Clustered data models - Exercises 2

1. (8.2 in Agresti) Using $\mathrm{E}(y) = \mathrm{E}[\mathrm{E}(y|x)]$ and $\mathrm{Var}(y) = \mathrm{E}[\mathrm{Var}(y|x)] + \mathrm{Var}[\mathrm{E}(y|x)]$, derive the mean and variance of the beta-binomial distribution.

Solution. $y \sim$ Neg-Bin if $\pi \sim \mathrm{Beta}(\alpha_1, \alpha_2)$ and $y|\pi \sim \mathrm{Bin}(n, \pi)$. We have that

$$\begin{aligned}
\mathrm{E}(\pi) &= \frac{\alpha_1}{\alpha_1+\alpha_2} =: \mu, \\
\mathrm{Var}(\pi) &= \rho\mu(1-\mu), \text{ where } \rho := \frac{1}{\alpha_1+\alpha_2+1}.
\end{aligned}$$

Now,

$$\mathrm{E}(y) = \mathrm{E}[\mathrm{E}(y|\pi)] = \mathrm{E}(\pi) = \mu,$$

and

$$\begin{aligned}
\mathrm{Var}(y) &= \mathrm{E}\{\mathrm{Var}(y|x)\} + \mathrm{Var}\{E(y|\pi)\} \\
&= \mathrm{E}\{n\pi(1-\pi)\} + \mathrm{Var}\{n\pi\} \\
&= n\{\mathrm{E}(\pi) - \mathrm{E}(\pi^2)\} + n^2\mathrm{Var}(\pi) \\
&= n\{\mu - [\mathrm{Var}(\pi) + (\mathrm{E}\pi)^2]\} + n^2\rho\mu(1-\mu) \\
&= n\{\mu - \rho\mu(1-\mu) - \mu^2\} + n^2\rho\mu(1-\mu) \\
&= n[1 + (n-1)\rho]\mu(1-\mu).
\end{aligned}$$

2. (8.3 in Agresti) Let y_1 and y_2 be independent and identically distributed negative binomial variates with dispersion parameter γ. (See the definition of γ in Section 7.3.2 of Agresti).

a) Show that $y_1 + y_2$ is negative binomial with dispersion parameter $\gamma/2$.

Solution. It is easier to work with the following parametrization: $y \sim \text{Neg-Bin}(k, p)$ if y is the number of failures before obtaining k successes in a sequence of independent Bernoulli-trials where the probability of success is p.

Solution 1. We use the fact that the negative-binomial distribution is a gamma-mixture of Poisson distributions. We have that $y_1 \sim \text{Neg-Bin}(k_1, p)$ and $y_2 \sim \text{Neg-Bin}(k_2, p)$ if

$$\begin{aligned}
y_1|\lambda_1 \sim \mathrm{Poi}(\lambda_1), \quad \lambda_1 \sim \mathrm{Gamma}(k_1, \frac{p}{1-p}), \\
y_2|\lambda_2 \sim \mathrm{Poi}(\lambda_2), \quad \lambda_2 \sim \mathrm{Gamma}(k_2, \frac{p}{1-p}).
\end{aligned}$$

The sum of independent Poisson random variables is Poisson-distributed, so that

$$\{y_1 + y_2|\lambda_1, \lambda_2\} \sim \mathrm{Poi}(\lambda_1 + \lambda_2).$$

Further, the sum of two independent Gamma variables with the same rate parameter is also Gamma-distributed:

$$\lambda_1 + \lambda_2 \sim \mathrm{Gamma}(k_1 + k_2, \frac{p}{1-p}).$$

So we have that $y|\lambda \sim \mathrm{Poi}(\lambda)$ and $\lambda \sim \Gamma(k_1 + k_2, p/(1-p))$, where $y := y_1 + y_2$ and $\lambda := \lambda_1 + \lambda_2$. This implies that $y \sim \text{Neg-Bin}(k_1 + k_2, p)$. Further, if y_1 and y_2 are identically distributed, $k_1 + k_2 = 1/\gamma + 1/\gamma = 2/\gamma$, so that the dispersion parameter for y is $\gamma/2$.

Solution 2. Let us denote $y := y_1 + y_2$. Now, we may apply the moment generating function as follows:

$$M_y(t) = \mathrm{E}e^{t(y_1+y_2)} = \left(\mathrm{E}e^{ty_1}\right)\left(\mathrm{E}e^{ty_2}\right) = \left(\frac{p}{1-(1-p)e^t}\right)^{k_1}\left(\frac{p}{1-(1-p)e^t}\right)^{k_2} = \left(\frac{p}{1-(1-p)e^t}\right)^{k_1+k_2}.$$

The second equality follows from y_1 and y_2 being independent. The result implies that $y \sim$ Neg-Bin(k_1+k_2, p). Further, if y_1 and y_2 are identically distributed, $k_1 + k_2 = 1/\gamma + 1/\gamma = 2/\gamma$, so that the dispersion parameter for y is $\gamma/2$.

b) Conditional on $y_1 + y_2$, show that y_1 has a beta-binomial distribution.

Solution 1. Given $y := y_1 + y_2$ and λ_1, λ_2, the conditional distribution of y_1 is binomial:

$$\{y_1|y, \lambda_1, \lambda_2\} \sim \mathrm{Bin}(y, \frac{\lambda_1}{\lambda_1+\lambda_2}).$$

This is a special case of the more general result concerning the connection between Poisson and Multinomial distributions (see section 7.2.1 in Agresti, 2015). Further, there is a connection between Gamma and Beta distributions:

$$\frac{\lambda_1}{\lambda_1+\lambda_2} \sim \mathrm{Beta}(k_1, k_2),$$

if λ_1 and λ_2 are independent, and $\lambda_i \sim \mathrm{Gamma}(k_i, \beta)$, for $i = 1, 2$, where β is a shared rate parameter. Combining these results, we have that

$$\{y_1|y\} \sim \text{Beta-Bin}(y, k_1, k_2).$$

Solution 2. (Only the idea is given here). Denoting $y := y_1 + y_2$ we have that

$$p(y_1|y) = \frac{p(y_1, y)}{p(y)} = \frac{p(y_1, y_2)}{p(y)} = \frac{p(y_1)p(y_2)}{p(y)}.$$

By substituting in the formula the point mass functions of y_1, y_2 and y, we obtain the point mass function of a beta-binomial random variable.

c) State the multicategory extension of (b) that yields a Dirichlet-multinomial distribution. Explain the analogy with the corresponding Poisson-multinomial result (Section 7.2.1. in Agresti).

Solution. If the Poisson variables $y_i \sim \mathrm{Poi}(\lambda_i),\ i = 1, ..., d$, are independent, then

$$\{(y_1, ..., y_d)|y, \boldsymbol{\lambda}\} \sim \mathrm{Multin}(y, (\lambda_1/\lambda, ..., \lambda_d/\lambda)),$$

where $y = \sum_{i=1}^d y_i$, $\lambda = \sum_{i=1}^d \lambda_i$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$. Further, if $\lambda_i,\ i = 1, ..., d$, are independent and $\lambda_i \sim \mathrm{Gamma}(k_i, \beta)$, where β is a shared rate parameter, then

$$\boldsymbol{\lambda} \sim \mathrm{Dirichlet}(k_1, ..., k_d).$$

It follows that the marginal distribution of $(y_1, ..., y_d)$, given y, is Dirichlet-Multinomial:

$$(y_1, ..., y_d)|y \sim \text{Dirichlet-Multin}(y, (k_1, ..., k_d)).$$

Thus, we see that that the connection between the Poisson and Dirichlet-Multinomial distributions is similar to that between the Poisson and Multinomial distributions.

3. (8.6 in Agresti) Motivation for the quasi-score equations (Equation 8.2 in Agresti): suppose we replace $\nu(\mu_i)$ by known variance ν_i. Show that the equations result from the weighted least squares approach of minimizing $\sum_i[(y_i - \mu_i)^2/\nu_i]$.

Solution. By differentiating with respect to $\boldsymbol{\beta}$ the weighted sum of squares we obtain that

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=1}^n \frac{(y_i-\mu_i)^2}{\nu_i} = -2\sum_{i=1}^n \frac{(y_i-\mu_i)}{\nu_i}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}.$$

This equals $\mathbf{0}$ if

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\nu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

which corresponds the QL equation in (8.2) with $\nu_i = \nu(\mu_i)$. (Note that in Agresti's notation, $\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$ is a row vector, so that $\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T$ is a column vector.)

4. (7.31 in Agresti) The table below (Table 7.5 in Agresti; the data available at www.stat.ufl.edu/~aa/glm/data), summarizes responses of 1308 subjects to the question: within the past 12 months, how many people have you known personally that were victims of homicide? The table shows responses by race, for those who identified their race as white or as black.

a) Let y_i denote the response for subject i and let $x_i = 1$ for blacks and $x_i = 0$ for whites. Fit the Poisson GLM $\log \mu_i = \beta_0 + \beta x_i$ and interpret $\hat{\beta}$.

b) Describe factors of heterogeneity such that a Poisson GLM may be inadequate. Fit the corresponding negative binomial GLM, and estimate how the variance depends on the mean. What evidence does this model fit provide that the Poisson GLM had overdispersion? (Table 7.5 also shows the fits for these two models.)

c) Show that the Wald 95% confidence interval for the ratio of means for blacks and whites is (4.2, 7.5) for the Poisson GLM but (3.5, 9.0) for the negative binomial GLM. Which do you think is more reliable? Why?

```r
homic <- read.table("../data/Homicides.dat", header = TRUE)
homic$race <- factor(homic$race, labels = c("white","black"))

# Frequency table
n <- table(homic$race)
tb <- table(homic$count, homic$race)


# Fitted values for the Poisson distribution
mod.pois <- glm(count ~ race, family = poisson, homic)
mu.pois.black <- exp(sum(mod.pois$coef))
mu.pois.white <- exp(mod.pois$coef[1])
tb.pois <- round(cbind(n[1]* dpois(0:6, mu.pois.white),
                       n[2] * dpois(0:6, mu.pois.black)), 1)

# Fitted values for the negative binomial distributin
library(MASS)
mod.nb <- glm.nb(count ~ race, homic)
mu.nb.black <- exp(sum(mod.nb$coef))
mu.nb.white <- exp(mod.nb$coef[1])
tb.nb <- round(cbind(n[1]* dnbinom(0:6, mu = mu.nb.white, size = mod.nb$theta),
                     n[2] * dnbinom(0:6, mu = mu.nb.black, size = mod.nb$theta)), 1)
dimnames(tb.nb) <- dimnames(tb.pois) <- dimnames(tb)

# Compare Table 7.5 in Agresti
cbind(tb, tb.pois, tb.nb)
##   white black  white black  white black
## 0  1070   119 1047.7  94.3 1064.9 122.8
## 1    60    16   96.7  49.2   67.5  17.9
## 2    14    12    4.5  12.9   12.7   7.8
## 3     4     7    0.1   2.2    2.9   4.1
## 4     0     3    0.0   0.3    0.7   2.4
## 5     0     2    0.0   0.0    0.2   1.4
## 6     1     0    0.0   0.0    0.1   0.9
```

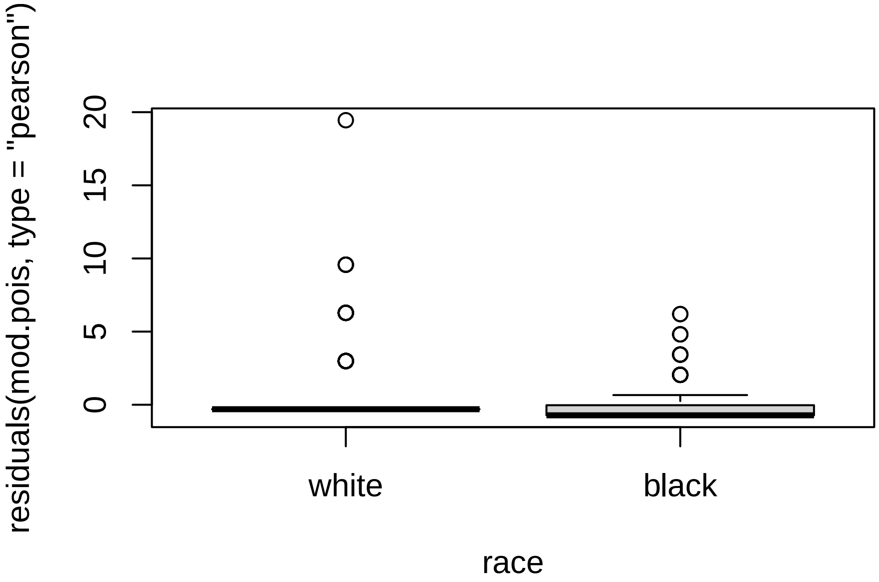
Solution. a) On average, blacks knew about 5.7 times as many homicide victims as whites.

```
summary(mod.pois)
##
## Call:
## glm(formula = count ~ race, family = poisson, data = homic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
## raceblack    1.73314    0.14657   11.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 962.80  on 1307  degrees of freedom
## Residual deviance: 844.71  on 1306  degrees of freedom
## AIC: 1122
##
## Number of Fisher Scoring iterations: 6
exp(mod.pois$coef[2])
## raceblack
##  5.658419
```

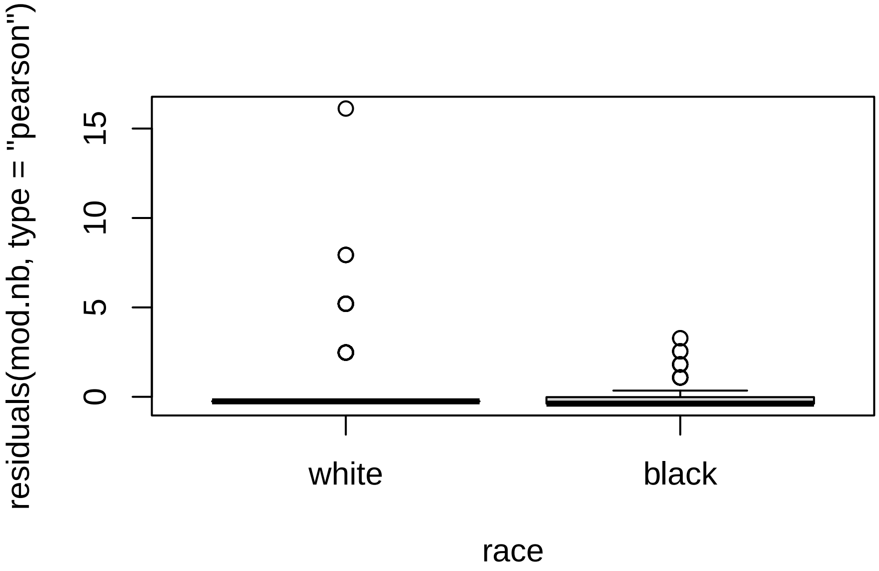
b) Reasons for heterogeneity: There may be clustering in the dataset due to differences between areas; in some areas the homicides may be more common. Further, there may be differences on an individual level; different people might know different numbers of people, that is, the sizes of their social networks differ. Some people may be informed about the causes of deaths more likely than others, or they may just be more interested to know.

The variance function is $\mathrm{Var}(y)) = \mu + \gamma\mu^2$, and the $\hat{\gamma} = 1/0.2023 = 4.943$. This estimate, the larger standard errors of the estimates, and the smaller AIC value (1002 vs. 1122) indicate that the Poisson model is inadequate.

```
summary(mod.nb)
##
## Call:
## glm.nb(formula = count ~ race, data = homic, init.theta = 0.2023119205,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
## raceblack     1.7331     0.2385   7.268 3.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)
##
##     Null deviance: 471.57  on 1307  degrees of freedom
## Residual deviance: 412.60  on 1306  degrees of freedom
## AIC: 1001.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2023
##           Std. Err.:  0.0409
##
##  2 x log-likelihood:  -995.7980
plot(residuals(mod.pois, type = "pearson") ~ race, homic)
```



```
plot(residuals(mod.nb, type = "pearson") ~ race, homic)
```



c)

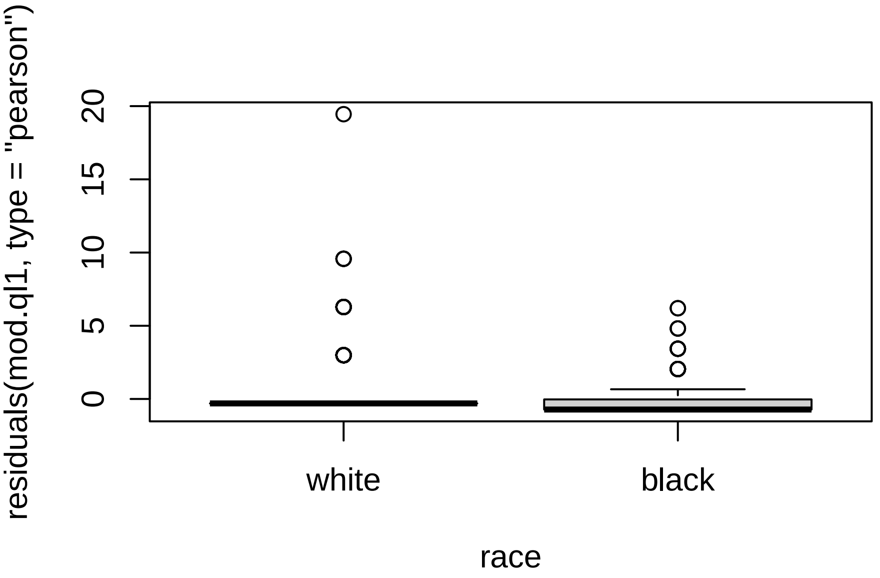
```
# Wald interval for exp(beta) in Poisson case
exp(c(1.733-1.96*0.14657,1.733+1.96*0.14657))
```

```
## [1] 4.244919 7.540415
# Wald interval for exp(beta) in Poisson case
exp(c(1.733-1.96*0.2385,1.733+1.96*0.2385))
## [1] 3.545007 9.029166
# The latter is more reliable because the variance function,
# and consequently the standard errors, is more realistic
```

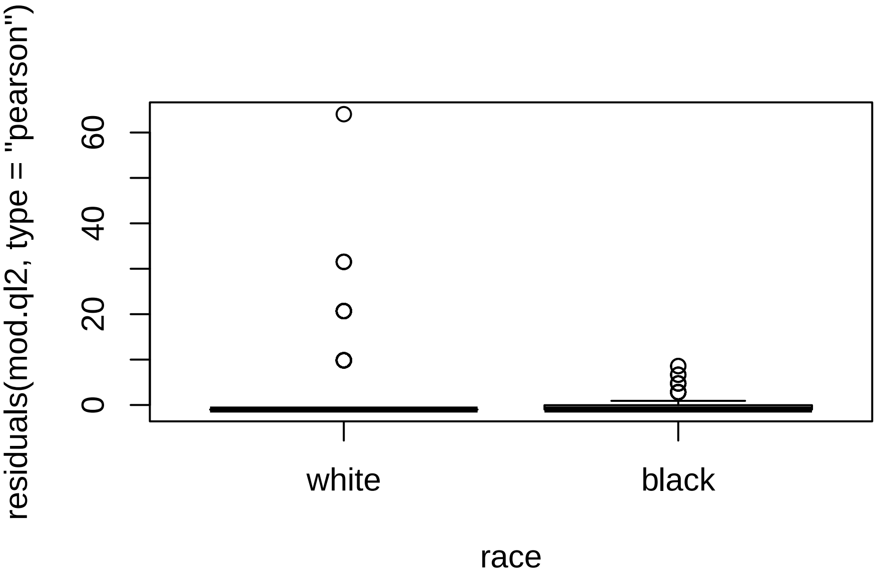
5. (8.14 in Agresti) Use QL methods to analyze Table 7.5 on counts of homicide victims. Interpret, and compare results with Poisson and negative binomial GLMs.

Solution The standard errors in quasi-likelihood cases are similar to those in negative-binomial GLM. Race is still a highly significant predictor. There are no changes in the parameter estimates. The residual plots are not very useful here, because the distributions are so skewed. However, the QL1 model appears to be better because the residual variance is more homoscedastic.

```
mod.ql1 <- glm(count ~ race, family = quasi(link="log", variance = "mu"), homic)
summary(mod.ql1)
##
## Call:
## glm(formula = count ~ race, family = quasi(link = "log", variance = "mu"),
##     data = homic)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.3832     0.1283  -18.57   <2e-16 ***
## raceblack     1.7331     0.1937    8.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 1.745693)
##
##     Null deviance: 962.80  on 1307  degrees of freedom
## Residual deviance: 844.71  on 1306  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
plot(residuals(mod.ql1, type = "pearson") ~ race, homic)
```



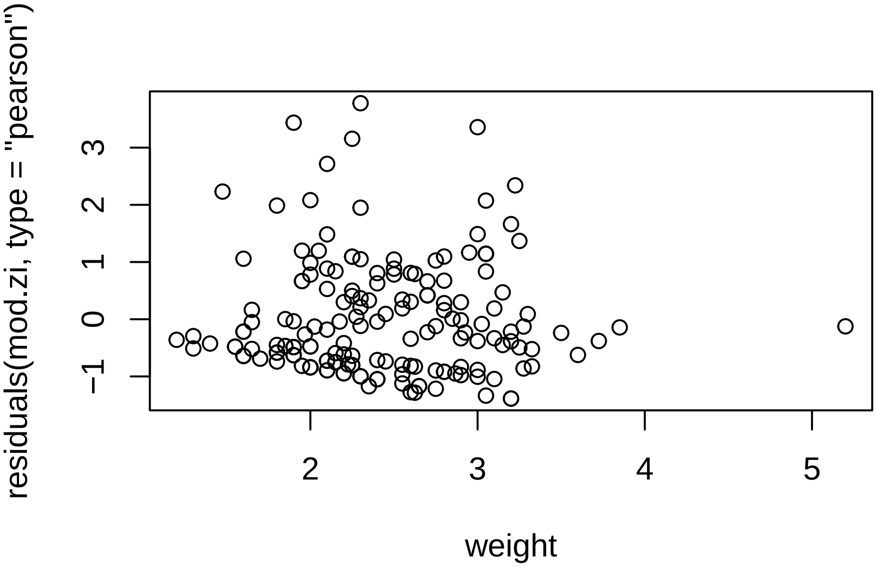
```
mod.ql2 <- glm(count ~ race, family = quasi(link="log", variance = "mu^2"), homic)
summary(mod.ql2)
##
## Call:
## glm(formula = count ~ race, family = quasi(link = "log", variance = "mu^2"),
##     data = homic)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.3832     0.1200 -19.861  < 2e-16 ***
## raceblack     1.7331     0.3442   5.036 5.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 16.54457)
##
##     Null deviance: 1839.0  on 1307  degrees of freedom
## Residual deviance: 1870.9  on 1306  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 10
plot(residuals(mod.ql2, type = "pearson") ~ race, homic)
```



6. (8.13 in Agresti) Use QL methods to construct a model for the horseshoe crab satellite counts, using weight, color, and spine condition as explanatory variables. Compare results with those obtained with zero-inflated GLMs (Section 7.5 in Agresti).

Solution. Clearly, the ZINF model fits the best. Color explains the probability of the zero count but it has no significance in the other model types. The QL fit with a variance function proportional to μ (mod.ql1) appears to be better than that proportional to μ^2 (mod.ql2) because the standardized residuals are more homoscedastic. Spin condition was not significant in any of the tested models and is not included here:

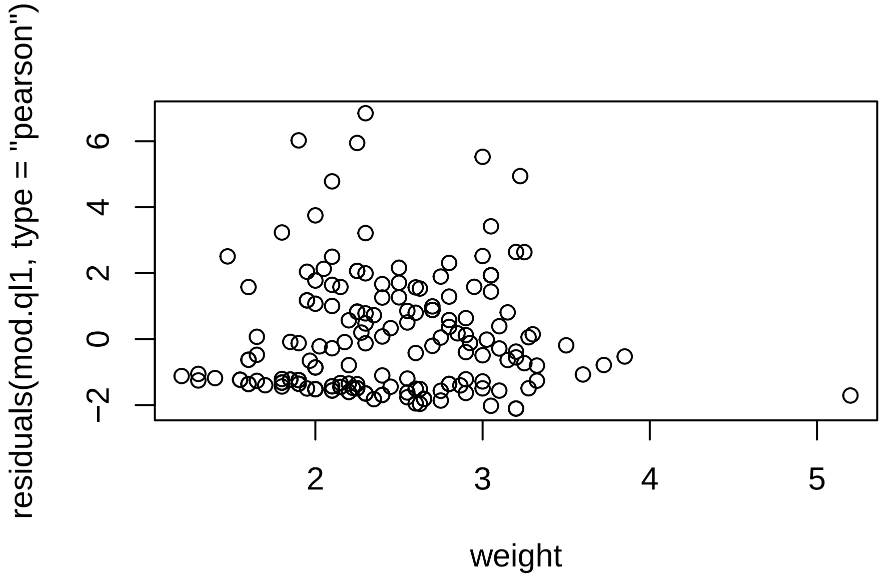
```
library(pscl)
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
Crabs <- read.table("../data/Crabs.dat", header=TRUE)
attach(Crabs)
mod.zi <- zeroinfl(y ~ weight  | weight + color , dist = "negbin", data = Crabs)
summary(mod.zi)
##
## Call:
## zeroinfl(formula = y ~ weight | weight + color, data = Crabs, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3864 -0.7506 -0.2666  0.5298  3.7767
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8961     0.3070   2.919  0.00351 **
## weight        0.2169     0.1125   1.928  0.05383 .
## Log(theta)    1.5802     0.3574   4.422 9.79e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8662     1.2415   1.503    0.133
## weight       -1.7531     0.4429  -3.958 7.55e-05 ***
## color         0.5985     0.2572   2.326    0.020 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 4.8558
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -349.9 on 6 Df
plot(residuals(mod.zi,type="pearson") ~ weight, data = Crabs)
```



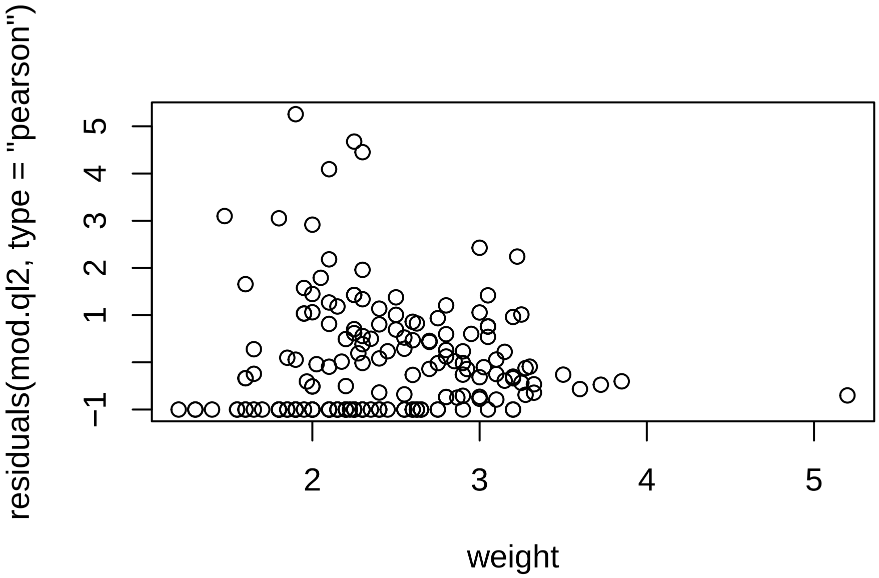
```
# plot(residuals(mod.zi,type="pearson") ~ fitted(mod.zi))

mod.ql1 <- glm(y ~ weight + color, family = quasi(link="log", variance = "mu"), Crabs)
summary(mod.ql1)
##
## Call:
## glm(formula = y ~ weight + color, family = quasi(link = "log",
##     variance = "mu"), data = Crabs)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08855    0.45424   0.195    0.846
## weight       0.54588    0.12049   4.531  1.1e-05 ***
## color       -0.17282    0.10988  -1.573    0.118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 3.18719)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 552.79  on 170  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
plot(residuals(mod.ql1,type="pearson") ~ weight, Crabs)
```



```
mod.ql2 <- glm(y ~ weight + color, family = quasi(link="log", variance = "mu^2"), Crabs)
summary(mod.ql2)
##
## Call:
## glm(formula = y ~ weight + color, family = quasi(link = "log",
##     variance = "mu^2"), data = Crabs)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.4710     0.5586  -0.843    0.400
## weight        0.7620     0.1649   4.621 7.51e-06 ***
## color        -0.1693     0.1186  -1.427    0.155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 1.459119)
##
##     Null deviance: 80.907  on 172  degrees of freedom
## Residual deviance: 91.583  on 170  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 19
plot(residuals(mod.ql2,type="pearson") ~ weight, Crabs)
```



7. (8.16 in Agresti) For the teratology study analyzed in Section 8.2.4, analyze the data using only the group indicators as explanatory variables (i.e., ignoring hemoglobin). Interpret results. Is it sufficient to use the simpler model having only the placebo indicator for the explanatory variable?

Solution. Interpretation: The probabilities of fetuses dying in different groups have been computed below. It's sufficient to use the placebo indicator as an explanatory variable because the likelihood ratio test against the larger model (that with group indicators) is not rejected.

```
Rats <- read.table("../data/Rats.dat", header=TRUE)
attach(Rats)
placebo <- ifelse(group == 1, 1, 0)
Rats$group <- factor(Rats$group)
library(VGAM)
fit.bb1 <- vglm(cbind(s,n-s) ~ group, betabinomial(zero = 2, irho = 0.2), data = Rats)
summary(fit.bb1)
##
## Call:
## vglm(formula = cbind(s, n - s) ~ group, family = betabinomial(zero = 2,
##     irho = 0.2), data = Rats)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.3458     0.2441   5.514 3.52e-08 ***
## (Intercept):2  -1.1459     0.3241  -3.536 0.000406 ***
## group2         -3.1143     0.5183  -6.009 1.87e-09 ***
## group3         -3.8679     0.8632  -4.481 7.44e-06 ***
## group4         -3.9225     0.6835  -5.739 9.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(mu), logitlink(rho)
##
## Log-likelihood: -93.4567 on 111 degrees of freedom
##
## Number of Fisher scoring iterations: 8
```

```
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## 'group3', 'group4'
# Coefficients in the linear part
beta <- coef(fit.bb1)[-2]
# Prob of death in 1st group
plogis(beta[1])
## (Intercept):1
##     0.7934496
# Prob of death in other groups
plogis(beta[1]+beta[-1])
##     group2     group3     group4
## 0.14573448 0.07432378 0.07065549


fit.bb2 <- vglm(cbind(s,n-s) ~ placebo, betabinomial(zero = 2, irho = 0.2), data = Rats)
anova(fit.bb2, fit.bb1, type = 1)
## Analysis of Deviance Table
##
## Model 1: cbind(s, n - s) ~ placebo
## Model 2: cbind(s, n - s) ~ group
##   Resid. Df  LogLik Df 2 * LogLik Diff. Pr(>Chi)
## 1       113 -94.274
## 2       111 -93.457  2           1.6347   0.4416
```