Covariance structure of observations

04.06.22

[Click here for PDF of slides](#)

Announcements

- Final project - optional draft due **Fri, Apr 15**, final report due **Wed, Apr 27**

Learning goals

- Use the model to understand the covariance structure among observations

Data: Housing prices in Southampton

The data includes the price and characteristics for 918 houses sold between 1986 and 1991 in Southampton, England. The data were originally collected from a local real estate agency and were analyzed in the 1991 article "Specifying and Estimating Multi-Level Models for Geographical Research" by Kelvyn Jones. The primary variables of interest are

- **price**: Sales price in thousands of £

- **Age**: Age of the house

- **Bedrooms**: Number of bedrooms

- **House Type**: (semi-detached, detached, bungalow, terrace, flat)

- **Central heating**: Whether house has central heating (0: yes, 1: no)

- **Garage**: Number of garages (none, single, double)

- **Districts**: one of 34 districts (baseline:)

- **Half-years**: Half-year periods beginning the second half of 1986

Data structure

(b) 3-level hierarchy

Level

3 District — Ocean Village, Shirley

2 Time — 1 ... 9, 1 ... 9
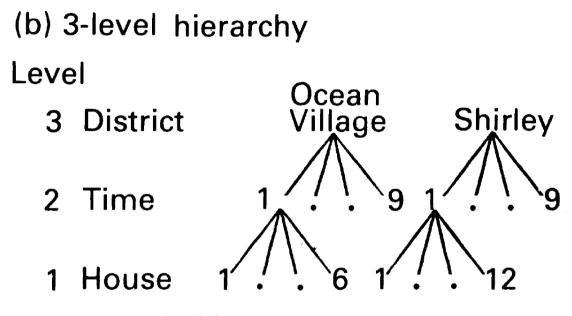
1 House — 1 ... 6, 1 ... 12

FIGURE 2. Hierarchical data structures

Adapted from Figure 2b from Jones (1991)

Note: The paper uses different symbols to represent parameters than what is in the textbook. The slides will follow the textbook.

Model A (unconditional means model)

$$Y_{ijk} = \alpha_0 + \tilde{u}_i + u_{ij} + \epsilon_{ijk}$$

$$\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2), \quad u_{ij} \sim N(0, \sigma_u^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Click here for model details.

Model B

$$Y_{ijk} = \alpha_0 + \beta_1 \text{age}_{ijk} + \beta_2 \text{detached}_{ijk} + \beta_3 \text{bungalow}_{ijk} + \beta_4 \text{terrace}_{ijk}$$
$$+ \beta_5 \text{flat}_{ijk} + \beta_6 \text{bedrooms}_{ijk} + \beta_7 \text{central heating}_{ijk}$$
$$+ \beta_8 \text{single}_{ijk} + \beta_9 \text{double}_{ijk}$$
$$+ \left[\tilde{u}_i + u_{ij} + \epsilon_{ijk} \right]$$

$$\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2), \quad u_{ij} \sim N(0, \sigma_u^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Click here for model details.

STA 310

Model C

$$Y_{ijk} = \alpha_0 + \beta_1 \text{age}_{ijk} + \beta_2 \text{detached}_{ijk} + \beta_3 \text{bungalow}_{ijk} + \beta_4 \text{terrace}_{ijk}$$
$$+ \beta_5 \text{flat}_{ijk} + \beta_6 \text{bedrooms}_{ijk} + \beta_7 \text{central heating}_{ijk}$$
$$+ \beta_8 \text{single}_{ijk} + \beta_9 \text{double}_{ijk}$$
$$+ [\tilde{u}_i + \tilde{v}_i \text{bedrooms}_{ijk} + u_{ij} + \epsilon_{ijk}]$$

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\tilde{u}}^2 & \sigma_{\tilde{u}\tilde{v}} \\ \sigma_{\tilde{u}\tilde{v}} & \sigma_{\tilde{v}}^2 \end{bmatrix} \right), \quad u_{ij} \sim N(0, \sigma_u^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Click here for model details.

Covariance structure of observations

What we've done

So far we have discussed...

- the covariance structure between error terms at a given level, e.g. the distribution of between $\tilde{u}_i$ and $\tilde{v}_i$ from a Level Two model:

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\tilde{u}}^2 & \sigma_{\tilde{u}\tilde{v}} \\ \sigma_{\tilde{u}\tilde{v}} & \sigma_{\tilde{v}}^2 \end{bmatrix} \right)$$

- how to use the intraclass correlation coefficient to get an idea of the average correlation among observations nested in the same Level Two and/or Level Three observations

Questions we want to answer

Now we want to be able to answer more specific questions about the covariance (and correlation) structure of *observations* at different levels.

- What is the covariance (and correlation) of houses in the same district sold in the same time period? $(Y_{ijk}, Y_{ijk'})$

- What is the covariance (and correlation) of houses in the same district sold in different time periods? $(Y_{ijk}, Y_{ij'k'})$

- What is the covariance structure of houses in the same district?

- What is the covariance structure of houses in the same district sold in the same time period?

How do these values differ between Model B and Model C in Jones (1991)?

Calculating variance and covariance

Suppose $Y_1 = a_1 X_1 + a_2 X_2 + a_3$ and $Y_2 = b_1 X_1 + b_2 X_2 + b_3$ where X_1 and X_2 are random variables and a_i and b_i are constants for $i = 1, 2, 3$, then we know from probability theory that:

$$Var(Y_1) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + 2a_1 a_2 Cov(X_1, X_2)$$

$$Cov(Y_1, Y_2) = a_1 b_1 Var(X_1) + a_2 b_2 Var(X_2) + (a_1 b_2 + a_2 b_1) Cov(X_1, X_2)$$

Note: This extends beyond two random variables

We will use these properties to define the covariance structure of the observations in the model.

from BMLR: Section 9.7.5

Calculating covariances from Model B

$$Y_{ijk} = \alpha_0 + \beta_1 \text{age}_{ijk} + \beta_2 \text{detached}_{ijk} + \beta_3 \text{bungalow}_{ijk} + \beta_4 \text{terrace}_{ijk}$$
$$+ \beta_5 \text{flat}_{ijk} + \beta_6 \text{bedrooms}_{ijk} + \beta_7 \text{central heating}_{ijk}$$
$$+ \beta_8 \text{single}_{ijk} + \beta_9 \text{double}_{ijk}$$
$$+ [\tilde{u}_i + u_{ij} + \epsilon_{ijk}]$$

$$\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2), \quad u_{ij} \sim N(0, \sigma_u^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Only use the random effects portion of the model in the variance and covariance derivations.

Use Model B to write the derivation of $Var(Y_{ijk})$, the variance of an individual observation. Then calculate $\hat{Var}(Y_{ijk})$.

$$\hat{\sigma}^2 = 236 \quad \hat{\sigma}^2_u = 139 \quad \hat{\sigma}^2_{\tilde{u}} = 201$$

Use Model B to write the derivation of $Cov(Y_{ijk}, Y_{ijk'})$, the covariance of houses in the same district sold in the same time period. Then calculate $\hat{Cov}(Y_{ijk}, Y_{ijk'})$.

$$\hat{\sigma}^2 = 236 \quad \hat{\sigma}^2_u = 139 \quad \hat{\sigma}^2_{\tilde{u}} = 201$$

Use Model B to write the derivation of $Cov(Y_{ijk}, Y_{ij'k'})$, the covariance of houses in the same district sold in different time periods. Then calculate $\hat{Cov}(Y_{ijk}, Y_{ij'k'})$.

$$\hat{\sigma}^2 = 236 \quad \hat{\sigma}_u^2 = 139 \quad \hat{\sigma}_{\tilde{u}}^2 = 201$$

Click here to submit your response.

STA 310

04:00

17

Correlation between observations

Recall,

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}$$

1. What is the estimated correlation between houses in the same district sold in the same time period?

2. What is the estimated correlation between houses in the same district sold in different time periods?

STA 310

Covariance structure

Write $Cov(\mathbf{Y}_{ij})$, the covariance matrix for observations from houses sold in the same time period that are in the same district.

$$Cov(\mathbf{Y}_{ij}) = \begin{bmatrix} \tau_1 & & & & \\ \tau_{12} & \tau_2 & & & \\ \tau_{13} & \tau_{23} & \tau_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tau_{1n_{ij}} & \tau_{2n_{ij}} & \tau_{3n_{ij}} & \cdots & \tau_{n_{ij}} \end{bmatrix}$$

where $\tau_k^2 = Var(Y_{ijk})$, $\tau_{kk'} = Cov(Y_{ijk}, Y_{ijk'})$, and n_{ij} is the number of houses sold in the ith district in the jth time period.

Correlation matrix

Write the correlation matrix for houses from the same district that were sold in the same time period.

Covariance of observations in Model C

Model C

$$Y_{ijk} = \alpha_0 + \beta_1 \text{age}_{ijk} + \beta_2 \text{detached}_{ijk} + \beta_3 \text{bungalow}_{ijk} + \beta_4 \text{terrace}_{ijk}$$
$$+ \beta_5 \text{flat}_{ijk} + \beta_6 \text{bedrooms}_{ijk} + \beta_7 \text{central heating}_{ijk}$$
$$+ \beta_8 \text{single}_{ijk} + \beta_9 \text{double}_{ijk}$$
$$+ [\tilde{u}_i + \tilde{v}_i \text{bedrooms}_{ijk} + u_{ij} + \epsilon_{ijk}]$$

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\tilde{u}} & \sigma_{\tilde{u}\tilde{v}} \\ \sigma_{\tilde{u}\tilde{v}} & \sigma^2_{\tilde{v}} \end{bmatrix} \right), \quad u_{ij} \sim N(0, \sigma^2_u), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

Use Model C to write the derivation of $Var(Y_{ijk})$, the variance of an individual observation. Then calculate $\hat{Var}(Y_{ijk})$ for a house with two bedrooms.

$$\hat{\sigma}^2 = 215 \quad \hat{\sigma}_u^2 = 127$$

$$\hat{\sigma}_{\tilde{u}}^2 = 213 \quad \hat{\sigma}_{\tilde{v}}^2 = 58 \quad \hat{\sigma}_{\tilde{u}\tilde{v}} = 111$$

Use Model C to write the derivation of $Cov(Y_{ijk}, Y_{ijk'})$, the covariance of houses in the same district sold in the same time period. Then calculate $\hat{Cov}(Y_{ijk}, Y_{ijk'})$ for two houses that each have two bedrooms.

$$\hat{\sigma}^2 = 215 \quad \hat{\sigma}^2_u = 127$$

$$\hat{\sigma}^2_{\tilde{u}} = 213 \quad \hat{\sigma}^2_{\tilde{v}} = 58 \quad \hat{\sigma}_{\tilde{u}\tilde{v}} = 111$$

Use Model C to write the derivation of $Cov(Y_{ijk}, Y_{ij'k'})$, houses in the same district sold in different time periods. Then calculate $\hat{Cov}(Y_{ijk}, Y_{ij'k'})$.

$$\hat{\sigma}^2 = 215 \quad \hat{\sigma}_u^2 = 127$$

$$\hat{\sigma}_{\tilde{u}}^2 = 213 \quad \hat{\sigma}_{\tilde{v}}^2 = 58 \quad \hat{\sigma}_{\tilde{u}\tilde{v}} = 111$$

04:00

Correlation between observations

Recall,

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}$$

1. What is the estimated correlation between two houses each with two bedrooms in the same district sold in the same time period?

2. What is the correlation between a house with two bedrooms and one with three bedrooms in the same district sold in the same time period?

STA 310

1. What is the estimated correlation between two houses each with two bedrooms in the same district sold in <u>different</u> time periods?

2. What is the correlation between a house with two bedrooms and one with three bedrooms in the same district sold in <u>different</u> time periods?

<u>Click here</u> to submit your response.

06:00

Which covariance structure do you think is most reasonable for the data? The variance structure imposed by Model B or the one imposed by Model C?

Alternative covariance structures

Different covariance structures

The standard covariance structure calculated from the multilevel model is useful in most situations. Sometimes, however, there may be a different covariance structure that better fits the data. A few alternatives are

Unstructured: Every variance and covariance term for observations with each level is a separate parameter and is uniquely estimated. No patterns among variances or correlations are assumed. Very flexible but requires the estimation of many parameters.

Compound Symmetry: Assume variance is constant across all Level One observations and correlation is constant across all pairs of Level One observations. Restrictive but few parameters to estimate.

Different covariance structures

Autoregressive: Assume constant variance across all time points, but correlation reduces in a systematic way such that closer time points are more correlated than those further apart.

Toeplitz: Similar to autoregressive but there is no imposed structure on the decreased correlation for time periods further apart.

Heterogeneous variances: Allows for equal variances across time points. Requires additional parameters to be estimated to allow for the unequal variances.

See "Fitting Linear Mixed Models in R" for details on R packages and code for multilevel models with a predetermined covariance structure, and Section 9.7.2 Alternative Covariance Structures in BMLR for more details.

Acknowledgements

- Jones, K. (1991). Specifying and estimating multi-level models for geographical research. Transactions of the institute of British geographers, 148-159.

- "Fitting Linear Mixed Models in R" by Brice Ozenne

- Beyond Multiple linear regression

  - Section 9.7: Covariance structure among observations

  - Section 10.9: Covariance structure