Chapter 1

Varieties of Count Data

Some Points of Discussion

- What are counts? What are count data?
- What is a linear statistical model?
- What is the relationship between a probability distribution function (PDF) and a statistical model?
- What are the parameters of a statistical model? Where do they come from, and can we ever truly know them?
- How does a count model differ from other regression models?
- What are the basic count models, and how do they relate with one another?
- What is overdispersion, and why is it considered to be the fundamental problem when modeling count data?

1.1 What Are Counts?

When discussing the modeling of count data, it's important to clarify exactly what is meant by a count, as well as "count data" and "count variable." The word "count" is typically used as a verb meaning to enumerate units, items, or events. We might count the *number* of road kills observed on a stretch of highway, *how many* patients died at a particular hospital within 48 hours of having a myocardial infarction, or *how many* separate sunspots were observed in March 2013. "Count data," on the other hand, is a plural noun referring to observations made about events or items that are enumerated. In statistics, count data refer to observations that have only nonnegative integer values ranging from zero to some greater undetermined value. Theoretically, counts can range from zero to infinity, but they are always limited to some lesser distinct value – generally the maximum value of the count data being modeled. When the data being modeled consist of a large number of distinct values, even if they are positive integers, many statisticians prefer to model the counts as if they were continuous data. We address this issue later in the book.

A "count variable" is a specific list or array of count data. Again, such observations can only take on nonnegative integer values. However, in a statistical model, a response variable is understood as being a random variable, meaning that the particular set of enumerated values or counts could be other than they are at any given time. Moreover, the values are assumed to be independent of one another (i.e., they show no clear evidence of correlation). This is an important criterion for count model data, and it stems from the fact that the observations of a probability distribution are independent. On the other hand, predictor values are fixed; that is, they are given as facts, which are used to better understand the response.

We will be primarily concerned with four types of count variables in this book. They are:

1. A count or enumeration of events
2. A count of items or events occurring within a period of time or over a number of periods
3. A count of items or events occurring in a given geographical or spatial area or over various defined areas
4. A count of the number of people having a particular disease, adjusted by the size of the population at risk of contracting the disease

Understanding how count data are modeled, and what modeling entails, is discussed in the following section. For readers with little background in linear models, I strongly suggest that you read through Chapter 1 even though various points may not be fully understood. Then re-read the chapter carefully. The essential concepts and relationships involved in modeling should then be clear. In Chapter 1, I have presented the fundamentals of modeling, focusing on normal and count model estimation from several viewpoints, which should at the end provide the reader with a sense of how the modeling process is to be understood when applied to count models. If certain points are still unclear, I am confident that any problem areas regarding the assessment of fit will be clear by the time you read through Chapter 4, on assessing model fit. Those who have taken a statistics course in which linear regression is examined should have no problem following the presentation.

1.2 Understanding a Statistical Count Model

1.2.1 Basic Structure of a Linear Statistical Model

Statistics may be generically understood as the science of collecting and analyzing data for the purpose of classification, prediction, and of attempting to quantify and understand the uncertainty inherent in phenomena underlying data.

A statistical model describes the relationship between one or more variables on the basis of another variable or variables. For the purpose of the models we discuss in this book, a statistical model can be understood as the mathematical explanation of a count variable on the basis of one or more explanatory variables.[1] Such statistical models are stochastic, meaning that they are based on probability functions. The traditional linear regression model is based on the normal or Gaussian probability distribution and can be formalized in the most simple case as

$$Y = \beta_0 + \beta X + \varepsilon \tag{1.1}$$

where Y is called the response, outcome, dependent, or sometimes just the y variable. We use the term "response" or y when referring to the variable being modeled. X is the explanatory or predictor variable that is used to explain the occurrence of y. β is the coefficient for X. It is a slope describing the rate of change in the response based on a one-unit change in X, holding other predictor values constant (usually at their mean values). β_0 is the intercept, which provides a value to fitted y, or $\hat{y}$, when, or if, X has the value of 0. ε (epsilon) is the error term, which reflects the fact that the relationship between X and Y is not exact, or deterministic. For the normal or linear regression model, the errors are Gaussian or normally distributed, which is the most

[1] A model may consist of only the response variable, unadjusted by explanatory variables. Such a model is estimated by modeling the response on the intercept. For example, using R: *lm(y ~ 1)*; using Stata: *reg y*.

well-used and basic probability distribution in statistics. ε is also referred to as the residual term.

When a linear regression has more than one predictor, it may be schematized by giving a separate *beta* and *X* value for each predictor, as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{1.2}$$

Statisticians usually convert equation (1.2) to one that has the left-hand side being the predicted or expected mean value of the response, based on the sum of the predictors and coefficients. Each associated coefficient and predictor is called a regression *term*:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1.3}$$

or

$$\hat{\mu} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1.4}$$

Notice that the error became part of the expected or predicted mean response. "$\frown$", or *hat* over *y* and μ (*mu*), indicates that this is an estimated value. From this point on, I use the symbol μ to refer to the predicted value, without a *hat*. Understand, though, that when we are estimating a parameter or a statistic, a *hat* should go over it. The true unknown parameter, on the other hand, has no *hat*. You will also at times see the term E(*y*) used to mean "estimated *y*." I will not use it here.

In matrix form, where the individual terms of the regression are expressed in a single term, we have

$$\mu = \beta X \tag{1.5}$$

with β*X* being understood as the summation of the various terms, including the intercept. As you may recall, the intercept is defined as $\beta_0(1)$, or simply β_0. It is therefore a term that can be placed within the single matrix term β*X*. When models become complicated, viewing them in matrix form is the only feasible way to see the various relationships involved. I should mention that sometimes you see the term β*X* expressed as *x*β. I reserve this symbol for another part of the model, which we discuss a bit later in this section.

Let's look at example data (**smoking**). Suppose that we have a six-observation model consisting of the following variables:

sbp: systolic blood pressure of subject
male: 1 = male; 0 = female
smoker: 1 = history of smoking; 0 = no history of smoking
age: age of subject

Using Stata statistical software, we display a linear regression of *sbp* on *male*, *smoker*, and *age*, producing the following (*nohead* suppresses the display of header statistics).

```
STATA CODE
. regress sbp male smoker age, nohead
------------------------------------------------------------------------
   sbp |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
-------+----------------------------------------------------------------
  male |   4.048601   .2507664    16.14   0.004      2.96964    5.127562
smoker |   6.927835   .1946711    35.59   0.001     6.090233    7.765437
   age |   .4698085     .02886    16.28   0.004     .3456341     .593983
 _cons |   104.0059   .7751557   134.17   0.000     100.6707    107.3411
------------------------------------------------------------------------
```

Continuing with Stata, we may obtain the predicted value, μ, which is the estimated mean systolic blood pressure, and display the predictor values together with μ (mu) as

```
. predict mu
. l              // 'l' is an abbreviation for list
     +------------------------------------+
     | sbp   male  smoker  sge      mu     |
     |------------------------------------|
  1. | 131        1    1      34   130.9558 |
  2. | 132        1    1      36   131.8954 |
  3. | 122        1    0      30   122.1488 |
  4. | 119        0    0      32   119.0398 |
  5. | 123        0    1      26   123.1488 |
  6. | 115        0    0      23   114.8115 |
     +------------------------------------+
```

To see exactly what this means, we sum the terms of the regression. The intercept term is also summed, but its values are set at 1. The _b[] term captures the coefficient from the results saved by the software. For the intercept, _b[_*cons*] adds the intercept term, slope[1], to the other values. The term *xb* is also commonly referred to as the *linear predictor*.

```
. gen xb =  _b[male]*male + _b[smoker]*smoker + _b[age]*age + _b[_cons]
. l
     +-----------------------------------------------+
     | sbp   male   smoker  age     mu          xb    |
     |-----------------------------------------------|
  1. | 131      1      1     34   130.9558   130.9558 |
  2. | 132      1      1     36   131.8954   131.8954 |
  3. | 122      1      0     30   122.1488   122.1488 |
  4. | 119      0      0     32   119.0398   119.0398 |
  5. | 123      0      1     26   123.1488   123.1488 |
  6. | 115      0      0     23   114.8115   114.8115 |
     +-----------------------------------------------+
```

The intercept is defined correctly; check by displaying it. The value is indeed 1,

```
. di _cons
1
```

whereas _b[_*cons*] is the constant slope of the intercept as given in the preceding regression output:

```
. di _b[_cons]    /* intercept slope */
104.00589
```

Using R, we may obtain the same results with the following code:

```
R CODE
> sbp    <- c(131,132,122,119,123,115)
> male   <- c(1,1,1,0,0,0)
> smoker <- c(1,1,0,0,1,0)
> age    <- c(34,36,30,32,26,23)
> summary(reg1 <- lm(sbp~ male+smoker+age))
           <results not displayed>
```

Predicted values may be obtained by

```
> mu <- predict(reg1)
> mu
       1        2        3        4        5        6
130.9558 131.8954 122.1487 119.0398 123.1487 114.8115
```

As was done with the Stata code, we may calculate the linear predictor, which is the same as μ, by first abstracting the coefficient

```
> cof <- reg1$coef
> cof
(Intercept)        male      smoker         age
104.0058910   4.0486009   6.9278351   0.4698085
```

and then the linear predictor, *xb*. Each coefficient can be identified with [ ]. The values are identical to *mu*.

```
> xb <- cof[1] + cof[2]*male + cof[3]*smoker + cof[4]*age
> xb
[1] 130.9558 131.8954 122.1487 119.0398 123.1487 114.8115
```

Notice the closeness of the observed response and predicted values. The differences are

```
> diff <- sbp - mu
> diff
          1           2           3           4           5           6
 0.04418262  0.10456554 -0.14874816 -0.03976436 -0.14874816  0.18851252
```

When the values of the linear predictor are close to the predicted or expected values, we call the model *well fitted*.

1.2.2 Models and Probability

One of the points about statistical modeling rarely discussed is the relationship of the data to a probability distribution. All parametric statistical models are based on an underlying probability distribution. I mentioned before that the normal or linear regression model is based on the Gaussian, or normal, probability distribution (see example in Figure 1.1). It is what defines the error terms. When we are attempting to estimate a least squares regression or more sophisticated maximum likelihood model, we are estimating the parameters of the underlying probability distribution that characterize the data. These two foremost methods of estimation are described in the next section of this opening chapter. The important point here is always to remember that when modeling count data we are really estimating the parameters of a probability distribution that we believe best represents the data we are modeling. We are never able to knowingly determine the true parameters

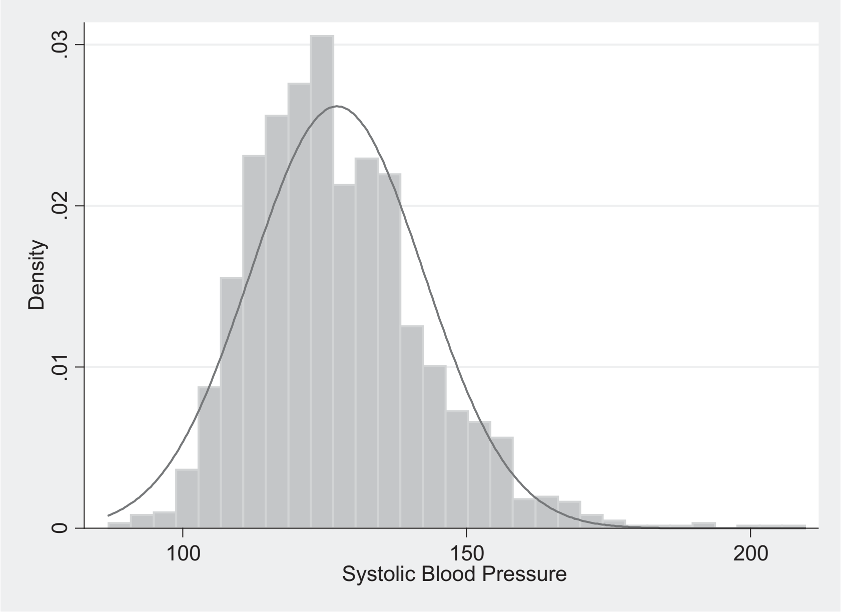

FIGURE 1.1. Gaussian distribution approximated by blood pressure data.

of the probability distribution function, which we shall refer to as the PDF, but we attempt to obtain the best unbiased estimate possible. The parameters are what provide the shape of the PDF we are using to describe the data. By knowing the estimated parameter or parameters, we can use them to predict data from inside the sample of data from which we are modeling and in special cases data from outside the sample.

This is also an important point to keep in mind. We assume that the data being modeled are a random sample from a greater population of data. The PDF whose parameters we are attempting to estimate is assumed to describe the population data, not only the sample from it that we are actually modeling. This way of looking at statistics and data is commonly referred to as frequency-based statistical modeling. Bayesian models look at the relationship of data to probability distributions in a different manner, which we discuss in the final chapter. However, the standard way of modeling is based on this frequency interpretation, which was championed by Ronald Fisher in the early twentieth century and has dominated statistics since. I might say here, though, that many statisticians are turning to Bayesian estimation when modeling certain types of data. Again, we'll address this situation in the final chapter, proposing several predictions in the process.

1.2.3 Count Models

The majority of count models discussed in this book are based on two probability distributions – the Poisson and negative binomial PDFs. I add three additional models in this volume that I consider important when initially evaluating count data – the Poisson inverse Gaussian model, or PIG, Greene's three-parameter negative binomial P, or NB-P, and generalized Poisson (GP) models. These five distributions are closely related. The Poisson distribution has a single parameter to be estimated, μ, or the mean, which is also sometimes referred to as the location parameter. The unique feature of the Poisson distribution is that the mean and variance are the same. The higher the value of the mean of the distribution, the greater the variance or variability in the data. For instance, if we are modeling the number of cars failing to properly stop at two different stop signs per day over a period of a month, and if the average number of failures to stop per day at Site A is 4 and at Site B is 8, we automatically know that the variance of the distribution of failures at Site A is also 4 and at Site B is 8. No other measurements need be done – that is, if the true distribution at each site is Poisson. Recall from algebra that the variance is the square of the standard deviation. The mean and standard deviation of the counts of failures at Site A are 4 and 2, respectively, and at Site B are 8 and $2\sqrt{2}$.

This criterion of the Poisson distribution is referred to as the equidispersion criterion. The problem is that when modeling real data, the equidispersion criterion is rarely satisfied. Analysts usually must adjust their Poisson model in some way to account for any under- or overdispersion that is in the data. Overdispersion is by far the foremost problem facing analysts who use Poisson regression when modeling count data.

I should be clear about the meaning of overdispersion since it is central to the modeling of count data and therefore plays an important role in this book. Overdispersion almost always refers to excess variability or correlation in a Poisson model, but it also needs to be considered when modeling other count models as well. Keep in mind, however, that when the term "overdispersion" is used, most analysts are referring to Poisson overdispersion (i.e., overdispersion in a Poisson model).

Simply put, Poisson overdispersion occurs in data where the variability of the data is greater than the mean. Overdispersion also is used to describe data in a slightly more general sense, as when the observed or "in fact" variance of the count response is greater than the variance of the predicted or expected counts. This latter type of variance is called expected variance. Again, if the observed variance of the response is greater than the expected variance, the data are overdispersed. A model that fails to properly adjust for overdispersed data is called an overdispersed model. As such, its standard errors are biased and cannot be trusted. The standard errors associated with model predictors may appear from the model to significantly contribute to the understanding of the response, but in fact they may not. Many analysts have been deceived into thinking that they have developed a well-fitted model.

Unfortunately, statistical software at times fails to provide an analyst with the information needed to determine if a Poisson model is overdispersed or underdispersed. We discuss in some detail exactly how we can determine whether a model is overdispersed. More properly perhaps, this book will provide guidelines to help you decide whether a Poisson model is equidispersed.

Probably the most popular method of dealing with apparent Poisson overdispersion is to model the data using a negative binomial model. The negative binomial distribution has an extra parameter, referred to as the negative binomial dispersion parameter. Some books and articles call the dispersion parameter the *heterogeneity* parameter or *ancillary* parameter. These are appropriate names as well. The dispersion parameter is a measure of the adjustment needed to accommodate the extra variability, or heterogeneity, in the data. However, the term *dispersion* parameter has become the standard name for the second parameter of the negative binomial distribution.

The negative binomial, which we discuss in more detail later, allows more flexibility in modeling overdispersed data than does a single-parameter Poisson model. The negative binomial is derived as a Poisson-gamma mixture model, with the dispersion parameter being distributed as gamma shaped. The gamma PDF is pliable and allows for a wide variety of shapes. As a consequence, most overdispersed count data can be appropriately modeled using a negative binomial regression. The advantage of using the negative binomial rests with the fact that when the dispersion parameter is zero (0), the model is Poisson.[2] Values of the dispersion parameter greater than zero indicate that the model has adjusted for correspondingly greater amounts of

[2] I term this the direct parameterization of the negative binomial. Unlike most commercial statistical software, R's **`glm`** and **`glm.nb`** functions employ an inverted relationship of the dispersion parameter, theta, so that a Poisson model results when theta approaches infinity. Most subsequent R functions have followed **`glm`** and **`glm.nb`**. I maintain the direct relationship for all count models in this volume and discuss the differences between the two parameterizations in some detail later in the book.