Medical Statistics – Lab 8

SPSS version

Welcome to lab 8. In this lab, we will build prediction models using backward elimination and automated procedures, and we will practice reasoning with causal diagrams (DAGs).

Part 1: Building prediction models using backward elimination

In this part of the lab, we will build a prediction model for hospital length of stay (`los`) in patients with acute myocardial infarction. The dataset comes from the Worcester Heart Attack Study (WHAS) and includes data from 500 patients admitted in Worcester, Massachusetts in 1997, 1999, and 2001.

Key variables in the dataset include:

- `los`: Length of hospital stay (days, continuous outcome)
- `age`: Age at hospital admission (years)
- `gender`: Gender (0 = Male, 1 = Female)
- `hr`: Initial heart rate (beats per minute)
- `sysbp` and `diasbp`: Initial systolic and diastolic blood pressure (mmHg)
- `bmi`: Body mass index (kg/m^2)
- `cvd`: Presence of cardiovascular disease (0 = No, 1 = Yes)
- `sho`: Presence of cardiogenic shock (0 = No, 1 = Yes)

Step 1: Fit the initial linear regression model

Download the dataset from the Datasets menu (`whas500.sav`) and open it in SPSS.

Create an initial model for hospital length of stay (`los`) using the following predictors: `age`, `gender`, `hr`, `sysbp`, `diasbp`, `bmi`, `cvd`, `sho`. Run/summarize the model to inspect coefficients and p-values.

Use the **General Linear Model** procedure: `Analyze → General Linear Model → Univariate.`

Step 2: Eliminate the least significant predictor

To identify the least significant predictor, we use the Type III ANOVA table:

- Significance threshold: $p > 0.10$
- Remove the predictor with the largest p-value above this threshold.

The Type III ANOVA table is generated in the General Linear Model output.

Step 3: Repeat the steps

Iteratively remove the least significant predictor until all predictors have $p < 0.10$. At each step:

- Rerun the regression model
- Generate the Type III ANOVA table
- Remove the least significant predictor

Step 4: Final model

Present the final linear regression model:

- Report the final model, including regression coefficients and 95% CIs.
- Create residual plots to assess the model assumptions (normality, homoscedasticity, linearity).

> 95% CIs for regression coefficients
>
> In **Analyze → General Linear Model → Univariate**, click **Options** and request **Parameter estimates** (95% CIs are included automatically).

Part 2: Automated procedures for building prediction models (logistic regression)

In this part, we explore automated procedures for predictor selection in **logistic regression** prediction models. We use the same WHAS dataset but now focus on predicting in-hospital death (`dstat`: alive/dead) from candidate predictors.

SPSS: Backward selection (LR tests)

- Go to **Analyze → Regression → Binary Logistic**.
- Put `dstat` in **Dependent**.
- Add your continuous candidate predictors (e.g., `age`, `hr`, `sysbp`, `diasbp`, `bmi`) to **Covariates**.
- Also add any categorical candidate predictors (e.g., `gender`, `cvd`, `sho`) to **Covariates**, then click **Categorical...** and move them into **Categorical Covariates** so SPSS treats them as factors.
- Under **Method**, select **Backward: LR** (backward selection based on likelihood ratio tests).
- Click **Options...** and tick **CI for exp(B)** to obtain 95% CIs for the odds ratios.

Note (categorical predictors): Make sure you correctly specify categorical predictors via **Categorical...** (otherwise SPSS may treat their numeric codes as continuous). If you include categorical predictors with more than two categories and specify them via **Categorical...**, SPSS will handle the dummy coding internally and the **Backward: LR** procedure evaluates the predictor as a factor (i.e., removal is based on the overall significance of the factor, not separately for each dummy coefficient).

> **i** Related note (linear regression)
>
> For stepwise selection with a continuous outcome, SPSS implements this under **Analyze → Regression → Linear** (Method: Forward/Backward/Stepwise), but this procedure does not have a dedicated "factors" box—categorical predictors must be dummy-coded manually. In that situation, it is possible for some dummy variables to enter/leave the model while others do not. The **General Linear Model → Univariate** procedure does not offer the same automated stepwise selection.

> Question
>
> Inspect your final selected logistic regression model. Which predictors are retained in the prediction model?

Part 3: Causal diagrams

For each of the exercises below:

- Try solving the diagrams by hand by using the recipe from the lecture (see lecture slides on Brightspace)
- Check your answer using the DAGitty webtool

Exercise 1

In the graph depicted below, for which variables do you need to adjust to assess the unconfounded effect of E on O (there may be several possibilities)?

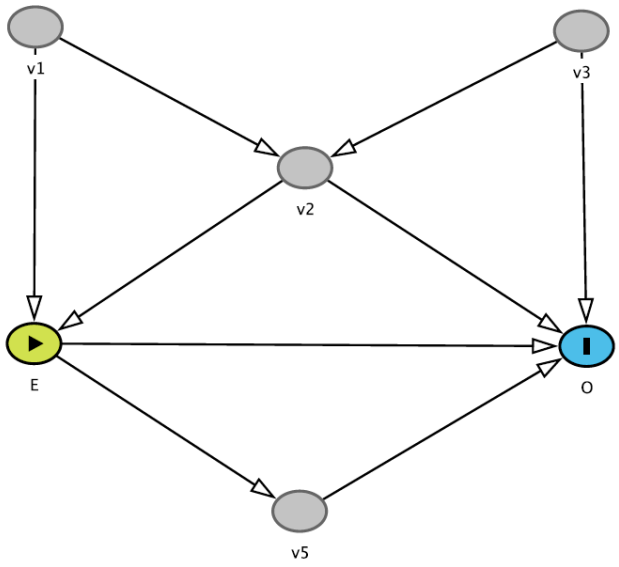

Figure 1: DAG exercise 1

Exercise 2

In the graph depicted below, what happens when you additionally adjust for **v5**?

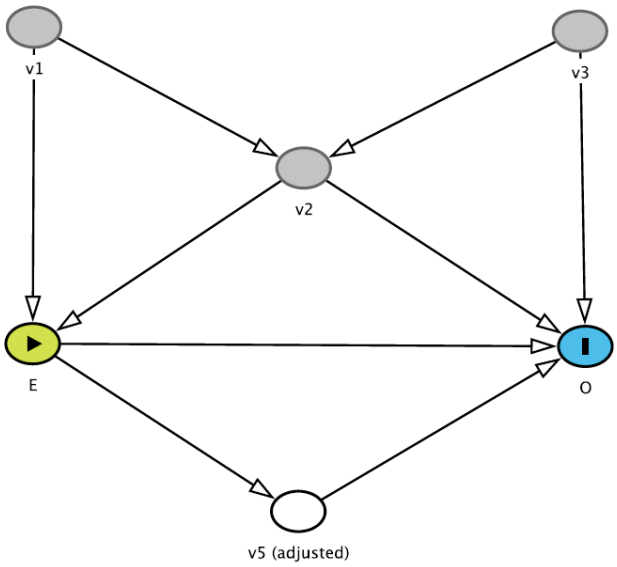

Figure 2: DAG exercise 2

Exercise 3

This diagram is slightly different: **v1** now is the exposure. For which variables do you need toadjust to assess the unconfounded effect of **v1** on **O**?

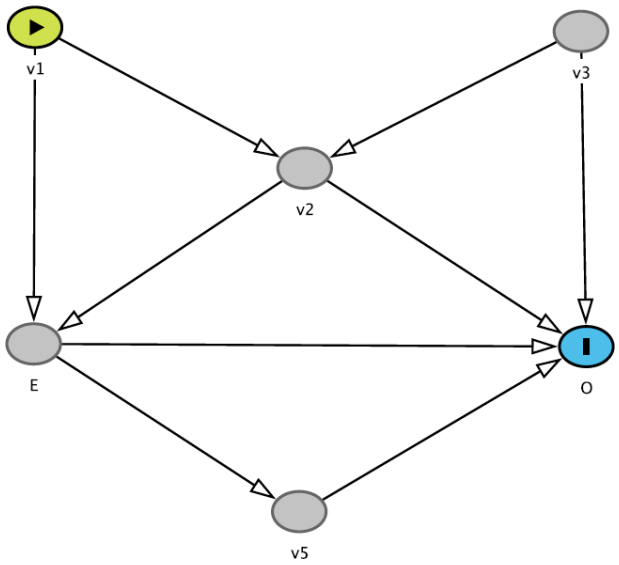

Figure 3: DAG exercise 3

Exercise 4

Now, **v2** is the exposure. For which variables do you need to adjust to assess the total unconfounded effect of **v2** on **O**?

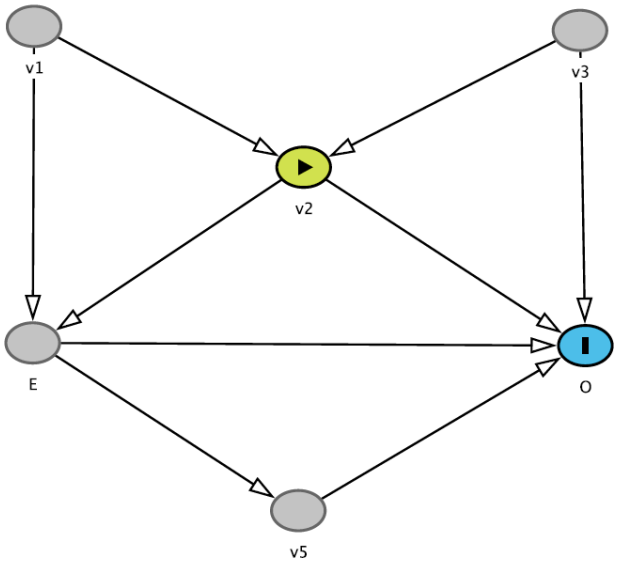

Figure 4: DAG exercise 4

Exercise 5

Back to the first DAG. However, **v2** is now unmeasured. Can we still obtain an unconfounded estimate of the effect of **E** on **O**?

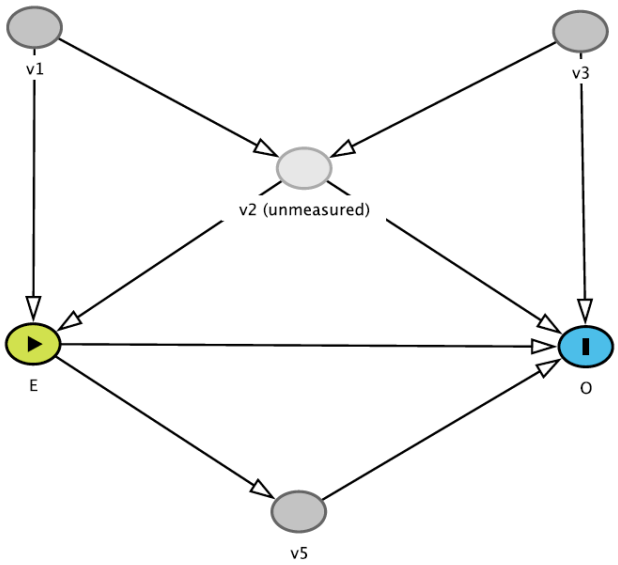

Figure 5: DAG exercise 5

Exercise 6

See the DAG below: you adjusted for **v5**. What would be the consequence of this action?

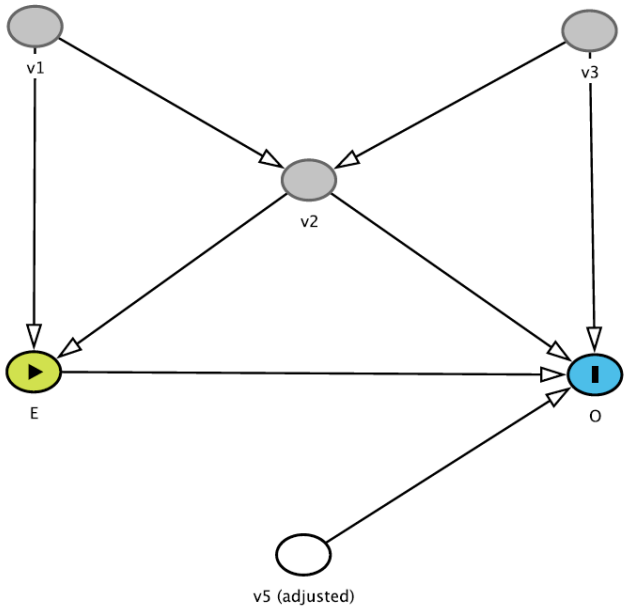

Figure 6: DAG exercise 6