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).

   (a) Extract the Pearson residuals $\frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$ from the fitted model using the `residuals(glmfit, type="pearson")`, then use them to estimate the dispersion parameter ϕ. Check that your answer agrees with the summary output from your model. You can find “Dispersion parameter for Gamma family taken to be . . .” in the summary output.

   (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.

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` as possible predictor variables. Select a model using stepwise model selection based on the AIC. Considering the deviance of the selected model, does this model fit the data? (i.e., is this model adequate?)

   (b) Extract the response residuals $y_i - \hat{\mu}_i$ from the fitted model using the `residuals(glmfit, type="response")`, then plot the response residuals against the fitted values. Why are there lines of observations on the plot?

   (c) Starting from the Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` as possible predictor variables, reduce the model as much as possible using backward elimination with a critical p-value of 5%.

   (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.