10.10 Practical 10

Dataset required: `lbw.dta` (Stata webuse)

Introduction

In this practical we will use a number of different analysis strategies to build models with birth weight as the outcome using the same dataset introduced in practical 4. There are two parts to this session.

- In Part A we will consider ***causal*** investigation
- In Part B we will conduct a ***prediction*** investigation

We will use the same dataset in both parts. This The variables we will consider are described in the table below.

Variable	Description
`bwt`	Baby's birth weight (grams)
`age`	Mother's age (years)
`lwt`	Mother's weight at last menstrual period (in lbs)
`race`	Mother's race (1=white; 2=black; 3=other)
`smoke`	Mother's smoking status during pregnancy
`ptl`	Number of previous premature labours
`ht`	Mother's history of hypertension (0=None, 1=previous hypertension)
`ui`	Presence of uterine irritability in mother (1=No, 1=Yes)
`ftv`	Number of visits to physician during first trimester

Aims

The aims of this session are to:

1 Understand why different modelling approaches are needed according to the aims of the research question

2 Understand how to select variables for possible inclusion in a causal investigation

3 Understand how to select variables for possible inclusion in a prediction investigation

4 Be able to use stepwise, change in estimate, and MSE approaches to modelling

Part A. Causal investigation

First, imagine that our aim is to estimate the effect of maternal hypertension on birthweight adjusting for all relevant confounders. So we have a single exposure of interest and a number of other variables which could potentially confound the relationship between hypertension and birthweight. Clinical input suggests that presence of uterine irritability (`ui`) and number of visits to the physician (`ftv`) could be on the causal pathway between maternal hypertension and birthweight, so we will not consider these variables as potential confounders.

1 The dataset is available directly from Stata:

```
webuse lbw, clear
```

(a) Explore the dataset. Are there any variables with missing values?

One issue to look for is where there are very few observations in a particular level of a categorical variable. If there are, we can merge categories to avoid running into problems during regression analysis. Are there any variables you would consider re-categorising in this dataset?

(b) Investigate the associations of the potential confounders with the exposure and outcome variables, using tables and plots as you see fit.

(c) Compare the mean birthweight in the two groups of women defined by hypertension status.

2 (a) Fit a simple linear regression model with birthweight as the outcome and history of hypertension as the only explanatory variable.

(b) Next fit the full model including all covariates (re-categorised if necessary), except the two deemed inappropriate to include.

Discuss: Compare the coefficient for history of hypertension in the two models. What conclusions do you draw?

3 (a) Perform backwards selection, forcing history of hypertension to be included in the model and using a threshold of 0.2 for exclusion from the model.

Notice that Stata treats categorical variables as a series of dummy variables and considers them separately when deciding whether or not to include them in the model. This means that, to take the race variable as an example, that white women could be included in the model and black women excluded. This is clearly undesirable. To force Stata to include or exclude all categories together you should enclose the variable in parentheses in the command statement.

(b) Repeat the above using forwards selection, with a threshold of 0.2 for inclusion in the model.

Discuss: Compare the results from the forwards and backwards methods. What impact does changing the p-value criteria for retaining and adding variables make to the final selection?

4 (a) Use the change in estimates method (section 10.4.3 of the Notes) to perform the model building, using the backwards approach and with a threshold of a 10% change in the value of the coefficient for hypertension.

(b) Investigate the effect of changing the threshold. Do you always select the same variables for your final model? If you're working in a group, each person in the group could investigate a different threshold.

Discuss: Compare the results from the different variable selection methods. Working with a few colleagues decide which model you prefer. Then present the results from the unadjusted model and your preferred adjusted model in an appropriate table. Write a paragraph summarising your conclusions concerning the effect of maternal hypertension on birthweight, including noting any assumptions or caveats. If online post the paragraph in the Zoom chat.

Part B. Prediction investigation

Now suppose that we instead want to build a predictive model for birthweight.

5 In the investigation into the causal effect of maternal hypertension on birthweight, we excluded two variables because it was thought that they could be on the causal pathway between the exposure the outcome. What will affect your choice of the set of variables you will consider for a prediction investigation?

6 (a) We will include all of the variables listed in the table above. Run a model with all of the variables included.

```
regress bwt i.ht age lwt i.race i.smoke i.premature i.ui i.visits
```

(b) Investigate stepwise methods for building your prediction model, as above, investigating the impact of different thresholds and the difference between forwards and backwards approaches.

Discuss: With your colleagues discuss the models and agree on a preferred model from the stepwise investigations. Contrast the results obtained from this model with those obtained from the alternative strategy of including all predictor variables. Which model do you prefer?

Bonus question

7 If you have time, return to the casual investigation and use the MSE method (section 10.5 of the Notes) to perform the model building. Code is provided in the solutions to this Practical.