HW1

Jinhong Du - 12243476

2020/01/03

Contents

1.1. Suppose that y_i has a $N(\mu_i, \sigma^2)$ distribution, $i = 1, \ldots, n$. Formulate the normal linear model as a special case of a GLM, specifying the random component, linear predictor, and link function.

The GLM is given by

$$g(\mathbb{E}(\mathbf{y})) = \mathbf{X}\beta$$

where $\mathbf{y} = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^\top$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix and $\beta \in \mathbb{R}^p$ is the vector of coefficients. The random component is $\mathbf{y}$, the linear predictor is $\eta = \mathbf{X}\beta$ and the link function is the identity $g(\mu) = \mu$.

The densities of y_i's are in the exponential dispersion family,

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]$$

where

$$\theta_i = \mu_i, \qquad b(\theta_i) = \frac{\theta_i^2}{2}, \qquad \phi = \sigma^2, \qquad c(y_i, \phi) = -\frac{y_i^2}{2\phi} - \log(2\pi\phi).$$

1.2. Link function of a GLM:

a. Describe the purpose of the link function g.

1. The link function is a monotonic and differentiable function. It is used to links $\mu_i = \mathbb{E}y_i$ to explanatory variables through the formula $g(\mu_i) = x_i^\top\beta$. Such relationship between the response and explanatory variables need not be of the simple linear form. For example, it can be log, exp and so on.
2. For some distribution, like binomial distribution, the response cannot range from $-\infty$ to $+\infty$. A link function, like logit function can map $x_i^\top\beta$ to $(-1, 1)$.
3. Without a properly specified link, the variances of the residuals will not be constant (a required assumption for inference with an Ordinary Least Squares estimate) or handled correctly.

b. The identity link is the standard one with normal responses but is not often used with binary or count responses. Why do you think this is?

Every entry of linear predictor $\eta = \mathbf{X}\beta$ ranges from $-\infty$ to $+\infty$. Since for the normality assumption, the responses range from $-\infty$ to $+\infty$, the identity link maps $\mathbb{R}$ to $\mathbb{R}$, which is reasonable. While the binary or count responses only takes discerete (at most countable) values, the identity link seems to be unsuitable.

1.22. Refer to the analyses in Section 1.5.3 for the horseshoe crab satellites.

a. With color alone as a predictor, why are standard errors much smaller for a Poisson model than for a normal model? Out of these two very imperfect models, which do you trust more for judging significance of the estimates of the color effects? Why?

From the Figure 1.2. in the textbook, we can see that the Variance increases as y increases. A normal model assumes constant Variance, which is not satisfied in this setting. While a Poisson model has Variance equals to mean μ_i. So the Variance increases as μ_i/y increases, which is similar to our data. Therefore, the standard errors are much smaller for a Poisson model than for a normal model.

Since the responses are count data, a normal model is less suitable for the data than a Poisson model. So the Poisson model is more trustable for judging significance of the estimates of the color effects.

b. Download the data (file `Crabs.dat`) from www.stat.ufl.edu/~aa/glm/data. When weight is also a predictor, identify an outlying observation. Refit the model with color and weight predictors without that observation. Compare results, to investigate the sensitivity of the results to this outlier.

As we can see below, the outlier is the 141-th sample, whose weight is 5.2. The estimated coefficients with this outlier are

$$\beta_{\text{intercept}} = -0.8232, \beta_{\text{color:2}} = -0.6181, \beta_{\text{color:3}} = -1.2404, \beta_{\text{color:4}} = -1.1882, \beta_{\text{weight}} = 1.8662.$$

Without the outlier, the estimated coefficients are

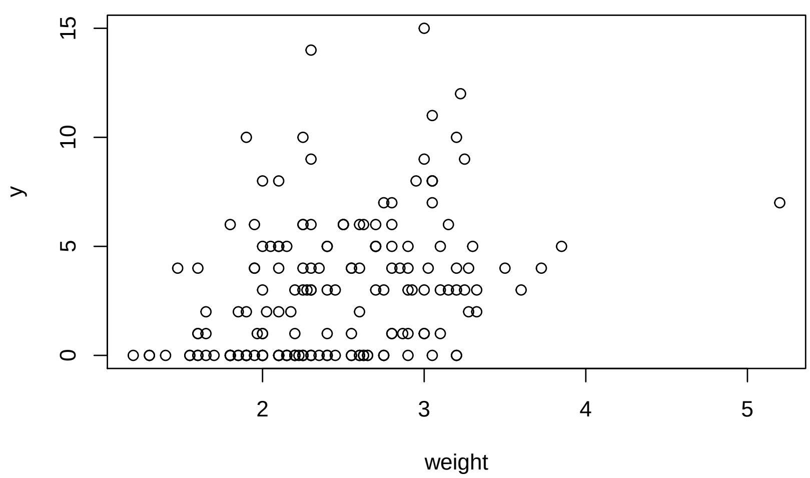
$$\beta'_{\text{intercept}} = -1.0160, \beta'_{\text{color:2}} = -0.5959, \beta'_{\text{color:3}} = -1.2162, \beta'_{\text{color:4}} = -1.1548, \beta'_{\text{weight}} = 1.9395.$$

We have the differences between these two groups of estimated coefficients are

$$\Delta\beta_{\text{intercept}} = 0.1928, \Delta\beta_{\text{color:2}} = -0.0222, \Delta\beta_{\text{color:3}} = -0.0242, \Delta\beta_{\text{color:4}} = -0.0334, \Delta\beta_{\text{weight}} = -0.0733.$$

As we can see, $\beta_{\text{intercept}}$ changes a lot. Therefore, the results are sensitive to this outlier.

```
Crabs <- read.table("Crabs.dat", header=T)
plot(y~weight, Crabs)
```




```
fit1 <- glm(y ~ factor(color)+weight, data=Crabs)
summary(fit1)
```

```
##
## Call:
## glm(formula = y ~ factor(color) + weight, data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5305  -2.1354  -0.6072   1.5223  11.1491
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.8232     1.3549  -0.608    0.544
## factor(color)2  -0.6181     0.9011  -0.686    0.494
## factor(color)3  -1.2404     0.9662  -1.284    0.201
## factor(color)4  -1.1882     1.0704  -1.110    0.269
```

```
## weight           1.8662     0.4018   4.645 6.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.637005)
##
##     Null deviance: 1704.9  on 172  degrees of freedom
## Residual deviance: 1451.0  on 168  degrees of freedom
## AIC: 870.88
##
## Number of Fisher Scoring iterations: 2
```

```
Crabs[Crabs$weight>5,]
```

```
##     crab y weight width color spine
## 141  141 7    5.2  33.5     2     1
```

```
fit2 <- glm(y ~ factor(color)+weight, data=Crabs[Crabs$weight<5,])
summary(fit2)
```

```
##
## Call:
## glm(formula = y ~ factor(color) + weight, data = Crabs[Crabs$weight <
##     5, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5946  -2.1318  -0.6038   1.5115  11.1510
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.0160     1.4204  -0.715    0.475
## factor(color)2  -0.5959     0.9045  -0.659    0.511
## factor(color)3  -1.2162     0.9699  -1.254    0.212
## factor(color)4  -1.1548     1.0754  -1.074    0.284
## weight           1.9395     0.4327   4.482 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.67758)
##
##     Null deviance: 1688.1  on 171  degrees of freedom
## Residual deviance: 1449.2  on 167  degrees of freedom
## AIC: 866.69
##
## Number of Fisher Scoring iterations: 2
```

```
fit1$coefficients - fit2$coefficients
```

```
##    (Intercept) factor(color)2 factor(color)3 factor(color)4         weight
##     0.19278163    -0.02221016    -0.02419089    -0.03338475    -0.07332423
```

4.3. Show that the t distribution is not in the exponential dispersion family. (Although GLM theory works out neatly for family (4.1), in practice it is sometimes useful to use other distributions, such as the Cauchy special case of the t.)

The density of t distribution with degree of freedom ν, is given by

$$\begin{aligned} f_\nu(t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ &= \exp\left[-\frac{\nu+1}{2}\ln\left(1+\frac{t^2}{\nu}\right)+\ln\Gamma(\frac{\nu+1}{2})-\frac{1}{2}\ln(\nu\pi)-\ln\Gamma(\frac{\nu}{2})\right] \end{aligned}$$

If the t distribution is in the exponential dispersion family, then

$$f_\nu(t) = \exp\left(t\theta - b(\theta)\right) f_0(t),$$

i.e.,

$$-\frac{\nu+1}{2}\ln\left(1+\frac{t^2}{\nu}\right) + d(\theta) = t\theta - b(\theta) + \ln f_0(t)$$

where $d(\theta) = \ln\Gamma(\frac{\nu+1}{2}) - \frac{1}{2}\ln(\nu\pi) - \ln\Gamma(\frac{\nu}{2})$. Then

$$\ln f_0(t) = [b(\theta)+d(\theta)] - \left[\frac{\nu+1}{2}\ln\left(1+\frac{t^2}{\nu}\right) + t\theta\right].$$

Since $\frac{\nu+1}{2}\ln\left(1+\frac{t^2}{\nu}\right) + t\theta$ is a function of t and it is nonzero for some $t \in \mathbb{R}$, it cannot equal to the constant $b(\theta)+d(\theta)$ for all t. So $f_0(t)$ is dependent on θ, which is a contradiction. Therefore, the t distribution is not in the exponential dispersion family.

4.6. Suppose y_i has a Poisson distribution with $g(\mu_i) = \beta_0 + \beta_1 x_i$, where $x_i = 1$ for $i = 1, \ldots, n_A$ from group A and $x_i = 0$ for $i = n_A+1, ..., n_A+n_B$ from group B, and with all observations being independent. Show that for the log-link function, the GLM likelihood equations imply that the fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

Since $y_i \sim \text{Poisson}(\mu_i)$, we have

$$\begin{aligned} f(y_i;\mu_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \exp\left[y_i\log\mu_i - \mu_i - \log(y_i!)\right] \\ &= \exp\left[y_i\theta_i - \exp\theta_i - \log(y_i!)\right], \qquad y_i = 0, 1, \ldots \end{aligned}$$

where $\theta_i = \log\mu_i$. Then $b(\theta_i) = \exp\theta_i = \mu_i$, $a(\phi) = 1$ and $c(y_i,\phi) = -\log(y_i!)$. Also, we have

$$\mathbb{E}y_i = \text{Var } y_i = \exp\theta_i = \mu_i.$$

Let $n = n_A + n_B$. The log-likelihood is given by

$$L(\beta) = \sum_{i=1}^n \log f(y_i;\mu_i) = \sum_{i=1}^n \left[y_i\theta_i - \exp\theta_i - \log(y_i!)\right]$$

When using the log-link function $\mu_i = \exp(\eta_i) = \exp(\beta_0 + \beta_1 x_i)$, then the likelihood equations are given by

$$\begin{aligned}
\frac{\partial L}{\partial \beta_1} &= \sum_{i=1}^{n} \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_1} \\
&= \sum_{i=1}^{n} (y_i - \mu_i)\frac{1}{\mu_i}\exp(\beta_0 + \beta_1 x_i) x_i \\
&= \sum_{i=1}^{n} (y_i - \mu_i) x_i = 0 \\
&= \sum_{i=1}^{n_A} (y_i - \mu_i) = 0 \\
\frac{\partial L}{\partial \beta_0} &= \sum_{i=1}^{n} (y_i - \mu_i)\frac{1}{\mu_i}\exp(\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} (y_i - \mu_i) = 0.
\end{aligned}$$

Therefore,

$$\begin{cases} \sum_{i=1}^{n_A} (y_i - \mu_i) = 0 \\ \sum_{i=n_A+1}^{n_A+n_B} (y_i - \mu_i) = 0 \end{cases},$$

i.e.,

$$\begin{cases} \hat{\mu}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} \mu_i = \frac{1}{n_A}\sum_{i=1}^{n_A} y_i \\ \hat{\mu}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B} \mu_i = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B} y_i \end{cases}.$$

4.7. Refer to the previous exercise. Using the likelihood equations, show that the same result holds for (a) any link function for this Poisson model, (b) any GLM of the form $g(\mu_i) = \beta_0 + \beta_1 x_i$ with a binary indicator predictor.

(a) The likelihood equations are given by

$$\begin{cases} \sum_{i=1}^{n} (y_i - \mu_i)\frac{1}{\mu_i}\frac{\partial \mu_i}{\partial \eta_i} x_i = 0 \\ \sum_{i=1}^{n} (y_i - \mu_i)\frac{1}{\mu_i}\frac{\partial \mu_i}{\partial \eta_i} = 0 \end{cases}$$

i.e.,

$$\begin{cases} \sum_{i=1}^{n_A} (y_i - \mu_i)\frac{\partial \ln g^{-1}(\eta_i)}{\partial \eta_i} = 0 \\ \sum_{i=n_A+1}^{n_A+n_B} (y_i - \mu_i)\frac{\partial \ln g^{-1}(\eta_i)}{\partial \eta_i} = 0 \end{cases}. \tag{1}$$

Notice that $\frac{\partial \ln g^{-1}(\eta_i)}{\partial \eta_i} = \frac{\partial \ln g^{-1}(\eta)}{\partial \eta}\Big|_{\eta=\eta_i}$ and $\eta_i = \beta_0 + \beta_1 x_i = \begin{cases} \beta_0 & , \ x_i = 0 \\ \beta_0 + \beta_1 & , \ x_i = 1 \end{cases}$, we have

$$\begin{aligned}
\frac{\partial \ln g^{-1}(\eta_1)}{\partial \eta_1} &= \cdots = \frac{\partial \ln g^{-1}(\eta_{n_A})}{\partial \eta_{n_A}}, \\
\frac{\partial \ln g^{-1}(\eta_{n_A+1})}{\partial \eta_{n_A+1}} &= \cdots = \frac{\partial \ln g^{-1}(\eta_{n_A+n_B})}{\partial \eta_{n_A+n_B}}.
\end{aligned}$$

As long as g is a proper link function, the above terms are non-zero. Therefore, (1) can be reduced to

$$\begin{cases} \sum_{i=1}^{n_A} (y_i - \mu_i) = 0 \\ \sum_{i=n_A+1}^{n_A+n_B} (y_i - \mu_i) = 0 \end{cases},$$

i.e.,

$$\begin{cases} \hat{\mu}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}\mu_i = \frac{1}{n_A}\sum_{i=1}^{n_A} y_i \\ \hat{\mu}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B}\mu_i = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B} y_i \end{cases}.$$

(b) For any GLM of the form $g(\mu_i) = \beta_0 + \beta_1 x_i$, we have

$$\begin{aligned} \frac{\partial L}{\partial \beta_1} &= \sum_{i=1}^n \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_1} \\ &= \sum_{i=1}^n (y_i - b'(\theta_i))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} x_i \\ &= \sum_{i=1}^n (y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} x_i, \\ \frac{\partial L}{\partial \beta_0} &= \sum_{i=1}^n (y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}. \end{aligned}$$

Then the likelihood equations are given by

$$\begin{cases} \sum_{i=1}^{n_A}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = 0 \\ \sum_{i=n_A+1}^{n_A+n_B}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = 0 \end{cases}.$$

Notice that θ_i is a function of μ_i in the sense that $\theta_i = h(\mu_i)$ for $i = 1, \ldots, n$. Since $\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial h(\mu)}{\mu}\Big|_{\mu_i}$ and $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial g^{-1}(\eta)}{\eta}\Big|_{\eta_i}$, and $\mu_i = g^{-1}(\eta_i)$ and $\eta_i = \begin{cases} \beta_0 & , \ x_i = 0 \\ \beta_0 + \beta_1 & , \ x_i = 1 \end{cases}$ are fixed, we have

$$\begin{aligned} \frac{\partial \theta_1}{\partial \mu_1}\frac{\partial \mu_1}{\partial \eta_1} &= \cdots = \frac{\partial \theta_{n_A}}{\partial \mu_{n_A}}\frac{\partial \mu_{n_A}}{\partial \eta_{n_A}}, \\ \frac{\partial \theta_{n_A+1}}{\partial \mu_{n_A+1}}\frac{\partial \mu_{n_A+1}}{\partial \eta_{n_A+1}} &= \cdots = \frac{\partial \theta_{n_A+n_B}}{\partial \mu_{n_A+n_B}}\frac{\partial \mu_{n_A+n_B}}{\partial \eta_{n_A+n_B}}. \end{aligned}$$

As long as h is nontrivial and g is a proper link function, the above terms are non-zero. Therefore,

$$\begin{cases} \sum_{i=1}^{n_A}(y_i - \mu_i) = 0 \\ \sum_{i=n_A+1}^{n_A+n_B}(y_i - \mu_i) = 0 \end{cases},$$

i.e.,

$$\begin{cases} \hat{\mu}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}\mu_i = \frac{1}{n_A}\sum_{i=1}^{n_A} y_i \\ \hat{\mu}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B}\mu_i = \frac{1}{n_B}\sum_{i=n_A+1}^{n_A+n_B} y_i \end{cases}.$$

4.9. Consider the expression for the weight matrix W in $\mathrm{Var}(\hat{\beta}) = (X^\top WX)^{-1}$ for a GLM. Find W for the ordinary normal linear model, and show how $\mathrm{Var}(\hat{\beta})$ follows from the GLM formula.

Since $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ and $Y \sim N(X\beta, \sigma^2)$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}[(X^\top X)^{-1}X^\top Y] \\ &= (X^\top X)^{-1}X^\top \mathrm{Var}(Y)X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}.\end{aligned}$$

Then $W = \sigma^{-2}I$,

$$\mathrm{Var}(\hat{\beta}) = (X^\top WX)^{-1} = (X^\top(\sigma^{-2}I)X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

8. $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim}$ Poisson(μ). Describe the distributions of $\overline{Y}$ and $S = \sum_i Y_i$, and say what are the exponential family quantities $(\theta, y, b(\cdot), \mu, V)$ in both cases.

Since for $k = 0, 1, 2, \ldots$,

$$\begin{aligned}\mathbb{P}(Y_1 + Y_2 = k) &= \sum_{i=0}^{k} \mathbb{P}(Y_1 = i, Y_2 = k - i) \\ &= \sum_{i=0}^{k} \frac{\mu^i}{i!}e^{-\mu}\frac{\mu^{k-i}}{(k-i)!}e^{-\mu} \\ &= \frac{\mu^k}{k!}e^{-2\mu}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!} \\ &= \frac{\mu^k}{k!}e^{-2\mu}\cdot 2^k \\ &= \frac{(2\mu)^k}{k!}e^{-2\mu},\end{aligned}$$

we have $Y_1 + Y_2 \sim$ Poisson(2μ). Therefore, $S = \sum_{i=1}^n Y_i \sim$ Poisson($n\mu$) and $\forall\ y \in \mathbb{N}$,

$$\begin{aligned}p_S(y) &= \mathbb{P}(S = y) \\ &= \frac{(n\mu)^y}{y!}e^{-n\mu} \\ &= \exp\left[y\log(n\mu) - n\mu\right]\frac{1}{y!},\end{aligned}$$

which yields $\theta_S = \log(n\mu)$, $b_S(\theta_S) = e^{\theta_S}$, $f_{0,S}(y,\phi) = \frac{1}{y!}$, $\mu_S = b'_S(\theta_S) = e^{\theta_S} = n\mu$ and $V_S = b''_S(\theta_S) = e^{\theta_S} = n\mu$.

Then the density of $\overline{Y}$ is given by

$$\begin{aligned}p_{\overline{Y}} &= \mathbb{P}(S = ny) \\ &= \frac{(n\mu)^{ny}}{(ny)!}e^{-n\mu} \\ &= \exp\left\{y[n\log(n\mu)] - n\mu\right\}\frac{1}{(ny)!},\end{aligned}$$

which yields $\theta_{\overline{Y}} = n\log(n\mu)$, $b_{\overline{Y}}(\theta_{\overline{Y}}) = e^{\frac{\theta_{\overline{Y}}}{n}}$, $f_{0,\overline{Y}}(y) = \frac{1}{(ny)!}$, $\mu_{\overline{Y}} = b'_{\overline{Y}}(\theta_{\overline{Y}}) = \frac{1}{n}e^{\frac{\theta_{\overline{Y}}}{n}}$ and $V_{\overline{Y}} = b''_{\overline{Y}}(\theta_{\overline{Y}}) = \frac{1}{n^2}e^{\frac{\theta_{\overline{Y}}}{n}}$.