STAR: Simultaneous Transformation and Rounding for Modeling Integer-Valued Data

Daniel R. Kowal

5/11/2019

Background: Integer-Valued Data

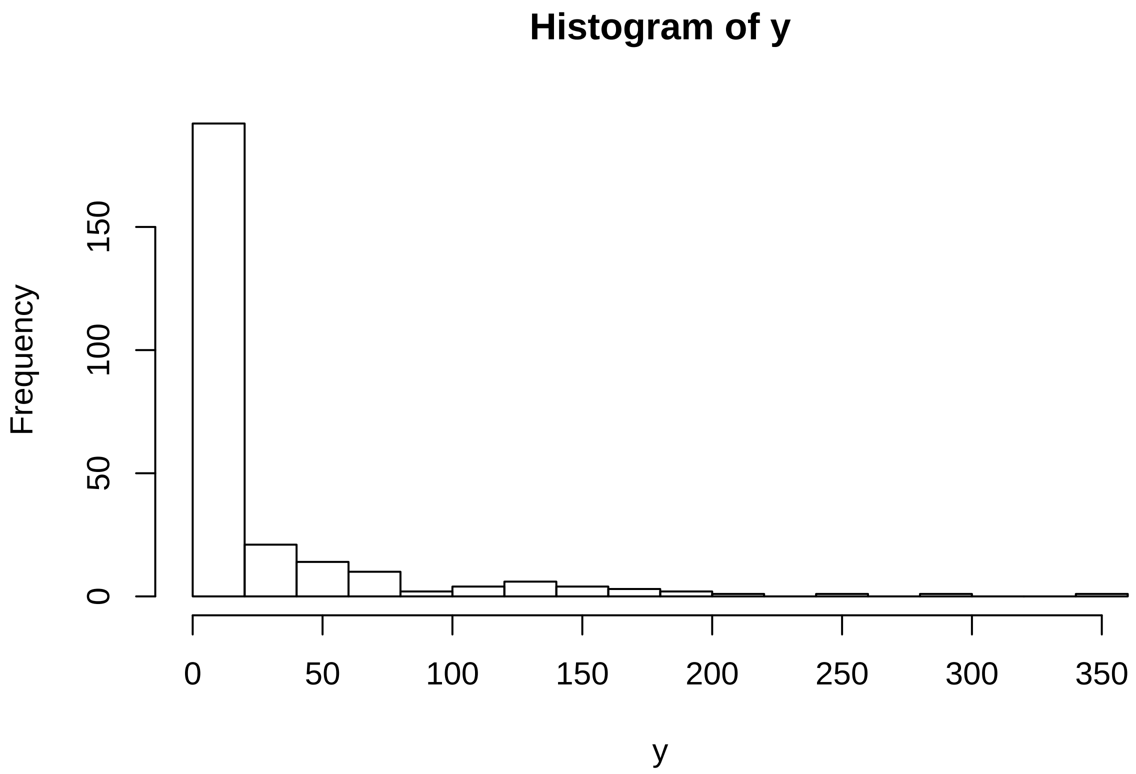
Integer-valued or count data are common in many fields. Frequently, integer-valued data are observed jointly with predictors, over time intervals, or across spatial locations. Integer-valued data also exhibit a variety of distributional features, including zero-inflation, skewness, over- and underdispersion, and in some cases may be bounded or censored. Flexible and interpretable models for *integer-valued processes* are therefore highly useful in practice.

As an illustration, consider the `roaches` data from Gelman and Hill (2006). The response variable, y_i, is the number of roaches caught in traps in apartment i, with $i = 1, \ldots, n = 262$.

```r
# Source: http://mc-stan.org/rstanarm/articles/count.html
data(roaches)

# Roaches:
y = roaches$y

# Histogram:
hist(y, breaks = 'scott')
```



There are several notable features in the data:

1. Zero-inflation: $\#\{y_i = 0\} = 94$, which accounts for 36% of the observations.
2. (Right-) Skewness, which is clear from the histogram and common for (zero-inflated) count data.
3. Overdispersion: the sample mean is $\bar{y} = 26$ and the sample variance is $\hat{s}^2 = 2585$.

A pest management treatment was applied to a subset of 158 apartments, with the remaining 104 apartments receiving a control. Additional data are available on the pre-treatment number of roaches, whether the apartment building is restricted to elderly residents, and the number of days for which the traps were exposed. We are interested in modeling how the roach incidence varies with these predictors.

```r
# Construct a design matrix:
X = model.matrix(y ~ roach1 + treatment + senior + log(exposure2),
                 data = roaches)

# Rename:
colnames(X)[2]= 'Pre-treat #Roaches'

# Dimensions:
n = nrow(X); p = ncol(X)
```

Consider the Gaussian linear regression model:

$$g(y_i) = x_i'\beta + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

STAR Models

$$y = h(y^*)$$

and

$$z^* = g(y), \quad z^* \sim \Pi_\theta$$

Note: probably should use signal-plus-noise and then show we can generalize

`rSTAR` package

Outline:

1. Plot some count data; show why Gaussian does not make sense
2. Introduce STAR (definition, inference, properties, estimation?)

- Note: we set $y_{min} = 0$ WLOG and specify $y_{max} = \infty$ by default.
- Importantly, we need
$$g(y_{min}) = -\infty, \quad g(y_{max}) = \infty$$

3. Introduce the rSTAR package
4. Other usage: BART, Additive models, linear models