Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.10 - Variable Selection Procedures

SS 2021 || Prof. Dr. Buchwitz

Wir geben Impulse

Outline

1. **Organizational Information**
2. Variable Selection Problem
3. Evaluation Criteria
4. Variable Selection Procedures
5. Example Processes

Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

Outline

1 Organizational Information

2 **Variable Selection Problem**

3 Evaluation Criteria

4 Variable Selection Procedures

5 Example Processes

Introduction

- So far we assumed that the variables in our equation were chosen in advance and we focused on examining the resulting equation and whether the specified functional form was correct.
- In most practical applications the regression model is not predetermined and it is often **the first part** of the analysis to select these variables.
- Given that theoretical considerations determine the variables to be included in the model, the seleciton problem does not arise. However, often there is no clear-cut theory and the variable selection problem gains importance.

Variable Selection Problem

- The problem of variable selection and finding an adequate funcitnoal specification of the equation are linked to each other.
  - Which variables should be included?
  - In which form should they be included $(X, X^2, log(X))$?
- Ideally the two problems should be solved **simultaneously**. For the sake of simplicity we cover them sequentially by first determining which variables should be included in the model and then investigate their exact form in which they enter it.

Variable Selection Problem

$$y_i \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i$$

- We have the response Y and q predictor variables $X_1, X_2, \ldots, X_q$, but instead of dealing with the full set of variables (especially when q is large), we **delete a number of variables** and construct the equation with the remaining subset.
- We denote the retained variables by $X_1, X_2, \ldots, X_p$ and those deleted by $X_{p+1}, X_{p+2}, \ldots, X_q$.

Variable Selection Problem

- We examine the effect of variable deletion under two conditions:
  1. The model that connects Y and the X's has **all** $\beta_1, \dots, \beta_q$ **nonzero**.
  2. The model has $\beta_1, \dots, \beta_p$ nonzero, but $\beta_{p+1}, \dots, \beta_q$ zero.
- What are the effects of including variables in an equaition when they should be properly left out (because their population regression coefficients are zero)?
- What are consequences of omitting variables that should be included (because their population regression coefficients are not zero)?

We will examine the effect of deletion of variables on the estimates of the parameters and the predicted values of Y.

Consequences of Variable Deletion

Uses of Regression Equation

Outline

1. Organizational Information
2. Variable Selection Problem
3. **Evaluation Criteria**
4. Variable Selection Procedures
5. Example Processes

Criteria for Evaluating Equations

Mallows C_p

Information Criteria

Outline

1 Organizational Information

2 Variable Selection Problem

3 Evaluation Criteria

4 **Variable Selection Procedures**

5 Example Processes

Outline

1. Organizational Information
2. Variable Selection Problem
3. Evaluation Criteria
4. Variable Selection Procedures
5. **Example Processes**