MAST30027: Modern Applied Statistics

Week 4 Lab

1. The `cornnit` dataset in the `faraway` package contains data on the effect of nitrogen on the yield of corn. Fit a gamma regression to this data, usnig the `glm` command. You will need to pay attention to the choice of link function (inverse, identity or log), and consider transforming the predictor variable (your first step should be to plot the data).

   Note: For a linear model A nested within linear model B, under the null hypothesis that model A is correct we have

   $$\frac{(D^A - D^B)/s}{X^2/(n-p)} \sim F_{s,n-p}$$

   where we have n observations, A has $p - s$ paramters and B has p parameters.

   In R X^2 (Pearson's chi-squared) is calculated using the fitted model B.

   For other GLM's this distributional result only holds approximately, but it can still be used for comparing models. In particular it can be used to compare gamma models.

   (a) Extract the Pearson residuals from the fitted model using the `residuals` function, then use them to estimate the dispersion parameter. Check that your answer agrees with the summary output from your model.

   (b) Suppose your fitted model is `gmod`, then the command `anova(gmod, test="F")` will compare your model against the null model, using an F test. Using the deviances and dispersion estimates reported by `summary(gmod)`, check that the F statistic reported by the `anova` function is correct.

   (c) Now do some diagnostic plots. Can you identify a potential outlier? [Solve this problem in week 5 after learning diagnostics for GLMs.]

   (d) Fit a linear model to the `cornnit` data.
   Which do you prefer, the linear model or the gamma model, and why?

2. The `dvisits` data in the `faraway` package comes from the Australian Health Survey of 1977–78 and consist of 5190 observations on single adults, where young and old have been oversampled.

   (a) Build a Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

   (b) Plot the response residuals against the fitted values. Why are there lines of observations on the plot?

   (c) Use backward elimination with a critical p-value of 5% to reduce the model as much as possible.

   (d) What sort of person would be predicted to visit the doctor the most under your selected model?

   (e) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

   (f) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how the Gaussian and Poisson models differ.