12.7 Practical 12

Dataset required: `nhanesglm.dta`

Introduction

There are three sections to this practical session.

The first section is a data analysis, looking at the predictive power of logistic regression models for hypertension.

The second section uses simulated data to explore the relationship between odds ratios, pseudo R^2 and the area under the ROC curve.

In the final optional section we ask you to prove a result which was stated in the notes, relating to the value of the AUC.

Aims

Understand how to assess the performance of a logistic regression model using

- model predictions
- pseudo R^2
- area under the ROC curve

Part A. Analysis of hypertension

1 Load the nhanesglm.dta dataset. As in the previous practical we will analyse hypertension as a binary variable, so first create a new variable for hypertension, defined as having systolic blood pressure of 140 mmHg or above. Tabulate the new variable.

2 Fit a logistic regression model for hypertension with BMI entered as a single linear covariate.

(a) What is the reported pseudo R^2?

(b) Use the predict command to generate a new variable containing the fitted probabilities from this model.

(c) Use the `estat gof` command to obtain observed and predicted values, according to deciles of predicted risk.

Discuss: How would you summarise the association between BMI and hypertension? What do you conclude about the model's performance from the predicted values and the pseudo R^2?

3 Add gender and age to your model, and again generate a variable containing the predicted probabilities from this model.

Plot the two predicted probabilities against each other.

4 Essentially the same comparison of the ranges of the fitted probabilities from the two models can be made by constructing predictiveness curves (section 12.5 in notes) for each of the two models, displaying both on the same graph. First create the new variables:

```
egen rank1 = rank(pr)
egen rank2 = rank(pr2)
egen num = max(rank1)
gen cent1=100*(rank1-0.5)/num
gen cent2=100*(rank2-0.5)/num
```

And then create the plot:

```
#delimit ;
twoway (line pr cent1, sort) (line pr2 cent2, sort lpattern(dash))
,
xtitle("Risk percentile") ytitle("Hypertension risk")
legend(order(1 "prediction using BMI alone"
2 "prediction using BMI, age and gender") rows(2))
;
#delimit cr
```

(The "`#delimit ;`" tells Stata that the end of each command is marked by a semi-colon. The "`#delimit cr`" tells Stata to again accept a carriage return as the end of line marker. This allows us to split a long command over several lines, to make it easier to read).

Discuss: What do you conclude about the two models from these plots?

5 Plot the ROC curve for the model including age and gender as well as BMI.

Discuss: What do you conclude regarding the model's discrimination ability from the ROC curve?

6 In the session 11 practical we investigated more complex models for the risk of hypertension. Return to the final model from that practical and use the techniques explored above to investigate the improvements in predictiveness and discrimination from using this model.

Part B. Simulation

To further explore the concept of explained variation and discrimination in binary regression models we will simulate some data, so that we know what the true conditional distribution of the outcome given the covariates is.

7 Use the following code to clear Stata's memory, set the number of observations (to 10,000), set Stata's random number seed mechanism (so we all obtain the same answers), and generate a covariate x from a Bernoulli distribution with probability of 1 equal to 0.5:

```
clear
set obs 10000
set seed 91413
gen x=(runiform()<0.5)
```

8 Next, add to your do file code to randomly generate Y , such that

$$P(Y = 1|x = 0) = 0.3$$

and

$$P(Y = 1|x = 1) = 0.7$$

Then fit the logistic regression model for y on x, and note the pseudo R^2 and AUC values.

9 Re-run your code from the previous question, but with

$$P(Y = 1|x = 0) = 0.1$$

and

$$P(Y = 1|x = 1) = 0.9$$

Again, note the pseudo R^2 and AUC values.

10 Re-run one more time with

$$P(Y = 1|x = 0) = 0.01$$

and

$$P(Y = 1|x = 1) = 0.99$$

Discuss: From these results, what do you conclude about explained variation in logistic regression models, and the pseudo R2 and AUC measures in particular, as the strength of association increases?

Part C. Theory (optional)

Prove the result stated in the lecture notes in section 12.4.3 regarding the area under the ROC curve, AUC:

$$AUC = P(\pi_i > \pi_j | y_i = 1 \ \& \ y_j = 0)$$

for two randomly selected subjects, one with $y_i = 1$ and the other with $y_j = 0$.