Homework 7: Diabetes

36-402, Advanced Data Analysis

Due at the start of class, 22 March 2011

A classic data set for classification problems, logistic regression and related methods comes from a study of the correlates of diabetes among the Pima Indians of Arizona, collected as part of a long-term study to understand why the Pima, like many other Native American groups, suffer from a much higher rate of diabetes than other populations in the US. (For background on the study, and the issue, see `http://diabetes.niddk.nih.gov/dm/pubs/pima/`.) Our version of the data is the data set `pima` in the package `faraway`.[1] It contains information of 768 adult Pima women, some but not all of whom have diabetes. See `help(pima)` for a description of the variables. Note that the column named `diabetes` indicates how much of a history of diabetes there was *in the woman's family*; it is the last column, `test`, which indicates whether the or not the woman herself is diabetic.

1. (10 points) Make graphic and numerical summaries of the data. If there are any obvious irregularities in the data, describe them, say why you think they are irregularities, and remove them as appropriate.

2. (20 points) Fit a logistic regression model to predict diabetes, using all the other variables as inputs. What are the estimated coefficients?

3. (10 points) What is the probability of having diabetes for a woman who has been pregnant twice, has a glucose concentration of 99, a diastolic pressure of 64, 22 mm of tricep thickness, an insulin level of 76, a BMI of 26, a diabetes "pedigree function" of 0.25, and is 30 years old. Give a 95% confidence interval for this prediction, assuming the model is correctly specified.

4. (10 points) How do the odds of having diabetes change for a woman who moves from the third quartile of the BMI distribution to the first quartile, with all else held constant? Give a 95% confidence interval for the difference in odds, assuming the model is correct specified.

5. (20 points) Do women with diabetes have higher diastolic blood pressure than women without diabetes? Is the blood pressure coefficient significant in your model? Explain why the answers to these two questions are actually compatible.

[1] This homework is in fact based on problem 3 in chapter 2 of Faraway's textbook.

6. (10 points) Describe how you can check whether this model fits the data.

7. (20 points) Does the model fit the data?

8. (10 points, extra credit) Use bootstrapping to find confidence intervals for the coefficients from question (2), the predicted probability in question (3), and the difference in odds in question (4). Compare them to your earlier answers, and explain how this relates to your findings in question (7).