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) = 2.91

> 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 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.

> 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.

---

8SAS (PROC GENMOD) reports $\hat{\gamma}$ as having $SE = 0.22$.

9Such models can also be fitted with the vglm function in the VGAM package.
EXAMPLE: MODELING COUNT DATA

> 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  21.4400   < 2e-16
Log(theta)  1.49525    0.34916   4.2823  1.85e-05

Zero-inflation model coefficients (binomial with logit link):
  Estimate  Std. Error   z value  Pr(>|z|)
(Intercept) -0.72790    0.18323  -3.9725  7.10e-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:

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

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

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:

```r
> 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

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

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

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

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

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

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
```

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 exercise, you can fit models without that observation to investigate how the results change.