Homework3

BIOS6643 Fall 2021

10/18/2021

By far the three most common cases for generalized linear models are for normal, binary, and count outcomes. Since you have seen so many examples of normal (standard linear models) and binary (logistic regression), these questions involve count outcomes. (Where appropriate include brief and annotated SAS or R output)

Question 1

The Cereal2.csv contains data from a nutrition study where several members of each of a number of families recorded the number of servings of breakfast cereal they ate each week starting at baseline and continuing for 14 weeks. Some families were in an experimental program (cond=1), the others did not receive anything special (cond=0). Family members were coded 1 = Mom, 2 = Dad, 3 = $Kid1$, 4 = $Kid2$, 5 = $Kid3$, 6 = $Kid4$, 7 = $Kid5$. Sex is coded 0 = $Male$, 1 = $Female$. Weight at baseline has been recoded in units of $100lbs$ (so $1.5 = 150lbs$) to avoid numerical problems in some procedures.

Note: To get agreement between GENMOD and NLMIXED you need to use a new variable Cond2 = 1 - Cond for the condition variable. This just gives a different parameterization.

In this question you will analyze only the week 1 (baseline) data, C1. There was some question as to whether the experimental and control groups were comparable at this time because some aspects of the intervention may have been done before week 1. Carry out the following analyses to compare the groups at baseline, using data only from $Kid1$ ($FamMem = 3$) - families were sampled based on this child. For each model, write the model equation, and write a sentence describing the results, understandable by dietitians.

a. Use a Poisson GzLM (i.e. Poisson regression) to estimate the association between condition and number of breakfast servings, adjusting for sex and weight.

Given some subject i, the Poisson regression with canonical link function ($g() = log()$):

$$
\begin{aligned}
Y_i &\sim \boldsymbol{Poisson}(\mu_i) \\
\mu_i &= E[Y_i] \\
\eta_i &= log(\mu_i) = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i \\
E[Y_i] = \mu_i &= \exp(\eta_i) = \exp(\beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i)
\end{aligned}
$$

Notes: First, Poisson distribution only contians one parameter, namely μ_i. Second, Poisson regression is built based on the exponential family distribution with canonical link function: the model is for the log transformed mean value of the (counting/integer) outcomes. Thus, **there is no residual** or error terms included in the model. Third, in longitudinal data analysis we need to specify both the mean features and the covariance features. Thus, later the model will be specified accordingly with both mean and covariance parts.

R code

```
cereal <- here("data", "Cereal2.csv") %>%
  read.csv() %>%
  janitor::clean_names() %>%
  filter(fam_mem == 3) %>%
  mutate(sex = as.factor(sex),
         cond = as.factor(cond))

## Poisson model with log link function
mod1a <- glm(c1 ~ cond + sex + wt1,
             data = cereal,
             family = poisson(link = "log"))
(tidy1a <- broom::tidy(mod1a))
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.998    0.199       5.02 5.14e- 7
## 2 cond1          0.865    0.135       6.40 1.50e-10
## 3 sex1          -0.199    0.0822     -2.42 1.56e- 2
## 4 wt1            0.197    0.132       1.49 1.36e- 1
```

SAS code

```
proc import DATAFILE='C:/Users/Goodgolden5/Desktop/Cereal2.csv'
DBMS=csv out=cereal replace; run;
/* Filter the family member 3 */
data cereal; set cereal; where fammem = 3; run;
/* Fit with Poisson using GENMOD */
PROC GENMOD data = cereal;
MODEL C1 = Cond Sex Wt1  / dist = poisson link=log;

ods select ParameterEstimates;
   ods trace on; ods show;
RUN;
```

```
##                             The GENMOD Procedure
##
##               Analysis Of Maximum Likelihood Parameter Estimates
##
##                                  Standard   Wald 95% Confidence         Wald
##    Parameter    DF    Estimate      Error          Limits         Chi-Square
```

```
##
##    Intercept     1     0.9979     0.1988     0.6084     1.3875        25.21
##    Cond          1     0.8645     0.1350     0.6000     1.1291        41.02
##    Sex           1    -0.1988     0.0822    -0.3600    -0.0377         5.85
##    Wt1           1     0.1973     0.1323    -0.0620     0.4567         2.22
##    Scale         0     1.0000     0.0000     1.0000     1.0000
##
##                              Analysis Of Maximum
##                              Likelihood Parameter
##                                   Estimates
##
##                             Parameter   Pr > ChiSq
##
##                             Intercept       <.0001
##                             Cond            <.0001
##                             Sex             0.0156
##                             Wt1             0.1359
##                             Scale
##
## NOTE: The scale parameter was held fixed.
```

At baseline, kids in the experimental group ate more servings of cereal per day, by a factor of exp(0.8645) = 2.3738, 95%CI (1.822, 3.093), p<0.0001.

b. Repeat **(a)** allowing for overdispersion by using quasilikelihood with the Poisson GzLM. Use the Pearson method for estimating the scale parameter. Show algebraically the relation between the QL SEs for the betas and those from the Poisson model in **(a)**.

$$
\begin{aligned}
Y_i &\sim \boldsymbol{\mathcal{P}oisson}(\mu_i) \\
\mu_i &= E[Y_i] \\
\eta_i &= log(\mu_i) = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i
\end{aligned}
$$

Thus, model will have the mean function as $E[Y_i] = \mu_i = \exp(\beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i)$ and the variance function as $Var[Y_i] = \phi\mu_i$.

R code ——————————————————————————————

```
## Quasi-Poisson with log link function
mod1b <- glm(c1 ~ cond + sex + wt1,
             data = cereal,
             family = quasipoisson(link = "log"))
(tidy1b <- broom::tidy(mod1b))
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.998     0.341     2.93  0.00430
```

```
2 cond1          0.865     0.232     3.73  0.000323
3 sex1          -0.199     0.141    -1.41  0.162
4 wt1            0.197     0.227     0.869 0.387
```

```
## to pull out the dispersion parameter
summary(mod1b)$dispersion
```

```
[1] 2.944663
```

```
sqrt(summary(mod1b)$dispersion)
```

```
[1] 1.716002
```

```
## to check the scale
## dispersion parameter = scale^2
standard_error(mod1b)[, 2] / standard_error(mod1a)[, 2]
```

```
[1] 1.716002 1.716002 1.716002 1.716002
```

SAS code

```
/* Fit with Quasi-Poisson using GENMOD */
proc genmod data=cereal;
model C1 = Cond Sex Wt1 / dist = poisson link = log pscale;
ods select ParameterEstimates;
   ods trace on; ods show;
run;
```

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	0.9979	0.3411	0.3295	1.6664	8.56
Cond	1	0.8645	0.2316	0.4106	1.3185	13.93
Sex	1	-0.1988	0.1411	-0.4754	0.0777	1.99
Wt1	1	0.1973	0.2271	-0.2477	0.6424	0.76
Scale	0	1.7160	0.0000	1.7160	1.7160	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.0034
Cond	0.0002
Sex	0.1588
Wt1	0.3849

```
                    Scale

NOTE: The scale parameter was estimated by the square root of Pearson's
      Chi-Square/DOF.
```

$$\hat{\phi} = \frac{1}{n-k}\sum \frac{\left(Y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i} = \frac{1}{n-k}\sum \frac{\left(Y_i - g^{-1}(\hat{\eta}_i)\right)^2}{g^{-1}(\hat{\eta}_i)} = \frac{\chi^2}{df}$$

Note that $\hat{\eta}_i$ is the fitted expected mean from linear regression; plug-in $\hat{\eta}_i$ to get the fitted expection of Y_i. From above formulation, we can see $\hat{\phi}$ is just a Pearson χ^2 divided by the residual degrees of freedom. Thus, the scale parameters are square root of Chi-square values/df, and SE's are Poisson SE's times scale parameter. Note CI's are much wider, and p-values larger, than for the Poisson model since these models account for the over-dispersion relative to Poisson.

Notes: Due to different parametrization, SAS[PROC NLMIXED] provides the $\sqrt{\phi}$ and R[glm(quasipoisson)] provides the ϕ. Since the QuasiPoisson model is not based on a probability model, the AIC is undefined. For the same reason, quantile residuals cannot be computed for the quasiPoisson glm since no probability model is defined

c. Repeat **(a)** allowing for overdispersion by adding a random normal error to the linear predictor in the Poisson GzLM and using maximum likelihood estimation. I am not aware of an algebraic relation between these ML SE's and those from the models in **(a)** or **(b)**.

$$\begin{aligned} Y_i|\epsilon_i &\sim \boldsymbol{Poisson}(\mu_i) \\ \epsilon_i &\overset{iid}{\sim} \boldsymbol{Normal}(0,\ \sigma^2) \\ \log(\mu_i) = \eta_i + \epsilon_i &= \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i + \epsilon_i \end{aligned}$$

From the conditional $E[Y_i|\epsilon_i] = \mu_i = e^{\epsilon_i}\exp(\beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i)$ equation, we can separate product into the two terms e^{ϵ_i} and $\exp(\eta_i)$ (the linear regressor term). We call this type of distribution as mixture distribution of Poisson-log-Normal. Because the ϵ_i follows Normal distribution, the e^{ϵ_i} follows a log-Normal distribution; the $\exp(\eta_i)$ is built for Poisson regression. Hence this a Poisson-log-Normal mixture distribution. The conditional variance function is $Var[Y_i|\epsilon_i] = \mu_i + \frac{Var[e^{\epsilon_i}]}{E[e^{\epsilon_i}]^2}\mu_i^2$. For the marginal mean function $E[Y_i] = E[e^{\epsilon_i}]\exp(\beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i)$, the function is still the same as in Q1a and Q1b (because $E[e^{\epsilon_i}] = 1$).

SE's are again larger than for the Poisson model, and vaguely similar to those from quasi-likelihood. Parameter estimates are not identical to those of the Poisson and quasilikelihood models.

Notes: Here we parametrize on the variance/SD for the Normal without restriction of non-positive ϵ_i. Also be aware of the difference between mixed model (which is a model) and

mixture distribution (which is a distribution). The term mixture distribution can refer to a different type of structure imposed in probability distributions involving a convolution of distribution, but here we will not extend to that perspective.

R code

```
## The Poisson-log-Normal mixture lme4::glmer
mod1c <- glmer(c1 ~ cond + sex + wt1 + (1|fam_idno),
                      data = cereal,
                      family = poisson(link = "log"))
broom.mixed::tidy(mod1c)
```

```
# A tibble: 5 x 7
  effect   group    term            estimate std.error statistic    p.value
  <chr>    <chr>    <chr>              <dbl>     <dbl>     <dbl>      <dbl>
1 fixed    <NA>     (Intercept)        0.882     0.324     2.72   0.00653
2 fixed    <NA>     cond1              0.855     0.197     4.34   0.0000143
3 fixed    <NA>     sex1              -0.153     0.143    -1.07   0.284
4 fixed    <NA>     wt1                0.164     0.236     0.693  0.488
5 ran_pars fam_idno sd__(Intercept)    0.543    NA        NA      NA
```

```
## The Poisson-log-Normal mixture glmmML
mod2c <- glmmML::glmmML(c1 ~ cond + sex + wt1,
                        family = poisson,
                        data = cereal,
                        cluster = fam_idno,
                        method = "ghq")
summary(mod2c)
```

```
Call:  glmmML::glmmML(formula = c1 ~ cond + sex + wt1, family = poisson,      data = cereal, cl

               coef se(coef)       z Pr(>|z|)
(Intercept)  0.8814   0.3256  2.7070 6.79e-03
cond1        0.8549   0.1978  4.3216 1.55e-05
sex1        -0.1531   0.1440 -1.0630 2.88e-01
wt1          0.1634   0.2373  0.6887 4.91e-01

Scale parameter in mixing distribution:  0.5469 gaussian
Std. Error:                              0.06857

        LR p-value for H_0: sigma = 0:  8.879e-19

Residual deviance: 219.1 on 94 degrees of freedom   AIC: 229.1
```

SAS code

```
/* Fit with Poisson-log-Normal */
proc nlmixed data=cereal;
model C1 ~ poisson(exp( b0 + b1 * Cond + b2 * Sex + b3 * Wt1 + eps) );
random eps ~ normal( 0, sig*sig ) subject=FamIDNO;
```

```
ods select ParameterEstimates;
   ods trace on; ods show;
run;
```


The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	0.9979	0.3411	0.3295	1.6664	8.56
Cond	1	0.8645	0.2316	0.4106	1.3185	13.93
Sex	1	-0.1988	0.1411	-0.4754	0.0777	1.99
Wt1	1	0.1973	0.2271	-0.2477	0.6424	0.76
Scale	0	1.7160	0.0000	1.7160	1.7160	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.0034
Cond	0.0002
Sex	0.1588
Wt1	0.3849
Scale	

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

The NLMIXED Procedure

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits	
b0	0.8815	0.3256	98	2.71	0.0080	0.2354	1.5275
b1	0.8549	0.1978	98	4.32	<.0001	0.4624	1.2474
b2	-0.1531	0.1440	98	-1.06	0.2903	-0.4388	0.1327
b3	0.1634	0.2372	98	0.69	0.4925	-0.3073	0.6341

```
sig          0.5468    0.06852     98      7.98     <.0001     0.4108     0.6827

                          Parameter Estimates

                          Parameter  Gradient

                          b0         2.103E-7
                          b1         -2.19E-7
                          b2         2.951E-7
                          b3         2.063E-6
                          sig        6.525E-7
```

d. Repeat **(a)** allowing for overdispersion by using a Negative Binomial GzLM estimated with maximum likelihood. I am not aware of an algebraic relation between these ML SE's and those from the models in **(a)**, **(b)**, or **(c)**.

One way to model over-dispersion for $Y_i \sim \boldsymbol{Poisson}(\mu_i)$ is through a hierarchical model with a mixture distribution (which Q1c can be generalized in a similar way). Here we add a second hierarchy for variability by allowing μ_i to be a random variable.

$$\begin{aligned} Y_i &\sim \boldsymbol{Poisson}(\lambda_i) \\ \lambda_i &\sim \boldsymbol{Gamma}(\mu_i, \phi) \\ E[Y_i] &= \mu_i \\ Var[Y_i] &= \mu_i + \phi\mu_i^2 \\ \eta_i &= log(\mu_i) = \beta_0 + \beta_1 Cond_i + \beta_2 Sex_i + \beta_3 Wt_i \end{aligned}$$

With this assumption above, we can rewrite it as $Y_i \sim \boldsymbol{GammaPoisson}(\mu_i, \phi)$, which is equivalent to $Y_i \sim \boldsymbol{NegativeBinomial}(\mu_i, k)$ with reparametrization with $k = 1/\phi$.

$$Y_i \sim \boldsymbol{NegativeBinomial}(k, p = \frac{\mu_i}{\mu_i + k})$$

$$p(Y_i; \mu_i, k) = \frac{\Gamma(y_i + k)}{\Gamma(y_i + k)\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + k}\right)^{k}$$

The estimation of k introduces an extra layer of uncertainty into a negative binomial GLM (generalized linear model). However the maximum likelihood estimator $\hat{k}$ of k is uncorrelated with the $\hat{\beta}$, according to the usual asymptotic approximations. Hence the GLM fit tends to be relatively stable with respect to estimation of k. Negative binomial give larger SE than the corresponding Poisson, depending on the size of $k = 1/\phi$. On the other hand, the coefficient estimates from a negative binomial may be similar to those produced from the corresponding Poisson. The negative binomial gives less weight to observations with large μ_i than does the Poisson, and relatively more weight to observations with small μ_i, so the coefficients.

R code

```r
## the glm.nb() function is in the MASS
## does not need to specify the theta
mod1d <- MASS::glm.nb(c1 ~ cond + sex + wt1,
                      data = cereal)
## how to calculate the theta
theta <- MASS::theta.ml(cereal$c1, fitted(mod1d))
broom::tidy(mod1d)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)   1.11       0.323     3.42  0.000617
2 cond1         0.857      0.195     4.39  0.0000116
3 sex1         -0.171      0.144    -1.19  0.234
4 wt1           0.0994     0.234     0.425 0.671
```

```r
summary(mod1d)$family
```

```
Family: Negative Binomial(3.1644)
Link function: log
```

```r
## use glm::negative.binomial
mod2d <- glm(c1 ~ cond + sex + wt1,
             family = negative.binomial(theta),
             data = cereal)
broom::tidy(mod2d)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)   1.11       0.314     3.52  0.000673
2 cond1         0.857      0.190     4.50  0.0000190
3 sex1         -0.171      0.140    -1.22  0.224
4 wt1           0.0994     0.228     0.436 0.663
```

```r
summary(mod2d)$family
```

```
Family: Negative Binomial(3.1644)
Link function: log
```

```r
# summary(mod2d)$dispersion
```

SAS code

```
/* Negative Binomial 1 */
proc nlmixed data=cereal;
  mu = exp( b0 + b1 * cond + b2 * sex + b3 * wt1 );
  ll = c1*log(k*mu) - (c1+1/k)*log(1+k*mu) + log(gamma(c1+1/k) / (gamma(1/k)*gamma(c1+1)));
  model c1 ~ general( ll );
ods select ParameterEstimates;
```

```
   ods trace on; ods show;
run;

proc genmod data=cereal;
model C1 = Cond Sex Wt1 / dist = negbin link = log;
ods select ParameterEstimates;
   ods trace on; ods show;
run;
```

Notes: Due to different parametrization, SAS[PROC NLMIXED] provides the ϕ and R[glm(negative binomial)] provides the $k = 1/\phi$.

Notes: The negative binomial distribution may be written/parameterized in several ways. No points are taken off if you parametrize on the wrong parameter, but in a real life study, be careful about which form of Negative-Binomial/Gamma-Poisson you are using. It is a general strategy to use hierarchical model and mixture distribution for Poisson. For example, if the Gamma/inverse-Gamma distribution can be approximated exactly by the log-Normal, then Q1c and Q1d will provide the same estimates and confidence intervals. More details in Dunn(2018) and Faraway(2016)

Question 2

As before use data for $Kid1$ only, at baseline only (C1), as in **question 1** above. You don't need to fit any new models.

a. Use the chart below to summarize the models you fit in **question 2** comparing Conditions and adjusting for Sex and Weight. In each entry except 'Intercept' and 'Other param' give the estimate of the rate ratio and a 95% CI (these are easy to calculate from the beta and SE). For 'Intercept' give the beta and its SE. For 'Other param', give the Scale parameter for QL, the SD for the normal error model, or the dispersion parameter for NB. $QL = Quasilikelihood$, $NB = Negative\ Binomial$.

	"Poisson Regression"	"Poisson QL"	"Poisson + Normal error"	"NB NLMIXED"
Intercept	1 (0.61, 1.39)	1 (0.33, 1.67)	0.88 (0.24, 1.53)	1.11 (0.47, 1.75)
Cond	2.37 (1.82, 3.09)	2.37 (1.51, 3.74)	2.35 (1.59, 3.48)	2.36 (1.6, 3.47)
Sex	0.82 (0.7, 0.96)	0.82 (0.62, 1.08)	0.86 (0.64, 1.14)	0.84 (0.64, 1.11)
Wt	1.22 (0.94, 1.58)	1.22 (0.78, 1.9)	1.18 (0.74, 1.89)	1.1 (0.69, 1.77)
Other	NA	\$ =\$2.94 or 1.72	\$ =\$0.54	\$k=\$3.16 or 0.316

b. Write a short paragraph summarizing the results of comparing conditions, e.g. which condition gives higher consumption and by how much.

All point estimates look roughly similar for regression coefficients. At baseline, kids in the experimental group ate more servings of cereal per day, by a factor of exp(0.86) = 2.37, 95% CI (1.51, 3.74), p<0.0001. I woudl report the results from other model than Poisson since it is the more restrictive.

c. Write a short paragraph comparing the model fits, e.g. differences between parameter estimates across models.

Roughly lognormal and NB are quite similar for all parameters and SEs. For condition, point estimates are quite similar (except for minus signs) but Poisson SEs are too small for all predictors (main point), QL SE is bigger for Cond but I don't think this is a general result. In practice I would be fine with any of the three overdispersed models and it would depend on what else I wanted to do with the models. Therefore, the model fit of Possion regression is not as good as the other three approaches, with underestimation of standard errors of estimates.