MAST90104: A First Course in Statistical Learning

Week 11 Lab and Workshop

Practical questions

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, using the `glm` command and store the model fit as `gmod`, using the canonical link function. Hint: **consider transforming the predictor variable first**.

   (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) The command `anova(gmod, test="F")` will compare your model against the intercept-only 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?

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

2. The `Articles` dataset in the `Rchoice` package contains data on the publication counts (`art`) of research scientists and their respective gender (`fem`), marital status (`mar`), number of children (`kid5`), prestige of graduate program (`phd`), and the number of articles published by their mentors (`ment`).

   (a) Fit a Poisson regression model with `art` as the response variable using the canonical link function.

   (b) Perform stepwise selection using AIC criterion starting from the full Poisson regression model with all predictors. Write down the equation of your your final regression model.

   (c) The `glm.nb` command in the `MASS` library fits the following model: $y_i \sim \text{NegBin}(\mu_i, k)$, where $\mu_i > 0$, $k > 0$, and

   $$p(y_i = y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^y \left(\frac{k}{\mu_i + k}\right)^k, \quad y = 0, 1, 2, \ldots$$

   Here, $E(y_i) = \mu_i$ and $\text{Var}(y_i) = \mu_i + \mu_i^2/k$. The default log link function is $g(\mu) = \log(\mu)$. Using `glm.nb`, fit a Negative Binomial regression model with `art` as the response variable using the log link function.

   (d) Perform stepwise selection using AIC criterion starting from the full Negative Binomial regression model with all predictors. Write down the equation of your your final regression model.

   (e) Which model would you prefer – the Poisson or Negative Binomial? Justify your answer with a suitable residual plot.

Workshop questions

1. Refer to Q2(c) from practical. By consider k as fixed, show that the negative binomial distribution belongs to the exponential family.

2. Prove the Lemma in page 17 of Lecture 8.

3. Refer to Q2 from practical. The following R output details the fit of two Poisson regression models (with some details redacted).

```
> mod.pois.workshop <- glm(art ~ fem + ment, family = poisson,data = Articles)
> summary(mod.pois.workshop)

Call:
glm(formula = art ~ fem + ment, family = poisson, data = Articles)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.34909    0.04191   8.329  < 2e-16 ***
fem         -0.18445    0.05235  -3.523 0.000426 ***
ment         0.02510    0.00193  13.005  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1657.0  on 912  degrees of freedom
AIC: 3330.7

Number of Fisher Scoring iterations: 5

> mod.pois.workshop2 <- glm(art ~ fem, family = poisson(link = "inverse"),data = Articles)
> summary(mod.pois.workshop2)

Call:
glm(formula = art ~ fem, family = poisson(link = "inverse"),
data = Articles)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.53118    0.01742  30.499  < 2e-16 ***
fem          0.14895    0.03241   4.595 4.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: <Redacted>  on 914  degrees of freedom
Residual deviance: 1794.4  on 913  degrees of freedom
AIC: 3466.1

Number of Fisher Scoring iterations: 7
> qchisq(0.95,1)
[1] 3.841459
```

(a) Write down the equation of the two fitted regression models, including the MLEs of their respective coefficients.

(b) Can we use a likelihood ratio test to compare `mod.pois.workshop2` against `mod.pois.workshop`? If yes, compute the test statistic, write down its null distribution, and state your conclustion. If no, suggest an alternative approach to compare the two models based on the above output.

(c) Compare `mod.pois.workshop2` against the intercept-only model.