Chapter 15. Supplemental Text Material

S15-1. The Form of a Transformation

In Section 3-4.3 of the textbook we introduce transformations as a way to stabilize the variance of a response and to (hopefully) induce approximate normality when inequality of variance and nonnormality occur jointly (as they often do). In Section 15-1.1 of the book the Box-Cox method is presented as an elegant analytical method for selecting the form of a transformation. However, many experimenters select transformations empirically by trying some of the simple power family transformations in Table 3-9 of Chapter 3 ($\sqrt{y}, \ln(y)$, or $1/y$, for example) or which appear on the menu of their computer software package.

It is possible to give a theoretical justification of the power family transformations presented in Table 3-9. For example, suppose that y is a response variable with mean $E(y) = \mu$ and variance $V(y) = \sigma^2 = f(\mu)$. That is, the variance of y is a function of the mean. We wish to find a transformation $h(y)$ so that the variance of the transformed variable $x = h(y)$ is a constant unrelated to the mean of y. In other words, we want $V[h(y)]$ to be a constant that is unrelated to $E[h(y)]$.

Expand $x = h(y)$ in a Taylor series about μ, resulting in

$$x = h(y)$$
$$= h(\mu) + h'(\mu)(y - \mu) + R$$
$$\cong h(\mu) + h'(\mu)(y - \mu)$$

where R is the remainder in the first-order Taylor series, and we have ignored the remainder. Now the mean of x is

$$E(x) = E[h(\mu) + h'(\mu)(y - \mu)]$$
$$= h(\mu)$$

and the variance of x is

$$V(x) = E\left[x - E(x)\right]^2$$
$$= E\left[h(\mu) + h'(\mu)(y - \mu) - h(\mu)\right]^2$$
$$= E\left[h'(\mu)(y - \mu)\right]^2$$
$$= \sigma^2\left[h'(\mu)\right]^2$$

Since $\sigma^2 = f(\mu)$, we have

$$V(x) = f(\mu)\left[h'(\mu)\right]^2$$

We want the variance of x to be a constant, say c^2. So set

$$c^2 = f(\mu)\left[h'(\mu)\right]^2$$

and solve for $h'(y)$, giving

$$h'(\mu) = \frac{c}{\sqrt{f(\mu)}}$$

Thus, the form of the transformation that is required is

$$h(\mu) = c\int \frac{dt}{\sqrt{f(t)}}$$
$$= cG(\mu) + k$$

where k is a constant.

As an example, suppose that for the response variable y we assumed that the mean and variance were equal. This actually happens in the Poisson distribution. Therefore,

$$\mu = \sigma^2 \text{ implying that } f(t) = t$$

So

$$h(\mu) = c\int \frac{dt}{\sqrt{t}}$$
$$= c\int t^{-1/2} dt + k$$
$$= c\frac{t^{-(1/2)+1}}{-(1/2)+1} + k$$
$$= 2c\sqrt{t} + k$$

This implies that taking the square root of y will stabilize the variance. This agrees with the advice given in the textbook (and elsewhere) that the square root transformation is very useful for stabilizing the variance in Poisson data or in general for count data where the mean and variance are not too different.

As a second example, suppose that the square root of the mean is approximately equal to the variance; that is, $\mu^{1/2} = \sigma^2$. Essentially, this says that

$$\mu = \left[\sigma^2\right]^2 \text{ which implies that } f(t) = t^2$$

Therefore,

$$h(\mu) = c\int \frac{dt}{\sqrt{t^2}}$$
$$= c\int \frac{dt}{t} + k$$
$$= c\log(t) + k, \text{ if } t > 0$$

This implies that for a positive response where $\mu^{1/2} = \sigma^2$ the log of the response is an appropriate variance-stabilizing transformation.

S15-2. Selecting λ in the Box-Cox Method

In Section 15-1.1 of the Textbook we present the Box-Cox method for analytically selecting a response variable transformation, and observe that its theoretical basis is the method of maximum likelihood. In applying this method, we are essentially maximizing

$$L(\lambda) = -\frac{1}{2}n\ln\left[SS_E(\lambda)\right]$$

or equivalently, we are minimizing the error sum of squares with respect to λ. An approximate 100(1-α) percent confidence interval for λ consists of those values of λ that satisfy the inequality

$$L(\hat{\lambda}) - L(\lambda) \le \frac{1}{2}\chi^2_{\alpha,1}/n$$

where n is the sample size and $\chi^2_{\alpha,1}$ is the upper α percentage point of the chi-square distribution with one degree of freedom. To actually construct the confidence interval we would draw on a plot of $L(\hat{\lambda})$ versus λ a horizontal line at height

$$L(\hat{\lambda}) - \frac{1}{2}\chi^2_{\alpha,1}$$

on the vertical scale. This would cut the curve of $L(\hat{\lambda})$ at two points, and the locations of these two points on the λ axis define the two end points of the approximate confidence interval for λ. If we are minimizing the residual or error sum of squares (which is identical to maximizing the likelihood) and plotting $SS_E(\lambda)$ versus λ, then the line must be plotted at height

$$SS^* = SS_E(\hat{\lambda})e^{\chi^2_{\alpha,1}/n}$$

Remember that $\hat{\lambda}$ is the value of λ that minimizes the error sum of squares.

Equation (14-20 in the textbook looks slightly different than the equation for SS^* above. The term $\exp(\chi^2_{\alpha,1}/n)$ has been replaced by $1+(t^2_{\alpha/2,v})/v$, where v is the number of degrees of freedom for error. Some authors use $1+(\chi^2_{\alpha/2,v})/v$ or $1+(z^2_{\alpha/2})/v$ instead, or sometimes $1+(t^2_{\alpha/2,n})/n$ or $1+(\chi^2_{\alpha/2,n})/n$ or $1+(z^2_{\alpha/2})/n$. These are all based on the expansion of $\exp(x) = 1 + x + x^2/2! + x^3/3! + \cdots \approx 1 + x$, and the fact that $\chi^2_1 = z^2 \approx t^2_v$, unless the number of degrees of freedom v is too small. It is perhaps debatable whether we should use n or v, but in most practical cases, there will be little difference in the confidence intervals that result.

S15-3. Generalized Linear Models

Section 15-1.2 considers an alternative approach to data transformation when the "usual" assumptions of normality and constant variance are not satisfied. This approach is based

on the generalized linear model or GLM. Examples 15-2, 15-3, and 15-4 illustrated the applicability of the GLM to designed experiments.

The GLM is a unification of nonlinear regression models and nonnormal response variable distributions, where the response distribution is a member of the **exponential family**, which includes the normal, Poisson, binomial, exponential and gamma distributions as members. Furthermore, the normal-theory linear model is just a special case of the GLM, so in many ways, the GLM is a unifying approach to empirical modeling and data analysis.

We begin our presentation of these models by considering the case of **logistic regression**. This is a situation where the response variable has only two possible outcomes, generically called "success" and "failure" and denoted by 0 and 1. Notice that the response is essentially qualitative, since the designation "success" or "failure" is entirely arbitrary. Then we consider the situation where the response variable us a count, such as the number of defects in a unit of product (as in the grille defects of Example 14-2), or the number of relatively rare events such as the number of Atlantic hurricanes than make landfall on the United States in a year. Finally, we briefly show how all these situations are unified by the GLM.

S15-3.1. Models with a Binary Response Variable

Consider the situation where the response variable from an experiment takes on only two possible values, 0 and 1. These could be arbitrary assignments resulting from observing a qualitative response. For example, the response could be the outcome of a functional electrical test on a semiconductor device for which the results are either a "success", which means the device works properly, or a "failure", which could be due to a short, an open, or some other functional problem.

Suppose that the model has the form

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i$$

where $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \cdots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \beta_2, \cdots, \beta_k]$, $\mathbf{x}_i'\beta$ is called the **linear predictor,** and the response variable y_i takes on the values either 0 or 1. We will assume that the response variable y_i is a **Bernoulli random variable** with probability distribution as follows:

y_i	Probability
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Now since $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(y_i) = 1(\pi_i) +)(1 - \pi_i)$$
$$= \pi_i$$

This implies that

$$E(y_i) = \mathbf{x}_i'\beta = \pi_i$$

This means that the expected response given by the response function $E(y_i) = \mathbf{x}_i'\beta$ is just the probability that the response variable takes on the value 1.

There are some substantive problems with this model. First, note that if the response is binary, then the error term ε_i can only take on two values, namely

$$\varepsilon_i = 1 - \mathbf{x}_i'\beta \text{ when } y_i = 1$$
$$\varepsilon_i = -\mathbf{x}_i'\beta \text{ when } y_i = 0$$

Consequently, the errors in this model cannot possibly be normal. Second, the error variance is not constant, since

$$\sigma_{y_i}^2 = E\{y_i - E(y_i)\}^2$$
$$= (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i)$$
$$= \pi_i(1 - \pi_i)$$

Notice that this last expression is just

$$\sigma_{y_i}^2 = E(y_i)[1 - E(y_i)]$$

since $E(y_i) = \mathbf{x}_i'\beta = \pi_i$. This indicates that the variance of the observations (which is the same as the variance of the errors because $\varepsilon_i = y_i - \pi_i$, and π_i is a constant) is a function of the mean. Finally, there is a constraint on the response function, because

$$0 \le E(y_i) = \pi_i \le 1$$

This restriction causes serious problems with the choice of a **linear response function**, as we have initially assumed.

Generally, when the response variable is binary, there is considerable evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) *S*-shaped (or reverse *S*-shaped) function is usually employed. This function is called the **logistic response function**, and has the form

$$E(y) = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}$$

or equivalently,

$$E(y) = \frac{1}{1 + \exp(-\mathbf{x}'\beta)}$$

The logistic response function can be easily linearized. Let $E(y) = \pi$ and make the transformation

$$\eta = \ln\left(\frac{\pi}{1 - \pi}\right)$$

Then in terms of our **linear predictor** $\mathbf{x}'\beta$ we have

$$\eta = \mathbf{x}'\beta$$

This transformation is often called the **logit transformation** of the probability π, and the ratio $\pi/(1-\pi)$ in the transformation is called the odds. Sometimes the logit transformation is called the log-odds.

There are other functions that have the same shape as the logistic function, and they can also be obtained by transforming π. One of these is the *probit* transformation, obtained by transforming π using the cumulative normal distribution. This produces a *probit regression model*. The probit regression model is less flexible than the logistic regression model because it cannot easily incorporate more than one predictor variable. Another possible transformation is the **complimentary log-log transformation** of π, given by $\ln[-\ln(1-\pi)]$. This results in a response function that is not symmetric about the value $\pi = 0.5$.

S15-3.2. Estimating the Parameters in a Logistic Regression Model

The general form of the logistic regression model is

$$y_i = E(y_i) + \varepsilon_i$$

where the observations y_i are independent Bernoulli random variables with expected values

$$E(y_i) = \pi_i$$
$$= \frac{\exp(\mathbf{x}'_i\beta)}{1 + \exp(\mathbf{x}'_i\beta)}$$

We will use the method of **maximum likelihood** to estimate the parameters in the linear predictor $\mathbf{x}'_i\beta$.

Each sample observation follows the Bernoulli distribution, so the probability distribution of each sample observation is

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i^{1-y_i}), i = 1, 2, \cdots, n$$

and of course each observation y_i takes on the value 0 or 1. Since the observations are independent, the likelihood function is just

$$L(y_1, y_2, \cdots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i)$$
$$= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i^{1-y_i})$$

It is more convenient to work with the log-likelihood

$$\ln L(y_1, y_2, \cdots, y_n, \beta) = \ln \prod_{i=1}^{n} f_i(y_i)$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

Now since $1-\pi_i = [1+\exp(\mathbf{x}_i'\beta)]^{-1}$ and $\eta_i = \ln[\pi_i / (1-\pi_i)] = \mathbf{x}_i'\beta$, the log-likelihood can be written as

$$\ln L(\mathbf{y}, \beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i'\beta - \sum_{i=1}^{n} \ln[1+\exp(\mathbf{x}_i'\beta)]$$

Often in logistic regression models we have repeated observations or trials at each level of the x variables. This happens frequently in designed experiments. Let y_i represent the number of 1's observed for the ith observation and n_i be the number of trials at each observation. Then the log-likelihood becomes

$$\ln L(\mathbf{y}, \beta) = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i)$$

Numerical search methods could be used to compute the maximum likelihood estimates (or MLEs) $\hat{\beta}$. However, it turns out that we can use iteratively reweighted least squares (IRLS) to actually find the MLEs. To see this recall that the MLEs are the solutions to

$$\frac{\partial L}{\partial \beta} = 0$$

which can be expressed as

$$\frac{\partial L}{\partial \pi_i} \frac{\partial \pi_i}{\partial \beta} = 0$$

Note that

$$\frac{\partial L}{\partial \pi_i} = \sum_{i=1}^{n} \frac{n_i}{\pi_i} - \sum_{i=1}^{n} \frac{n_i}{1-\pi_i} + \sum_{i=1}^{n} \frac{y_i}{1-\pi_i}$$

and

$$\frac{\partial \pi_i}{\partial \beta} = \left\{ \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)} - \left[\frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)} \right]^2 \right\} \mathbf{x}_i$$

$$= \pi_i(1-\pi_i)\mathbf{x}_i$$

Putting this all together gives

$$\frac{\partial L}{\partial \beta} = \left[\sum_{i=1}^{n} \frac{n_i}{\pi_i} - \sum_{i=1}^{n} \frac{n_i}{1-\pi_i} + \sum_{i=1}^{n} \frac{y_i}{1-\pi_i} \right] \pi_i (1-\pi_i) \mathbf{x}_i$$

$$= \sum_{i=1}^{n} \left[\frac{y_i}{\pi_i} - \frac{n_i}{1-\pi_i} + \frac{y_i}{1-\pi_i} \right] \pi_i (1-\pi_i) \mathbf{x}_i$$

$$= \sum_{i=1}^{n} (y_i - n_i \pi_i) \mathbf{x}_i$$

Therefore, the maximum likelihood estimator solves

$$\mathbf{X}'(\mathbf{y} - \mu) = \mathbf{0}$$

where $\mathbf{y}' = [y_1, y_2, \cdots, y_n]$ and $\mu' = [n_1 \pi_1, n_2 \pi_2 \cdots, n_n \pi_n]$. This set of equations is often called the **maximum likelihood score equations**. They are actually the same form of the normal equations that we have seen previously for linear least squares, because in the linear regression model, $E(\mathbf{y}) = \mathbf{X}\beta = \mu$ and the normal equations are

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

which can be written as

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$$
$$\mathbf{X}'(\mathbf{y} - \mu) = \mathbf{0}$$

The **Newton-Raphson** method is actually used to solve the score equations. This procedure observes that in the neighborhood of the solution, we can use a first-order Taylor series expansion to form the approximation

$$p_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \beta} \right)' (\beta^* - \beta) \tag{1}$$

where

$$p_i = \frac{y_i}{n_i}$$

and β^* is the value of β that solves the score equations. Now $\eta_i = \mathbf{x}_i'\beta$, and

$$\frac{\partial \eta_i}{\partial \beta} = \mathbf{x}_i$$

so

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

By the chain rule

$$\frac{\partial \pi_i}{\partial \beta} = \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{\partial \pi_i}{\partial \eta_i} \mathbf{x}_i$$

Therefore, we can rewrite (1) above as

$$p_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \eta_i}\right) \mathbf{x}_i'(\beta^* - \beta)$$

$$p_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \eta_i}\right) (\mathbf{x}_i'\beta^* - \mathbf{x}_i'\beta) \tag{2}$$

$$p_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \eta_i}\right) (\eta_i^* - \eta_i)$$

where η_i^* is the value of η_i evaluated at β^*. We note that

$$(y_i - n_i\pi_i) = (n_ip_i - n_i\pi_i) = n_i(p_i - \pi_i)$$

and since

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

we can write

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} - \left[\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\right]^2$$

$$= \pi_i(1 - \pi_i)$$

Consequently,

$$y_i - n_i\pi_i \approx [n_i\pi_i(1 - \pi_i)](\eta_i^* - \eta_i)$$

Now the variance of the linear predictor $\eta_i^* = \mathbf{x}_i'\beta^*$ is, to a first approximation,

$$V(\eta_i^*) \approx \frac{1}{n_i\pi_i(1 - \pi_i)}$$

Thus

$$y_i - n_i\pi_i \approx \left[\frac{1}{V(\eta_i^*)}\right](\eta_i^* - \eta_i)$$

and we may rewrite the score equations as

$$\sum_{i=1}^{n} \left[\frac{1}{V(\eta_i)}\right](\eta_i^* - \eta_i) = 0$$

or in matrix notation,

$$\mathbf{X}'\mathbf{V}^{-1}(\eta^* - \eta) = \mathbf{0}$$

where $\mathbf{V}$ is a diagonal matrix of the weights formed from the variances of the η_i. Because $\eta = \mathbf{X}\beta$ we may write the score equations as

$$\mathbf{X}'\mathbf{V}^{-1}(\eta^* - \mathbf{X}\beta) = \mathbf{0}$$

and the maximum likelihood estimate of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\eta^*$$

However, there is a problem because we don't know η^*. Our solution to this problem uses equation (2):

$$p_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \eta_i}\right)(\eta_i^* - \eta_i)$$

which we can solve for η_i^*,

$$\eta_i^* \approx \eta_i + (p_i - \pi_i)\frac{\partial \eta_i}{\partial \pi_i}$$

Let $z_i = \eta_i + (p_i - \pi_i)\dfrac{\partial \eta_i}{\partial \pi_i}$ and $\mathbf{z}' = [z_1, z_2, \cdots, z_n]$. Then the Newton-Raphson estimate of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{z}$$

Note that the random portion of z_i is

$$(p_i - \pi_i)\frac{\partial \eta_i}{\partial \pi_i}$$

Thus

$$V\left[(p_i - \pi_i)\frac{\partial \eta_i}{\partial \pi_i}\right] = \left[\frac{\pi_i(1-\pi_i)}{n_i}\right]\left(\frac{\partial \eta_i}{\partial \pi_i}\right)^2$$

$$= \left[\frac{\pi_i(1-\pi_i)}{n_i}\right]\left(\frac{1}{\pi_i(1-\pi_i)}\right)^2$$

$$= \frac{1}{n_i\pi_i(1-\pi_i)}$$

So $\mathbf{V}$ is the diagonal matrix of weights formed from the variances of the random part of $\mathbf{z}$.

Thus the IRLS algorithm based on the Newton-Raphson method can be described as follows:

1. Use ordinary least squares to obtain an initial estimate of β, say $\hat{\beta}_0$;

2. Use $\hat{\beta}_0$ to estimate $\mathbf{V}$ and π;

3. Let $\eta_0 = \mathbf{X}\hat{\beta}_0$;

4. Base $\mathbf{z}_1$ on η_0;

5. Obtain a new estimate $\hat{\beta}_1$, and iterate until some suitable convergence criterion is satisfied.

If $\hat{\beta}$ is the final value that the above algorithm produces and if the model assumptions are correct, then we can show that asymptotically

$$E(\hat{\beta}) = \beta \quad \text{and} \quad V(\hat{\beta}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

The fitted value of the logistic regression model is often written as

$$\hat{\pi}_i = \frac{\exp(\mathbf{x}_i'\hat{\beta})}{1+\exp(\mathbf{x}_i'\hat{\beta})}$$

$$= \frac{1}{1+\exp(-\mathbf{x}_i'\hat{\beta})}$$

S15-3.3. Interpreting the Parameters in a Logistic Regression Model

It is relatively easy to interpret the parameters in a logistic regression model. Consider first the case where the linear predictor has only a single predictor, so that the fitted value of the model at a particular value of x, say x_i, is

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The fitted value at $x_i + 1$ is

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1(x_i + 1)$$

and the difference in the two predicted values is

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1$$

Now $\hat{\eta}(x_i)$ is just the log-odds when the regressor variable is equal to x_i, and $\hat{\eta}(x_i + 1)$ is just the log-odds when the regressor is equal to $x_i + 1$. Therefore, the difference in the two fitted values is

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \ln(odds_{x_i+1}) - \ln(odds_{x_i})$$

$$= \ln\left(\frac{odds_{x_i+1}}{odds_{x_i}}\right)$$

$$= \hat{\beta}_1$$

If we take antilogs, we obtain the **odds ratio**

$$\hat{O}_R \equiv \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$

The odds ratio can be interpreted as the estimated increase in the probability of success associated with a one-unit change in the value of the predictor variable. In general, the estimated increase in the odds ratio associated with a change of d units in the predictor variable is $\exp(d\hat{\beta}_1)$.

The interpretation of the regression coefficients in the multiple logistic regression model is similar to that for the case where the linear predictor contains only one regressor. That is, the quantity $\exp(\hat{\beta}_j)$ is the odds ratio for regressor x_j, assuming that all other predictor variables are constant.

S15-3.4. Hypothesis Tests on Model Parameters

Hypothesis testing in the GLM is based on the general method of **likelihood ratio tests**. It is a large-sample procedure, so the test procedures rely on asymptotic theory. The likelihood ratio approach leads to a statistic called **deviance**.

Model Deviance

The deviance of a model compares the log-likelihood of the fitted model of interest to the log-likelihood of a saturated model; that is, a model that has exactly n parameters and which fits the sample data perfectly. For the logistic regression model, this means that the probabilities π_i are completely unrestricted, so setting $\hat{\pi}_i = y_i$ (recall that $y_i = 0$ or 1) would maximize the likelihood. It can be shown that this results in a maximum value of the likelihood function for the saturated model of unity, so the maximum value of the log-likelihood function is zero.

Now consider the log- likelihood function for the fitted logistic regression model. When the maximum likelihood estimates $\hat{\beta}$ are used in the log- likelihood function, it attains its maximum value, which is

$$\ln L(\hat{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}' \hat{\beta}_i - \sum_{i=1}^{n} \ln[1 + \exp(\mathbf{x}_i' \hat{\beta})]$$

The value of the log-likelihood function for the fitted model can never exceed the value of the log-likelihood function for the saturated model, because the fitted model contains fewer parameters. The deviance compares the log-likelihood of the saturated model with the log-likelihood of the fitted model. Specifically, **model deviance** is defined as

$$\lambda(\beta) = 2\ln L(\text{saturated model}) - 2\ln L(\hat{\beta})$$
$$= 2[\ell((\text{saturated model}) - \ell(\hat{\beta})]$$

(3)

where ℓ denotes the log of the likelihood function. Now if the logistic regression model is the correct regression function and the sample size n is large, the model deviance has an approximate chi-square distribution with $n - p$ degrees of freedom. Large values of

the model deviance would indicate that the model is not correct, while a small value of model deviance implies that the fitted model (which has fewer parameters than the saturated model) fits the data almost as well as the saturated model. The formal test criteria would be as follows:

$$\text{if } \lambda(\beta) \leq \chi^2_{\alpha,n-p} \text{ conclude that the fitted model is adequate}$$

$$\text{if } \lambda(\beta) > \chi^2_{\alpha,n-p} \text{ conclude that the fitted model is not adequate}$$

The deviance is related to a very familiar quantity. If we consider the standard normal-theory linear regression model, the deviance turns out to be the error or residual sum of squares divided by the error variance σ^2.

Testing Hypotheses on Subsets of Parameters using Deviance

We can also use the deviance to test hypotheses on subsets of the model parameters, just as we used the difference in regression (or error) sums of squares to test hypotheses in the normal-error linear regression model case. Recall that the model can be written as

$$\eta = \mathbf{X}\beta$$
$$= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$$

where the *full model* has p parameters, β_1 contains $p - r$ of these parameters, β_2 contains r of these parameters, and the columns of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ contain the variables associated with these parameters. Suppose that we wish to test the hypotheses

$$H_0 : \beta_2 = \mathbf{0}$$
$$H_1 : \beta_2 \neq \mathbf{0}$$

Therefore, the *reduced model* is

$$\eta = \mathbf{X}_1\beta_1$$

Now fit the reduced model, and let $\lambda(\beta_1)$ be the deviance for the reduced model. The deviance for the reduced model will always be larger than the deviance for the full model, because the reduced model contains fewer parameters. However, if the deviance for the reduced model is not much larger than the deviance for the full model, it indicates that the reduced model is about as good a fit as the full model, so it is likely that the parameters in β_2 are equal to zero. That is, we cannot reject the null hypothesis above. However, if the difference in deviance is large, at least one of the parameters in β_2 is likely not zero, and we should reject the null hypothesis. Formally, the difference in deviance is

$$\lambda(\beta_2|\beta_1) = \lambda(\beta_1) - \lambda(\beta)$$

and this quantity has $n - (p - r) - (n - p) = r$ degrees of freedom. If the null hypothesis is true and if n is large, the difference in deviance has a chi-square distribution with r degrees of freedom. Therefore, the test statistic and decision criteria are

$$\text{if } \lambda(\beta_2|\beta_1) \ge \chi^2_{\alpha,r} \text{ reject the null hypothesis}$$

$$\text{if } \lambda(\beta_2|\beta_1) < \chi^2_{\alpha,r} \text{ do not reject the null hypothesis}$$

Sometimes the difference in deviance $\lambda(\beta_2|\beta_1)$ is called the **partial deviance**. It is a likelihood ratio test. To see this, let $L(\hat{\beta})$ be the maximum value of the likelihood function for the full model, and $L(\hat{\beta}_1)$ be the maximum value of the likelihood function for the reduced model. The **likelihood ratio** is

$$\frac{L(\hat{\beta}_1)}{L(\hat{\beta})}$$

The test statistic for the likelihood ratio test is equal to minus two times the log-likelihood ratio, or

$$\chi^2 = -2\ln\frac{L(\hat{\beta}_1)}{L(\hat{\beta})}$$

$$= 2\ln L(\hat{\beta}) - 2\ln L(\hat{\beta}_1)$$

However, this is exactly the same as the difference in deviance. To see this, substitute from the definition of the deviance from equation (3) and note that the log-likelihoods for the saturated model cancel out.

Tests on Individual Model Coefficients

Tests on individual model coefficients, such as

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \ne 0$$

can be conducted by using the difference in deviance method described above. There is another approach, also based on the theory of maximum likelihood estimators. For large samples, the distribution of a maximum likelihood estimator is approximately normal with little or no bias. Furthermore, the variances and covariances of a set of maximum likelihood estimators can be found from the second partial derivatives of the log-likelihood function with respect to the model parameters, evaluated at the maximum likelihood estimates. Then a t-like statistic can be constructed to test the above hypothesis. This is sometimes referred to as **Wald inference.**

Let **G** denote the $p \times p$ matrix of second partial derivatives of the log-likelihood function; that is

$$G_{ij} = \frac{\partial^2 \ell(\beta)}{\partial \beta_i \partial \beta_j}, i, j = 0,1,\cdots,k$$

G is called the **Hessian matrix**. If the elements of the Hessian are evaluated at the maximum likelihood estimators $\beta = \hat{\beta}$, the large-sample approximate covariance matrix of the regression coefficients is

$$V(\hat{\beta}) \equiv \hat{\Sigma} = -\mathbf{G}(\hat{\beta})^{-1}$$

The square roots of the diagonal elements of this matrix are the large-sample standard errors of the regression coefficients, so the test statistic for the null hypothesis in

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

is

$$Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

The reference distribution for this statistic is the standard normal distribution. Some computer packages square the Z_0 statistic and compare it to a chi-square distribution with one degree of freedom. It is also straightforward to use Wald inference to construct confidence intervals on individual regression coefficients.

S15-3.5. Poisson Regression

We now consider another regression modeling scenario where the response variable of interest is not normally distributed.   In this situation the response variable represents a count of some relatively rare event, such as defects in a unit of manufactured product, errors or "bugs" in software, or a count of particulate matter or other pollutants in the environment. The analyst is interested in modeling the relationship between the observed counts and potentially useful regressor or predictor variables. For example, an engineer could be interested in modeling the relationship between the observed number of defects in a unit of product and production conditions when the unit was actually manufactured.

We assume that the response variable y_i is a count, such that the observation $y_i = 0, 1, \cdots,$. A reasonable probability model for count data is often the Poisson distribution

$$f(y) = \frac{e^{-\mu} \mu^y}{y!}, y = 0, 1, \cdots$$

where the parameter $\mu > 0$. The Poison is another example of a probability distribution where the mean and variance are related. In fact, for the Poisson distribution it is straightforward to show that

$$E(y) = \mu \text{ and } V(y) = \mu$$

That is, both the mean *and* variance of the Poisson distribution are equal to the parameter μ.

The Poisson regression model can be written as

$$y_i = E(y_i) + \varepsilon_i, i = 1, 2, \cdots, n$$

We assume that the expected value of the observed response can be written as

$$E(y_i) = \mu_i$$

and that there is a function g that relates the mean of the response to a linear predictor, say

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
$$= \mathbf{x}_i' \beta$$

The function g is usually called the **link function**. The relationship between the mean and the linear predictor is

$$\mu_i = g^{-1}(\mathbf{x}_i' \beta)$$

There are several link functions that are commonly used with the Poisson distribution. One of these is the **identity link**

$$g(\mu_i) = \mu_i = \mathbf{x}_i' \beta$$

When this link is used, $E(y_i) = \mu_i = \mathbf{x}_i' \beta$ since $\mu_i = g^{-1}(\mathbf{x}_i' \beta) = \mathbf{x}_i' \beta$. Another popular link function for the Poisson distribution is the **log link**

$$g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i' \beta$$

For the log link, the relationship between the mean of the response variable and the linear predictor is

$$\mu_i = g^{-1}(\mathbf{x}_i' \beta)$$
$$= e^{\mathbf{x}_i' \beta}$$

The log link is particularly attractive for Poisson regression because it ensures that all of the predicted values of the response variable will be nonnegative.

The method of maximum likelihood is used to estimate the parameters in Poisson regression. The development follows closely the approach used for logistic regression. If we have a random sample of n observations on the response y and the predictors $\mathbf{x}$, then the likelihood function is

$$L(\mathbf{y}, \beta) = \prod_{i=1}^{n} f_i(y_i)$$
$$= \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$
$$= \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp(-\sum_{i=1}^{n} \mu_i)}{\prod_{i=1}^{n} y_i!}$$

where $\mu_i = g^{-1}(\mathbf{x}_i' \beta)$. Once the link function is specified, we maximize the log-likelihood

$$\ln L(\mathbf{y}, \beta) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!)$$

Iteratively reweighted least squares can be used to find the maximum likelihood estimates of the parameters in Poisson regression, following an approach similar to that used for logistic regression. Once the parameter estimates $\hat{\beta}$ are obtained, the fitted Poisson regression model is

$$\hat{y}_i = g^{-1}(\mathbf{x}_i'\hat{\beta})$$

For example, if the identity link is used, the prediction equation becomes

$$\hat{y}_i = g^{-1}(\mathbf{x}_i'\hat{\beta})$$
$$= \mathbf{x}_i'\hat{\beta}$$

and if the log link is specified, then

$$\hat{y}_i = g^{-1}(\mathbf{x}_i'\hat{\beta})$$
$$= \exp(\mathbf{x}_i'\hat{\beta})$$

Inference on the model and its parameters follows exactly the same approach as used for logistic regression. That is, model deviance is an overall measure of goodness of fit, and tests on subsets of model parameters can be performed using the difference in deviance between the full and reduced models. These are likelihood ratio tests. Wald inference, based on large-sample properties of maximum likelihood estimators, can be used to test hypotheses and construct confidence intervals on individual model parameters.

S15-3.6. The Generalized Linear Model

All of the regression models that we have considered in this section belong to a *family* of regression models called the **generalized linear model**, or the **GLM**. The GLM is actually a unifying approach to regression and experimental design models, uniting the usual normal-theory linear regression models and nonlinear models such as logistic and Poisson regression.

A key assumption in the GLM is that the response variable distribution is a member of the exponential family of distributions, which includes the normal, binomial, Poisson, inverse normal, exponential and gamma distributions. Distributions that are members of the exponential family have the general form

$$f(y_i, \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + h(y_i, \phi)\}$$

where ϕ is a scale parameter and θ_i is called the natural location parameter. For members of the exponential family,

$$\mu = E(y) = \frac{db(\theta_i)}{d\theta_i}$$

$$V(y) = \frac{d^2 b(\theta_i)}{d\theta_i^2} a(\phi)$$

$$= \frac{d\mu}{d\theta_i} a(\phi)$$

Let

$$\text{var}(\mu) = \frac{V(y)}{a(\phi)} = \frac{d\mu}{d\theta_i}$$

where var(μ) denotes the dependence of the variance of the response on its mean. As a result, we have

$$\frac{d\theta_i}{d\mu} = \frac{1}{\text{var}(\mu)}$$

It is easy to show that the normal, binomial and Poisson distributions are members of the exponential family.

The Normal Distribution

$$f(y_i, \theta_i, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y-\mu)^2 \right)$$

$$= \exp\left[-\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \right]$$

$$= \exp\left[\frac{1}{\sigma^2}\left(-\frac{y^2}{2} + y\mu - \frac{\mu^2}{2} \right) - \frac{1}{2}\ln(2\pi\sigma^2) \right]$$

$$= \exp\left[\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2} \right) - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \right]$$

Thus for the normal distribution, we have

$$\theta_i = \mu$$

$$b(\theta_i) = \frac{\mu^2}{2}$$

$$a(\phi) = \sigma^2$$

$$h(y_i, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \mu, \text{ and } V(y) = \frac{d^2 b(\theta_i)}{d\theta_i^2} a(\phi) = \sigma^2$$

The Binomial Distribution

$$f(y_i, \theta_i, \phi) = \binom{n}{m} \pi^y (1-\pi)^{n-y}$$

$$= \exp\left\{ \ln\binom{n}{y} + y \ln \pi + (n-y)\ln(1-\pi) \right\}$$

$$= \exp\left\{ \ln\binom{n}{y} + y \ln \pi + n\ln(1-\pi) - y\ln(1-\pi) \right\}$$

$$= \exp\left\{ y \ln\left[\frac{\pi}{1-\pi}\right] + n\ln(1-\pi) + \ln\binom{n}{y} \right\}$$

Therefore, for the binomial distribution,

$$\theta_i = \ln\left[\frac{\pi}{1-\pi}\right] \text{ and } \pi = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$$

$$b(\theta_i) = -n\ln(1-\pi)$$

$$a(\phi) = 1$$

$$h(y_i, \phi) = \ln\binom{n}{y}$$

$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \frac{db(\theta_i)}{d\pi}\frac{d\pi}{d\theta_i}$$

We note that

$$\frac{d\pi}{d\theta_i} = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} - \left[\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right]^2$$

$$= \pi(1-\pi)$$

Therefore,

$$E(y) = \left(\frac{n}{1-\pi}\right)\pi(1-\pi)$$

$$= n\pi$$

We recognize this as the mean of the binomial distribution. Also,

$$V(y) = \frac{dE(y)}{d\theta_i}$$

$$= \frac{dE(y)}{d\pi}\frac{d\pi}{d\theta_i}$$

$$= n\pi(1-\pi)$$

This last expression is just the variance of the binomial distribution.

The Poisson Distribution

$$f(y_i,\theta_i,\phi) = \frac{\lambda^y e^{-\lambda}}{y!}$$
$$= \exp[y \ln \lambda - \lambda - \ln(y!)]$$

Therefore, for the Poisson distribution, we have

$$\theta_i = \ln(\lambda) \text{ and } \lambda = \exp(\theta_i)$$
$$b(\theta_i) = \lambda$$
$$a(\phi) = 1$$
$$h(y_i,\phi) = -\ln(y!)$$

Now

$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \frac{db(\theta_i)}{d\lambda} \frac{d\lambda}{d\theta_i}$$

However, since

$$\frac{d\lambda}{d\theta_i} = \exp(\theta_i) = \lambda$$

the mean of the Poisson distribution is

$$E(y) = 1 \cdot \lambda = \lambda$$

The variance of the Poisson distribution is

$$V(y) = \frac{dE(y)}{d\theta_i} = \lambda$$

S15-3.7. Link Functions and Linear Predictors

The basic idea of a GLM is to develop a linear model for an appropriate **function** of the expected value of the response variable. Let η_i be the **linear predictor** defined by

$$\eta_i = g[E(y_i)] = g(\mu_i) = \mathbf{x}_i'\beta$$

Note that the expected response is just

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\beta)$$

We call the function g the **link function**. Recall that we introduced the concept of a link function in our description of Poisson regression in Section S15-3.5 above. There are many possible choices of the link function, but if we choose

$$\eta_i = \theta_i$$

we say that η_i is the **canonical link**. Table 1 shows the canonical links for the most common choices of distributions employed with the GLM.

Table 1. Canonical Links for the Generalized Linear Model

Distribution	Canonical Link
Normal	$\eta_i = \mu_i$ (identity link)
Binomial	$\eta_i = \ln\left(\dfrac{\pi_i}{1-\pi_i}\right)$ (logistic link)
Poisson	$\eta_i = \ln(\lambda)$ (log link)
Exponential	$\eta_i = \dfrac{1}{\lambda_i}$ (reciprocal link)
Gamma	$\eta_i = \dfrac{1}{\lambda_i}$ (reciprocal link)

There are other link functions that could be used with a GLM, including:

1. The probit link,

$$\eta_i = \Phi^{-1}[E(y_i)]$$

where Φ represents the cumulative standard normal distribution function.

2. The complimentary log-log link,

$$\eta_i = \ln\{\ln[1 - E(y_i)]\}$$

3. The power family link,

$$\eta_i = \begin{cases} E(y_i)^\lambda, & \lambda \neq 0 \\ \ln[E(y_i)], & \lambda = 0 \end{cases}$$

A very fundamental idea is that there are two components to a GLM; the response variable distribution, and the link function. We can view the selection of the link function in a vein similar to the choice of a transformation on the response. However, unlike a transformation, the link function takes advantage of the *natural* distribution of the response. Just as not using an appropriate transformation can result in problems with a fitted linear model, improper choices of the link function can also result in significant problems with a GLM.

S15-3.8. Parameter Estimation in the GLM

The method of maximum likelihood is the theoretical basis for parameter estimation in the GLM. However, the actual implementation of maximum likelihood results in an algorithm based on iteratively reweighted least squares (IRLS). This is exactly what we saw previously for the special case of logistic regression.

Consider the method of maximum likelihood applied to the GLM, and suppose we use the canonical link. The log-likelihood function is

$$\ell(\mathbf{y}, \beta) = \sum_{i=1}^{n} [y_i \theta_i - b(\theta_i)] / a(\phi) + h(y_i, \phi)$$

For the canonical link, we have $\eta_i = g[E(y_i)] = g(\mu_i) = \mathbf{x}_i'\beta$; therefore,

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta}$$

$$= \frac{1}{a(\phi)} \sum_{i=1}^{n} \left[y_i - \frac{db(\theta_i)}{d\theta_i} \right] \mathbf{x}_i$$

$$= \frac{1}{a(\phi)} \sum_{i=1}^{n} (y_i - \mu_i) \mathbf{x}_i$$

Consequently, we can find the maximum likelihood estimates of the parameters by solving the system of equations

$$\frac{1}{a(\phi)} \sum_{i=1}^{n} (y_i - \mu_i) \mathbf{x}_i = 0$$

In most cases, $a(\phi)$ is a constant, so these equations become:

$$\sum_{i=1}^{n} (y_i - \mu_i) \mathbf{x}_i = 0$$

This is actually a *system* of $p = k + 1$ equations, one for each model parameter. In matrix form, these equations are

$$\mathbf{X}'(\mathbf{y} - \mu) = \mathbf{0}$$

where $\mu' = [\mu_1, \mu_2, \cdots, \mu_p]$. These are called the maximum likelihood score equations, and they are just the same equations that we saw previously in the case of logistic regression, where $\mu' = [n_1 \pi_1, n_2 \pi_2 \cdots, n_n \pi_n]$.

To solve the score equations, we can use IRLS, just as we did in the case of logistic regression. We start by finding a first-order Taylor series approximation in the neighborhood of the solution

$$y_i - \mu_i \approx \frac{d\mu_i}{d\eta_i} (\eta_i^* - \eta_i)$$

Now for a canonical link $\eta_i = \theta_i$, and

$$y_i - \mu_i \approx \frac{d\mu_i}{d\theta_i} (\eta_i^* - \eta_i) \qquad (4)$$

Therefore, we have

$$\eta_i^* - \eta_i \approx (y_i - \mu_i) \frac{d\theta_i}{d\mu_i}$$

This expression provides a basis for approximating the variance of $\hat{\eta}_i$.

In maximum likelihood estimation, we replace η_i by its estimate, $\hat{\eta}_i$. Then we have

$$V(\eta_i^* - \eta_i) \approx V\left[(y_i - \mu_i)\frac{d\theta_i}{d\mu_i}\right]$$

Since η_i^* and μ_i are constants,

$$V(\hat{\eta}_i) \approx \left[\frac{d\theta_i}{d\mu_i}\right]^2 V(y_i)$$

But

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{\text{var}(\mu_i)}$$

and $V(y_i) = \text{var}(\mu_i)a(\phi)$. Consequently,

$$V(\hat{\eta}_i) \approx \left[\frac{1}{\text{var}(\mu_i)}\right]^2 \text{var}(\mu_i)a(\phi)$$

$$\approx \frac{1}{\text{var}(\mu_i)}a(\phi)$$

For convenience, define $\text{var}(\eta_i) = [\text{var}(\mu_i)]^{-1}$, so we have

$$V(\hat{\eta}_i) \approx \text{var}(\eta_i)a(\phi).$$

Substituting this into Equation (4) above results in

$$y_i - \mu_i \approx \frac{1}{\text{var}(\eta_i)}(\eta_i^* - \eta) \tag{5}$$

If we let $\mathbf{V}$ be an $n \times n$ diagonal matrix whose diagonal elements are the $\text{var}(\eta_i)$, then in matrix form, Equation (5) becomes

$$\mathbf{y} - \mu \approx \mathbf{V}^{-1}(\eta^* - \eta)$$

We may then rewrite the score equations as follows:

$$\mathbf{X}'(\mathbf{y} - \mu) = \mathbf{0}$$
$$\mathbf{X}'\mathbf{V}^{-1}(\eta^* - \eta) = \mathbf{0}$$
$$\mathbf{X}'\mathbf{V}^{-1}(\eta^* - \mathbf{X}\beta) = \mathbf{0}$$

Thus, the maximum likelihood estimate of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\eta^*$$

Now just as we saw in the logistic regression situation, we do not know η^*, so we pursue an iterative scheme based on

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$$

Using iteratively reweighted least squares with the Newton-Raphson method, the solution is found from

$$\hat{\beta} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}z}$$

Asymptotically, the random component of $\mathbf{z}$ comes from the observations y_i. The diagonal elements of the matrix $\mathbf{V}$ are the variances of the z_i's, apart from $a(\phi)$.

As an example, consider the logistic regression case:

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\frac{d\eta_i}{d\mu_i} = \frac{d\eta_i}{d\pi_i} = \frac{d\ln\left(\frac{\pi_i}{1-\pi_i}\right)}{d\pi_i}$$

$$= \frac{1-\pi_i}{\pi_i}\left[\frac{\pi_i}{1-\pi_i} + \frac{\pi_i}{(1-\pi_i)^2}\right]$$

$$= \frac{(1-\pi_i)}{\pi_i(1-\pi_i)}\left[1 + \frac{\pi_i}{1-\pi_i}\right]$$

$$= \frac{1}{\pi_i}\left[\frac{1-\pi_i+\pi_i}{1-\pi_i}\right]$$

$$= \frac{1}{\pi_i(1-\pi_i)}$$

Thus, for logistic regression, the diagonal elements of the matrix $\mathbf{V}$ are

$$\left(\frac{d\eta_i}{d\mu_i}\right)^2 V(y_i) = \left[\frac{1}{\pi_i(1-\pi_i)}\right]^2\frac{\pi_i(1-\pi_i)}{n_i}$$

$$= \frac{1}{n_i\pi_i(1-\pi_i)}$$

which is exactly what we obtained previously.

Therefore, IRLS based on the Newton-Raphson method can be described as follows:

1. Use ordinary least squares to obtain an initial estimate of β, say $\hat{\beta}_0$;

2. Use $\hat{\beta}_0$ to estimate $\mathbf{V}$ and μ;

3. Let $\eta_0 = \mathbf{X}\hat{\beta}_0$;

4. Base $\mathbf{z}_1$ on η_0;

5. Obtain a new estimate $\hat{\beta}_1$, and iterate until some suitable convergence criterion is satisfied.

If $\hat{\beta}$ is the final value that the above algorithm produces and if the model assumptions, including the choice of the link function, are correct, then we can show that asymptotically

$$E(\hat{\beta}) = \beta \quad \text{and} \quad V(\hat{\beta}) = a(\phi)\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

If we don't use the canonical link, then $\eta_i \neq \theta_i$, and the appropriate derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \beta} = \frac{d\ell}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta}$$

Note that:

1. $\dfrac{d\ell}{d\theta_i} = \dfrac{1}{a(\phi)}\left[y_i - \dfrac{db(\theta_i)}{d\theta_i}\right] = \dfrac{1}{a(\phi)}(y_i - \mu_i)$

2. $\dfrac{d\theta_i}{d\mu_i} = \dfrac{1}{\text{var}(\mu_i)}$ and

3. $\dfrac{\partial \eta_i}{\partial \beta} = \mathbf{x}_i$

Putting this all together yields

$$\frac{\partial \ell}{\partial \beta} = \frac{y_i - \mu_i}{a(\phi)} \frac{1}{\text{var}(\mu_i)} \frac{d\mu_i}{d\eta_i} \mathbf{x}_i$$

Once again, we can use a Taylor series expansion to obtain

$$y_i - \mu_i \approx \frac{d\mu_i}{d\eta_i}(\eta_i^* - \eta_i)$$

Following an argument similar to that employed before,

$$V(\hat{\eta}_i) \approx \left[\frac{d\theta_i}{d\mu_i}\right]^2 V(y_i)$$

and eventually we can show that

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{\eta_i^* - \eta_i}{a(\phi) \, \text{var}(\eta_i)} \mathbf{x}_i$$

Equating this last expression to zero and writing it in matrix form, we obtain

$$\mathbf{X}' \mathbf{V}^{-1} (\eta^* - \eta) = \mathbf{0}$$

or, since $\eta = \mathbf{X}\beta$,

$$\mathbf{X}' \mathbf{V}^{-1} (\eta^* - \mathbf{X}\beta) = \mathbf{0}$$

The Newton-Raphson solution is based on

$$\hat{\beta} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{z}$$

where

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i) \frac{d\eta_i}{d\mu_i}$$

Just as in the case of the canonical link, the matrix $\mathbf{V}$ is a diagonal matrix formed from the variances of the estimated linear predictors, apart from $a(\phi)$.

Some important observations about the GLM:

1. Typically, when experimenters and data analysts use a transformation, they use ordinary least squares or OLS to actually fit the model in the transformed scale.

2. In a GLM, we recognize that the variance of the response is not constant, and we use weighted least squares as the basis of parameter estimation.

3. This suggests that a GLM should outperform standard analyses using transformations when a problem remains with constant variance after taking the transformation.

4. All of the inference we described previously on logistic regression carries over directly to the GLM. That is, model deviance can be used to test for overall model fit, and the difference in deviance between a full and a reduced model can be used to test hypotheses about subsets of parameters in the model. Wald inference can be applied to test hypotheses and construct confidence intervals about individual model parameters.

S15-3.9. Prediction and Estimation with the GLM

For any generalized linear model, the estimate of the mean response at some point of interest, say $\mathbf{x}_0$, is

$$\hat{y}_0 = \hat{\mu}_0 = g^{-1}(\mathbf{x}_0' \hat{\beta})$$

where g is the link function and it is understood that $\mathbf{x}_0$ may be expanded to "model form" if necessary to accommodate terms such as interactions that may have been included in the linear predictor. An approximate confidence interval on the mean response at this point can be computed as follows. The variance of the linear predictor

$\mathbf{x}_0'\hat{\beta}$ is $\mathbf{x}_0'\hat{\Sigma}\mathbf{x}_0$, where $\hat{\Sigma}$ is the estimated of the covariance matrix of $\hat{\beta}$. The $100(1-\alpha)\%$ confidence interval on the true mean response at the point $\mathbf{x}_0$ is

$$L \le \mu(\mathbf{x}_0) \le U$$

where

$$L = g^{-1}(\mathbf{x}_0'\hat{\beta} - Z_{\alpha/2}\mathbf{x}_0'\hat{\Sigma}\mathbf{x}_0) \text{ and } U = g^{-1}(\mathbf{x}_0'\hat{\beta} + Z_{\alpha/2}\mathbf{x}_0'\hat{\Sigma}\mathbf{x}_0)$$

This method is used to compute the confidence intervals on the mean response reported in SAS PROC GENMOD. This method for finding the confidence intervals usually works well in practice, because $\hat{\beta}$ is a maximum likelihood estimate, and therefore any function of $\hat{\beta}$ is also a maximum likelihood estimate. The above procedure simply constructs a confidence interval in the space defined by the linear predictor and then transforms that interval back to the original metric.

It is also possible to use Wald inference to derive approximate confidence intervals on the mean response. Refer to Myers and Montgomery (1997) for the details.

S15-3.10. Residual Analysis in the GLM

Just as in any model-fitting procedure, analysis of residuals is important in fitting the GLM. Residuals can provide guidance concerning the overall adequacy of the model, assist in verifying assumptions, and give an indication concerning the appropriateness of the selected link function.

The ordinary or **raw residuals** from the GLM are just the differences between the observations and the fitted values,

$$\begin{aligned}e_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\mu}_i\end{aligned}$$

It is generally recommended that residual analysis in the GLM be performed using **deviance residuals**. The ith deviance residual is defined as the square root of the contribution of the ith observation to the deviance, multiplied by the sign of the raw residual, or

$$r_{Di} = \sqrt{d_i}\, sign(y_i - \hat{y}_i)$$

where d_i is the contribution of the ith observation to the deviance. For the case of logistic regression (a GLM with binomial errors and the logit link), we can show that

$$d_i = y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\left[\frac{1-(y_i/n_i)}{1-\hat{\pi}_i}\right], i = 1,2,\cdots,n$$

where

$$\hat{\pi}_i = \frac{1}{1+e^{-\mathbf{x}_i'\hat{\beta}}}$$

Note that as the fit of the model to the data becomes better, we would find that $\hat{\pi}_i \cong y_i / n_i$, and the deviance residuals will become smaller, close to zero. For Poisson regression with a log link, we have

$$d_i = y_i \ln\left(\frac{y_i}{e^{\mathbf{x}_i'\hat{\beta}}}\right) - (y_i - e^{\mathbf{x}_i'\hat{\beta}}), i = 1, 2, \cdots, n$$

Once again, notice that as the observed value of the response y_i and the predicted value $\hat{y}_i = e^{\mathbf{x}_i'\hat{\beta}}$ become closer to each other, the deviance residuals approach zero.

Generally, deviance residuals behave much like ordinary residuals do in a standard normal theory linear regression model. Thus plotting the deviance residuals on a normal probability scale and versus fitted values are logical diagnostics. When plotting deviance residuals versus fitted values, it is customary to transform the fitted values to a constant information scale. Thus,

1. for normal responses, use $\hat{y}_i$

2. for binomial responses, use $2\sin^{-1}\sqrt{\hat{\pi}_i}$

3. for Poisson responses, use $2\sqrt{\hat{y}_i}$

4. for gamma responses, use $2\ln(\hat{y}_i)$

S15-4. Unbalanced Data in a Factorial Design

In this chapter we have discussed several approximate methods for analyzing a factorial experiment with unbalanced data. The approximate methods are often quite satisfactory, but as we observed, exact analysis procedure are available. These exact analyses often utilize the connection between ANOVA and regression. We have discussed this connection previously, and the reader may find it helpful to review Chapters 3 and 5, as well as the Supplemental Text Material for these chapters.

We will use a modified version of the battery life experiment of Example 5-1 to illustrate the analysis of data from an unbalanced factorial. Recall that there are three material types of interest (factor *A*) and three temperatures (factor *B*), and the response variable of interest is battery life. Table 2 presents the modified data. Notice that we have eliminated certain observations from the original experimental results; the smallest observed responses for material type 1 at each of the three temperatures, and one (randomly selected) observation from each of two other cells.

S15-4.1. The Regression Model Approach

One approach to the analysis simply formulates the ANOVA model as a regression model and uses the general regression significance test (or the "extra sum of squares method' to perform the analysis. This approach is easy to apply when the unbalanced design has *all cells filled*; that is, there is at least **one** observation in each cell.

Table 2. Modified Data from Example 5-1

Material types	Temperature		
	15	70	125
1	130,155, 180	40,80,75	70,82,58
2	150,188, 159,126	136,122, 106,115	25,70,45
3	138,110, 168,160	120,150, 139	96,104, 82,60

Recall that the regression model formulation of an ANOVA model uses indicator variables. We will define the indicator variables for the design factors material types and temperature as follows:

Material type	X_1	X_2
1	0	0
2	1	0
3	0	1

Temperature	X_3	X_4
15	0	0
70	1	0
125	0	1

The regression model is

$$
\begin{aligned}
y_{ijk} = \beta_0 &+ \beta_1 x_{ijk1} + \beta_2 x_{ijk2} + \beta_3 x_{ijk3} + \beta_4 x_{ijk4} \\
&+ \beta_5 x_{ijk1} x_{ijk3} + \beta_6 x_{ijk1} x_{ijk4} + \beta_7 x_{ijk2} x_{ijk3} + \beta_8 x_{ijk2} x_{ijk4} + \varepsilon_{ijk}
\end{aligned}
\tag{6}
$$

where $i, j = 1,2,3$ and the number of replicates $k = 1,2,\cdots,n_{ij}$, where n_{ij} is the number of replicates in the ijth cell. Notice that in our modified version of the battery life data, we have $n_{11} = n_{12} = n_{13} = n_{23} = n_{32} = 3$, and all other $n_{ij} = 4$.

In this regression model, the terms $\beta_1 x_{ijk1} + \beta_2 x_{ijk2}$ represent the main effect of factor A (material type), and the terms $\beta_3 x_{ijk3} + \beta_4 x_{ijk4}$ represent the main effect of temperature. Each of these two groups of terms contains two regression coefficients, giving two degrees of freedom. The terms $\beta_5 x_{ijk1} x_{ijk3} + \beta_6 x_{ijk1} x_{ijk4} + \beta_7 x_{ijk2} x_{ijk3} + \beta_8 x_{ijk2} x_{ijk4}$ represent the AB interaction with four degrees of freedom. Notice that there are four regression coefficients in this term.

Table 3 presents the data from this modified experiment in regression model form. In Table 3, we have shown the indicator variables for each of the 31 trials of this experiment.

Table 3. Modified Data from Example 5-1 in Regression Model Form

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
130	0	0	0	0	0	0	0	0
150	1	0	0	0	0	0	0	0
136	1	0	1	0	1	0	0	0
25	1	0	0	1	0	1	0	0
138	0	1	0	0	0	0	0	0
96	0	1	0	1	0	0	0	1
155	0	0	0	0	0	0	0	0
40	0	0	1	0	0	0	0	0
70	0	0	0	1	0	0	0	0
188	1	0	0	0	0	0	0	0
122	1	0	1	0	1	0	0	0
70	1	0	0	1	0	1	0	0
110	0	1	0	0	0	0	0	0
120	0	1	1	0	0	0	1	0
104	0	1	0	1	0	0	0	1
80	0	0	1	0	0	0	0	0
82	0	0	0	1	0	0	0	0
159	1	0	0	0	0	0	0	0
106	1	0	1	0	1	0	0	0
58	0	0	0	1	0	0	0	0
168	0	1	0	0	0	0	0	0
150	0	1	1	0	0	0	1	0
82	0	1	0	1	0	0	0	1
180	0	0	0	0	0	0	0	0
75	0	0	1	0	0	0	0	0
126	1	0	0	0	0	0	0	0
115	1	0	1	0	1	0	0	0
45	1	0	0	1	0	1	0	0
160	0	1	0	0	0	0	0	0
139	0	1	1	0	0	0	1	0
60	0	1	0	1	0	0	0	1

We will use this data to fit the regression model in Equation (6). We will find it convenient to refer to this model as the **full model**. The Minitab output is:

```
Regression Analysis
The regression equation is
Y = 155 + 0.7 X1 - 11.0 X2 - 90.0 X3 - 85.0 X4 + 54.0 X5 - 24.1 X6
    + 82.3 X7 + 26.5 X8

Predictor          Coef         StDev            T          P
Constant         155.00         12.03        12.88      0.000
X1                 0.75         15.92         0.05      0.963
X2               -11.00         15.92        -0.69      0.497
X3               -90.00         17.01        -5.29      0.000
X4               -85.00         17.01        -5.00      0.000
X5                54.00         22.51         2.40      0.025
X6               -24.08         23.30        -1.03      0.313
X7                82.33         23.30         3.53      0.002
X8                26.50         22.51         1.18      0.252

S = 20.84       R-Sq = 83.1%      R-Sq(adj) = 76.9%

Analysis of Variance

Source           DF            SS           MS          F          P
Regression        8       46814.0       5851.8      13.48      0.000
Residual Error   22        9553.8        434.3
Total            30       56367.9
```

We begin by testing the hypotheses associated with interaction. Specifically, in terms of the regression model in Equation (6), we wish to test

$$H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_1: \text{ at least one } \beta_j \neq 0, j = 5,6,7,8$$

(7)

We may test this hypothesis by using the general regression significance test or "extra sum of squares" method. If the null hypothesis of no-interaction is true, then the **reduced model** is

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk1} + \beta_2 x_{ijk2} + \beta_3 x_{ijk3} + \beta_4 x_{ijk4} + \varepsilon_{ijk}$$

(8)

Using Minitab to fit the reduced model produces the following:

```
Regression Analysis
The regression equation is
Y = 138 + 12.5 X1 + 23.9 X2 - 41.9 X3 - 82.1 X4

Predictor          Coef         StDev            T          P
Constant         138.02         11.02        12.53      0.000
X1                12.53         11.89         1.05      0.302
X2                23.92         11.89         2.01      0.055
X3               -41.91         11.56        -3.62      0.001
X4               -82.14         11.56        -7.10      0.000
```

```
S = 26.43        R-Sq = 67.8%      R-Sq(adj) = 62.8%

Analysis of Variance

Source            DF          SS          MS          F          P
Regression         4      38212.5      9553.1      13.68      0.000
Residual Error    26      18155.3       698.3
Total             30      56367.9
```

Now the regression or model sum of squares for the full model, which includes the interaction terms, is $SS_{Model}(FM) = 46,814.0$ and for the reduced model [Equation (8)] it is $SS_{Model}(RM) = 38,212.5$. Therefore, the increase in the model sum of squares due to the interaction terms (or the extra sum of squares due to interaction) is

$$SS_{Model}(\text{Interaction}|\text{main effects}) = SS_{Model}(FM) - SS_{Model}(RM)$$
$$= 46,814.0 - 38,212.5$$
$$= 8601.5$$

Since there are 4 degrees of freedom for interaction, the appropriate test statistic for the no-interaction hypotheses in Equation (7) is

$$F_0 = \frac{SS_{Model}(\text{Interaction}|\text{main effects})/4}{MS_E(FM)}$$
$$= \frac{8601.5/4}{434.3}$$
$$= 4.95$$

The *P*-value for this statistic is approximately 0.0045, so there is evidence of interaction.

Now suppose that we wish to test for a material type effect. In terms of the regression model in Equation (6), the hypotheses are

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: \beta_1 \text{ and / or } \beta_2 \neq 0$$

(9)

and the reduced model is

$$y_{ijk} = \beta_0 + \beta_3 x_{ijk3} + \beta_4 x_{ijk4}$$
$$+ \beta_5 x_{ijk1} x_{ijk3} + \beta_6 x_{ijk1} x_{ijk4} + \beta_7 x_{ijk2} x_{ijk3} + \beta_8 x_{ijk2} x_{ijk4} + \varepsilon_{ijk}$$

(10)

Fitting this model produces the following:

```
Regression Analysis
The regression equation is
Y = 151 - 86.3 X3 - 81.3 X4 + 54.8 X5 - 23.3 X6 + 71.3 X7 + 15.5 X8

Predictor       Coef        StDev          T          P
Constant      151.273       6.120       24.72      0.000
X3            -86.27        13.22       -6.53      0.000
X4            -81.27        13.22       -6.15      0.000
X5             54.75        15.50        3.53      0.002
X6            -23.33        16.57       -1.41      0.172
X7             71.33        16.57        4.30      0.000
X8             15.50        15.50        1.00      0.327

S = 20.30      R-Sq = 82.5%      R-Sq(adj) = 78.1%

Analysis of Variance

Source             DF          SS          MS          F          P
Regression          6      46480.6      7746.8      18.80      0.000
Residual Error     24       9887.3       412.0
Total              30      56367.9
```

Therefore, the sum of squares for testing the material types main effect is

$$SS_{Model}(\text{Material types}) = SS_{Model}(FM) - SS_{Model}(RM)$$
$$= 46,814.0 - 46,480.6$$
$$= 333.4$$

The *F*-statistic is

$$F_0 = \frac{SS_{Model}(\text{Material types})/2}{MS_E(FM)}$$
$$= \frac{333.4/2}{434.3}$$
$$= 0.38$$

which is not significant. The hypotheses for the main effect of temperature is

$$H_0: \beta_3 = \beta_4 = 0$$
$$H_1: \beta_3 \text{ and } / \text{ or } \beta_4 \neq 0$$

(11)

and the reduced model is

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk1} + \beta_2 x_{ijk2}$$
$$+ \beta_5 x_{ijk1} x_{ijk3} + \beta_6 x_{ijk1} x_{ijk4} + \beta_7 x_{ijk2} x_{ijk3} + \beta_8 x_{ijk2} x_{ijk4} + \varepsilon_{ijk}$$

(12)

Fitting this model produces:

```
Regression Analysis
The regression equation is
Y = 96.7 + 59.1 X1 + 47.3 X2 - 36.0 X5 - 109 X6 - 7.7 X7 - 58.5 X8

Predictor          Coef        StDev            T          P
Constant          96.67        10.74         9.00      0.000
X1                59.08        19.36         3.05      0.005
X2                47.33        19.36         2.45      0.022
X5               -36.00        22.78        -1.58      0.127
X6              -109.08        24.60        -4.43      0.000
X7                -7.67        24.60        -0.31      0.758
X8               -58.50        22.78        -2.57      0.017

S = 32.21      R-Sq = 55.8%      R-Sq(adj) = 44.8%

Analysis of Variance

Source              DF           SS           MS            F          P
Regression           6        31464         5244         5.05      0.002
Residual Error      24        24904         1038
Total               30        56368
```

Therefore, the sum of squares for testing the temperature main effect is

$$SS_{Model}(\text{Temperature}) = SS_{Model}(FM) - SS_{Model}(RM)$$
$$= 46,814.0 - 31,464.0$$
$$= 15,350.0$$

The F-statistic is

$$F_0 = \frac{SS_{Model}(\text{Temperature})/2}{MS_E(FM)}$$
$$= \frac{15,350.0/2}{434.3}$$
$$= 17.67$$

The P-value for this statistic is less than 0.0001. Therefore, we would conclude that the main effect of temperature has an effect on battery life. Since both the main effect of temperature and the materials type-temperature interaction are significant, we would likely reach the same conclusions for this data that we did from the original balanced-data factorial in the textbook.

S15-4.2. The Type 3 Analysis

Another approach to the analysis of an unbalanced factorial is to directly employ the **Type 3 analysis** procedure discussed previously. Many computer software packages will directly perform the Type 3 analysis, calculating Type 3 sums of squares or "adjusted" sums of squares for each model effect. The Minitab **General Linear Model** procedure will directly perform the Type 3 analysis. Remember that this procedure is only appropriate when there are no empty cells (i.e., $n_{ij} > 0$, for all i, j).

Output from the Minitab General Linear Model routine for the unbalanced version of Example 5-1 in Table 3 follows:

General Linear Model

Factor   Type   Levels Values

Mat   fixed   3   1 2 3

Temp   fixed   3   15 70 125

Analysis of Variance for Life, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Mat	2	2910.4	**3202.4**	1601.2	3.69	0.042
Temp	2	35302.1	**36588.7**	18294.3	42.13	0.000
Mat*Temp	4	8601.5	**8601.5**	2150.4	4.95	0.005
Error	22	9553.8	**9553.8**	434.3		
Total	30	56367.9				

The "Adjusted" sums of squares, shown in boldface type in the above computer output, are the Type 3 sums of squares. The *F*-tests are performed using the Type 3 sums of squares in the numerator. The hypotheses that are being tested by a type 3 sum of squares is essentially equivalent to the hypothesis that would be tested for that effect if the data were balanced. Notice that the error or residual sum of squares and the interaction sum of squares in the Type 3 analysis are identical to the corresponding sums of squares generated in the regression-model formulation discussed above.

When the experiment is unbalanced, but there is at least one observation in each cell, the Type 3 analysis is generally considered to be the correct or "standard" analysis. A good reference is Freund, Littell and Spector (1988). Various SAS/STAT users' guides and manuals are also helpful.

S15-4.3. Type 1, Type 2, Type 3 and Type 4 Sums of Squares

At this point, a short digression on the various types of sums of squares reported by some software packages and their uses is warranted. Many software systems report Type 1 and Type 3 sums of squares; the SAS software system reports *four* types, called (originally enough!!) Types 1, 2, 3 and 4. For an excellent detailed discussion of this topic, see the technical report by Driscoll and Borror (1999).

As noted previously, Type 1 sums of squares refer to a sequential or "effects-added-in-order" decomposition of the overall regression or model sum of squares. In sequencing the factors, interactions should be entered only after all of the corresponding main effects, and nested factors should be entered in the order of their nesting.

Type 2 sums of squares reflect the contribution of a particular effect to the model after all other effects have been added, except those that contain the particular effect in question. For example, an interaction contains the corresponding main effects. For unbalanced data, the hypotheses tested by Type 2 sums of squares contain, in addition to the parameters of interest, the cell counts (i.e., the n_{ij}). These are not the same hypotheses that would be tested by the Type 2 sums of squares if the data were balanced, and so most analysts have concluded that other definitions or types of sums of squares are necessary. In a regression model (i.e., one that is **not overspecified**, as in the case of an ANOVA model), Type 2 sums of squares are perfectly satisfactory, so many regression programs (such as SAS PROC REG) report Type 1 and Type 2 sums of squares.

Type 3 and Type 4 sums of squares are often called partial sums of squares. For balanced experimental design data, Types 1, 2, 3, and 4 sums of squares are identical. However, in unbalanced data, differences can occur, and it is to this topic that we now turn.

To make the discussion specific, we consider the two-factor fixed-effects factorial model. For proportional data, we will find that for the main effects the relationships between the various types of sums of squares is Type 1 = Type 2, and Type 3 = Type 4, while for the interaction it is Type 1 = Type 2 = Type 3 = Type 4. Thus the choice is between Types 1 and 4. If the cell sample sizes are representative of the population from which the treatments were selected, then an analysis based on the Type 1 sums of squares is appropriate. This, in effect, makes the factor levels have important that is proportional to the sample sizes. If this is not the case, then the Type 3 analysis is appropriate.

With unbalanced data having at least one observation in each cell, we find that for the main effects that Types 1 and 2 will generally not be the same for factor *A*, but Type 1 = Type 2 for factor *B*. This is a consequence of the order of specification in the model. For both main effects, Type 3 = Type 4. For the interaction, Type 1 = Type 2 = Type 3 = Type 4. Generally, we prefer the Type 3 sums of squares for hypothesis testing in these cases.

If there are empty cells, then *none* of the four types will be equal for factor *A*, while Type 1 = Type 2 for factor *B*. For the interaction, Type 1 = Type 2 = Type 3 = Type 4. In general, the Type 4 sums of squares should be used for hypothesis testing in this case, but it is not always obvious exactly *what* hypothesis is being tested. When cells are empty, certain model parameters will not exist and this will have a significant impact on which functions of the model parameters are estimable. Recall that only estimable functions can be used to form null hypotheses. Thus, when we have missing cells the exact nature of the hypotheses being tested is actually a function of which cells are missing. There is a process in SAS PROC GLM where the estimable functions can be determined, and the specific form of the null hypothesis involving fixed effects determined for any of the four types of sum of squares. The procedure is described in Driscoll and Borror (1999).

S15-4.4. Analysis of Unbalanced Data using the Means Model

Another approach to the analysis of unbalanced data that often proves very useful is to abandon the familiar effects model, say

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,b \\ k = 1,2,\cdots,n_{ij} \end{cases}$$

and employ instead the means model

$$y_{jik} = \mu_{ij} + \varepsilon_{ijk} \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,b \\ k = 1,2,\cdots,n_{ij} \end{cases}$$

where of course $\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}$. This is a particularly useful approach when there are empty cells; that is, $n_{ij} = 0$ for some combinations of i and j. When the ijth cell is empty, this means that the treatment combination τ_i and β_j is not observed.

Sometimes this happens by design and sometimes it is the result of chance. The analysis employing the means model is often quite simple, since the means model can be thought of as a **single-factor model** with $ab - m$ treatments, where m is the number of empty cells. That is, each factor level or treatment in this one-way model is actually a *treatment combination* from the original factorial.

To illustrate, consider the experiment shown in Table 4. This is a further variation of the battery life experiment (first introduced in text Example 5-1), but now in addition to the missing observations in cells (1,1), (1,2), (1,3), (2,3) and (3,2), the (3,3) cell is empty. In effect, the third material was never exposed to the highest temperature, so we have no information on those treatment combinations.

Table 4. Modified Data from Example 5-1 with an Empty Cell

Material types	Temperature		
	15	70	125
1	130,155, 180	40,80,75	70,82,58
2	150,188, 159,126	136,122, 106,115	25,70,45
3	138,110, 168,160	120,150, 139	

It is easy to analyze the data of Table 4 as a single-factor experiment with $ab - m = (3)(3) - 1 = 8$ treatment combinations. The Minitab one-way analysis of variance output follows. In this output, the factor levels are denoted $m_{11}, m_{12}, \cdots, m_{23}$.

One-way Analysis of Variance

```
Analysis of Variance for BattLife
Source      DF        SS        MS        F        P
Cell         7     43843      6263    14.10    0.000
Error       19      8439       444
Total       26     52282
         Individual Confidence Intervals Based on Pooled Std Dev.
Level       N      Mean     StDev  ------+---------+---------+--------+
m11         3    155.00     25.00                          (----*----)
m12         3     65.00     21.79       (----*----)
m13         3     70.00     12.00         (----*----)
m21         4    155.75     25.62                            (---*----)
m22         4    119.75     12.66                   (---*---)
m23         3     46.67     22.55   (----*----)
m31         4    144.00     25.97                        (----*---)
m32         3    136.33     15.18                     (----*----)
                                   ------+---------+---------+--------+
Pooled StDev =    21.07               50       100       150       200
Fisher's pairwise comparisons
Family error rate = 0.453
Individual error rate = 0.0500


Critical value = 2.093


Confidence Intervals for (column level mean) - (row level mean)

                 m11        m12        m13        m21        m22        m23

    m12        53.98
              126.02

    m13        48.98     -41.02
              121.02      31.02

    m21       -34.44    -124.44    -119.44
               32.94     -57.06     -52.06

    m22         1.56     -88.44     -83.44       4.81
               68.94     -21.06     -16.06      67.19

    m23        72.32     -17.68     -12.68      75.39      39.39
              144.35      54.35      59.35     142.77     106.77

    m31       -22.69    -112.69    -107.69     -19.44     -55.44    -131.02
               44.69     -45.31     -40.31      42.94       6.94     -63.64

    m32       -17.35    -107.35    -102.35     -14.27     -50.27    -125.68
               54.68     -35.32     -30.32      53.11      17.11     -53.65


                 m31

    m32       -26.02
               41.36
```

First examine the F-statistic in the analysis of variance. Since $F = 14.10$ and the P-value is small, we would conclude that there are significant differences in the treatment means. We also used Fisher's LSD procedure in Minitab to test for differences in the individual treatment means. There are significant differences between seven pairs of means:

$$\mu_{11} \neq \mu_{12}, \mu_{11} \neq \mu_{13}, \mu_{11} \neq \mu_{22}, \mu_{11} \neq \mu_{23}$$
$$\mu_{21} \neq \mu_{22}, \mu_{21} \neq \mu_{23}, \text{ and } \mu_{22} \neq \mu_{23}$$

Furthermore, the confidence intervals in the Minitab output indicate that the longest lives are associated with material types 1,2 and 3 at low temperature and material types 2 and 3 at the middle temperature level.

Generally, the next step is to form and comparisons of interest (contrasts) in the cell means. For example, suppose that we are interested in testing for interaction in the data. If we had data in all 9 cells there would be 4 degrees of freedom for interaction. However, since one cell is missing, there are only 3 degrees of freedom for interaction. Practically speaking, this means that there are only three linearly independent *contrasts* that can tell us something about interaction in the battery life data. One way to write these contrasts is as follows:

$$C_1 = \mu_{11} - \mu_{13} - \mu_{21} + \mu_{23}$$
$$C_2 = \mu_{21} - \mu_{22} - \mu_{31} + \mu_{32}$$
$$C_3 = \mu_{11} - \mu_{12} - \mu_{31} + \mu_{32}$$

Therefore, some information about interaction is found from testing

$$H_0 : C_1 = 0, H_0 : C_2 = 0, \text{ and } H_0 : C_3 = 0$$

Actually there is a way to *simultaneously* test that all three contrasts are equal to zero, but it requires knowledge of linear models beyond the scope of this text, so we are going to perform t-tests. That is, we are going to test

$$H_0 : \mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} = 0$$
$$H_0 : \mu_{21} - \mu_{22} - \mu_{31} + \mu_{32} = 0$$
$$H_0 : \mu_{11} - \mu_{12} - \mu_{31} + \mu_{32} = 0$$

Consider the first null hypothesis. We estimate the contrast by replacing the cell means by the corresponding cell averages. This results in

$$\hat{C}_1 = \bar{y}_{11.} - \bar{y}_{13.} - \bar{y}_{21.} + \bar{y}_{32.}$$
$$= 155.00 - 70.00 - 155.75 + 46.67$$
$$= -24.08$$

The variance of this contrast is

$$V(\hat{C}_1) = V(\bar{y}_{11.} - \bar{y}_{13.} - \bar{y}_{21.} + \bar{y}_{32.})$$

$$= \sigma^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{13}} + \frac{1}{n_{21}} + \frac{1}{n_{32}} \right)$$

$$= \sigma^2 \left(\frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{3} \right)$$

$$= \sigma^2 \left(\frac{5}{4} \right)$$

From the Minitab ANOVA, we have $MS_E = 444$ as the estimate of σ^2, so the t-statistic associated with the first contrast C_1 is

$$t_0 = \frac{\hat{C}_1}{\sqrt{\hat{\sigma}^2 (5/4)}}$$

$$= \frac{-24.08}{\sqrt{(444)(5/4)}}$$

$$= -1.02$$

which is not significant. It is easy to show that the t-statistics for the other two contrasts are for C_2

$$t_0 = \frac{\hat{C}_2}{\sqrt{\hat{\sigma}^2 (13/12)}}$$

$$= \frac{28.33}{\sqrt{(444)(13/12)}}$$

$$= 1.29$$

and for C_3

$$t_0 = \frac{\hat{C}_3}{\sqrt{\hat{\sigma}^2 (5/4)}}$$

$$= \frac{82.33}{\sqrt{(444)(5/4)}}$$

$$= 3.49$$

Only the t-statistic for C_3 is significant ($P = 0.0012$). However, we would conclude that there is some indication between material types and temperature.

Notice that our conclusions are similar to those for the balanced data in Chapter 5. There is little difference in materials at low temperature, but at the middle level of temperature only materials types 2 and 3 have the same performance – material type 1 has significantly lower life. There is also some indication of interaction, implying that not all materials perform similarly at different temperatures. In the original experiment we had information about the effect of all three materials at high temperature, but here we do not. All we can say is that there is no difference between material types 1 and 2 at high

temperature, and that both materials provide significantly reduced life performance at the high temperature than they do at the middle and low levels of temperature.

S15-5. Computer Experiments

There has been some interest in recent years in applying statistical design techniques to **computer experiments**. A computer experiment is just an experiment using a computer program that is a model of some system. There are two types of computer models that are usually encountered. The first of these is where the response variable or output from the computer model is a random variable. This often occurs when the computer model is a Monte Carlo or computer simulation model. These models are used extensively in many areas, including machine scheduling, traffic flow analysis, and factory planning. When the output of a computer model is a random variable, often we can use the methods and techniques described in the book with little modification. The response surface approach has been shown to be quite useful here. What we are doing then, is to create a model of a model. This is often called a **metamodel.**

In some computer simulation models the output is observed over *time*, so the output response of interest is actually a *time series*. Many books on computer simulation discuss the analysis of simulation output. Several specialized analysis techniques have been developed.

The other type of computer model is a **deterministic** computer model. That is, the output response has no random component, and if the model is run several times at exactly the same settings for the input variables, the response variable observed is the same on each run. Deterministic computer models occur often in engineering as the result of using finite element analysis models, computer-based design tools for electrical circuits, and specialized modeling languages for specific types of systems (such as Aspen for modeling chemical processes).

The design and analysis of deterministic computer experiments is different in some respects from the usual types of experiments we have studied. First, statistical inference (tests and confidence intervals) isn't appropriate because the observed response isn't a random variable. That is, the system model is

$$y = f(x_1, x_2, \cdots, x_k)$$

and **not**

$$y = f(x_1, x_2, \cdots, x_k) + \varepsilon$$

where ε is the usual random error component. Often the experimenter want to find a model that passes very near (or even exactly through!) each sample point generated, and the sample points cover a very broad range of the inputs. In other words, the possibility of fitting an empirical model (low-order polynomial) that works well in a *region of interest* is ignored. Many types of fitting functions have been suggested. Barton (1992) gives a nice review.

If a complex metamodel is to be fit, then the design must usually have a fairly large number of points, and the designs dominated by boundary points that we typically use with low-order polynomial are not going to be satisfactory. **Space-filling designs** are often suggested as appropriate designs for deterministic computer models. A **Latin hypercube design** is an example of a space-filling design. In a Latin hypercube design, the range of each factor is divided into n equal-probability subdivisions. Then an experimental design is created by randomly matching each of the factors. One way to perform the matching is to randomly order or shuffle each of the n divisions of each factor and then take the resulting order for each factor. This ensures that each factor is sampled over its range. An example for two variables and $n = 16$ is shown below.

Latin Hypercube Design
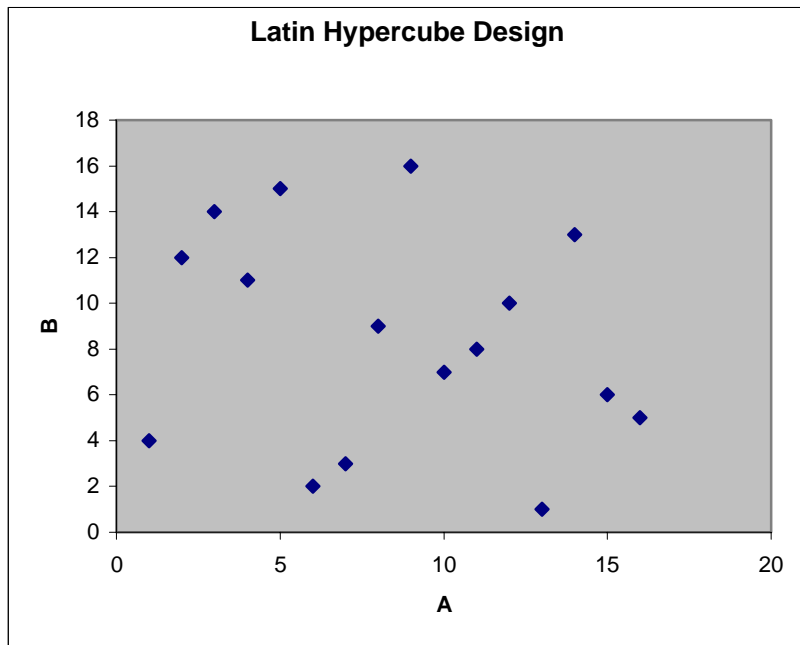
The design points for this Latin hypercube are shown in the Table 5. For more information on computer experiments and Latin hypercube designs, see Donohue (1994), McKay, Beckman and Conover (1979), Welch and Yu (1990), Morris (1991), Sacks, Welch and Mitchell (1989), Stein, M. L. (1987), Owen (1994) and Pebesma and Heuvelink (1999).

Table 5. A Latin Hypercube Design

A	B
8	9
11	8
9	16

13	1
16	5
6	2
12	10
14	13
5	15
4	11
7	3
1	4
10	7
15	6
2	12
3	14

Supplemental References

Barton, R. R. (1992). "Metamodels for Simulation Input-Output Relations", *Proceedings of the Winter Simulation Conference*, pp. 289-299.

Donohue, J. M. (1994). "Experimental Designs for Simulation", *Proceedings of the Winter Simulation Conference*, pp. 200-206.

Driscoll, M. F. and Borror, C. M. (1999). *Sums of Squares and Expected Mean Squares in SAS*, Technical Report, Department of Industrial Engineering, Arizona State University, Tempe AZ.

Freund, R. J., Littell, R. C., and Spector, P. C. (1988). *The SAS System for Linear Models*, SAS Institute, Inc., Cary, NC.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code", *Technometrics*, Vol. 21, pp. 239-245.

Morris, M. D. (1991). "Factorial Sampling Plans for Preliminary Computer Experiments", *Technometrics*, Vol. 33, pp. 161-174.

Owen, A. B. (1994), "Controlling Correlations in Latin Hypercube Sampling", *Journal of the American Statistical Association*, Vol. 89, pp. 1517-1522

Pebesma, E. J. and Heuvelink, G. B. M. (1999), "Latin Hypercube Sampling of Gaussian Random Fields", *Technometrics*, Vol. 41, pp. 303-312.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). "Design and Analysis of Computer Experiments", *Statistical Science*, Vol. 4, pp. 409-435.

Stein, M. L. (1987), Large Sample Properties of Simulations using Latin Hypercube Sampling", *Technometrics*, Vol. 29, pp. 1430151.

Welch, W. J and Yu, T. K. (1990). "Computer Experiments for Quality Control by Parameter Design" *Journal of Quality Technology*, Vol. 22, pp. 15-22.