Medical Statistics – Lab 8

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 R or SPSS. When using R, make sure that the categorical variables are correctly coded as factors.

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.

R instructions: use the `lm()` function to fit the model.

SPSS instructions: use the **General Linear Model** procedure, which can be accessed via `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.

R instructions: use the `Anova()` function from the `car` package to obtain the Type III ANOVA table (see previous lab).

SPSS instructions: 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:

- Summarize the remaining predictors and their coefficients
- Discuss how each variable contributes to predicting hospital length of stay

Create residual plots to assess the model assumptions (normality, homoscedasticity, linearity).

Part 2: Automated procedures for building prediction models

In this part, we explore automated procedures for predictor selection in regression models. These procedures can be useful when dealing with a large number of predictors. Both R and SPSS provide tools for these procedures, though SPSS has specific limitations with categorical variables in linear regression models.

SPSS: Stepwise selection

- SPSS offers tools for forward, backward, or stepwise selection:
  - Access these procedures via **Analyze → Regression → Linear**
  - Under **Method**, choose:
    * **Forward** for forward selection
    * **Backward** for backward elimination
    * **Stepwise** for a combination of both approaches
- SPSS automatically includes or excludes predictors based on significance levels

Important note for SPSS users: SPSS does not automatically handle categorical variables with more than two categories:

- You must manually create dummy variables for each category (excluding the reference category) using **Transform → Recode into Different Variables**
- SPSS treats each dummy variable as a separate predictor during stepwise procedures
- This means the overall contribution of the original categorical variable cannot be evaluated as a whole
- As a result, some dummy variables may be included or excluded independently, breaking the connection to the original variable

R: Automated model selection

In R, the `stepAIC` function from the `MASS` package allows for automated selection based on AIC (Akaike Information Criterion). Backward, forward, or stepwise selection can be specified using the `direction` argument:

```
library(MASS)
fit <- lm(los ~ age + gender + hr + sysbp + diasbp + bmi + cvd + sho, data = whas)
step_model <- stepAIC(fit, direction = "backward")
summary(step_model)
```

Exercise: Automated procedures vs manual model

1. Apply the automated backward elimination procedure:
   - **R users**:
     - Use the `stepAIC` function with `direction = "backward"`
   - **SPSS users**:
     - Use **Analyze → Regression → Linear**

   - Select **Backward** under **Method**

2. Compare the final model obtained from the automated procedure to the manually created model in Part 1:

   - Are the same predictors included in both models?
   - If not, what differences do you observe, and what might explain them?

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)?

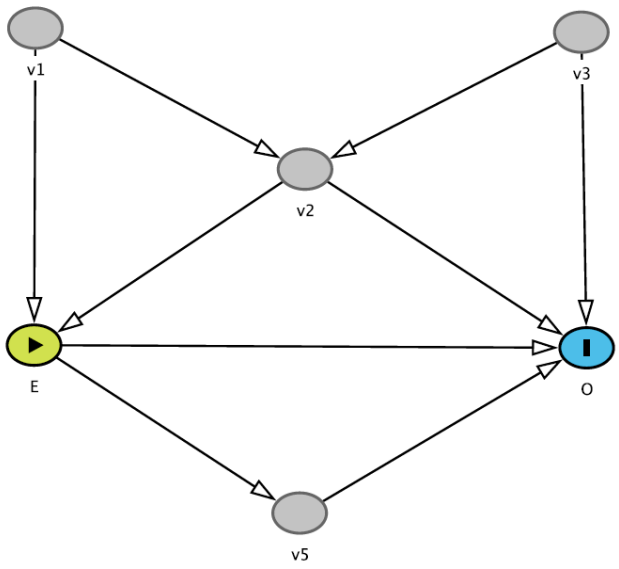

Figure 1: DAG exercise 1

Exercise 2

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

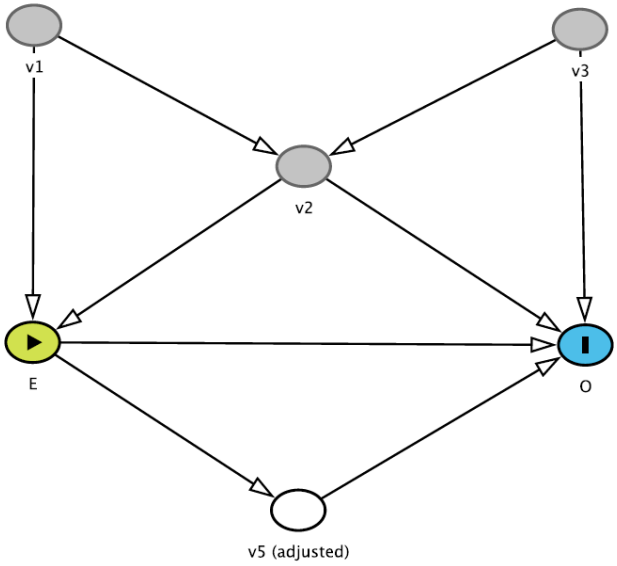

Figure 2: DAG exercise 2

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**?

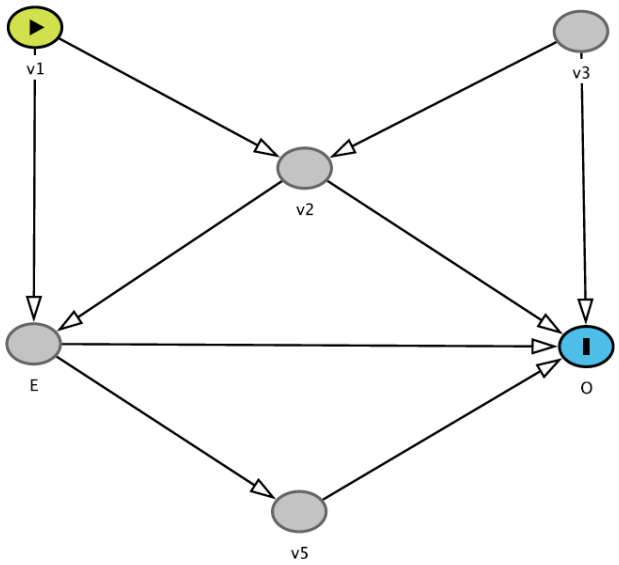

Figure 3: DAG exercise 3

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**?

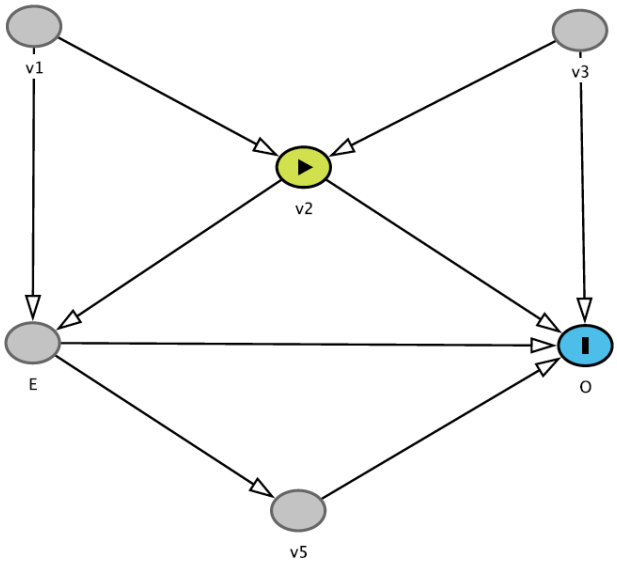

Figure 4: DAG exercise 4

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**?

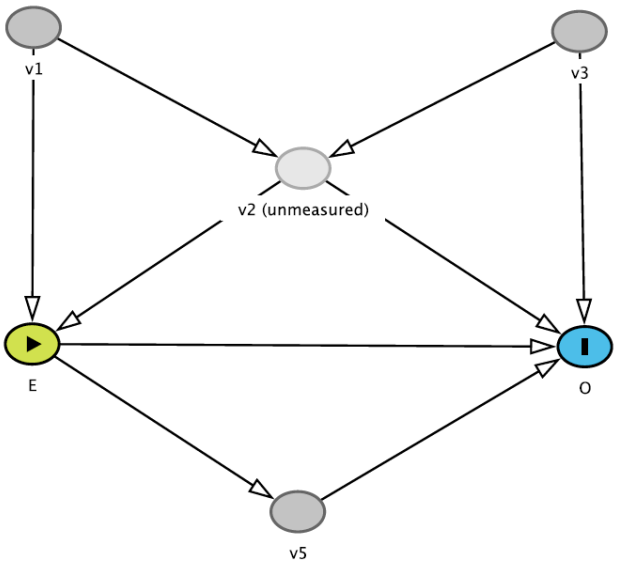

Figure 5: DAG exercise 5

Exercise 6

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

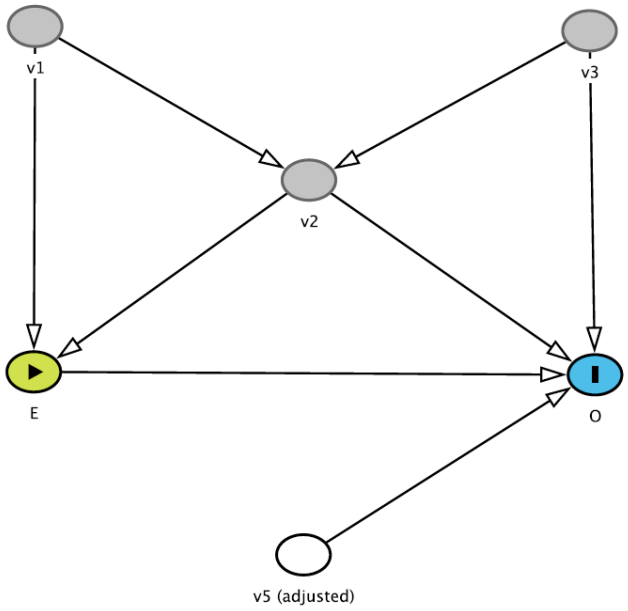

Figure 6: DAG exercise 6