SS9155 - Assignment 4 - 250620601

Ravin Lathigra

2019-03-26

Chapter 8 | Question 3

```
### Import Data
data(gala,package="faraway")

data <- gala
```

Q3. Part a

Fit a Poisson model to the species response with the five geographic variables as predictors. Do not use the endemics variable. Report the values of the coefficients and the deviance.

Table 1 and Table 2 show the coefficients and deviance for a poisson model with geographical features used to model the response, Speciesm respectively.

```
gala_pois <- glm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, family = poisson, data= data )

gala_pois.coef <- as_tibble(data.frame(gala_pois$coefficients) %>% rownames_to_column() %>%
                                        rename(Feature = 'rowname',
                                               Coefficient = gala_pois.coefficients))

gala_pois.resdev <- as_tibble(data.frame(gala_pois$deviance) %>% rownames_to_column() %>%
                                        rename(Deviance = gala_pois.deviance) %>%
                                        mutate(DF = 24) %>%
                                        dplyr::select(-rowname))

gala_pois.nulldev <- gala_pois$null.deviance
gala_pois.nulldf <- gala_pois$df.null



gala_pois.deviance <- rbind(gala_pois.resdev,data.frame(Deviance = gala_pois.nulldev, DF = gala_pois.nul
  rename(Type = rowname) %>%
  mutate(Type = ifelse(Type == 1,"Residual","Null"))

kable(gala_pois.coef,
      align = rep("c",ncol(gala_pois.coef)),
      booktabs = T,
      caption = "Poisson Coefficients") %>%
  kable_styling(position = "center", latex_options = "hold_position")
```

```
kable(gala_pois.deviance,
      align = rep("c",ncol(gala_pois.deviance)),
      booktabs = T,
      caption = "Poisson Deviance") %>%
  kable_styling(position = "center", latex_options = "hold_position")
```

Table 1: Poisson Coefficients

Feature	Coefficient
(Intercept)	3.1548079
Area	-0.0005799
Elevation	0.0035406
Nearest	0.0088256
Scruz	-0.0057094
Adjacent	-0.0006630

Table 2: Poisson Deviance

Type	Deviance	DF
Residual	716.8458	24
Null	3510.7286	29

Q3. Part b

For a Poisson GLM, derive , d /d , V( ) and the weights to be used in an iteratively fit GLM. What is the form of the adjusted dependent variable here?

Considering a response that is poisson distributed its general form is:

$$f(y|\theta,\phi) = \frac{e^{-\mu}\mu^{y}}{y!}$$

where,

$$\theta = log(\mu), \phi = 1, a(\phi) = 1, b(\theta) = e^{\theta}, c(y,\theta) = -log(y)$$

$$\eta = log\mu$$

$$\frac{d\eta}{d\mu} = \frac{d(log\mu)}{d\mu} = \frac{1}{\mu}$$

$$V(u) = \frac{b\theta}{w} = \mu$$

$$w = (\frac{d(log\mu)}{d\mu})^{2}V(\mu) = \mu$$

adjusted dependent variable takes the form:

$$z = \hat{\eta} + (y - \hat{\mu})\frac{d(log\mu)}{d\mu}$$

```
y <- data$Species

#use y for initial guess of mu
mu <- y

eta <- log(mu)

vu <- mu

z <- eta + (y-mu)/(mu)

w <- mu
```

Q3. Part c

Using the observed response as initial values, compute the first stage of the iteration, stopping after the first linear model fit. Compare the coefficients of this linear model to those found in the GLM fit. How close are they?

After the first stage of iteration, the coefficients of the linear model are as follows:

```
lmod <- lm(z~Area+Elevation+Nearest+Scruz+Adjacent, weights=w, data=data)
coef(lmod)
```

```
  (Intercept)          Area     Elevation       Nearest         Scruz      Adjacent
 3.5191545412 -0.0005298484  0.0031643557  0.0025188990 -0.0037899780 -0.0006623523
```

The question is *"how are these coefficients to those from the glm fit?"* the following output expresses the ratio of the coefficients from the linear model to those from the glm.

```
coef(lmod)/coef(gala_pois)
```

```
(Intercept)        Area   Elevation     Nearest       Scruz    Adjacent
  1.1154893   0.9136218   0.8937358   0.2854092   0.6638111   0.9989762
```

Notice that while some coefficients are similar i.e. ratios are close to 1, a few have drastically different coefficients.

Q3. Part d

Continue the iteration to get the next and . Use this to compute the current value of the deviance. How close is this to the deviance from the GLM?

After the next stage of iteration, the coefficients of the linear model are as follows:

```
eta <- lmod$fit

mu <- exp(eta)

vu <- mu

z <- eta + (y-mu)/(mu)

w <- mu
```

```
lmod2 <- lm(z~Area+Elevation+Nearest+Scruz+Adjacent, weights=w, data=data)
coef(lmod2)
```

```
  (Intercept)          Area     Elevation       Nearest         Scruz      Adjacent
 3.2102594447 -0.0005651969  0.0034606226  0.0077171134 -0.0052400871 -0.0006604828
```

"How are these coefficients to those from the glm fit?" the following output expresses the ratio of the coefficients from the second interation of the linear model to those from the glm.

```
coef(lmod2)/coef(gala_pois)
```

```
(Intercept)        Area   Elevation     Nearest       Scruz    Adjacent
  1.0175768   0.9745734   0.9774130   0.8744038   0.9177964   0.9961567
```

Notice that the ratio of the coefficients have improved materially from the previous iteration which is as we would expect i.e. quick convergence.

Q3. Part e

Compute one more iteration of the GLM fit, reporting the next calculation of the coefficients and deviance. How close are these to target now?

After the next stage of iteration, the coefficients of the linear model are as follows:

```
eta <- lmod2$fit

mu <- exp(eta)

vu <- mu

z <- eta + (y-mu)/(mu)

w <- mu



lmod3 <- lm(z~Area+Elevation+Nearest+Scruz+Adjacent, weights=w, data=data)
coef(lmod3)
```

```
  (Intercept)          Area     Elevation       Nearest         Scruz      Adjacent
 3.1562582546 -0.0005793855  0.0035379237  0.0087861184 -0.0056868875 -0.0006630167
```

"How are these coefficients to those from the glm fit?" the following output expresses the ratio of the coefficients from the third interation of the linear model to those from the glm.

```
coef(lmod3)/coef(gala_pois)
```

```
(Intercept)        Area   Elevation     Nearest       Scruz    Adjacent
  1.0004597   0.9990389   0.9992458   0.9955296   0.9960531   0.9999783
```

Notice that the ratio of the coefficients are essentially equal after this iteration.

Q3. Part f

Repeat these iterations a few more times,computing the deviance in each time. Stop when the deviance does not change much. Compare your final estimated coefficients to that produced by the GLM fit.

The following output shows the ratio of coefficients from futher iterations and those from the glm model as well as the deviance. After the 5th iteration, the deviance converges and the coefficients are equal to those from the glm model.

```
lmod4 <- lmod3

for (i in 4:6){

  eta <- lmod4$fit

  mu <- exp(eta)

  vu <- mu

  z <- eta + (y-mu)/(mu)

  w <- mu

  deviance = round(2*sum(data$Species*log(data$Species/mu)-(data$Species-mu)),10)

  lmod4 <- lm(z~Area+Elevation+Nearest+Scruz+Adjacent, weights=w, data=data)

  cat("Iteration: ", i, "|", "Ratio of Coefficients: ", coef(lmod4)/coef(gala_pois), "|", "Deviance:", d

}
```

```
Iteration:  4 | Ratio of Coefficients:  1 0.9999988 0.9999992 0.9999928 0.9999926 1 | Deviance: 716.8488
Iteration:  5 | Ratio of Coefficients:  1 1 1 1 1 1 | Deviance: 716.8458
Iteration:  6 | Ratio of Coefficients:  1 1 1 1 1 1 | Deviance: 716.8458
```

```
final_model <- lmod4
```

Q3. Part g

Use your final iterated linear model fit to produce standard errors for the coefficients. How close are these to that produced by the direct GLM fit?

Table 3 shows the standard errors of the coefficients for both the model from the final iteration and the glm model. The standard errors from the iterated linear model are at least 5x greater than those from the GLM.

```
iteration_6 <- data.frame(summary(final_model)$coefficients[,2]) %>% rownames_to_column() %>%
  rename(`Standard Errors - Iteration 6` = summary.final_model..coefficients...2.,
         Feature = rowname)

GLM <- data.frame(summary(gala_pois)$coefficients[,2]) %>% rownames_to_column() %>%
  rename(`Standard Errors - GLM` = summary.gala_pois..coefficients...2.,
         Feature = rowname)

std_err_summary <- iteration_6 %>%
                      mutate(`Standard Errors - GLM` = round(GLM$`Standard Errors - GLM`,5),
                             `Standard Errors - Iteration 6` = round(`Standard Errors - Iteration 6`,5))
                      mutate(Ratio = `Standard Errors - Iteration 6`/`Standard Errors - GLM`)


kable(std_err_summary,
```

```
      align = rep("c",ncol(std_err_summary)),
      booktabs = T,
      caption = "Comparison of Standard Errors of Coefficients") %>%
  kable_styling(position = "center", latex_options = "hold_position")
```

Table 3: Comparison of Standard Errors of Coefficients

Feature	Standard Errors - Iteration 6	Standard Errors - GLM	Ratio
(Intercept)	0.29159	0.05175	5.634589
Area	0.00015	0.00003	5.000000
Elevation	0.00049	0.00009	5.444444
Nearest	0.01026	0.00182	5.637363
Scruz	0.00353	0.00063	5.603175
Adjacent	0.00017	0.00003	5.666667

Chapter 10 | Question 1

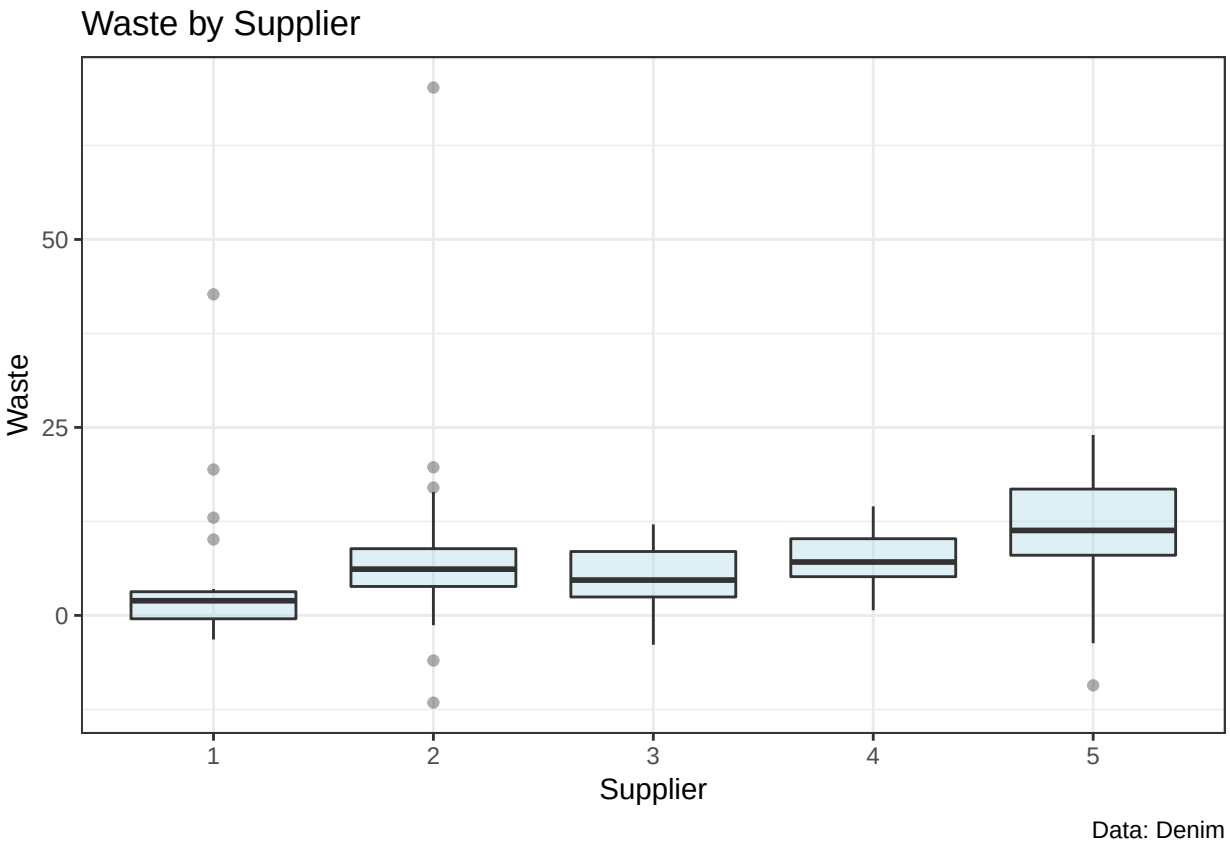
```
### Import Data
data2 <- denim
```

Q1. Part a

Plot the data and comment.

The following plot compares the percentage of waste relative to target across 5 suppliers. It is interesting that supplier 5 generally tends to have a larger proportion of waste than other suppiers. Suppliers 2,3 and 4 have similar behavior with respect to waste though supplier 2 has some extreme outliers worth investigating. Supplier 1 has the lowest waste, though similar to supplier 2, has outliers worth investigating.

```
ggplot(data2) +
  geom_boxplot(aes(x = supplier, y = waste), fill = "lightblue", alpha = .4)+
  ggtitle("Waste by Supplier")+
  labs(caption = "Data: Denim")+
  xlab("Supplier")+
  ylab("Waste")+
  theme_bw()
```



Q1. Part b

Fit the linear fixed effects model. Is the operator significant?

```
lmod <- aov(waste ~ supplier, data2)
summary(lmod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
supplier     4    451  112.73    1.16  0.334
Residuals   90   8749   97.21
```

The effect of the supplier is not statistically significant as shown by the p-value of 0.334.

Q1. Part d

Analyze the data with supplier as a random effect. What are the estimated standard deviations of the effects?

```
op1 <- options(contrasts=c("contr.sum", "contr.poly"))

mod <- lmer(waste ~ (1|supplier), denim)

summary(mod)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: waste ~ (1 | supplier)
   Data: denim
```

```
REML criterion at convergence: 702.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9095 -0.4363 -0.1669  0.3142  6.3817

Random effects:
 Groups   Name        Variance Std.Dev.
 supplier (Intercept)  0.6711  0.8192
 Residual             97.3350  9.8658
Number of obs: 95, groups:  supplier, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)    6.997      1.078    6.49
```

The random effect of supplier has a standard deviation of 0.8192.

Q1. Part e

Test the significance of the supplier term.

```
nullmod <- lm(waste ~ 1, data2)
lrtstat <- as.numeric(2*(logLik(mod)-logLik(nullmod)))
pvalue <- pchisq(lrtstat,1,lower=FALSE)
pvalue
```

```
[1] 0.1690049
```

The above output shows that the p-value is 0.1690 exceeds both 5% and 10% significance levels suggesting the supplier term not signifigant.

Q1. Part f

Compute confidence intervals for the random effect SDs.

```
set.seed(1993)
confint(mod, method= "boot")
```

```
                2.5 %    97.5 %
.sig01      0.000000  3.492744
.sigma      8.400979 11.317689
(Intercept) 4.814839  9.309345
```

The output above displays the confidence intervals for the random effect SDs [0,3.49].