CHAPTER 13

Special Types of Regression

CONTENTS

13.1 INTRODUCTION

The power and flexibility of the general linear model make it the single most useful technique in the statistical toolbox. However, there are situations where it is not appropriate. In this chapter, we will cover several situations where either the relation between the dependent and independent variables cannot be expressed linearly, or the dependent variable cannot be normally distributed, or both.

All of these techniques have a certain similarity to regression. Each is concerned with modeling the influence of one or more independent variables on the mean of a response variable. These influences are expressed through the regression coefficients. This gives their results a kind of familiarity, which eases the transition from the general linear model to these special adaptations.

13.1.1 Maximum Likelihood and Least Squares

The estimation of parameters for models in Chapters 4 through 11 rested on the principle of least squares. With this criterion, parameters are chosen to minimize the sums of squared errors (SSE). In mathematics, minimization usually is achieved by setting the derivatives with respect to the unknown parameters equal to zero

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00013-5

and solving the resulting set of normal equations. This is the origin of the normal equations given in Sections 7.3 and 8.2.

As long as the relationship between the dependent and independent variables is correctly described by a model that is linear in the parameters, least squares will lead to unbiased estimates of those parameters. If the error distribution is normal with constant variance, the least squares estimates will be the best possible; that is, they will have the smallest standard errors.

As we move away from the assumption that the dependent variable follows a normal distribution, **maximum likelihood estimation** can give better results than least squares. Briefly, the likelihood is the probability for the observed data set given a set of proposed values for the parameters. (For continuous dependent variables, we replace probability with probability density, but that does not affect the basic idea.) Likelihood is a relative rather than an absolute quantity. That is, saying one choice of parameter values gives a likelihood of 3 tells us nothing. However, knowing that another choice will give us a likelihood of 2 tells us that the first choice provides a better match of the parameters to the data.

The principle of maximum likelihood (ML) simply states that we should estimate parameters by choosing parameter values that give the largest possible likelihood. A vast body of statistical theory has been developed for ML (see Wackerly *et al.*, 1996), showing that in many situations these estimators are among the best possible. Most of this theory is asymptotic; that is, it requires large samples.

It turns out the principles of least squares and maximum likelihood are not competitors. For cases where the error terms are normally distributed, least squares and maximum likelihood yield the same results! This provides the mathematical justification for least squares when the usual regression assumptions are satisfied.

We will not attempt to develop ML theory here, as it generally requires training in calculus. However, some quantities are cited repeatedly in output from ML procedures. It will help you to be able to recognize and interpret these quantities.

- Likelihood, L, measures the fit of the parameters to the data. Large values (relative to other choices of the parameter values) denote good fit.
- Log-likelihood, $\ln(L)$, is calculated because it is easier to manipulate mathematically than L. Again, large values denote good fit.
- Negative log-likelihood, $-\ln(L)$, has the property that small values denote good fit. This quantity is calculated because it has the same interpretation as SSE, smaller is better. In fact, when errors are normally distributed, $-\ln(L)$ and SSE are equivalent.
- Likelihood ratio tests. Comparisons of a full and reduced model are based on the differences in their $-2 \times \ln(L)$ rather than their SSE. The resulting test is a χ^2 test with degrees of freedom equal to the number of parameters dropped from the full model. These tests are analogous to the F tests used to compare full and reduced models in the general linear model.

There are a few situations where ML yields easily manipulated normal equations similar to those for least squares theory. Unfortunately, those situations are rare. Computationally, ML parameters are found by numerical optimization procedures. Users of the major statistical packages will rarely have to worry about the details as long as the models are properly specified.

13.2 LOGISTIC REGRESSION

Logistic regression is possibly the most frequently used regression-like procedure. It is designed for the situation where the response variable, y, has only two possible outcomes. We say y is **dichotomous**, or **binary**. For example, y might represent

- whether a parolee does or does not violate parole during the first six months,
- whether a computer does or does not require servicing during its warranty period,
- whether an elderly person does or does not show signs of dementia, or
- whether a student succeeds or does not succeed in passing college algebra.

We can represent each individual y_i generically as a 0 (for failure) or a 1 (for success). We focus solely on the probability p of a success, since the probability of a failure is necessarily $1 - p$. This is the binomial situation with $n = 1$ (see Section 2.3).

Our interest is in whether the probability p is influenced by one or more independent variables $x_1, x_2, \ldots, x_m$. We will denote the value of p at some specific set of values for the independent variables as p_x. Using the properties of discrete probability distributions, we have that

$$\mathrm{E}(y_i) = \mu_{y|x} = p_x \quad \text{and}$$
$$\mathrm{Var}(y_i) = \sigma_i^2 = p_x(1 - p_x).$$

We might try an ordinary regression of the y_i on the $x_1, x_2, \ldots, x_m$, since regression is intended to model the impact of the independent variables on $\mu_{y|x}$, which is the same as the probability of success. We immediately encounter two problems. First, there is no way to force the fitted values $\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_{mi}$ to remain between 0 and 1. Since the fitted values are estimated probabilities of success, this is a fatal flaw. Second, even if we could find a method to restrict the fitted values, the distribution of the binary y_i is not even roughly normal.

The first problem is addressed by expressing the relationship between the p_x and the independent variables as a nonlinear function known as the logistic function. The second problem is solved by estimating parameters using maximum likelihood rather than least squares.

The logistic function is

$$p(x_1, \ldots, x_m) = p_x = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}.$$