Lecture 3
Generalized Least Squares and Autocovariance Functions

Dennis Sun
Stanford University
Stats 253

June 26, 2015

1. A Model for Correlated Data

2. (Auto)covariance Functions

1. A Model for Correlated Data

2. (Auto)covariance Functions

The Model

Because of omitted variables, we end up with a situation where the errors are correlated:

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

where $\mathrm{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\epsilon}|X] = \Sigma$.

For now, we will assume that Σ is known (pretty unrealistic).

Ordinary Least Squares

For estimating $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{OLS}$ is still unbiased.

$$\mathrm{E}[\hat{\boldsymbol{\beta}}^{OLS}] = \mathrm{E}[(X^T X)^{-1} X^T \mathbf{y}] = (X^T X)^{-1} X^T X \boldsymbol{\beta} = \boldsymbol{\beta}.$$

What is its variance?

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}^{OLS}] = \mathrm{Var}[(X^T X)^{-1} X^T \mathbf{y}] = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}.$$

As long as you use the correct standard errors, OLS is fine.

But can we do better?

Generalized Least Squares

Heuristic: Decorrelate the data.
First, find a matrix $\Sigma^{-1/2}$ such that $\Sigma^{-1} = (\Sigma^{-1/2})^T \Sigma^{-1/2}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{GLS} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\Sigma^{-1/2}(\mathbf{y} - X\boldsymbol{\beta})||^2 \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\Sigma^{-1/2}\mathbf{y} - \Sigma^{-1/2}X\boldsymbol{\beta}||^2 \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}.
\end{aligned}$$

Tip for the Quiz: Think about more formal justifications for $\hat{\boldsymbol{\beta}}^{GLS}$ like the ones we saw for $\hat{\boldsymbol{\beta}}^{OLS}$ in Lecture 2.

Matrix Square Root

How do you calculate $\Sigma^{-1/2}$?

Two ways:

- If $\Sigma = V\Lambda V^T$ is the eigendecomposition of Σ, then $\Sigma^{-1/2} = V\Lambda^{-1/2}V^T$.
- Compute a Cholesky decomposition of the matrix, i.e., $\Sigma = LL^T$ where L is lower triangular. Then $\Sigma^{-1/2} = L^{-1}$.

The decomposition is not unique!

Both require $O(n^3)$ operations. But Cholesky is more stable and not iterative. It also results in an upper triangular matrix, which is easier to solve.

Computational Tricks of the Trade

How do you actually compute $A^{-1}\mathbf{x}$?

Do you calculate A^{-1} and multiply by $\mathbf{x}$?

No! $\mathbf{z} = A^{-1}\mathbf{x}$ is the solution to the system $A\mathbf{z} = \mathbf{x}$.

- In general, a system of equations also requires $O(n^3)$ operations. But you'll save the $O(n^2)$ memory to store A^{-1}.
- For some matrices A, solving $A\mathbf{z} = \mathbf{x}$ can be more efficient. If A is triangular, we can use **back substitution** to solve the system in $O(n^2)$ operations.

How would you calculate $\Sigma^{-1/2}\mathbf{y} = L^{-1}\mathbf{y}$?

Computing the GLS Estimator

$$\hat{\boldsymbol{\beta}}^{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}$$

1. Compute the Cholesky decomposition of $\Sigma = LL^T$.
2. Solve $L\mathbf{z} = \mathbf{y}$ and $L\mathbf{w}_j = \mathbf{x}_j$ (for each column j of X) by back substitution. This gives us $\mathbf{z} = \Sigma^{-1/2}\mathbf{y}$ on $W = \Sigma^{-1/2}X$.
3. Obtain $\hat{\boldsymbol{\beta}}^{GLS}$ by linear regression of $\mathbf{z}$ on W.

1. A Model for Correlated Data

2. (Auto)covariance Functions

What if we don't know Σ?

$$\Sigma = \text{Var}(\boldsymbol{\epsilon}) = \begin{pmatrix} \text{Var}[\epsilon_1] & \text{Cov}[\epsilon_1, \epsilon_2] & \dots & \text{Cov}[\epsilon_n, \epsilon_n] \\ \text{Cov}[\epsilon_2, \epsilon_1] & \text{Var}[\epsilon_2] & \dots & \text{Cov}[\epsilon_n, \epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[\epsilon_n, \epsilon_1] & \text{Cov}[\epsilon_n, \epsilon_2] & \dots & \text{Var}[\epsilon_n] \end{pmatrix}$$

- Can we estimate it from the data $(\mathbf{x}_i, y_i), i = 1, ..., n$?
- No! Σ has n^2 entries (actually $\frac{n(n-1)}{2}$ unique entries) and we only have n observations.
- We have to make more assumptions if we hope to estimate it from the data.

Parametrizing Σ

- Assume that there is a **(auto)covariance function**:

$$\Sigma_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$$

  that tells us the covariance between any two observations.
- $\mathbf{x}$ represents predictors, which is often spatial ($\mathbf{s}$) or temporal (t) coordinates.
- The covariance matrix is obtained by evaluating the covariance function at the data points.

$$\Sigma_{ij} = \Sigma_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$$

Stationary Covariances

Many covariance functions only depend on the $\mathbf{x} - \mathbf{x}'$.

$$\Sigma_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \Sigma_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{x}')$$

Such covariance functions are called **stationary**.

Example

Suppose we have a time series y_t. Here $\mathbf{x}$ represents time t.
Stationarity means

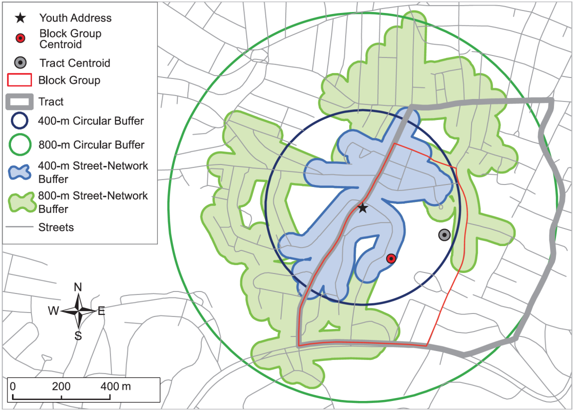
$$\mathrm{Cov}[y_1, y_5] = \Sigma_{\boldsymbol{\theta}}(5 - 1) = \mathrm{Cov}[y_3, y_7]$$

Isotropic Covariances

An even stronger assumption is that the covariance function depend only on the "distance" $d(\mathbf{x}, \mathbf{x}')$ between $\mathbf{x}$ and $\mathbf{x}'$.

Such covariance functions are called **isotropic**.

$d(\mathbf{x}, \mathbf{x}')$ can be Euclidean distance $||\mathbf{x} - \mathbf{x}'||$, but it can also be road distance, etc.



Common Covariance Functions

- Triangular: $\Sigma_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \max(\theta_1 - \theta_2 d(\mathbf{x}, \mathbf{x}'), 0)$.
- Exponential: $\Sigma_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_1 \exp\{-\theta_2 d(\mathbf{x}, \mathbf{x}')\}$.
- Gaussian: $\Sigma_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_1 \exp\{-\theta_2 d(\mathbf{x}, \mathbf{x}')^2\}$.

Next time, we'll talk about how to estimate $\boldsymbol{\theta}$ from the data.
This time, we'll focus on properties of covariance functions.

Valid Covariance Functions

- Covariance functions have to be **positive semidefinite**.
- That is, if we evaluate it at any set of n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the resulting covariance matrix is positive semidefinite.

$$\Sigma_{ij} = \Sigma_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$$

- What does it mean for a symmetric matrix A to be positive semidefinite? There are several equivalent definitions:
    - $\mathbf{x}^T A \mathbf{x} \geq 0$ for all $\mathbf{x}$.
    - The eigenvalues of A are all ≥ 0.
- **Test yourself:** Let X be any matrix. Is $X^T X$ positive definite?

Valid Covariance Functions

To check that a stationary covariance function $\Sigma_{\boldsymbol{\theta}}(\mathbf{h})$ is valid, we have **Bochner's theorem**, which says that it is valid if and only if

$$\Sigma_{\boldsymbol{\theta}}(\mathbf{h}) = \int_{\mathbb{R}^D} e^{i2\pi \mathbf{s}\cdot\mathbf{h}}\, d\mu(\mathbf{s})$$

for some measure $\mu \geq 0$.

To use Bochner's Theorem in practice: Take the Fourier transform of $\Sigma_{\boldsymbol{\theta}}(\mathbf{h})$ and check that the resulting function is positive.

In general, checking that a covariance function is valid is tricky, so it's best to stick to known covariance functions.

Obtaining New Covariance Functions from Old

Suppose Σ_1 and Σ_2 are two valid covariance functions. Then:

- $\Sigma(\mathbf{x}, \mathbf{x}') = \Sigma_1(\mathbf{x}, \mathbf{x}') + \Sigma_2(\mathbf{x}, \mathbf{x}')$ is also valid.
- $\Sigma(\mathbf{x}, \mathbf{x}') = \Sigma_1(\mathbf{x}, \mathbf{x}')\Sigma_2(\mathbf{x}, \mathbf{x}')$ is also valid.