Homework 11

Advanced Methods for Data Analysis (36-402)

Due Friday April 26, 2019 *at 3:00pm*

1 Diabetes amongst Native Americans

The data set we will use for this problem contains information on 768 female Pima people from Arizona. It is posted in the file `pima.csv`. There are nine variables, including a test for the presence of diabetes:

pregnant: Number of times pregnant

glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test

diastolic: Diastolic blood pressure (mm Hg)

triceps: Triceps skin fold thickness (mm)

insulin: 2-hour serum insulin (mu U/ml)

bmi: Body mass index (weight in kg/(height in metres squared))

diabetes: Diabetes pedigree function

age: Age (years)

test: test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

We will model the response `test` using logistic regression on all or some of the other measured variables.

The `glm` function in *R* will fit a logistic regression. For example, suppose that we want to use only `bmi` and `insulin` as predictors. Then,

```
glm(test~bim+insulin,family=binomial,data=pima)
```

will return a logistic regression fit with maximum likelihood estimates of the parameters.

(a) Compute summaries and create plots of the data to look for patterns. Comment on any patterns and/or relationships you see. In particular, you might notice that almost every variable has a minimum value of 0. For some variables (such as `test` or `pregnant`) a value of 0 is easy to understand. What could a value of 0 for `bmi` or `diastolic` or `triceps` mean? Replace all 0 values of the following variables by NA: `glucose`, `diastolic`, `triceps`, `insulin`, and `bmi`. Create a new `data.frame` containing only those observations that have no NA values. Use this reduced `data.frame` for the rest of the questions. There should be 392 observations in the resulting `data.frame`.

(b) Start by using all eight of the predictors in a logistic regression model. Call this Model1. Report a summary of the fit. Does it appear that all of the variables are contributing to the fit?

(c) Fit a model in which no predictors appear, that is, every observation has the same distribution for `test` (the `formula` for `glm` is `test~1`). Call this Model2. Test whether Model1 is a significant improvement on Model2.

(d) Do women with signs of diabetes have higher 2-hour serum insulin values? Is the `insulin` coefficient significant and positive in Model1? Explain why these answers are not contradictory.

(e) Fit a third model by running backward elimination on Model1. Model3 will choose to stop eliminating when *AIC* is minimized (the default for `step`.) To fit Model3, run the `step` function with the argument `direction="backward"` as in the demo called Demo_10_2_logreg.R. **NOTE:** For this part and all other parts of the assignment that involve the `step` command, *DON'T* include the spew of backward elimination output in your write-up. If some part of that output is useful to your analysis, extract it separately. The argument `trace=0` will suppress printing of the intermediate output. Compare Model3 to Model1 using a deviance test. Which model do the tests in this part and part (c) suggest describes the data best?

(f) One of the models compared in part (e) was chosen by backward elimination. The deviance test does not take into account the fact that one of the models being compared was chosen by backward elimination. As a result, the nominal asymptotic χ^2 distribution of the deviance test statistic under the null hypothesis might be incorrect. Perform a bootstrap analysis to determine the appropriate distribution of the deviance test statistic for comparing Model1 to the result of backward elimination with *AIC* penalty. Assume that the "true" model is the model that uses the predictors chosen by backward elimination in part (e) and called Model3. Use a parametric bootstrap in which bootstrap sample b is drawn as follows: For each predictor vector, x_i, let $Y_{b,i}$ have the Bernoulli distribution with probability of success equal to what the Model3 fit predicts. For each b, the Bernoulli random variables $Y_{b,1}, \ldots, Y_{b,n}$ are independent, one for each predictor vector. Repeat the sampling $B = 1000$ times. Make sure to use the argument `trace=0` in the `step` command to avoid seeing the intermediate output from 1000 backward eliminations. (The bootstrap will take a few minutes.)

(g) Consider a Pima woman who has been pregnant 3 times, has a glucose concentration of 103, diastolic pressure of 70, 29.2mm of tricep thickness, 2-hour serum insulin level of 160, body mass index of 32.4, diabetes pedigree function of 0.6, and is 32 years old. Predict her `test` result by computing the probability of positive `test` result based on Model3. Also give a 90% confidence interval for the Model3 probability of a positive test result.

(h) Consider a Pima woman whose predictors are all the same as those of the woman in part (h) except that her `diabetes` pedigree function is 0.25. How much different are the log-odds of $Y = 1$ for these two women? Give a point estimate as well as a 90% confidence interval for the difference. Base all answers on the fitted Model3.

(i) The tests in parts (c) and (e) give us some idea about how much better one model fits than another between the three models involved. But does any of them fit well or are they all garbage? One way to get at this question is by grouping observations by their probability (of $Y = 1$) estimates and seeing whether the fractions of $Y = 1$ cases are close to the estimated probabilities. A model is called *well-calibrated* if the estimated probabilities (also known as the `fitted.values`) are close to the fractions of $Y = 1$ cases. Model2 is trivially well-calibrated, since it has only one estimated probability for all observations, and that estimated probability equals the fraction of $Y = 1$ cases. (Verify that claim based on your earlier analysis.) The other two models have 392 different estimated probabilities spread from 0 to 1. Checking how well-calibrated are these models can be thought of as a smoothing exercise. One could use a spline, a kernel regression, and/or K-nearest neighbors to get a mapping from estimated probabilities to fraction of $Y = 1$ cases. Run all three of these smoothers to smooth `test` with `fitted.values(Model3)` as the predictor. (Let `npreg` use bandwidth 0.075. For the `smooth.spline` use `df=10`. Use $K = 35$ for nearest neighbors.) On a single set of axes, plot the three sets of fitted values against `fitted.values(Model3)` to see how well-calibrated Model3 is. Ideally the plots should all look like the line $y = x$. Of course they do not look like the line $y = x$. Comment on how well-calibrated Model3 seems to be. Do the same thing with Model1. Does either of these models appear to be noticeably better calibrated than the other?