{\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}, \ldots, \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\},$$

where $\boldsymbol{I} - \boldsymbol{P}$ is $kn \times kn$ and $\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}$ is $n \times n$.

Thus

$$\boldsymbol{E}_{xx} = diag\{\boldsymbol{x}_1'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{x}_k\} = diag\{\sum_j (x_{1j} - \bar{x}_{1\cdot})^2, \ldots, \sum_j (x_{kj} - \bar{x}_{k\cdot})^2\}.$$

Similarly, we have

$$\boldsymbol{e}_{xy} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \begin{pmatrix} \boldsymbol{x}_1'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{y}_1 \\ \boldsymbol{x}_2'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{x}_k'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{y}_k \end{pmatrix}.$$

The estimate of $\boldsymbol{\beta}$ is thus given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy}.$$

Thus, for the full model, the SSE is given by

$$SSE(Full) = \boldsymbol{e}_{yy} - \boldsymbol{e}_{xy}'\boldsymbol{E}_{xx}^{-1}\boldsymbol{e}_{xy}.$$

For the reduced model $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$, we can get SSE(Reduced) as given by $SSE_{y \cdot x}$ in (2). The resulting test statistics

$$F = \frac{[SSE(reduced) - SSE(Full)]/(k-1)}{SSE(Full)/(nk - 2k)},$$

which is distributed as $F(k-1, k(n-2))$ under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$.

3 Two-Way Model with One Covariate

Consider the model

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \delta_{ij} + \beta x_{ijk} + \epsilon_{ijk}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, c, \quad k = 1, 2, \ldots, n,$$

where α_i is the effect of factor A, γ_j is the effect of factor C, δ_{ij} is the AC interaction effect, and x_{ijk} is a covariate.

3.1 Tests for Main Effects and Interactions

We first consider the test for the hypothesis of no overall treatment effect, that is, no A effect, no C effect, and no interaction. Therefore, the reduced model is given by

$$y_{ijk} = \mu^* + \beta x_{ijk} + \epsilon_{ijk}.$$

The SSE for the reduced model is given by

$$SSE_{rd} = \sum_{i=1}^{a}\sum_{j=1}^{c}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\ldots})^2 - \frac{[\sum_{ijk}(x_{ijk} - \bar{x}_{\ldots})(y_{ijk} - \bar{y}_{\ldots})]^2}{\sum_{ijk}(x_{ijk} - \bar{x}_{\ldots})^2}.$$

The SSE for the full model is given by

$$SSE_{y \cdot x} = \sum_{i=1}^{a}\sum_{j=1}^{c}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2 - \frac{[\sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})(y_{ijk} - \bar{y}_{ij\cdot})]^2}{\sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})^2},$$

which has $ac(n-1) - 1$ degrees of freedom.

The overall sum of squares for treatment is

$$\begin{aligned} SS(\alpha, \gamma, \delta|\mu, \beta) &= SSE_{rd} - SSE_{y \cdot x} \\ &= \sum_{ij} \frac{y_{ij\cdot}^2}{n} - \frac{y_{\ldots}^2}{acn} + \frac{[\sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})(y_{ijk} - \bar{y}_{ij\cdot})]^2}{\sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})^2} \\ &\quad - \frac{[\sum_{ijk}(x_{ijk} - \bar{x}_{\ldots})(y_{ijk} - \bar{y}_{\ldots})]^2}{\sum_{ijk}(x_{ijk} - \bar{x}_{\ldots})^2}. \end{aligned}$$

Further, we can partition the term $\sum_{ij} \frac{y_{ij\cdot}^2}{n} - \frac{y_{\ldots}^2}{acn}$ as follows:

$$\begin{aligned} \sum_{ij} \frac{y_{ij\cdot}^2}{n} - \frac{y_{\ldots}^2}{acn} &= cn\sum_{i}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\ldots})^2 + an\sum_{j}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\ldots})^2 + n\sum_{ij}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\ldots})^2 \\ &:= SSA_y + SSC_y + SSAC_y. \end{aligned}$$

To conform with this notation, we define

$$SSE_y = \sum_{ijk}(y_{ijk} - \bar{y}_{ij\cdot})^2.$$

and

$$SSE_x = \sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})^2.$$

We have an analogous partitioning of the overall treatment sum of squares for x:

$$\begin{aligned}\sum_{ij}\frac{x_{ij\cdot}^2}{n} - \frac{x_{\cdots}^2}{acn} &= cn\sum_i(\bar{x}_{i\cdot\cdot} - \bar{x}_{\cdots})^2 + an\sum_j(\bar{x}_{\cdot j\cdot} - \bar{x}_{\cdots})^2 + n\sum_{ij}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x}_{\cdots})^2 \\ &:= SSA_x + SSC_x + SSAC_x.\end{aligned}$$

The "overall treatment sum of products" can be partitoned:

$$\begin{aligned}\sum_{ij}\frac{x_{ij\cdot}y_{ij\cdot}}{n} - \frac{x_{\cdots}y_{\cdots}}{acn} &= cn\sum_i(\bar{x}_{i\cdot\cdot} - \bar{x}_{\cdots})(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + an\sum_j(\bar{x}_{\cdot j\cdot} - \bar{x}_{\cdots})(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) \\ &+ n\sum_{ij}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x}_{\cdots})(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) \\ &:= SPA + SPC + SPAC.\end{aligned}$$

Also, we define

$$SPE = \sum_{ijk}(x_{ijk} - \bar{x}_{ij\cdot})(y_{ijk} - \bar{y}_{ij\cdot}).$$

We can now write $SSE_{y\cdot x}$ in the simplified form:

$$SSE_{y\cdot x} = SSE_y - \frac{(SPE)^2}{SSE_x}.$$

These terms can be summarized into Table 2.

We now proceed to develop hypothesis tests for factor A, factor C, and the interaction AC. The orthogonality of the balanced design is lost when adjustments are made for the covariate. We therefore obtain a "total" for each term (A, C, or AC) by adding the edrror SS or SP to the term SS or SP for each of x, y and xy. These totals are used to obtain sums of squares adjusted for the covariate in a manner analogous to that employed in the one-way model. For example, the adjusted sum of squares $\mathrm{SSA}_{y\cdot x}$ for factor A is obtained as follows:

$$\begin{aligned}SS(A+E)_{y\cdot x} &= SSA_y + SSE_y - \frac{(SPA + SPE)^2}{SSA_x + SSE_x}, \\ SSE_{y\cdot x} &= SSE_y - \frac{(SPE)^2}{SSE_x}, \\ SSA_{y\cdot x} &= SS(A+E)_{y\cdot x} - SSE_{y\cdot x}.\end{aligned}$$

Table 2: Sums of squares and products for x and y in a twi-way model

	SS and SP corrected for the mean		
Source	y	x	xy
A	SSA_y	SSA_x	SPA
C	SSC_y	SSC_x	SPC
AC	SSAC_y	SSAC_x	SPAC
Error	SSE_y	SSE_x	SPE
A+E	SSA_y+SSE_y	SSA_x +SSE_x	SPA+SPE
C+E	SSC_y+SSE_y	SSC_x+SSE_x	SPC + SPE
AC+E	SSAC_y+SSE_y	SSAC_x+SSE_x	SPAC + SPE

The statistic for testing $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a$, corresponding to the main effect of A, is then given by

$$F = \frac{SSA_{y \cdot x}/(a-1)}{SSE_{y \cdot x}/[ac(n-1)-1]},$$

which is distributed as $F(a-1, ac(n-1)-1)$ if H_0 is true. Tests for factor C and the interaction AC are developed in an analogous fashion.

3.2 Test for Slope

To test the hypothesis $H_0 : \beta = 0$, the F-statistic is given by

$$F = \frac{(SPE)^2/SSE_x}{SSE_{y \cdot x}/[ac(n-1)-1]},$$

which is distributed as $F(1, ac(n-1)-1)$.