Medical Statistics

Assignment part 2: Cleveland heart disease dataset

Introduction

In this assignment, you will continue exploring the Cleveland heart disease dataset. The assignment consists of three parts:

1. **Part A: ANCOVA (Analysis of Covariance):** Investigate sex differences in resting blood pressure while adjusting for age.
2. **Part B: Logistic Regression:** Use a DAG to choose confounders and compare unadjusted vs adjusted estimates for cholesterol and heart disease.
3. **Part C: Building a Prediction Model:** Build a prediction model for maximum heart rate using backward elimination.

Dataset description

The Cleveland heart disease dataset includes various demographic, clinical, and diagnostic variables. Key variables for this assignment include:

- `age`: Age in years
- `sex`: Sex (1 = female; 2 = male)
- `cp`: Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- `trestbps`: Resting blood pressure (mm Hg)
- `chol`: Serum cholesterol (mg/dl)
- `fbs`: Fasting blood sugar > 120 mg/dl (0 = no; 1 = yes)
- `thalach`: Maximum heart rate achieved
- `target`: Diagnosis of heart disease (1 = heart disease; 2 = no heart disease)

Individualized dataset

R users must use the same 200-patient subset that they used for Assignment 1. If you need to recreate it, ensure you use the same random seed (the numeric part of your student or staff number). SPSS users can use the whole original dataset (N = 303) for this second assignment without further sampling.

> ❗ Important
>
> **Reminder for R Users:**

```
library(haven)
library(dplyr)

heart_data <- read_sav("datasets/heart_disease_cleveland.sav")
heart_data <- heart_data |> mutate(across(where(is.labelled), as_factor))

set.seed(123456) # Replace with your student number
heart_data <- heart_data[sample(nrow(heart_data), 200), ]
```

Part A: ANCOVA (Analysis of Covariance)

In this part, you will investigate the relationship between demographic factors and **resting blood pressure** (`trestbps`).

A1. Research Question

Research Question: *Is there a difference in resting blood pressure between males and females, after accounting for age?*

A2. Fitting the ANCOVA Model

Fit an **additive ANCOVA model** with `trestbps` as the outcome, and `age` and `sex` as predictors (this assumes the effect of age is the same for males and females).

Report the estimated difference in resting blood pressure between males and females, holding age constant, together with a 95% CI, and use the regression p-value to determine if `sex` is significantly associated with `trestbps` after adjusting for `age`.

A3. Model Evaluation and Assumptions

Assess whether the additive ANCOVA model is reasonable.

- **Assumption checks (additive model):** For the fitted additive ANCOVA model, use the **Q-Q plot** and **Residuals vs. Fitted plot** to assess normality, homoscedasticity, and linearity of the errors.
- **Interaction check (new model):** Fit an interaction model (`age * sex`) and use a Type III ANOVA table to test the `age:sex` term; state whether this provides evidence against the additive assumption.

Part B: Logistic Regression

In this part, you will investigate whether **cholesterol** is associated with the presence of **heart disease** (`target`), after adjusting for confounders identified in a DAG.

> Coding reminder (R users)
>
> For logistic regression in R (`glm(..., family = binomial)`), the outcome should be coded as 0/1. If `target` is a factor, create a numeric version, e.g.:
>
> ```
> heart_data <- heart_data |>
>   mutate(target01 = ifelse(target == "Heart Disease", 1, 0))
> ```
>
> Then use `target01` as the outcome in your logistic regression models.

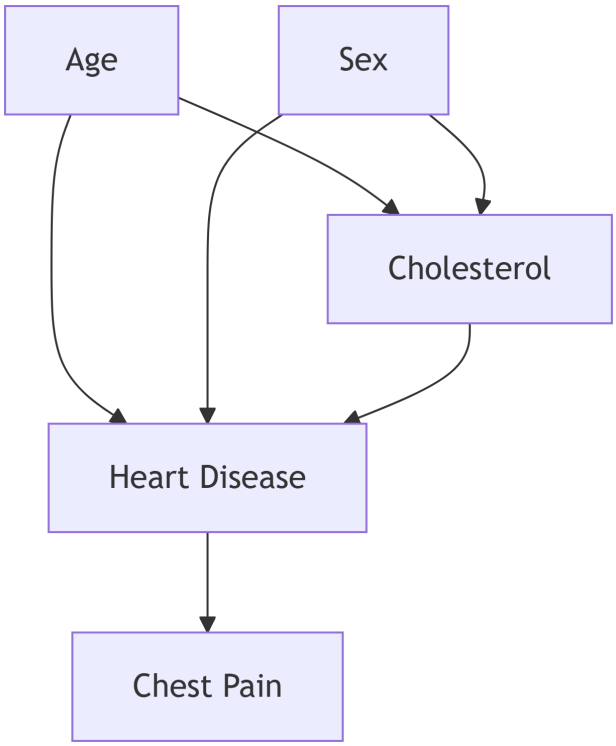
B1. Unadjusted Model

Fit a simple logistic regression model with `target` as the outcome and `chol` as the sole predictor.

- Report the **Odds Ratio (OR)** for a one-unit increase in cholesterol (or per 10 mg/dl if you rescale).
- Interpret the OR and its 95% confidence interval.

B2. Causal Modeling and Adjusted Model

Examine the following hypothetical causal diagram (DAG) for the relationship between cholesterol and heart disease, then use it to determine your adjustment set and fit the adjusted model.



Use the **"recipe" from Week 8** to analyze this diagram:

1. **Remove all arrows leaving the exposure (`chol`).** Are there any remaining paths between `chol` and `target`? (These are backdoor paths).
2. **Identify Confounders:** Which variable(s) must be adjusted for to "block" these backdoor paths?
3. **Adjusted Model:** Fit a multiple logistic regression model adjusting only for the confounder(s) you identified to estimate the **unconfounded effect** of cholesterol on heart disease.
4. **Comparison:** Put the unadjusted OR (B1) and adjusted OR (aOR) side by side. Say whether the OR changes, whether the CI widens or narrows, and what that suggests about confounding.

Part C: Building a Prediction Model

In this final section, you will develop a broader model to predict **maximum heart rate** (`thalach`), using `age`, `sex`, `cp`, `chol`, `fbs`, and `trestbps` as the candidate predictors.

Perform a manual **backward elimination** procedure:

1. Identify the predictor with the highest p-value that is above the significance threshold of **0.10** (using the Type III ANOVA table).
2. Remove that predictor and refit the model.
3. Repeat until all remaining predictors have a p-value < 0.10.

Present a table showing the removal steps (which variable was removed and its p-value at that step), and report the final model (coefficients, 95% CIs, and p-values).

Report instructions

Structure your report using the same headings as the assignment (**Part A**, **Part B**, **Part C**). Under each part, write a short results-focused summary and report the requested estimates (with 95% CIs and p-values) and any required figures/tables. Use the **Reporting Examples** below as a template for how to write up numerical results in sentences.

- **Part A (ANCOVA):** Report the adjusted male–female difference in `trestbps` (95% CI, p-value), and briefly state what the diagnostic plots and interaction test suggest.
- **Part B (Logistic regression + DAG):** State the adjustment set, report the unadjusted OR and adjusted OR for `chol` (95% CIs), and briefly comment on what the change suggests about confounding.
- **Part C (Prediction):** Include the backward-elimination table, report the final model (coefficients, 95% CIs, p-values), and briefly summarize which predictors remain.

Reporting Examples

Reporting an ANCOVA Result

"In an ANCOVA of exam score, we adjusted for study hours and compared instruction format. After accounting for study hours, online students scored 3.6 points lower than in-person students (95% CI: 1.1, 6.1), $p = 0.005$."

Reporting a Logistic Regression Result

"In the unadjusted model, students in the evening program had higher odds of course withdrawal than daytime students (OR = 1.85, 95% CI: 1.10, 3.10, $p = 0.020$). After adjusting for baseline GPA (identified as a confounder in the DAG), the association attenuated but remained (aOR = 1.50, 95% CI: 1.02, 2.40, $p = 0.041$)."

Submission Instructions

- **The report:** Submit your report as a Word or PDF document.
- **The analysis file:** Submit your R script (`.R`) or SPSS output/syntax files.

Downloads

Cleveland heart disease dataset (SPSS format)