11.10 Practical 11

Dataset required: `nhanesglm.dta`

Introduction

In this practical we will explore model development and model checking, making use of the residuals described in the lecture. We will continue analysing the NHANES alcohol data introduced in the lecture. In this practical we will develop a model for whether a participant in the study has hypertension, defined as having a systolic blood pressure of 140 mmHg or above.

In this session we give you most of the Stata code you will need; we ask that you spend time on interpreting the output from each command, and in discussing conclusions and model choices with your colleagues.

Variable	Description
`gender`	1 = male, 2 = female
`ageyrs`	Age in (whole number of) years
`bmi`	Body Mass Index (kg/m^2)
`sbp`	Systolic blood pressure (mmHg)
`ALQ130`	Average number of alcoholic drinks per day in past year
`alccat`	Average number of alcoholic drinks per day in past year: 1 = 1 drink per day 2 = 2-5 drinks per day 3 = 6+ drinks per day

Aims

The aims of this session are to:

- Understand how to assess model fit for logistic regression models
- Interpret residual plots to assess if explanatory variables are modelled correctly

Analysis

Load the data and explore the variables.

Before we start, we need to generate a new binary variable to indicate whether each observation has hypertension (1) or not (0).

```
gen ht = (sbp>=140)
```

Univariable model

1 Fit a logistic regression model relating the log odds of hypertension to age using the glm command.

```
glm hypertension ageyrs, family(binomial)
```

We will explore various options with the predict command to investigate different ways to check if the model provides a good fit to the data.

First, we will obtain the predicted probability of hypertension for each person in the study. This uses the option mu, the expected value of Y.

```
predict pr1, mu
```

Examine these probabilities. We can also look at the average predicted probability among people without hypertension and with hypertension. We would hope that the mean will be close to zero in the first group and close to one in the second group.

```
bysort ht:  sum pr1
```

2 To obtain the *i*ndividual Pearson standardized residuals we use the following:

```
predict sp1, pearson standardized
```

Examine the distribution of these residuals. What is their mean and variance? Are they normally distributed?

(a) The first plot we will look at is an index plot. This can be obtained using

```
gen id = _n
scatter sp1 id
```

Discuss: What do you notice from this plot? Are there any outliers you would want to investigate? Can you explain why there are no residuals with a value between 0 and about 0.8?

(b) To plot the residuals against the linear predictor we can use the following: predict xb, xb

```
lowess sp1 xb, yline(0)
```

(c) We can plot the residuals against age using:

```
lowess sp1 ageyrs, yline(0)
```

Discuss: Why does the plot appear to only show a maximum of two residuals at each age?

In seeking to understand this you may find it helpful to 'jitter' the points a little.

```
lowess sp1 ageyrs, yline(0) jitter(3)
```

3 Since this is a relatively large dataset and age is only recorded to the nearest year we can consider a grouped approach.

(a) There are a number of ways of converting the individual patient data to grouped data. One way is as follows.

```
egen n=count(ht), by(ageyrs)
egen r=sum(ht), by(ageyrs)
egen pickone=tag(ageyrs)
```

The first two lines generate the denominator and numerator, respectively, for each group by age. The third line selects one observation per group, which is marked by the variable pickone taking the value 1; for all other rows pickone takes the value of 0. Browse the data to make sure that you understand the new variables.

(b) Now refit the model to the data in the grouped form; when we do this we must restrict the dataset to just one row per group using "`if pickone==1`". Then

obtain the grouped standardised Pearson residuals from this model.

```
glm r ageyrs if pickone==1, family(binomial n)
predict grp1, pearson standardized
```

Plot these against age.

```
lowess grp1 ageyrs if pickone==1, yline(0)
```

4 If the model is fitted in Stata using the `logit` or `logistic` command then covariate pattern residuals can be obtained using the option "`rstandard`":

```
logit ht ageyrs
predict cp1, rstandard
lowess cp1 ageyrs if pickone==1, yline(0)
```

Note here that although we must fit the data on all observations, when we plot the residuals we must again restrict it to one per group to allow the lowess function to appropriately smooth the curve.

Discuss: Compare the covariate pattern residuals (and plots) to those obtained from the grouped analysis (in Q3) and the individual residuals in Q2. What do you notice?

Discuss: From looking at all of the plots, what do you conclude about the functional form in which age should be included into the logistic regression? In which plot do you find it easiest to discern patterns?

Multivariable models

5 Add the square of the age variable to the linear predictor for the logistic regression model. Use residual plots analogous to those above to explore the fit of this model.

Discuss: What do you conclude about the most appropriate way to include age in this model?

6 Use individual-level residuals from the model in 5 to examine whether BMI ought to be included in the model, and depending on what you find, continue with your previous model or add BMI. In the latter case, generate new residuals and assess if you have included BMI using the most appropriate functional form.

7 (optional) Explore whether gender should be included in the model, including whether or not the other covariates already included interact with gender in their effects on the log odds of hypertension.

Discuss: What is your final model? Have you included the same variables in the same way as your colleagues?

8 (optional) Based on your final model, use margins to calculate fitted probabilities for an individual aged 60 years, at BMI values from 20 to 40 in increments of 5, separately for men and women, and plot the resulting values.