Case Study: Applying Generalized Linear Models

Dr. Kempthorne

May 12, 2016

Contents

1 Generalized Linear Models of Semi-Quantal Biological Assay Data

1.1 Coal miners Pneumoconiosis Data

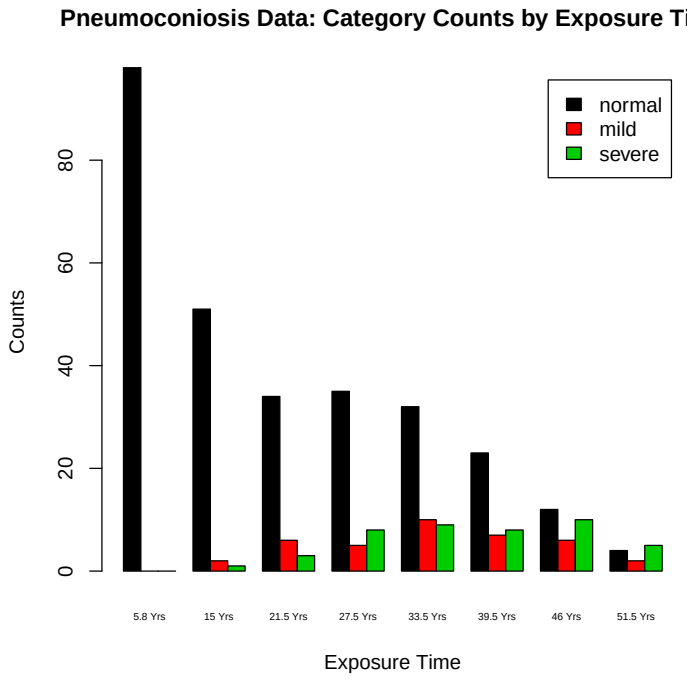
McCullagh and Nelder (1989) discuss the application of generalized linear models to modeling the incidence and severity of lung disease in coal miners as it relates to the degree of exposure to coal dust. They introduce the data as follows:

> The data, taken from Ashford (1959), concern the degree of pneumoconiosis in coalface workers as a function of exposure t measured in years. Severity of disease is measured radiologically and is, of necessity, qualitative. A four-category version of the ILO rating scale was used initially, but the two most severe categories were subsequently combined.
>
> McCullagh and Nelder (1989), p. 179.

Using R and Yee's (2010) R-package VGAM (Vector Generalized Linear and Additive Models), we load in the data set *pneumo*, compute summary statistics and plots.

```
> # 0.1 Load R packages ====
> require(stats)
> require(graphics)
> library("VGAM")
> # 1.1  Display and summarize dataset pneumo ====
> print(pneumo)

  exposure.time normal mild severe
1           5.8     98    0      0
2          15.0     51    2      1
3          21.5     34    6      3
4          27.5     35    5      8
5          33.5     32   10      9
6          39.5     23    7      8
7          46.0     12    6     10
8          51.5      4    2      5

> summary(pneumo)

 exposure.time       normal           mild           severe
 Min.   : 5.80   Min.   : 4.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:19.88   1st Qu.:20.25   1st Qu.: 2.00   1st Qu.: 2.50
 Median :30.50   Median :33.00   Median : 5.50   Median : 6.50
 Mean   :30.04   Mean   :36.12   Mean   : 4.75   Mean   : 5.50
 3rd Qu.:41.12   3rd Qu.:39.00   3rd Qu.: 6.25   3rd Qu.: 8.25
 Max.   :51.50   Max.   :98.00   Max.   :10.00   Max.   :10.00
```

```
>

> # 1.2  Plot data
>
> # Attaching the dataset allows access to column variables using their names.
>
> names(pneumo)

[1] "exposure.time" "normal"        "mild"          "severe"

> attach(pneumo)
> matrix.counts<-t(as.matrix(pneumo[,2:4]))
> dimnames(matrix.counts)[[2]]<-paste( as.character(pneumo$exposure.time)," Yrs",sep="")
> barplot(matrix.counts, beside=TRUE, col=(c(1,2,3)),
+   legend.text=(c("normal","mild","severe")),
+   cex.names=.5,
+   ylab="Counts",main="Pneumoconiosis Data: Category Counts by Exposure Time",
+   xlab="Exposure Time")
>
```



1.2 Multinomial Model for Incidence Counts

Let t_i denote the ith exposure time in the data set, $i = 1, \ldots, 8$ and define $y_{i,j}$ to be the incidence count for exposure time t_i of category j: normal(j=1), mild(j=2), severe(j=3). With this notation, define $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, y_{i,3})$ to be the multivariate random vector of counts for exposure time t_i.

Consider independent multinomial models for the $\boldsymbol{y}_i$ which allow the multinomial probabilities to vary with the exposure time t_i:

- $\boldsymbol{y}_i, i = 1, \ldots, 8$ are independent multinomial distributions
- For each exposure time t_i, let $m_i = y_{i,1} + y_{i,2} + y_{i,3}$ be the sample size of men with exposure time t_i.
- Let the multinomial distributions vary with i: $(\pi_1, \pi_2, \pi_3) = (\pi_{i,1}, \pi_{i,2}, \pi_{i,3})$

  $$\boldsymbol{y}_i = (Y_{i,1}, Y_{i,2}, Y_{i,3}) \sim Multinomial(m_i, \pi_{i,1}, \pi_{i,2}, \pi_{i,3}))$$

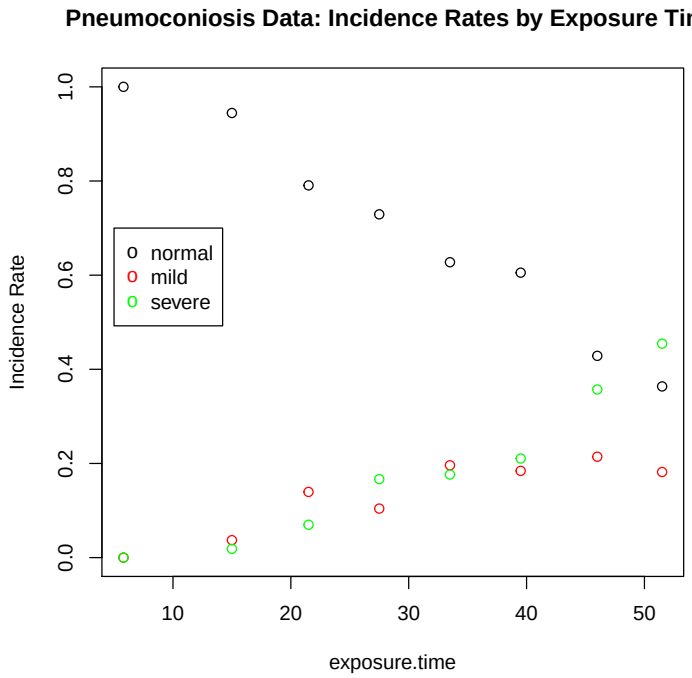
  with $\sum_{j=1}^{3} \pi_{i,j} = 1$ for each group i.

Simple estimates of the multinomial probabilities are obtained using the marginal distribution of each $Y_{i,j} \sim Binomial(m_i, \pi_{i,j})$:

$$\hat{\pi}_{i,j} = y_{i,j}/m_i$$

These estimates are the incidence rates of each category per exposure time. We plot these together for all exposure times.

```
> # Display data together as incidence rate per exposure time
> # for the 3 categories: normal, mild, and severe.
>
> m.count=normal+mild+severe
> par(mfcol=c(1,1))
> plot(exposure.time, normal/m.count, ylab="Incidence Rate", ylim=c(0,1))
> points(exposure.time, mild/m.count, col='red')
> points(exposure.time, severe/m.count, col='green')
> title(main="Pneumoconiosis Data: Incidence Rates by Exposure Time")
> legend(x=5, y=.7, legend=c("normal", "mild", "severe"),
+        pch=c("o","o","o"), col=c("black","red","green"))
```



When the categories ($j = 1, 2, 3$) are ordered, it is convenient to work with cumulative response probabilities:

$$\begin{aligned} \gamma_{i,1} &= \pi_{i,1} \\ \gamma_{i,2} &= \pi_{i,1} + \pi_{i,2} \\ \gamma_{i,3} &= 1 \end{aligned}$$

With these cumulative response probabilities, consider the log-odds of staying in category 1 (normal) as a function of exposure time, i.e,.

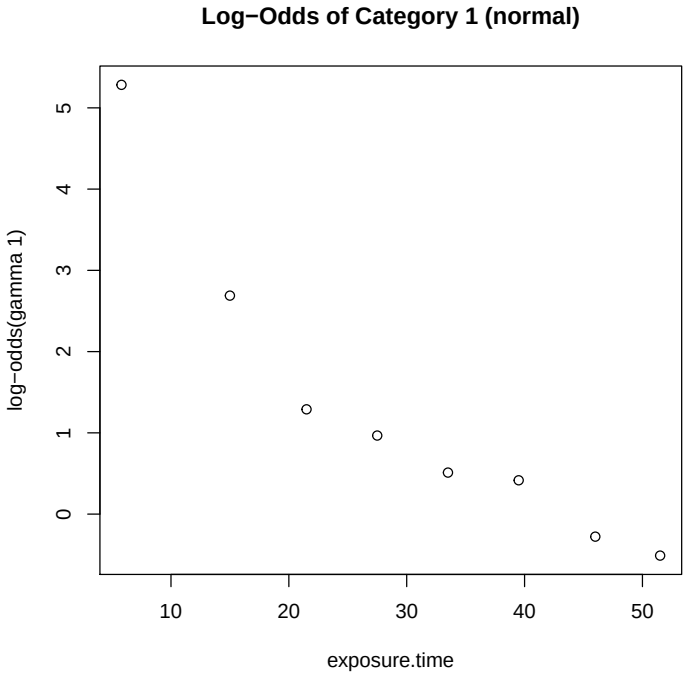
$$\log\left(\frac{\gamma_{i,1}}{1-\gamma_{i,1}}\right) \text{ vs. } t_i.$$

To allow for extreme count values (0 or m_i), estimate the log-odds with

$$\log\left(\frac{y_{i,1}+\frac{1}{2}}{m_i - y_{i,1}+\frac{1}{2}}\right)$$

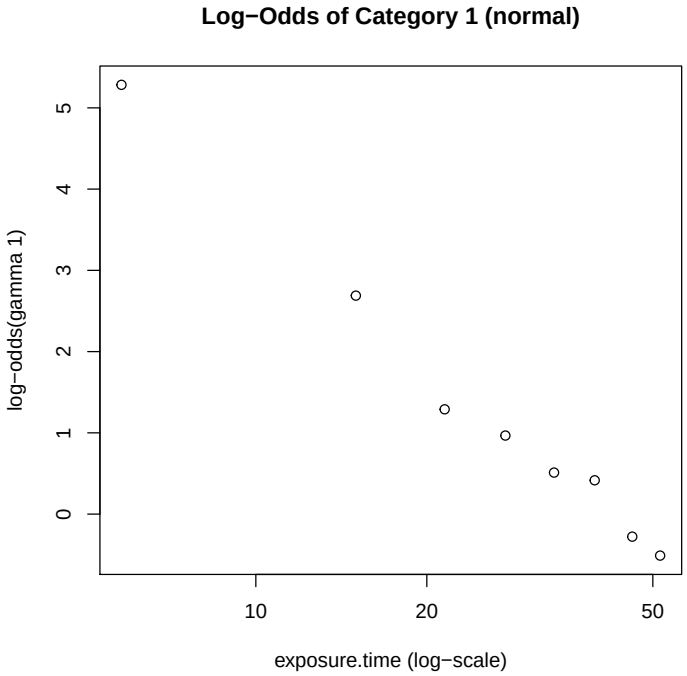
The relationship of the log-odds to exposure time can be displayed in a plot:

```
> logoddsGamma1<-log( (normal + 1/2)/(m.count - normal +1/2))
> plot(x=exposure.time, y=logoddsGamma1, ylab="log-odds(gamma 1)",
+      main="Log-Odds of Category 1 (normal)" )
```



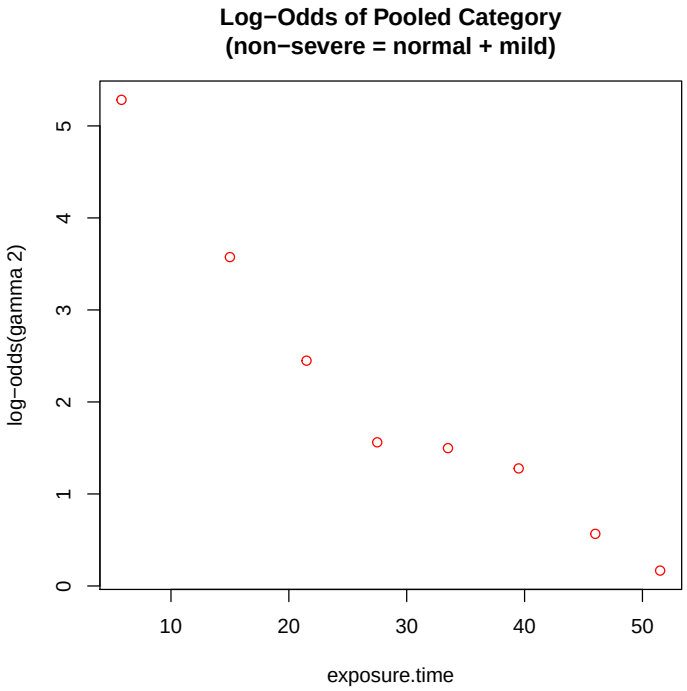
The relationship appears close to linear when we plot exposure time on the log scale.

```
> plot(x=exposure.time, y=logoddsGamma1, ylab="log-odds(gamma 1)",
+      xlab="exposure.time (log-scale)",log="x",
+      main="Log-Odds of Category 1 (normal)")
```



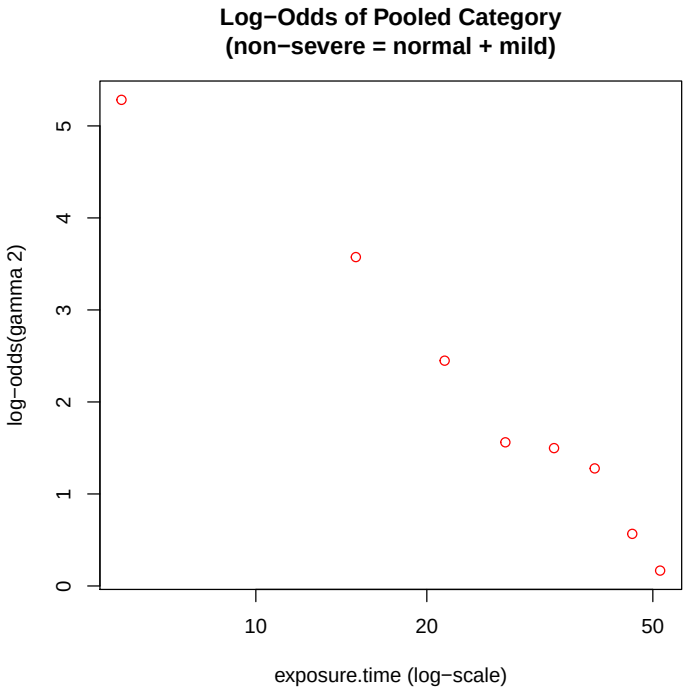
Analogous computations and plots are made for the log-odds of the pooled "non-severe" category (normal plus mild).
Using the following estimate for the log-odds, we plot the relationship:
with $\log\left(\frac{y_{i,1}+y_{i,2}+\frac{1}{2}}{m_i-y_{i,1}-y_{i,2}+\frac{1}{2}}\right)$

```
> logoddsGamma2<-log( (normal + mild +1/2)/(m.count - normal -mild +1/2))
> plot(x=exposure.time, y=logoddsGamma2, ylab="log-odds(gamma 2)",col=2,
+      main="Log-Odds of Pooled Category\n(non-severe = normal + mild)")
```



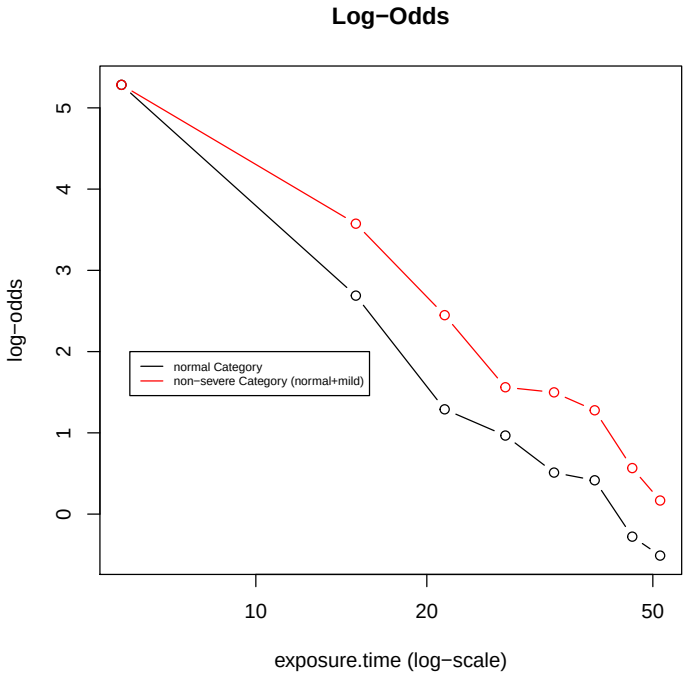
Again, the relationship appears close to linear when we plot exposure time on the log scale.

```
> plot(x=exposure.time, y=logoddsGamma2, ylab="log-odds(gamma 2)",
+      xlab="exposure.time (log-scale)", log="x",col=2,
+           main="Log-Odds of Pooled Category\n(non-severe = normal + mild)")
```



To compare these log-odds relationships with exposure time we plot them together:

```
> plot(x=exposure.time, y=logoddsGamma1, ylab="log-odds",
+      xlab="exposure.time (log-scale)",log="x",
+      main="Log-Odds", type="b")
> lines(x=exposure.time, y=logoddsGamma2,
+       type="b",col='red')
> legend(x=6,y=2,
+        legend=c("normal Category", "non-severe Category (normal+mild)"),
+        col=c('black','red'), lty=c(1,1),cex=.6)
```


Log−Odds
log−odds
0
1
2
3
4
5
10
20
50
exposure.time (log−scale)
normal Category
non−severe Category (normal+mild)

1.3 Proportional Odds Model: Parallel Linear Logit Model

McCullagh and Nelder comment that these plots of the transformed variables suggest considering the model:

$$\log[\gamma_{i,j}/(1-\gamma_{i,j})] = \theta_j - \beta \log t_i,\ j = 1, 2;\ i = 1, \ldots, 8.$$

Yee's (2010) R-package VGAM (Vector Generalized Linear and Additive Models) provides the function $vglm()$ to fit this model.

```
> pneumo <- transform(pneumo, log.expos.time = log(exposure.time))
> fit1<-vglm(cbind(normal, mild, severe) ~ log.expos.time,
+            cumulative(reverse=FALSE, parallel=TRUE),data = pneumo)
```

The R object $fit1$ (a class $vglm$ object) provides details of the fitted generalized linear model. First, print a summary of the fit:

```
> summary(fit1)

Call:
vglm(formula = cbind(normal, mild, severe) ~ log.expos.time,
    family = cumulative(reverse = FALSE, parallel = TRUE), data = pneumo)

Pearson residuals:
                  Min       1Q Median     3Q    Max
logit(P[Y<=1]) -1.248 -0.07164 0.1441 0.3086 0.7714
logit(P[Y<=2]) -1.044 -0.18415 0.3093 0.3353 0.5048

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1    9.6761     1.3241   7.308 2.72e-13 ***
(Intercept):2   10.5817     1.3454   7.865 3.69e-15 ***
log.expos.time  -2.5968     0.3811  -6.814 9.50e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: 5.0268 on 13 degrees of freedom

Log-likelihood: -25.0903 on 13 degrees of freedom

Number of iterations: 4

Exponentiated coefficients:
```

```
log.expos.time
    0.07451115
```

Important components of the summary are:

- Coefficients: maximum-likelihood estimates of the model parameters. In addition to the *Estimates*, estimates of their standard deviation (*Std. Error*), their ratio (*z value*), and the P-value for the (asymptotic) test of whether the underlying coefficient is zero.

  Note: the coefficients specify the parallel lines defining the log-odds as a function of the $log(exposure\ time)$.

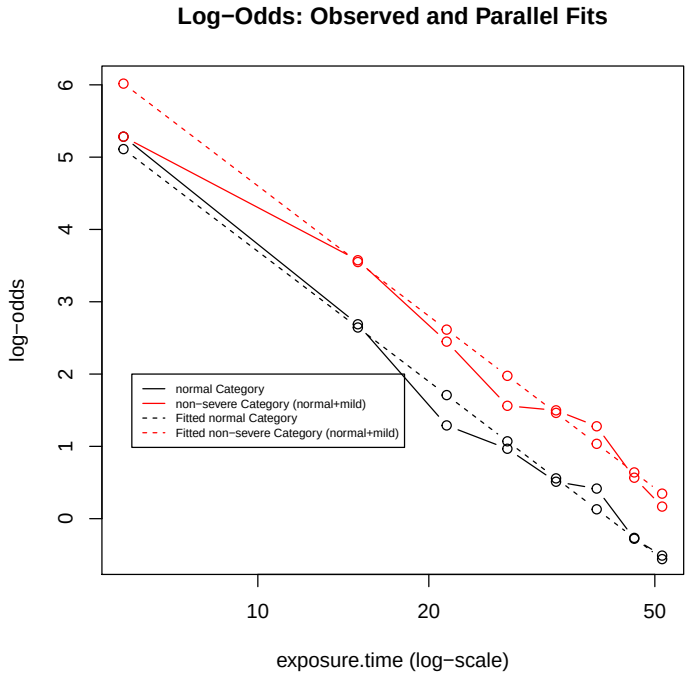
- Log-Likelihood: -25.0903 on 13 degrees of freedom.

  Note: the degrees of freedom are the total degrees of freedom $(8 \times (3-1))$ minus the number of estimated parameters 3.

- Residual deviance (see *Deviance* definition in lecture notes)

To the plot of observed log-odds vs exposure-time, we add the ML-Fitted log-odds according to the (parallel) cumulative log-odds model.

```
> plot(x=exposure.time, y=logoddsGamma1, ylab="log-odds",
+      xlab="exposure.time (log-scale)",log="x",
+      main="Log-Odds: Observed and Parallel Fits", type="b",ylim=c(min(logoddsGamma1), 6))
> lines(x=exposure.time, y=logoddsGamma2,
+       type="b",col='red')
> lines(exposure.time, y=fit1@predictors[,1],type="b",lty=2, col='black')
> lines(exposure.time, y=fit1@predictors[,2],type="b",lty=2, col='red')
> legend(x=6,y=2.,
+        legend=c("normal Category", "non-severe Category (normal+mild)",
+     "Fitted normal Category", "Fitted non-severe Category (normal+mild)"),
+        col=c('black','red','black','red'), lty=c(1,1,2,2),cex=.6)
```



The *vglm* object $fit1$ includes fitted values for the multinomial probabilities. These are printed out together with the observed frequencies:

```
> pneumo.rates<-data.frame(exposure.time, normal= normal/m.count,
+                      mild=mild/m.count, severe=severe/m.count)
> pneumo.fittedrates<-data.frame(cbind(exposure.time,fit1@fitted.values))
> print(cbind(pneumo.rates, pneumo.fittedrates),digits=3)

  exposure.time normal  mild severe exposure.time normal    mild  severe
1           5.8  1.000 0.000 0.0000           5.8  0.994 0.00356 0.00243
2          15.0  0.944 0.037 0.0185          15.0  0.934 0.03843 0.02794
3          21.5  0.791 0.140 0.0698          21.5  0.847 0.08509 0.06821
4          27.5  0.729 0.104 0.1667          27.5  0.745 0.13364 0.12181
5          33.5  0.627 0.196 0.1765          33.5  0.636 0.17615 0.18802
6          39.5  0.605 0.184 0.2105          39.5  0.532 0.20558 0.26210
7          46.0  0.429 0.214 0.3571          46.0  0.434 0.22078 0.34536
8          51.5  0.364 0.182 0.4545          51.5  0.364 0.22202 0.41430
```

1.4 General/Independent Linear Logit Models

The model of the previous section assumes parallel linear log-odds relationships on log exposure time. A more general model allows these lines to have different slopes.

The R-code below fits this model.

```
> #pneumo <- transform(pneumo, log.expos.time = log(exposure.time))
> fit2<-vglm(cbind(normal, mild, severe) ~ log.expos.time,
+            cumulative(reverse=FALSE, parallel=FALSE),data = pneumo)
```

The R object $fit2$ (a class $vglm$ object) provides details of the fitted generalized linear model. We print out the summary of this fit and focus on the coefficients corresponding to the slope parameters.

```
> summary(fit2)

Call:
vglm(formula = cbind(normal, mild, severe) ~ log.expos.time,
    family = cumulative(reverse = FALSE, parallel = FALSE), data = pneumo)

Pearson residuals:
                 Min      1Q Median     3Q    Max
logit(P[Y<=1]) -1.15 -0.1457 0.1249 0.3824 0.7288
logit(P[Y<=2]) -1.15 -0.0506 0.1886 0.2864 0.5659

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept):1      9.5933     1.3308   7.208 5.66e-13 ***
(Intercept):2     11.1048     1.8930   5.866 4.45e-09 ***
log.expos.time:1  -2.5713     0.3839  -6.698 2.11e-11 ***
log.expos.time:2  -2.7435     0.5323  -5.155 2.54e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1

Residual deviance: 4.8844 on 12 degrees of freedom

Log-likelihood: -25.0191 on 12 degrees of freedom

Number of iterations: 6
```

```
Exponentiated coefficients:
log.expos.time:1 log.expos.time:2
      0.07643613       0.06434155
```

Important components of the summary are:

- Coefficients: maximum-likelihood estimates of the model parameters. In addition to the *Estimates*, estimates of their standard deviation (*Std. Error*), their ratio (*z value*), and the P-value for the (asymptotic) test of whether the underlying coefficient is zero.

  Note: the coefficients specify two lines:

  $$logit(P[Y \leq 1]) = [(Intercept) : 1] + [log.expos.time : 1] \times log(Exposure.Time)$$

  $$logit(P[Y \leq 2]) = [(Intercept) : 2] + [log.expos.time : 2] \times log(Exposure.Time)$$

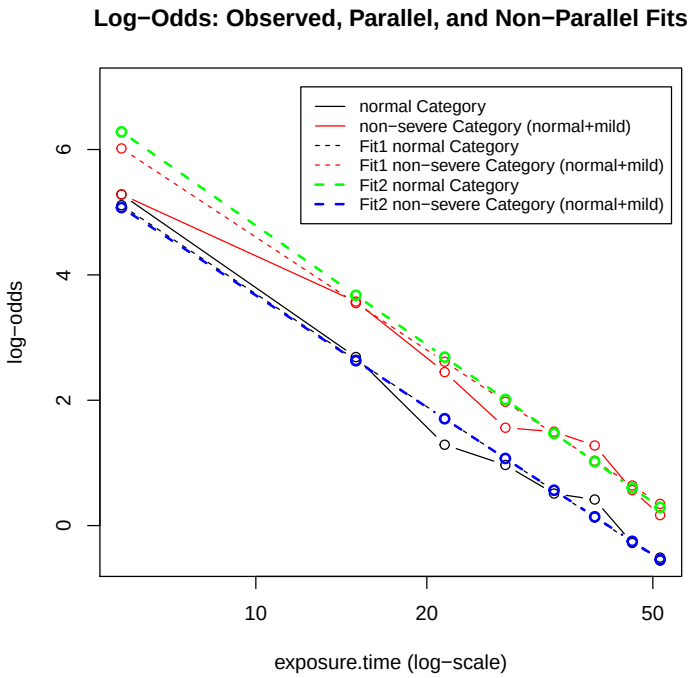
  The estimated slopes are very close -2.5713 versus -2.7435, and very similar to the slope of -2.5968 in the first model.

- Log-Likelihood: -25.0191 on 12 degrees of freedom.

  Note: the degrees of freedom are the total degrees of freedom $(8 \times (3-1))$ minus the number of estimated parameters 4 (two intercepts and two slopes).

The ML-Fitted log-odds according to this (non-parallel) cumulative log-odds model can be added to the plot given before:

```
> plot(x=exposure.time, y=logoddsGamma1, ylab="log-odds",
+      xlab="exposure.time (log-scale)",log="x",
+      main="Log-Odds: Observed, Parallel, and Non-Parallel Fits",
+      type="b",ylim=c(min(logoddsGamma1), 7))
> lines(x=exposure.time, y=logoddsGamma2,
+       type="b",col='red')
> lines(exposure.time, y=fit1@predictors[,1],type="b",lty=2, col='black')
> lines(exposure.time, y=fit1@predictors[,2],type="b",lty=2, col='red')
> lines(exposure.time, y=fit2@predictors[,1],type="b",lty=2, col='blue',lwd=2)
> lines(exposure.time, y=fit2@predictors[,2],type="b",lty=2, col='green',lwd=2)
> legend(x=12,y=7.,
+        legend=c("normal Category", "non-severe Category (normal+mild)",
+     "Fit1 normal Category", "Fit1 non-severe Category (normal+mild)",
+     "Fit2 normal Category", "Fit2 non-severe Category (normal+mild)"),
+        col=c('black','red','black','red','green','blue'), lty=c(1,1,2,2,2,2),
+     lwd=c(1,1,1,1,2,2),cex=.8)
```



This plot demonstrates that model $fit2$ with independent linear logit functions is very close to model $fit1$ with parallel linear logit functions.

1.5 Likelihood-Ratio Test of Proportional Odds

We use the $VGAM$-package function $lrtest\_vglm$ to conduct a likelihood ratio test comparing the two models.

```
> lrtest_vglm(fit2,fit1)

Likelihood ratio test

Model 1: cbind(normal, mild, severe) ~ log.expos.time
Model 2: cbind(normal, mild, severe) ~ log.expos.time
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -25.019
2  13 -25.090  1 0.1424     0.7059

>
```

Note that the likelihood ratio test statistic is

$$\begin{aligned} LR - Statistic &= -2 \times (Log - likelihood[fit1] - Log - Likelihood[fit2]) \\ &= -2 \times (-25.0903 - [-25.0191]) \\ &= -2 \times (+.0712) = +.1424 \end{aligned}$$

Under the null hypothesis of no improvement allowing the slopes of the log-odds functions to be different, the statistic is asymptotically distributed as a Chi-Square random variable with degrees of freedom equal to the difference in degrees of freedom of the two models (1 in this case). The large P-Value ($0.7059 >> 0.05$) indicates that improvement of model $fit2$ over model $fit1$ is not statistically significant.

1.6 References

Ashford (1959). An Approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics* **15**: 573-81.

McCullagh and Nelder (1989). Generalized Linear Models, 2nd Ed. Chapman and Hall, New York.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software,* **32:** 1-34. http://www.jstatsoft.org/v32/i10/.

18.655 Mathematical Statistics

Spring 2016