Medical Statistics

Lab 10: Cleveland heart disease dataset

Introduction

In this final lab of the Medical Statistics course, you will continue working with the Cleveland heart disease dataset. You will apply **linear regression** to build a **prediction model** for resting blood pressure and use **logistic regression** to test an **inference-focused** research question on heart disease status. Through both parts, you will:

- Practice **backward elimination** in linear regression to arrive at a simpler predictive model.
- Demonstrate the ability to interpret **odds ratios** in a logistic regression setting, distinguishing unadjusted from adjusted associations.

Dataset description

The Cleveland heart disease dataset originates from the Cleveland Clinic Foundation and focuses on heart disease diagnosis. Key variables include:

- `age` (years)
- `sex` (1=female; 2=male)
- `cp` (chest pain type; 4 categories)
- `trestbps` (resting blood pressure in mm Hg)
- `chol` (serum cholesterol in mg/dl)
- `fbs` (fasting blood sugar >120 mg/dl; binary)
- `target` (heart disease status; 1 = heart disease, 2 = no heart disease)

Research questions

1. *Which demographic and clinical variables best predict resting blood pressure (`trestbps`)?*
    - You will use **backward elimination** to identify a final model and generate a **patient-level prediction**.

2. *Is sex (male vs. female) significantly associated with heart disease (`target`) status, after accounting for age, chest pain type, fasting blood sugar, and cholesterol?*

    - You will fit **simple** and **multivariable** logistic models to see whether the association between sex and heart disease remains after adjusting for additional factors.

Part A: Linear regression

A1. Initial model setup

- **Outcome variable**: `trestbps` (resting blood pressure)
- **Candidate variables**:
  - `cp` (chest pain type)
  - `age`
  - `chol`
  - `sex` (1 = female; 2 = male)
  - `fbs` (fasting blood sugar >120 mg/dl)

Fit a multiple linear regression model including all five variables as potential contributors to resting blood pressure. Summarize the model (coefficients, p-values, R-squared, etc.).

A2. Backward elimination

1. **Remove** the least significant variable (highest p-value).
2. **Refit** the model.
3. **Repeat** until all remaining terms are below the 10% significance threshold for variable inclusion (i.e., $p < 0.10$ for all remaining predictors).

A3. Final model summary

- Present the final set of variables. Interpret them in a predictive context (e.g., "An increase of 10 mg/dl in cholesterol corresponds to a 1.2 mm Hg change in resting blood pressure, holding other factors constant.").
- Evaluate model goodness-of-fit by performing a residual analysis.

A4. Patient-level prediction

- **Hypothetical patient**:
  - Age = 62
  - CP = 1 (typical angina)
  - Chol = 240 mg/dl
  - Sex = 1 (female)
  - FBS = 0 (120 mg/dl)
- **Plug** these values into your final regression equation to calculate the predicted resting blood pressure. If you removed any variables during your model selection, simply exclude them from the calculation.
- **Interpret** the predicted value. For example, "The model estimates this patient's resting blood pressure to be around 135 mm Hg."

Part B: Logistic regression

B1. Simple model (unadjusted)

Begin by ensuring that heart disease status is coded appropriately, for example 0 for no heart disease and 1 for heart disease if your software or analysis requires it. Then, run a simple logistic regression using sex as the sole explanatory variable, focusing on the odds ratio comparing males to females. Finally, interpret that odds ratio alongside its confidence interval, explaining how much higher or lower the odds of heart disease are for one sex relative to the other.

B2. Expanded model (adjusted)

Next, fit a logistic regression that includes sex, age, chest pain type, fasting blood sugar, and cholesterol. Pay particular attention to how the odds ratio for sex shifts (in terms of magnitude and significance) once these additional factors are taken into account. In your write-up, note any change in its magnitude relative to the simpler model and discuss plausible clinical reasons for any observed differences.