Homework 2
ADVANCED METHODS IV 140.754

1 GLM Basics

(a) Link function: monotone differentiable function g that connects the random and systematic components.

If we have the a Y_i whose density looks like this:

$$f(y_i|\theta, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

then $g(\mu_i) = \theta_i$ is the canonical link.

The linear predictor, or systematic component, is $\eta_i = x_i'\beta$.

IRLS is an algorithm that is very fast at estimating the coefficients β. You first get one guess, and then use it to calculate an adjusted dependent variable z_i and some weights w_i. These are then used to get a new estimate of β by performing weighted linear regression.

(b) The main advantage of using the canonical link is to get an easy to interpret result of the coefficients. However, it might not be the best option when you observe a variable that is censored according to some cutoff as Parichoy presented in lab (slide 9).

(c) Gaussian for bell shaped and symmetric data, Poisson for count data, Gamma for positive continuous data, Binomial for two categories (success, failure), Inverse Gaussian for positive data.

(d) Poisson:

$$f(Y|\theta = \lambda, \phi) = \frac{1}{y!}\exp\left(\log\lambda \cdot y - \lambda\right) = \exp\left(\log\lambda \cdot y - \lambda - \log y!\right)$$
$$\theta = \lambda, \phi = 1, b(\theta) = \log\lambda, a(\phi) = 1, c(y, \phi) = -\log y!$$

Gamma:

$$f(Y|\theta, \phi) = \exp\left((\log y, y) \cdot (\alpha - 1, -b) - (\log\Gamma(\alpha) - \alpha\log b)\right)$$
$$\theta = (\alpha, b), \phi = 1, b(\theta) = \log\Gamma(\alpha) - \alpha\log b, a(\phi) = 1, c(y, \phi) = 0$$

Inverse Gaussian:

$$f(Y|\theta, \phi) = \exp\left((y, 1/y) \cdot (-\lambda/2\mu^2, -\lambda/2) - (-\lambda/\mu - 1/2\log\lambda) - 1/2\log(2\pi y^3)\right)$$
$$\theta = (\mu, \lambda), \phi = 1, b(\theta) = -\lambda/\mu - 1/2\log\lambda, a(\phi) = 1, c(y, \phi) = -1/2\log(2\pi y^3)$$

(e)

$$E[Y|\theta, \phi] = b'(\theta)$$
$$Var(Y|\theta, \phi) = b''(\theta)a(\phi)$$

(f) Poisson:

$$\log(\theta)$$

Gamma:

$$-\theta^{-1}$$

Inverse Gaussian:

$$\theta^{-2}$$

(g) No data, or at least I can't download it.

```
## From ?glm
clotting <- data.frame(u = c(5, 10, 15, 20, 30, 40, 60, 80, 100), lot1 = c(118, 58, 42, 35,
    27, 25, 21, 19, 18), lot2 = c(69, 35, 26, 21, 18, 16, 13, 12, 12))
fit <- glm(lot1 ~ log(u), data = clotting, family = Gamma)
summary(fit)

##
## Call:
## glm(formula = lot1 ~ log(u), family = Gamma, data = clotting)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.0401  -0.0376  -0.0264   0.0290   0.0864
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.016554   0.000928   -17.9  4.3e-07 ***
## log(u)       0.015343   0.000415    37.0  2.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.002446)
##
##     Null deviance: 3.51283  on 8  degrees of freedom
## Residual deviance: 0.01673  on 7  degrees of freedom
## AIC: 37.99
##
## Number of Fisher Scoring iterations: 3
```

(h)

(i)

```
library(MASS)
confint(fit, level = 0.9)
## Waiting for profiling to be done...

##                   5 %     95 %
## (Intercept) -0.01807 -0.01502
## log(u)       0.01466  0.01603
```

2 Interpreting coefficients

2.1 Model 1

β_0 is the log odds at baseline, β_1 is the log odds ratio holding Z_i constant, and γ_2 is the log odds ratio holding X_i constant.

2.2 Model 2

γ_0 is the log odds at baseline, γ_1 is the log odds ratio when Z_i is 0, and γ_2 is the log odds ratio when X_i is 0. Then γ_3 is the difference in the log odds ratio for a one unit increase in Z_i holding X_i constant.

2.3 Param vs non-param

In non-parametric, we have to say that it's the log ratio of the means instead of the mean instead of the log odds when we are assuming a parametric model; in particular a Binomial distribution.

2.4 Transformation

The inverse logit (called expit) might make things more interpretable for the non-statistician because the mean can be interpreted directly