Stat Summary CheatSheet

immediate

2024-11-06

Abstract

Start with literature review and idea about how to estimate moment function with DIFM data

0.1 Cheat Sheet 1: Linear and Non-Linear Regression Models

1. Linear Regression Model

- Equation:

$$Y = X\beta + \epsilon$$

  Where Y is the dependent variable, X is the matrix of independent variables, β is the coefficient vector, and ϵ represents the error term.
- Assumptions:
  1. Linearity: The relationship between Y and X is linear.
  2. Full Rank: The X matrix has full rank; multicollinearity is absent.
  3. No Endogeneity: X and ϵ are uncorrelated.
  4. Homoscedasticity: Constant variance of the error terms.
  5. No Autocorrelation: Errors are not correlated with one another.
  6. Normality of Errors: Errors are normally distributed for inference.
- Violation Impacts:
  - Multicollinearity: Leads to large standard errors for β, making coefficients imprecise.
  - Endogeneity: Causes bias in β estimates.
  - Heteroscedasticity: Leads to inefficient estimators; standard errors are incorrect, affecting hypothesis tests.
  - Autocorrelation: Leads to inefficient β estimates and unreliable standard errors.
- Remedies:
  - Multicollinearity: Drop collinear variables or use regularization techniques (e.g., Ridge/Lasso).
  - Endogeneity: Use instrumental variables (IV).
  - Heteroscedasticity: Use robust standard errors or GLS.
  - Autocorrelation: Use GLS or Newey-West standard errors.

2. Non-Linear Regression Model

   - Equation (Example - Logistic Regression):

     $$P(Y = 1|X) = \frac{1}{1 + e^{-X\beta}}$$

     The response variable is binary, and the model is nonlinear in parameters.
   - Key Assumptions:
     - Independent Errors: Observations are independent.
     - Correct Model Specification: The functional form is correctly specified.
   - Violation Impacts:
     - Misspecification: Leads to biased estimates.
     - Multicollinearity: Impacts the stability of estimated coefficients.
   - Remedies:
     - Misspecification: Use non-parametric techniques to verify functional form.
     - Multicollinearity: Use variable selection or regularization.

3. Bias and Efficiency

   - Unbiased Estimator: An estimator is unbiased if $E(\hat{\beta}) = \beta$. Violations like omitted variables or endogeneity cause bias.
   - Efficiency: An efficient estimator has the smallest variance among all unbiased estimators. Violations of homoscedasticity or autocorrelation typically lead to inefficiencies.

0.2 Cheat Sheet 2: Statistical Tests for Regression Models

1. Assumption Checks for Linear Regression

   - Multicollinearity:
     - Variance Inflation Factor (VIF): High VIF (> 10) indicates multicollinearity.
   - Homoscedasticity:
     - Breusch-Pagan Test: Tests if variance of errors is constant.
     - White Test: Tests for heteroscedasticity without assuming a specific form.
   - Normality of Errors:
     - Shapiro-Wilk Test: Tests normality of residuals.
     - Q-Q Plot: Visual inspection for normality.
   - No Autocorrelation:
     - Durbin-Watson Test: Checks for first-order autocorrelation in residuals.

2. Assumption Checks for Non-Linear Models

   - Model Fit:

    - Likelihood Ratio Test: Compares nested models to determine if added complexity improves fit.
    - Wald Test: Tests the significance of individual regression coefficients.
- Multicollinearity:
    - Condition Index: High values (> 30) indicate multicollinearity.
- Goodness of Fit:
    - Pseudo R^2 (e.g., McFadden's R^2): Used for logistic regression to measure model fit.

3. Model Feature Tests

    - Endogeneity:
        - Hausman Test: Compares IV and OLS to determine if an endogeneity problem exists.
    - Nonlinearity:
        - RESET Test: Tests if non-linear combinations of the fitted values help explain the response variable.

4. Hypothesis Testing

    - T-Test: Tests the significance of individual coefficients.
    - F-Test: Tests the joint significance of multiple coefficients.
    - Likelihood Ratio Test: Used for nested model comparison.

0.3 Summary

- Relaxation of Assumptions can cause bias (e.g., endogeneity leads to biased β) or inefficiency (e.g., autocorrelation affects standard errors).
- Tests help identify violations of key assumptions, and remedies such as using robust standard errors or instrumental variables can address these issues.