7.5 EXAMPLE: MODELING COUNT DATA

We illustrate models for discrete data using the horseshoe crab dataset introduced in Sec. 1.5.1. The response variable for the $n = 173$ mating female crabs is y = number of "satellites" — male crabs that group around the female and may fertilize her eggs. Explanatory variables are the female crab's color, spine condition, weight, and carapace width.

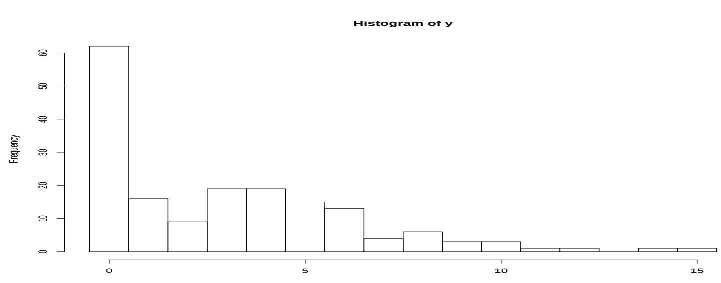
7.5.1 Fits to Marginal Distribution of Satellite Counts

To illustrate the Poisson, negative binomial, ZIP, and ZINB distributions introduced in this chapter, we first investigate the marginal distribution of satellite counts. From Sec. 1.5.1, the mean of 2.919 and variance of 9.912 suggest overdispersion relative to the Poisson.

```
---------------------------------------------------------------------
> attach(Crabs) # file Crabs.dat at www.stat.ufl.edu/~aa/glm/data
> hist(y, breaks=c(0:16)-0.5) # Histogram display with sufficient bins
---------------------------------------------------------------------
```

The histogram (Figure 7.2) shows a strong mode at 0 but slightly elevated frequencies for satellite counts of 3 through 6 before decreasing substantially. Because the distribution may not be unimodal, the negative binomial may not fit as well as a zero-inflated distribution.

Figure 7.2. Histogram for sample distribution of y = number of horseshoe crab satellites.



We fit the Poisson distribution and negative binomial distribution with quadratic variance (NB2) by fitting GLMs having only an intercept.

```
-------------------------------------------------------------------
> summary(glm(y ~ 1, family=poisson, data=Crabs)) # default link is log
           Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.0713      0.0445    24.07  <2e-16 # exp(1.0713) = 2.919
---
> logLik(glm(y ~ 1, family=poisson, data=Crabs))
'log Lik.' -494.045

> library(MASS)
> summary(glm.nb(y ~ 1, data=Crabs)) # default link is log
           Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.0713      0.0980    10.93  <2e-16
---
Theta: 0.758, Std. Err.: 0.126
> logLik(glm.nb(y ~ 1, data=Crabs))
'log Lik.' -383.705
-------------------------------------------------------------------
```

The estimated NB2 dispersion parameter[8] is $\hat{\gamma} = 1/0.758 = 1.32$. This estimate, the much larger SE (0.0980 vs. 0.0445) for the log mean estimate of $\log(2.919) = 1.071$, and the much larger log-likelihood also suggest that the Poisson distribution is inadequate.

Next, we consider zero-inflated models[9].

```
---------------------------------------------------------------
> library(pscl) # pscl package can fit zero-inflated distributions
> summary(zeroinfl(y ~ 1)) # uses log link
Count model coefficients (poisson with log link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.50385     0.04567    32.93    <2e-16

Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.6139      0.1619   -3.791   0.00015
---
Log-likelihood: -381.615 on 2 Df # 2 is model df, not residual df
---------------------------------------------------------------
```

The fitted ZIP distribution is a mixture with probability $e^{-0.6139}/[1 + e^{-0.6139}] = 0.351$ for the degenerate distribution at 0 and probability $1 - 0.351 = 0.649$ for a Poisson with mean $e^{1.50385} = 4.499$. The fitted value of $173[0.351 + 0.649e^{-4.499}] = 62.0$ for the 0 count reproduces the observed value of 62. The fitted value for the ordinary Poisson model is only $173e^{-2.919} = 9.3$. The log-likelihood increases substantially when we fit a zero-inflated negative binomial (ZINB) model.

```
------------------------------------------------------------------
```

[8] SAS (`PROC GENMOD`) reports $\hat{\gamma}$ as having $SE = 0.22$.
[9] Such models can also be fitted with the `vglm` function in the `VGAM` package.

```
> summary(zeroinfl(y ~ 1, dist="negbin")) # uses log link in pscl lib.
Count model coefficients (negbin with log link):
            Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  1.46527     0.06834   21.440   < 2e-16
Log(theta)   1.49525     0.34916    4.282  1.85e-05

Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.7279      0.1832   -3.973   7.1e-05
---
Theta = 4.4605     Log-likelihood: -369.352 on 3 Df
----------------------------------------------------------------------
```

This distribution is a mixture with probability $e^{-0.7279}/[1 + e^{-0.7279}] = 0.326$ for the degenerate distribution at 0 and probability 0.674 for a negative binomial with mean $e^{1.465} = 4.33$ and dispersion parameter estimate $\hat{\gamma} = 1/4.4605 = 0.22$.

To further investigate lack of fit, we grouped the counts into ten categories, using a separate category for each count from 0 to 8 and then combining counts of 9 and above into a single category. Comparing these with the ZINB fitted distribution of the 173 observations into these 10 categories, we obtained $X^2 = 7.7$ for $df = 10 - 3 = 7$ (since the model has three parameters), an adequate fit. For the other fits, $X^2 = 522.3$ for the Poisson model, 33.6 for the ordinary negative binomial model, and 31.3 for the ZIP model. Here are the fitted counts for the four models:

```
---------------------------------------------------------------------
count    observed    fit.p  fit.nb  fit.zip  fit.zinb
0           62        9.34   52.27   62.00    62.00
1           16       27.26   31.45    5.62    12.44
2            9       39.79   21.94   12.63    16.73
3           19       38.72   16.01   18.94    17.74
4           19       28.25   11.94   21.31    16.30
5           15       16.50    9.02   19.17    13.58
6           13        8.03    6.87   14.38    10.55
7            4        3.35    5.27    9.24     7.76
8            6        1.22    4.06    5.20     5.48
9 or more   10        0.55   14.16    4.51    10.43
---------------------------------------------------------------------
```

The ZIP model tends to be not dispersed enough, having fitted value that is too small for the counts of 1 and ≥ 9.

7.5.2 GLMs for Crab Satellite Numbers

We now consider zero-inflated negative binomial models with the explanatory variables from Table 1.3. Weight and carapace width have a correlation of 0.887, and we shall use only weight to avoid issues with collinearity. Darker-colored crabs tend to be older. Most crabs have both spines worn or broken (category 3). When we fit the ZINB main-effects model using weight, color, and spine condition for each component, with color and spine condition as qualitative factors, we find that weight is significant in each component but neither of color or spine condition are. Adding interaction terms does not yield an improved fit. Analyses using color in a quantitative manner with category scores $\{c_i = i\}$ gives relatively strong evidence that darker crabs tend to have more 0 counts. If we use weight w_i in both components of the model but quantitative color only in the zero-component, we obtain:

```
-------------------------------------------------------------------
> summary(zeroinfl(y ~ weight | weight + color, dist="negbin"))
Count model coefficients (negbin with log link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   0.8961      0.3070    2.919    0.0035
weight        0.2169      0.1125    1.928    0.0538 .
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
---
Theta = 4.8558     Log-likelihood: -349.865 on 6 Df
-------------------------------------------------------------------
```

The fitted distribution is a mixture with probability $\hat{\phi}_i$ of a negative binomial having mean $\hat{\mu}_i$ satisfying

$$\log \hat{\mu}_i = 0.896 + 0.217 w_i$$

with dispersion parameter estimate $\hat{\gamma} = 1/4.8558 = 0.21$, and a probability mass $1 - \hat{\phi}_i$ at 0 satisfying

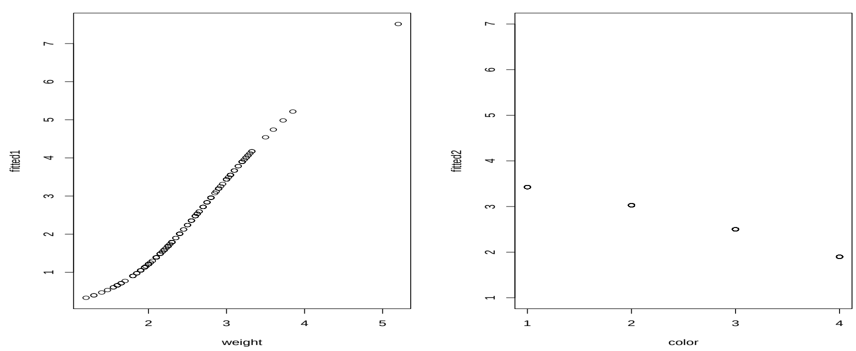
$$\text{logit}(1 - \hat{\phi}_i) = 1.866 - 1.753 w_i + 0.598 c_i.$$

The overall fitted mean response at a particular weight and color equals

$$\hat{E}(y_i) = \hat{\phi}_i \hat{E}(y_i \mid z_i = 1) = \left(\frac{1}{1 + e^{1.866 - 1.753 w_i + 0.598 c_i}} \right) e^{0.896 + 0.217 w_i}.$$

As weight increases for a particular color, the fitted probability mass at the 0 outcome decreases, and the fitted negative binomial mean increases. Figure 7.3 plots the overall fitted mean as a function of weight for the dark crabs (color 4) and as a function of color at the median weight of 2.35 kg.

Figure 7.3. Fitted mean number of horseshoe crab satellites for zero-inflated negative binomial model, plotted as a function of weight for dark crabs and as a function of color for median-weight crabs



If we drop color completely and exclude weight from the NB2 component of the model, the log-likelihood decreases to −354.7 but we obtain the simple expression for the overall fitted mean of $\exp(1.47094)/[1+\exp(3.927-1.985w_i)]$. This has a logistic shape for the increase in the fitted mean as a function of weight.

If we ignore the zero-inflation and fit an ordinary NB2 model with weight and quantitative color predictors, we obtain:

```
-------------------------------------------------------------------
> summary(glm.nb(y ~ weight + color))
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.1487      0.6424   -0.231     0.817
weight        0.7072      0.1612    4.387  1.15e-05
color        -0.1734      0.1199   -1.445     0.148
---
Theta: 0.956   2 x log-likelihood: -746.452 # L = -373.226
-------------------------------------------------------------------
```

This describes the tendency of the overall mean response to increase with weight and decrease with color (but not significantly). In not having a separate component to handle the zero count, the NB2 model has dispersion parameter estimate $\hat{\gamma} = 1/0.956 = 1.05$ that is much greater than $\hat{\gamma}$ for the NB2 component of ZINB models. The fit is similar to that of the geometric distribution, which is NB2 with $\gamma = 1$. But its log-likelihood of −373.2 is considerably worse than values obtained for ZINB models.

Unless previous research or theory suggests more-complex models, it seems adequate to use a zero-inflated NB2 model with weight as the primary predictor, adding color as a predictor of the mass at 0. In these analyses, however, we have ignored that the data set contains an outlier – an exceptionally heavy crab weighing 5.2 kg of medium color that had 7 satellites. As as exercise, you can fit models without that observation to investigate how the results change.