Predictive Checking

Readings GH Chapter 6-8

February 8, 2017

Model Choice and Model Checking

2 Questions:

1. Is my Model good enough? (no alternative models in mind)
2. Which Model is best? (comparison among a subset of models)

Note that 2 does not imply 1.

Focus on simulation based methods

HIV & Risk Behaviour Study

- The variables `couples` and `women_alone` code the **intervention**:
  - control - no counselling (both 0)
  - the couple was counselled together (couple=1, women_alone=0)
  - only the woman was counselled (couple=0, women_alone=1)
- `bs_hiv` indicates whether the member reporting sex acts was HIV-positive at "baseline" (the beginning of the study)
- `bupacts` - number of unprotected sex acts reported at "baseline"
- `fupacts` - number of unprotected sex acts reported at the end of the study
- `sex` - factor with levels "woman" and "man". This is the member of the couple that reports sex acts to the researcher

Data, Design & balance

```
##      sex          couples         women_alone
##  woman:217   Min.   :0.0000   Min.   :0.0000
##  man  :217   1st Qu.:0.0000   1st Qu.:0.0000
##              Median :0.0000   Median :0.0000
##              Mean   :0.3733   Mean   :0.3364
##              3rd Qu.:1.0000   3rd Qu.:1.0000
##              Max.   :1.0000   Max.   :1.0000
##       bs_hiv        bupacts          fupacts
##  negative:337   Min.   :  0.00   Min.   :  0.00
##  positive: 97   1st Qu.:  5.00   1st Qu.:  0.00
##                 Median : 15.00   Median :  5.00
##                 Mean   : 25.91   Mean   : 16.49
##                 3rd Qu.: 36.00   3rd Qu.: 21.00
##                 Max.   :300.00   Max.   :200.00
```

Model

```
hiv.glm = glm(fupacts ~ bs_hiv + log(bupacts +1) + sex +
                couples + women_alone, data=hiv,
                family=poisson(link="log"))
```

```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.10334    0.04706  23.445  < 2e-16 ***
## bs_hivpositive    -0.40556    0.03543 -11.445  < 2e-16 ***
## log(bupacts + 1)   0.66456    0.01217  54.596  < 2e-16 ***
## sexman            -0.08181    0.02368  -3.454 0.000551 ***
## couples           -0.30894    0.02799 -11.038  < 2e-16 ***
## women_alone       -0.50952    0.03031 -16.810  < 2e-16 ***
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13298.6  on 433  degrees of freedom
## Residual deviance:  9184.3  on 428  degrees of freedom
## AIC: 10521
##
```

Over-Dispersion, Goodness of Fit or Lack of Fit?

- deviance is -2 log(likelihood) evaluated at the MLE of the parameters in that model $\hat{\lambda}_i = \exp(\mathbf{x}_i^T \hat{\beta})$

$$-2 \sum_i (y_i \log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(y_i!))$$

- smaller is better (larger likelihood)
- null deviance is the deviance under the "Null" model, that is a model with just an intercept or $\lambda_i = \lambda$ and $\hat{\lambda} = \bar{Y}$
- saturated model deviance is the deviance of a model where each observation has there own unique λ_i and the MLE of $\hat{\lambda}_i = y_i$,
- the change in deviance has a Chi-squared distribution with degrees of freedom equal to the change in number of parameters in the models.

Residual Deviance

the residual deviance is the change in the deviance between the given model and the saturated model. Substituting the expressions for deviance, we have

$$
\begin{aligned}
D = & -2\sum_i \Big(y_i \log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(y_i!)\Big) - \\
& -2\sum_i \left(y_i \log(y_i) - y_i - \log(y_i!)\right) \\
= & 2\sum_i \Big(y_i(\log(y_i) - \log(\hat{\lambda}_i)) - (y_i - \hat{\lambda}_i))\Big) \\
= & 2\sum \Big(y_i(\log(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i))\Big) = \qquad \sum d_i^2
\end{aligned}
$$

This has a chi squared distribution with $n - (p + 1)$ degrees of freedom. ($p + 1$ is the number of parameters in the linear predictor)

Test in R

```
## Residual deviance:  9184.3  on 428  degrees of freedom
```

Estimate of overdispersion: Residual Deviance/ Residual df $= 21.46$

Overdispersion if greater than 1.

Formal Test

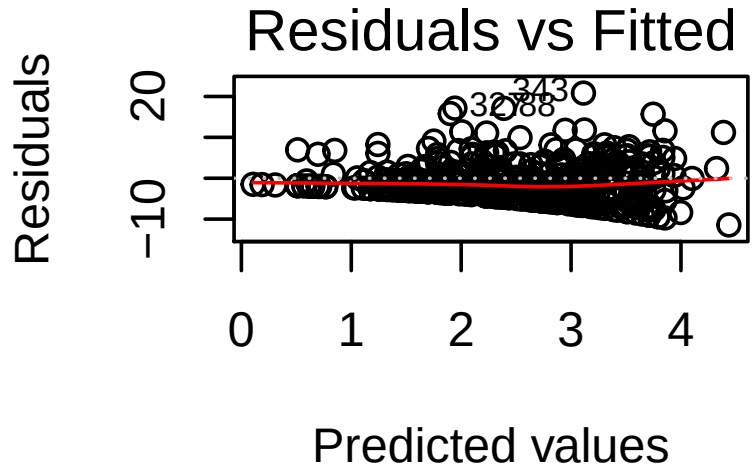
```
pchisq(hiv.glm$deviance, hiv.glm$df.residual, lower.tail=F)
```

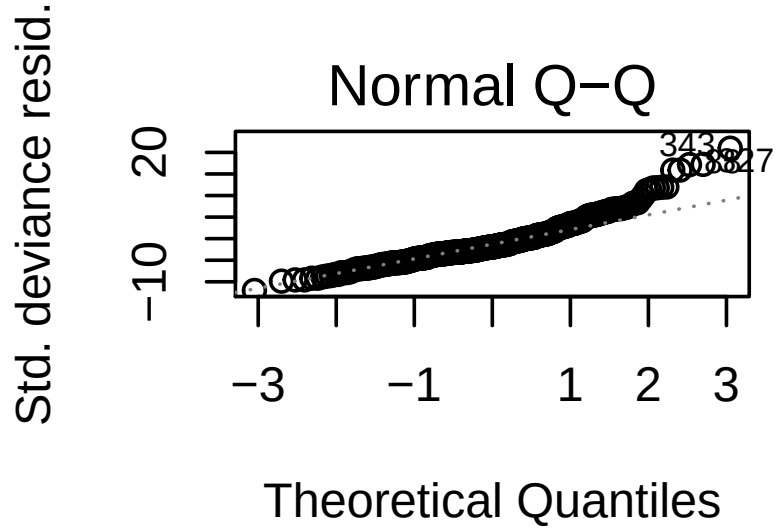
```
## [1] 0
```

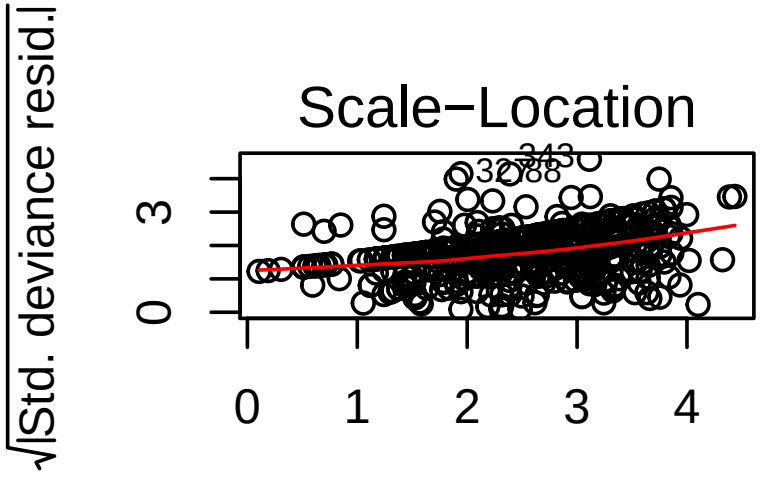
The above p-value suggests that a residual deviance as large or larger than what we observed under the model in `hiv.glm` is highly unlikely!

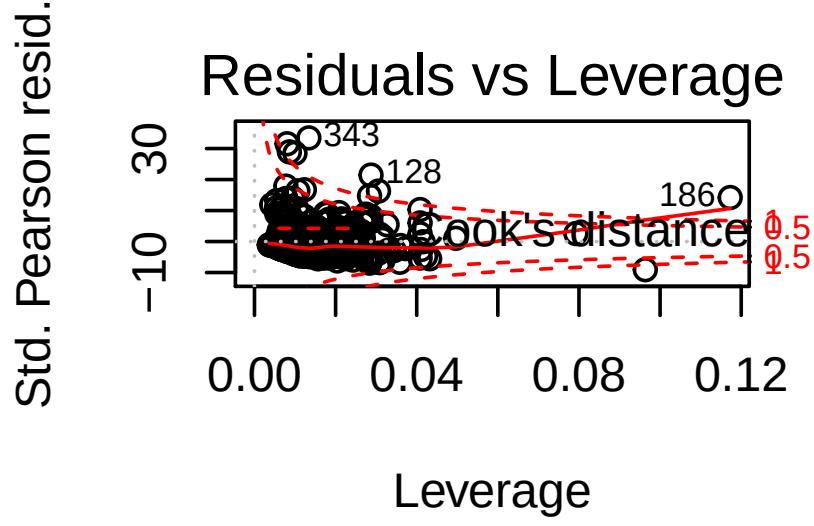
Suggests that the model is not adequate or lack of fit.

Diagnostics Plots

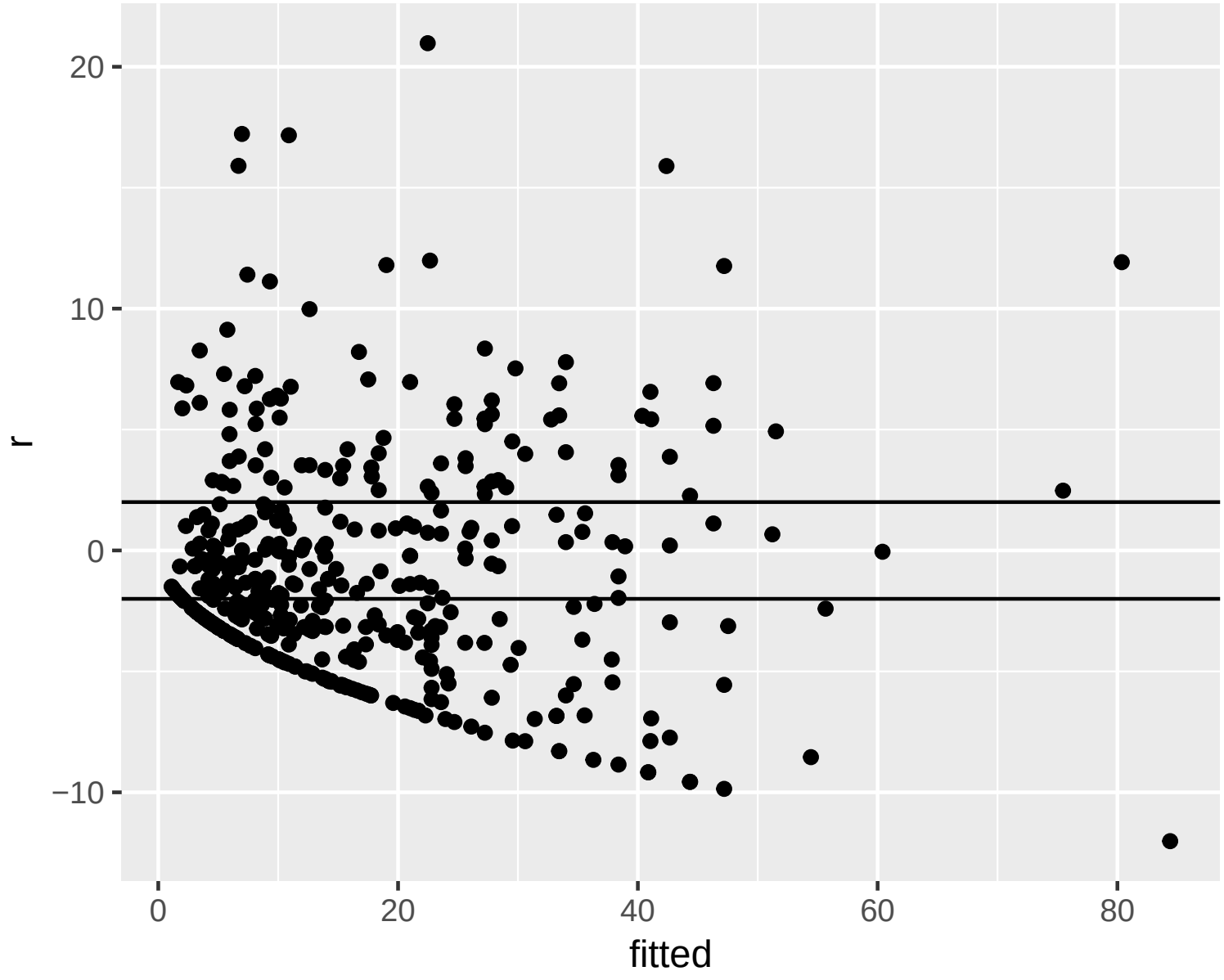




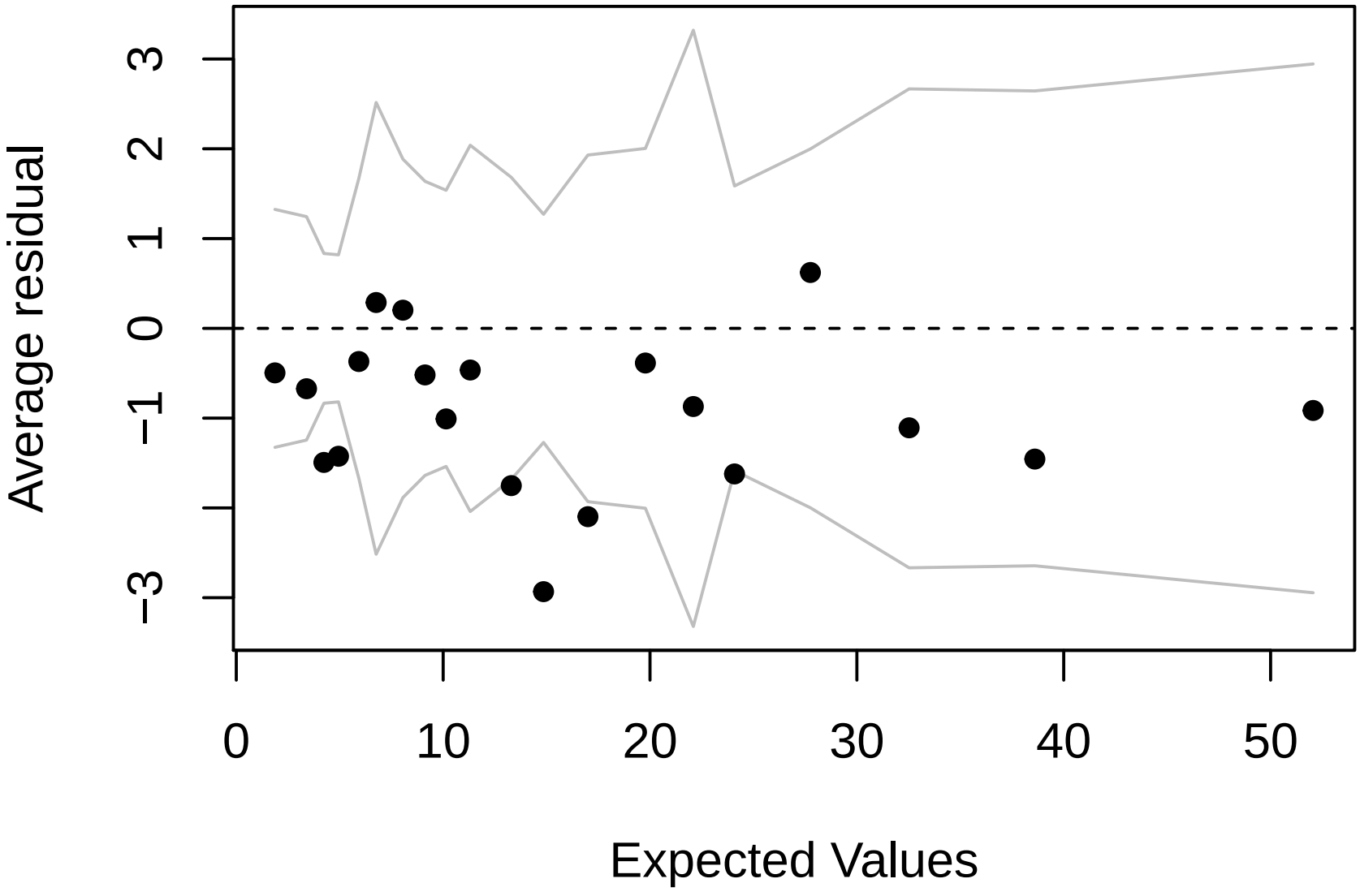




Standardized Residuals



Binned Residuals library(arm)



Overdispersed Poisson - QuasiLikelihood

```
hiv.glmod = glm(fupacts ~ bs_hiv + log(bupacts +1) + sex +
                couples + women_alone, data=hiv,
                family=quasipoisson(link="log"))
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.10334    0.24820   4.445 1.12e-05 *
## bs_hivpositive    -0.40556    0.18689  -2.170  0.03055 *
## log(bupacts + 1)   0.66456    0.06420  10.352  < 2e-16 *
## sexman            -0.08181    0.12491  -0.655  0.51283
## couples           -0.30894    0.14762  -2.093  0.03695 *
## women_alone       -0.50952    0.15986  -3.187  0.00154 *
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to b
##
##     Null deviance: 13298.6  on 433  degrees of freedom
## Residual deviance:  9184.3  on 428  degrees of freedom
## AIC: NA
```

Negative Binomial Distribution

- The formulation of the negative binomial distribution as a gamma mixture of Poissons can be used to model count data with overdispersion.

$$p(y \mid \mu, \theta) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y}$$

- The negative binomial distribution has two parameters:
  - μ is the mean or expected value of the distribution $\mu_i = \exp(\mathbf{x}_i^T \beta)$
  - a is the over dispersion parameter $V(Y) = \mu + \mu^2/\theta$
- When $\theta \to \infty$ the negative binomial distribution is the same as a Poisson distribution

Review of Mixtures

$$Y \mid \lambda \sim \text{Poi}(\lambda)$$
$$p(y \mid \lambda) = \frac{y^{\lambda} e^{-\lambda}}{y!}$$

$$\lambda \mid \mu, \theta \sim \text{Gamma}(\theta, \theta/\mu)$$
$$p(\lambda \mid \mu, \theta) = \frac{(\theta/\mu)^{\theta}}{\Gamma(\theta)} \lambda^{\theta-1} e^{-\lambda\theta/\mu}$$

$$p(Y \mid \mu, a) = \int p(Y \mid \lambda) p(\lambda \mid \theta, \theta/\mu) d\lambda$$
$$= \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y}$$
$$Y \mid \mu, \theta \sim \text{NegBin}(\mu, \theta)$$

Iterated Expectations Review

- expectation $\mathrm{E}[Y] = \mathrm{E}_\lambda[\mathrm{E}_Y[Y \mid \lambda]]$
- variance

$$\mathrm{Var}[Y] = \mathrm{Var}_\lambda[E_Y[Y \mid \lambda]] + \mathrm{E}_\lambda[\mathrm{Var}_Y[Y \mid \lambda]]$$

Variance(Expected Value) + Expected Value(Variance)

You should be able to derive the mean and variance of the NegBin using the above expressions (HW)

Fitting the Negative Binomial Model in R

```
library(MASS)
hiv.glm.nb = glm.nb(fupacts ~ bs_hiv + log(bupacts + 1) +
                      sex + couples + women_alone,
                      data=hiv)
```

associated `summary` and `plot` functions available

Model Summary (subset)

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.25829    0.24261   5.186 2.14e-07 ***
## bs_hivpositive    -0.51314    0.18384  -2.791 0.005251 **
## log(bupacts + 1)   0.61832    0.06470   9.557  < 2e-16 ***
## sexman             0.05974    0.14917   0.400 0.688796
## couples           -0.36679    0.18531  -1.979 0.047779 *
## women_alone       -0.64007    0.18901  -3.386 0.000708 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4358) fami
##
##     Null deviance: 603.09  on 433  degrees of freedom
## Residual deviance: 487.97  on 428  degrees of freedom
## AIC: 2953.3
##
```

Using Simulation to Check the Model

- Find a test statistic (meaningful quantity)
- simulate 1000 replicates of Y's from the model
- compute the test statistics for each set of replicate data
- estimate distribution of test statistics from the simulations
- compare observed statistics to the simulated data (predictive p-value)

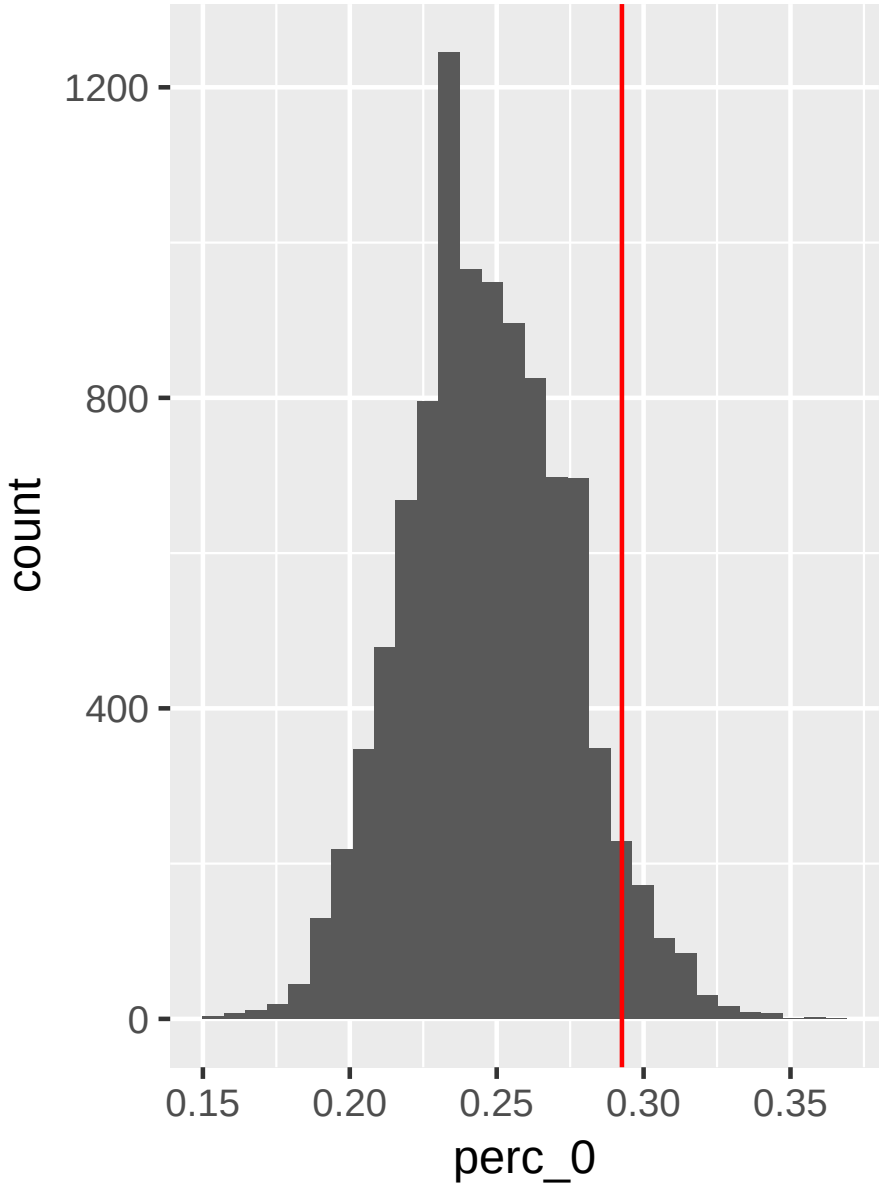
R Code

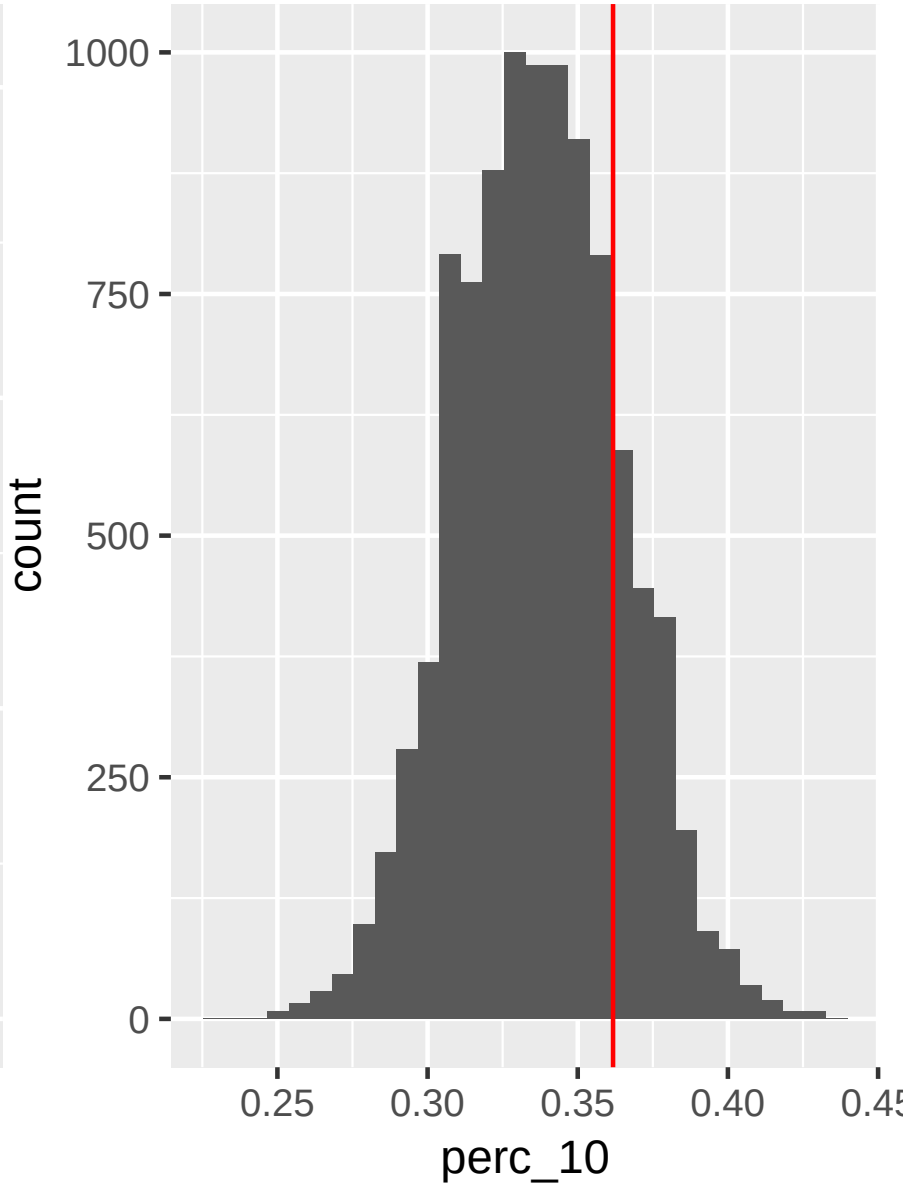
```
nsim = 10000
n = nrow(hiv)
X = model.matrix(hiv.glm.nb)
class(hiv.glm.nb) <- "glm"  # over-ride class of "glm.nb"
sim.hiv.nb = sim(hiv.glm.nb, nsim)  # use GLM to generate
sim.hiv.nb@sigma = rnorm(nsim, hiv.glm.nb$theta,
                         hiv.glm.nb$SE.theta) # add slot f
y.rep = array(NA, c(nsim, nrow(hiv)))

for (i in 1:nsim) {
  mu = exp(X %*% sim.hiv.nb@coef[i,])
  y.rep[i,] = rnegbin(n, mu=mu, theta=sim.hiv.nb@sigma[i])
}

perc_0 = apply(y.rep, 1, function(x) {mean(x == 0)})
perc_10 = apply(y.rep, 1, function(x) {mean( x > 10)})
```

Comparison


1200
800
400
0
count
0.15
0.20
0.25
0.30
0.35
perc_0


1000
750
500
250
0
count
0.25
0.30
0.35
0.40
perc_10

Confidence Intervals

Observed proportion at zero is 0.29 and proportion at 0, 95% CI from simulated replicates:

```
##  2.5% 97.5%
##   0.2   0.3
```

Observed proportion > 10 is 0.36 and 95% CI from simulated replicates

```
##  2.5% 97.5%
##  0.29  0.39
```

Observed data seem to have summaries in line with simulated replicated data based on Negative Binomial model

Model appears to capture these features adequately (may change with other summaries)

Estimates of Relative Risks

	RR	2.5	97.5
(Intercept)	3.52	2.21	5.69
bs_hivpositive	0.60	0.42	0.88
log(bupacts + 1)	1.86	1.64	2.10
sexman	1.06	0.79	1.43
couples	0.69	0.48	1.01
women_alone	0.53	0.36	0.77

- 1 = no change
- Values less than 1 imply decrease
- Values greater than 1 imply increase
- to obtain percent increase RR - 1 or CI - 1 and multiply by 100%
- to obtain percent decrease 1 - RR or 1 - CI and multiply by 100%

Conclusions

The intervention had a significant impact on reducing the number of unprotected sex acts:

In couples where only the woman took part in the counseling sessions, we estimated a significant decrease in unprotected sex acts of 47%; 95% CI: (23, 64)

When both partners were counseled unprotected acts are expected to decrease by 31% (although p.value $>$ .05)

There is no evidence to suggest that the sex of partner who reports to the researcher has an effect on the number of unprotected acts.

There is evidence to suggest that if the partner who reports is HIV positive there is a significant reduction of unprotected acts of 40%; 95% CI: (12, 58)

Energetic Student

- provide an interpretation of the coefficient for `log(bupacts + 1)`
- offsets are used to remove effects, so that the mean $=$ $\exp(\textit{offset} + \mathbf{X}\beta)$. In `R` this is expressed by
- adding `log(bupacts +1)` as an offset is equivalent to constraining its β to be 1
- does the previously fitted model support that this should be handled by an offset?
- Do any of the coefficient estimates change by using an offset?
- Which model seems to be better?
- What happens if you do not add 1 to `bupacts`?
- Do you think that how we handle zero values of `bupacts` has an impact on our predictive check for predicting at 0?