Summary of Day 9: Social Relations Model and Goodness of Fit using Simulations

Overview

This lecture focuses on using the Social Relations Model (SRM) for network analysis, particularly addressing how to assess the goodness of fit of models through simulation based approaches. The lecture also highlights the limitations of generalized linear models (GLMs) and introduces the SRM as a framework to account for and estimate these network structures.

Key Points

1. **Introduction to Network Modeling**
    - Key features in network data include homophily (tendency for similar nodes to connect), degree heterogeneity (variability in number of connections per node), and reciprocity (mutual connections).
2. **Limitations of GLMs in Network Data**
    - Traditional GLMs assume independence of observations, which is often violated in network data.
    - Ignoring dependencies can lead to biased parameter estimates and poor model performance.
3. **Social Relations Model (SRM)**
    - SRM helps in decomposing the variability in the observed data into distinct components, capturing both observed and unobserved factors that influence these interactions.
    - SRM accounts for sender and receiver effects, allowing for heterogeneity in how actors send and receive connections. We referred to these as first order dependencies.
    - It models dyadic interactions and includes random effects to capture unobserved heterogeneity. We referred to these as second order dependencies.
4. **Simulations for Goodness of Fit Assessment**
    - **Purpose:** To evaluate how well a model captures observed network structures.
    - **Process:** Simulate multiple sets of predictions from the model and calculate measures of fit, such as out-degrees (variation across rows) and in-degrees (variation across columns), as well as reciprocity.
    - **Example:** We went over an example using trade data, where an OLS model was fit and residuals were analyzed for structure, revealing dependencies not accounted for by the model.
5. **Practical Application and Analysis**

- Using simulations, the lecture demonstrated how to assess the goodness of fit by comparing simulated data against observed data.
- This approach helps identify whether the model correctly specifies the relationships and dependencies present in the network and is the dominant approach to studying goodness of fit in a network context. We will return to this approach when we discuss ERGMs and SAOMs as well.

Using Simulations for Goodness of Fit

Here are summary of the steps involved in using simulations to assess the goodness of fit of a model in network analysis:

Example: Trade Network Analysis

1. **Model Specification:**
   - Fit a model (e.g., OLS) using trade data, with predictors like polity scores, conflicts, distance, and shared international organizations.
   - Analyze the residuals to check for patterns indicating violations of independence assumptions.
2. **Simulating Predictions:**
   - Generate 1000 simulations of the model's parameters.
   - For each simulation, predict the network outcomes and organize these into adjacency matrices.
3. **Evaluating Fit:**
   - Compare the simulated network metrics (e.g., mean out-degrees and in-degrees) against the observed data metrics.
   - Check for structural patterns in the residuals, such as sender and receiver effects, that the model may not capture.

By using simulations, we can quantify how well the model fits the data and identify areas where the model may need improvement, such as including additional covariates or adjusting for network dependencies.

Detailed Explanation: Drawing Samples of Coefficients for Simulations

Understanding the Concept

When we fit a statistical model, such as a linear regression (or any of the "network" regression we will discuss in this class), to data, we estimate the coefficients (β) for each predictor variable. These coefficients represent the relationship between the predictors and the response variable. However, due to variability in the data and potential sampling errors, there is uncertainty associated with these estimated coefficients.

To account for this uncertainty in simulations, we don't just use the point estimates of the coefficients. Instead, we draw multiple samples of these coefficients from a distribution that reflects our uncertainty about their true values. This process helps us understand how the model's predictions might vary if the true coefficients differ slightly from our estimates.

Steps to Draw Samples of Coefficients

1. **Estimate Coefficients and Covariance Matrix:**
   - After fitting the model, we obtain the estimated coefficients ($\hat{\beta}$) and the variance-covariance matrix of these estimates. The variance-covariance matrix describes the variability of each coefficient estimate and the covariances between them.
2. **Multivariate Normal Distribution:**
   - We assume that the estimated coefficients follow a multivariate normal distribution, a common statistical assumption for parameter estimates in linear models.
   - The multivariate normal distribution is characterized by:
     - **Mean Vector (μ):** The vector of estimated coefficients ($\hat{\beta}$).
     - **Covariance Matrix (Σ):** The variance-covariance matrix obtained from the model fitting process.
3. **Sampling from the Distribution:**
   - We draw multiple samples (e.g., 1000) of the coefficients from the multivariate normal distribution. Each sample represents a possible set of true coefficients, considering the uncertainty captured by the variance-covariance matrix.
   - In R, this can be done using the `mvrnorm` function from the `MASS` package:

     ```
     library(MASS)
     betaDraws <- mvrnorm(n = 1000, mu = coef(ols), Sigma = vcov(ols))
     ```

     Here, `n = 1000` specifies the number of samples, `mu = coef(ols)` is the vector of estimated coefficients, and `Sigma = vcov(ols)` is the variance-covariance matrix.

Why Use This Approach?

- **Incorporating Uncertainty:** This method accounts for the uncertainty in the coefficient estimates, providing a range of possible outcomes rather than a single point estimate.
- **Assessing Robustness:** By generating a range of possible coefficients, we can see how sensitive the model's predictions are to changes in the parameter estimates. This helps in assessing the robustness of the model.

- **Goodness of Fit Evaluation:** Using these samples, we can simulate multiple datasets and compare them to the observed data, helping us evaluate the model's goodness of fit more comprehensively.

Common Student Confusions and Clarifications

Confusion 1: Why Use the Multivariate Normal Distribution?

- The multivariate normal distribution is used because it naturally extends the univariate normal distribution to multiple dimensions, capturing the relationships between variables. It's a standard approach when dealing with multiple correlated coefficients.

Confusion 2: Interpretation of the Covariance Matrix

- The covariance matrix contains the variances of each coefficient (diagonal elements) and the covariances between them (off-diagonal elements). Variance indicates how much a coefficient might vary, while covariance indicates how two coefficients might vary together.