Unbalanced Analysis of Variance, Design, and Regression
Ronald Christensen, fletcher@stat.unm.edu
Copyright 2013

Chapter 21

Logistic Regression: Predicting Counts

For the most part, this book concerns itself with measurement data and the corresponding analyses based on normal distributions. In this chapter and the next we consider data that consist of counts. Elementary count data were introduced in Chapter 5.

Frequently data are collected on whether or not a certain event occurs. A mouse dies when exposed to a dose of chloracetic acid or it does not. In the past, O-rings failed during a space shuttle launch or they did not. Men have coronary incidents or they do not. These are modeled as random events and we collect data on how often the event occurs. We also collect data on potential predictor (explanatory) variables. For example, we use the size of dose to estimate the probability that a mouse will die when exposed. We use the atmospheric temperature at launch time to estimate the probability that O-rings fail. We may use weight, cholesterol, and blood pressure to estimate the probability that men have coronary incidents. Once we have estimated the probability that these events will occur, we are ready to make predictions. In this chapter we investigate the use of logistic models to estimate probabilities. Logistic models (also known as logit models) are linear models for the log-odds that an event will occur. For a more complete discussion of logistic and logit models, see Christensen (1997).

Section 1 introduces models for predicting count data. Section 2 presents a simple model with one predictor variable where the data are the proportions of trials that display the event. It also discusses the output one typically obtains from running a logistic regression program. Section 3 discusses how to perform model tests with count data. Section 4 discusses how logistic models are fitted. Section 5 introduces the important special case in which each observation is a separate trial that either displays the event or does not. Section 6 explores the use of multiple continuous predictors. Section 7 examines ANOVA type models with Section 8 examining ACOVA type models.

21.1 Models for Binomial Data

Logistic regression is a method of modeling the relationships between probabilities and predictor variables. We begin with an example.

EXAMPLE 21.1.1. Woodward et al. (1941) reported data on 120 mice divided into 12 groups of 10. The mice in each group were exposed to a specific dose of chloracetic acid and the observations consist of the number in each group that lived and died. Doses were measured in grams of acid per kilogram of body weight. The data are given in Table 21.1, along with the proportions y_h of mice who died at each dose x_h.

We could analyze these data using the methods discussed earlier in Chapter 5. We have samples from twelve populations. We could test to see if the populations are the same. We don't think they are because we think survival depends on dose. More importantly though, we want to try to model the relationship between dose level and the probability of dying. This allows us to make predictions about the probability of dying for any dose level that is similar to those in the original data. □

In Section 3.1 we talked about models for measurement data y_h, $h = 1, \dots, n$ with $\mathrm{E}(y_h) \equiv \mu_h$

Table 21.1: *Lethality of chloracetic acid*

Dose (x_h)	Group (h)	Died	Survived	Total	Proportion (y_h)
.0794	1	1	9	10	.1
.1000	2	2	8	10	.2
.1259	3	1	9	10	.1
.1413	4	0	10	10	.0
.1500	5	1	9	10	.1
.1588	6	2	8	10	.2
.1778	7	4	6	10	.4
.1995	8	6	4	10	.6
.2239	9	4	6	10	.4
.2512	10	5	5	10	.5
.2818	11	5	5	10	.5
.3162	12	8	2	10	.8

and $\text{Var}(y_h) = \sigma^2$. For testing models, we eventually assumed

$$y_h\text{s} \quad \text{independent} \quad N(\mu_h, \sigma^2),$$

with some model for the μ_hs. In Section 3.9 we got more specific about models, writing

$$y_h\text{s} \quad \text{independent} \quad N[m(x_h), \sigma^2],$$

where x_h is the value of some predictor variable or vector and $m(\cdot)$ is the model for the means, i.e.,

$$\mu_h \equiv m(x_h).$$

We then discussed a variety of models $m(\cdot)$ that could be used for various types of predictor variables and exploited those models in subsequent chapters.

In this chapter, we discuss similar models for data that are binomial proportions. In Section 1.4 we discussed binomial sampling. In particular, if we have N independent trials of whether some event occurs (e.g., flipping a coin and seeing heads) and if each trial has the same probability p that the event occurs, then the number of occurrences is a binomial random variable W, say

$$W \quad \sim \quad \text{Bin}(N, p),$$

with

$$\text{E}(W) = Np \qquad \text{and} \qquad \text{Var}(W) = Np(1-p).$$

We will be interested in binomial proportions

$$y \equiv \frac{W}{N},$$

with

$$\text{E}(y) = p$$

and

$$\text{Var}(y) = \frac{p(1-p)}{N},$$

see Proposition 1.2.11. In applications, N is known and p is an unknown parameter to be modeled and estimated.

In general, we assume n independent binomial proportions y_h for which we know the number of trials N_h, i.e.,

$$N_h y_h \quad \text{independent} \quad \text{Bin}(N_h, p_h), \qquad h = 1, \ldots, n.$$

With $\mathrm{E}(y_h) = p_h$, much like we did for measurement data, we want to create a model for the p_hs that depends on a predictor x_h. In fact, we would like to use the same models, simple linear regression, multiple regression, one-way ANOVA and multifactor ANOVA, that we used for measurement data. But before we can do that, we need to deal with a problem.

We want to create models for $p_h = \mathrm{E}(y_h)$, but with binomial proportions this mean value is always a probability and probabilities are required to be between 0 and 1. If we wrote a simple linear regression model such as $p_h = \beta_0 + \beta_1 x_h$ for some predictor variable x, nothing forces the probabilities to be between 0 and 1. When modeling probabilities, it seems reasonable to ask that they be between 0 and 1.

Rather than modeling the probabilities directly, we model a function of the probabilities that is not restricted between 0 and 1. In particular, we model the log of the odds, rather than the actual probabilities. The odds O_h are defined to be the probability that the event occurs, divided by the probability that it does not occur, thus

$$O_h \equiv \frac{p_h}{1 - p_h}.$$

Probabilities must be between 0 and 1, so the odds can take any values between 0 and $+\infty$. Taking the log of the odds permits any values between $-\infty$ and $+\infty$, so we consider models

$$\log\left(\frac{p_h}{1-p_h}\right) = m(x_h), \tag{21.1.1}$$

where $m(\cdot)$ is any of the models that we considered earlier.

Two different names have been used for such models. If $m(x_h)$ corresponds to a one-sample, two-sample, one-way ANOVA, or multifactor ANOVA, these models have often been called *logit models*. The name stems from using the transformation

$$\eta = f(p) \equiv \log\left(\frac{p}{1-p}\right),$$

which is known as the *logit transform*. It maps the unit interval into the real line. On the other hand, if the model $m(x_h)$ corresponds to any sort of regression model, models like (1) are called *logistic regression* models. These models are named after the *logistic transform*, which is the inverse of the logit transform,

$$p = g(\eta) \equiv \frac{e^\eta}{1 + e^\eta}.$$

The functions are inverses in the sense that $g(f(p)) = p$ and $f(g(\eta)) = \eta$. To perform any worthwhile data analysis requires using both the logit transform and the logistic transform, so it really does not matter what you call the models. These days, any model of the form (1) is often called logistic regression, regardless of whether $m(x_h)$ corresponds to a regression model.

In Chapter 3, to perform tests and construct confidence intervals, we assumed that the y_h observations were independent, with a common variance σ^2, and normally distributed. In this chapter, to perform tests and construct confidence intervals similar to those used earlier, we need to rely on having large amounts of data. That can happen in two different ways. The best way is to have the N_h values large for every value of h. In the chloracetic acid data, each N_h is 10, which is probably large enough. Unfortunately, this best way to have the data may be the least common way of actually obtaining data. The other and more common way to get a lot of data is to have the number of proportions n reasonably large but the N_hs possibly small. Frequently, the N_hs all equal 1. When worrying about O-ring failure, each shuttle launch is a separate trial, $N_h = 1$, but we have $n = 23$ launches to examine. When examining coronary incidents, each man is a separate trial, $N_h = 1$, but we have $n = 200$ men to examine. In other words, if the N_hs are all large, we don't really care if n is large or not. If the N_hs are not all large, we need n to be large. A key point is that n needs to be large relative to the number of parameters we fit in our model. For the O-ring data, we will only fit two parameters, so $n = 23$ is probably reasonable. For the coronary incident data, we have many more predictors so we need many more subjects. In fact, we will need to resist the temptation to fit too many parameters to the data.

21.2 Simple Logistic Regression

In simple logistic regression we use a single measurement variable to predictor probabilities.

EXAMPLE 21.2.1. In Example 21.1.1 and Table 21.1 we presented the data of Woodward et al. (1941) on the slaughter of mice. These data are extremely well behaved in that they all have the same reasonably large number of trials $N_h = 10$, $h = 1, \ldots, 12$, and there is only one measurement predictor variable, the dose x_h.

A simple linear logistic regression model has

$$\log\left(\frac{p_h}{1-p_h}\right) = \beta_0 + \beta_1 x_h, \tag{21.2.1}$$

so our model fits a straight line in dose to the log-odds. Alternatively,

$$p_h = \frac{e^{\beta_0+\beta_1 x_h}}{1+e^{\beta_0+\beta_1 x_h}}.$$

Indeed, for an arbitrary dose x we can write

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{21.2.2}$$

Standard computer output involves a table of coefficients:

Table of Coefficients: Model (21.2.1)

Predictor	*Est*	SE	*t*	*P*
Constant	−3.56974	0.705330	−5.06	0.000
Dose	14.6369	3.33248	4.39	0.000

The validity of everything but the point estimates relies on having large amounts of data. Using the point estimates gives the *linear predictor*

$$\hat{\eta}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = -3.56974 + 14.6369x.$$

Applying the logistic transformation to the linear predictor gives the estimated probability for any x,

$$\hat{p}(x) = \frac{e^{\hat{\eta}(x)}}{1+e^{\hat{\eta}(x)}}.$$

This function is plotted in Figure 21.1. The approximate model is unlikely to fit well outside the range of the x_h values that actually occurred in Table 21.1.

The table of coefficients is used exactly like previous tables of coefficients, e.g., $\hat{\beta}_1 = 14.64$ is the estimated slope parameter and $\text{SE}(\hat{\beta}_1) = 3.326$ is its standard error. The t values are simply the estimates divided by their standard errors, so they provide statistics for testing whether the regression coefficient equals 0. The P values are based on large sample normal approximations, i.e., the t statistics are compared to a $t(\infty)$ distribution. Clearly, there is a significant effect for fitting the dose, so we reject the hypothesis that $\beta_1 = 0$. The dose helps explain the data.

Many computer programs expand the table of coefficients to include odds ratios, defined as $\xi_k \equiv e^{\beta_k}$, and a confidence interval for the odds ratio. The $(1-\alpha)$ confidence interval for ξ_k is typically found by exponentiating the limits of the confidence interval for β_k, i.e., it is (e^{L_k}, e^{U_k})

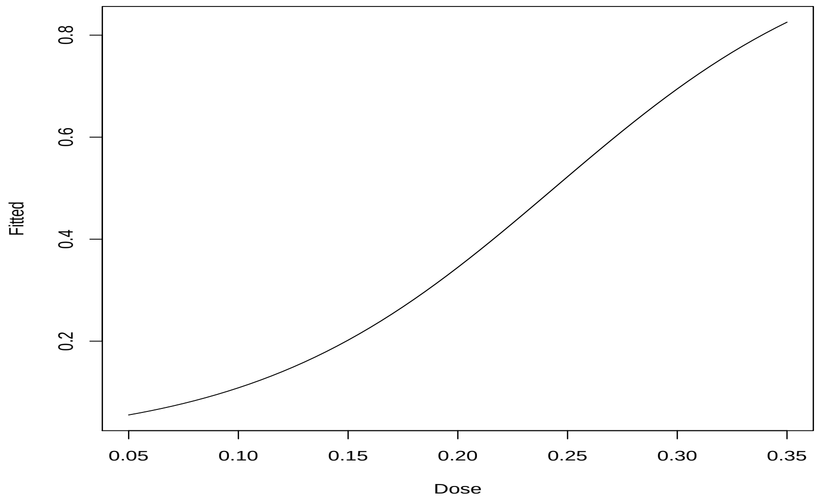

Figure 21.1: *Fitted probabilities as a function of dose.*

where $L_k \equiv \hat{\beta}_k - t(1-\alpha/2,\infty)\mathrm{SE}(\hat{\beta}_k)$ and $U_k \equiv \hat{\beta}_k + t(1-\alpha/2,\infty)\mathrm{SE}(\hat{\beta}_k)$ provide the $(1-\alpha)100\%$ confidence limits for β_k.

Additional standard output includes the Log-Likelihood $= -63.945$ (explained in Section 4) and a model based χ^2 test for $\beta_1 = 0$ that is explained in Section 3. The model based test for $\beta_1 = 0$ has $G^2 = 23.450$ with $df = 1$ and a P value of 0.000, obtained by comparing 23.450 to a $\chi^2(1)$ distribution. This test provides substantial evidence that O-ring failure is related to temperature. □

21.2.1 Goodness-of-Fit Tests

Computer programs written specifically for logistic regression frequently report goodness-of-fit tests. If a valid goodness-of-fit test is rejected, it suggests that the fitted model is wrong. Typical output is

Goodness-of-Fit Tests

Method	Chi-Square	df	P
Pearson	8.7421	10	0.557
Deviance	10.2537	10	0.419
Hosmer-Lemeshow	6.7203	4	0.151

There are a couple of problems with this output. First of all, as Hosmer, Lemeshow, and colleagues established in Hosmer et al. (1997), their χ^2 test isn't worth the toner it takes to print it. It amazes me that so many programs persist in computing it. (It is not a bad idea, but there was never any reason to think the statistic had a χ^2 distribution.) Second, the reported deviance is often problematic in many specialized programs for doing logistic regression. The deviance is well-defined for these particular data, because all the N_hs are large but as we will see later, specialized programs for logistic regression frequently pool cases together to increase the size of the N_hs, which can destroy the usefulness of the numbers reported as the deviance.

If the fitted model is correct and all of the N_h values are large, the Pearson and Deviance statistics should have χ^2 distributions with df degrees of freedom. So the question becomes, "Do $X^2 = 8.7421$ and $G^2 = 10.254$ look like they could reasonably come from a $\chi^2(10)$ distribution?" To answer that question, check whether the P values are small. Alternatively, we could compare the test statistics to values from a table of percentiles for the $\chi^2(10)$ distribution, see Appendix B.2.

However, since the mean of a $\chi^2(df)$ distribution is df, our values of 8.7421 and 10.254 are very close to the mean of the distribution which is 10, so it is pretty obvious that the data are consistent with the simple logistic regression model even if we did not have the P values given to us. The Pearson and Deviance statistics are computed as in Chapter 5 for the 12×2 table of 12 rows (dose groups) and 2 columns (Died and Survived) except to make the computations one must define the observed counts as $O_{h1} = N_h y_h$, $O_{h2} = N_h(1 - y_h)$ and define the (estimated) expected counts as $\hat{E}_{h1} = N_h \hat{p}_h$, and $\hat{E}_{h2} = N_h(1 - \hat{p}_h)$. The 10 degrees of freedom are the number of rows $n = 12$ minus the number of parameters we fit in the model, $p = 2$.

The reason that the deviance and Pearson tests work as advertised is because the fitted regression model provides reasonable estimates of the probabilities for each case, i.e., for $h = 1, \ldots, n$ model (21.2.1) provides good estimates of the *linear predictor*

$$\hat{\eta}_h \equiv \hat{\beta}_0 + \hat{\beta}_1 x_h$$

and

$$\hat{p}_h \equiv \frac{e^{\hat{\eta}_h}}{1 + e^{\hat{\eta}_h}},$$

but in addition the values y_h from Table 21.1 provide reasonable estimates for the twelve death probabilities without fitting any obvious model. The problem with the Pearson and Deviance goodness-of-fit tests is that when some or all of the N_hs are small, the y_hs no longer provide good estimates of the case probabilities, so the $\chi^2(df)$ is no longer an appropriate reference distribution for the Pearson and Deviance statistics.

As we will see in Section 5, in an attempt to get valid goodness-of-fit tests, many computer programs for logistic regression will redefine the N_hs to make them larger (and n smaller). They do this by pooling together any cases that have exactly the same predictor variables x. With continuous predictors I have never seen this pooling procedure get the N_hs large enough to validate a χ^2 distribution but it certainly is possible. Although the deviance G^2 may or may not provide a valid goodness of fit test, ideally the deviance is extremely useful for constructing model tests. Unfortunately, different models with different predictors typically have different poolings, which destroys the usefulness of the deviance as a tool for comparing models. When using logistic regression programs, one must compare models by constructing the likelihood ratio test statistic from the reported log-likelihoods. It is also possible to fit logistic regression models by using programs for fitting *generalized linear models*. ("Generalized linear models" are something distinct from "general linear models.") Generalized linear model programs rarely indulge in the pooling silliness that logistic regression programs often display, so their reported deviance values can be used to compare models.

21.2.2 Assessing Predictive Ability

We can measure the predictive ability of the model through R^2, which is the squared correlation between the y_h values and the $\hat{p}_h$ values. For these data $R^2 = 0.759$, which is quite high for a logistic regression. The high value is related to the fact that we have 10 observations in each binomial proportion. We are evaluating the model on its ability to predict the outcome of 10 trials, not the outcome of predicting one trial.

Frequently with *dose-response* data like the Woodward data, one uses the log dose as a predictor, i.e., the model becomes

$$\log\left(\frac{p_h}{1 - p_h}\right) = \beta_0 + \beta_1 \log(x_h).$$

For these data we get $R^2 = 0.760$ based on the log dose, which indicates that log dose is not much of an improvement over dose.

Note that the predictive ability of the model depends a great deal on where the predictor variables are located. At $x = -\beta_0/\beta_1$, from equation (21.2.2) the probability is 0.5. Nobody can predict well a

Table 21.2: *Diagnostics for rat data.*

Group	y_h	$\hat{p}_h$	r_h	$\tilde{r}_h$	Leverage	Cook
1	0.1	0.082597	0.19992	0.21824	0.160889	
2	0.2	0.108510	0.93020	1.01250	0.155954	
3	0.1	0.150978	−0.45026	−0.48593	0.141430	
4	0.0	0.182195	−1.49260	−1.60009	0.129843	
5	0.1	0.201942	−0.80301	−0.85751	0.123079	
6	0.2	0.223498	−0.17837	−0.18978	0.116595	
7	0.4	0.275420	0.88188	0.93280	0.106197	
8	0.6	0.343063	1.71151	1.81021	0.106081	
9	0.4	0.427383	−0.17504	−0.18754	0.128851	
10	0.5	0.526738	−0.16935	−0.18769	0.185894	
11	0.5	0.635281	−0.88874	−1.04399	0.275295	
12	0.8	0.742394	0.41655	0.52474	0.369843	

50:50 outcome like a coin toss. As x gets further from $-\beta_0/\beta_1$, $p(x)$ gets further from 0.5, so closer to 0 or 1. When the probability is close to 0 or 1, predicting the outcome is easy. Thus the predictive ability of the model depends on the spread of x away from $-\beta_0/\beta_1$. For these data $-\beta_0/\beta_1$ is called the LD_{50} which denotes the *lethal dose 50* and is defined to be the dose at which lethality is 50%. In other contexts this number might be called the *effective dose 50* denoted ED_{50}.

Many programs for fitting logistic regression report other values that can be used to assess the predictive ability of the model. Typical output includes:

```
                    Measures of Association
   Between the Response Variable and Predicted Probabilities
Pairs        Number   Percent || Summary Measures
Concordant     2326      73.6 || Somers' D                0.53
Discordant      636      20.1 || Goodman-Kruskal Gamma    0.57
Ties            197       6.2 || Kendall's Tau-a          0.24
Total          3159     100.0 ||
```

EXPLAIN

21.2.3 Case Diagnostics

Diagnostic quantities can be computed that are similar to those for standard regression. Raw residuals, $y_h - \hat{p}_h$, are not of much interest. The *Pearson residuals* are just the observations minus their estimated probability divided by the standard error of the observation, i.e.,

$$r_h = \frac{y_h - \hat{p}_h}{\sqrt{\hat{p}_h(1-\hat{p}_h)/N_h}}.$$

This SE does not really account for the process of fitting the model, i.e., estimating p_h. We can incorporate the fitting process by incorporating the leverage, say, a_h. A *standardized Pearson residual* is

$$\tilde{r}_h = \frac{y_h - \hat{p}_h}{\sqrt{\hat{p}_h(1-\hat{p}_h)(1-a_h)/N_h}}.$$

Leverages for logistic regression are similar in spirit to those discussed in Chapter 11, but rather more complicated to compute. Values near 1 are still high leverage points and the $2r/n$ and $3r/n$ rules of thumb can be applied where r is the number of (functionally distinct) parameters in the model. Table 21.2 contains diagnostics for the rat data. Nothing seems overly disturbing.

I prefer using the standardized Pearson residuals, but the Pearson residuals often get used because of their simplicity. When all N_hs are large, both residuals can be compared to a $N(0,1)$ distribution to asses whether they are consistent with the model and the other data. In this large N_h

case, much like the spirit of Chapter 5, we use the residuals to identify cases that cause problems in the goodness-of-fit test. Even with small N_hs, where no valid goodness-of-fit test is present, the residuals are used to identify potential problems.

With measurement data, residuals are used to check for outliers in the dependent variable, i.e., values of the dependent variable that do not seem to belong with the rest of the data. With count data it is uncommon to get anything that is really an outlier in the counts. The y_hs are proportions, so outliers would be values that are not between 0 and 1. With count data, large residuals really highlight areas where the model is not fitting the data very well. If you have a high dose of poison but very few rats die, something is wrong. The problem is often something that we have left out of the model.

21.2.4 Computer Commands

Minitab has programs for fitting logistic regression but none for generalized linear models.

SAS contains a powerful program for logistic regression, PROC LOGISTIC, but it pools cases for goodness-of-fit tests and diagnostics. The first line below controls printing. The next four lines involve defining and reading the data. The remaining lines specify the model and that a logistic regression is to be performed.

```
options ps=60 ls=72 nodate;
data mice;
   infile ’tab21-1.dat’;
   input x Ny N y;
proc logistic data=mice descending;
   model y=x ;
run;
```

Alternatively, to get usable deviance values from SAS, we can use the generalized linear model program PROC GENMOD. To perform logistic regression one must specify an appropriate link and distribution. One also specifies the number of deaths for each case (Ny) divided by the number of trials for each case (N).

```
options ps=60 ls=72 nodate;
data mice;
   infile ’tab21-1.dat’;
   input x Ny N y;
proc genmod data=mice ;
   model Ny/N = x / link=logit dist=binomial;
run;
```

21.3 Model Testing

Based on the results of a valid goodness-of-fit test, we already have reason to believe that a simple linear logistic regression fits the chloracetic acid data reasonably well, but for the purpose of illustrating the procedure for testing models, we will test the simple logistic model against a cubic polynomial logistic model. This section demonstrates the test. In the next section we discuss the motivation for it.

In Section 21.2 we gave the table of coefficients and the table of goodness-of-fit tests for the simple logistic regression model

$$\log\left(\frac{p_h}{1-p_h}\right) = \beta_0 + \beta_1 x_h. \tag{21.3.1}$$

The table of coefficients along with the deviance information follows.

Table of Coefficients: Model (21.3.1)

Predictor	*Est*	SE	*t*	*P*
Constant	-3.56974	0.705330	-5.06	0.000
Dose	14.6369	3.33248	4.39	0.000

Deviance: $G^2 = 10.254$ $df = 10$

Additional standard output includes the Log-Likelihood $= -63.945$ (explained in Section 4) and a model based test for $\beta_1 = 0$ that is also discussed in Section 4, for which the test statistic is $G^2 = 23.450$ with $df = 1$ and a P value of 0.000.

The cubic polynomial logistic regression is

$$\log\left(\frac{p_h}{1-p_h}\right) = \gamma_0 + \gamma_1 x_h + \gamma_2 x_h^2 + \gamma_3 x_h^3. \tag{21.3.2}$$

with table of coefficients

Table of Coefficients: Model (21.3.2)

Predictor	$\hat{\gamma}_k$	SE($\hat{\gamma}_k$)	*t*	*P*
Constant	-2.47396	4.99096	-0.50	0.620
dose	-5.76314	83.1709	-0.07	0.945
x^2	114.558	434.717	0.26	0.792
x^3	-196.844	714.422	-0.28	0.783

and goodness-of-fit tests

Goodness-of-Fit Tests: Model (21.3.2)

Method	Chi-Square	*df*	*P*
Pearson	8.7367	8	0.365
Deviance	10.1700	8	0.253
Hosmer-Lemeshow	6.3389	4	0.175

Additional standard output includes the Log-Likelihood $= -63.903$ and a model based test that all slopes are zero, i.e., $0 = \gamma_1 = \gamma_2 = \gamma_3$, that has $G^2 = 23.534$ with $df = 3$, and a P value of 0.000.

To test the full cubic model against the reduced simple linear model we compute the likelihood ratio test statistic from the log-likelihoods,

$$G^2 = -2[(-63.903) - (-63.945)] = 0.084.$$

There are 4 parameters in model (21.3.2) and only 2 parameters in model (21.3.1) so there are $4 - 2 = 2$ degrees of freedom associated with this test. When the total number of cases n is large compared to the number of parameters in the full model, we can compare $G^2 = 0.084$ to a $\chi^2(4-2)$ distribution. This provides no evidence that the cubic model fits better than the simple linear model. Note that the validity of this test does not depend on having the N_hs large.

For these data, we can also obtain G^2 by the difference in deviances reported for the two models,

$$G^2 = 10.254 - 10.1700 = 0.084.$$

The difference in the deviance degrees of freedom is $10 - 8 = 2$ which is also the correct degrees of freedom.

Although finding likelihood ratio tests by subtracting deviances and deviance degrees of freedom is our preferred computational tool, unfortunately, *subtracting the deviances and the deviance degrees of freedom cannot be trusted to give the correct* G^2 *and degrees of freedom* when using programs designed for fitting logistic models (as opposed to programs for fitting generalized linear models). As discussed in Section 5, many logistic regression programs pool cases with identical predictor variables prior to computing the deviance and when models use different predictors, the pooling often changes, which screws up the test. Subtracting the deviances and deviance degrees of freedom does typically give the correct result when using programs for generalized linear models.

The standard output for model (21.3.1) also included a model based test for $\beta_1 = 0$ with $G^2 = 23.450$, $df = 1$, and a P value of 0.000. This is the likelihood ratio test for comparing the full model (21.3.1) with the intercept only model

$$\log\left(\frac{p_h}{1-p_h}\right) = \delta_0. \tag{21.3.3}$$

Alas, many logistic regression programs do not like to fit model (21.3.3), so we take the program's word for the result of the test. (Programs for generalized linear models are more willing to fit model (21.3.3).) Finding the test statistic is discussed in Section 5.

The usual output for fitting model (21.3.2) has a model based test that all slopes are zero, i.e., that $0 = \gamma_1 = \gamma_2 = \gamma_3$, for which $G^2 = 23.534$ with $df = 3$ and a P value of 0.000. This is the likelihood ratio test for the full model (21.3.2) against the reduced model (21.3.3). Generally, when fitting a model these additional G^2 tests are for comparing the current model to the intercept only model (21.3.3).

21.4 Fitting Logistic Models

In this section we discuss the ideas behind our methods for estimating parameters and for testing models. First we define the likelihood function. Our point estimates are *maximum likelihood estimates (MLEs)* which are the parameter values that maximize the likelihood function. We compare models by comparing the maximum value that the likelihood function achieves under each model. Such tests are *(generalized) likelihood ratio tests* for binomial count data. While we did not present the likelihood function for normal data, least squares estimates are also MLEs and F tests are also equivalent to (generalized) likelihood ratio tests.

Our logistic models take the form

$$\log\left(\frac{p_h}{1-p_h}\right) = m(x_h), \tag{21.4.1}$$

where x_h is a vector of measurement or categorical variables and $m(\cdot)$ is any of the models that we have considered earlier for such predictor variables. The model $m(x_h)$ can correspond to a one-sample, two-sample, one-way ANOVA, or multifactor ANOVA model or any sort of regression model. We can solve (1) for p_h by writing

$$p_h = \frac{e^{m(x_h)}}{1+e^{m(x_h)}}. \tag{21.4.2}$$

Given the estimate $\hat{m}(x)$ we get

$$\hat{p}(x) = \frac{e^{\hat{m}(x)}}{1+e^{\hat{m}(x)}}.$$

For example, given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for a simple linear logistic regression, we get

$$\hat{p}(x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}. \tag{21.4.3}$$

In particular, this formula provides the $\hat{p}_h$s when doing predictions at the x_h's.

Estimates of coefficients are found by maximizing the likelihood function. The likelihood function is the probability of getting the data that were actually observed. It is a function of the unknown model parameters contained in $m(\cdot)$. Because the $N_h y_h$s are independent binomials, the likelihood function is

$$L(p_1,\ldots,p_n) = \prod_{h=1}^{n} \binom{N_h}{N_h y_h} p_h^{N_h y_h}(1-p_h)^{N_h - N_h y_h}. \tag{21.4.4}$$

For a particular proportion y_h, $N_h y_y$ is $\text{Bin}(N_h, p_h)$ and the probability from Section 1.4 is an individual term on the right. We multiply the individual terms because the $N_h y_h$s are independent.

If we substitute for the p_h's using (21.4.2) into the likelihood function (21.4.4), the likelihood becomes a function of the model parameters. For example, if $m(x_h) = \beta_0 + \beta_1 x_h$ the likelihood becomes a function of the model parameters β_0 and β_1 for known values of (x_h, y_h, N_h), $h = 1, \ldots, n$. Computer programs maximize this function of β_0 and β_1 to give maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ along with approximate standard errors. The estimates have approximate normal distributions for large sample sizes. For the large sample approximations to be valid, it is typically enough that the total number of trials in the entire data n be large relative to the number of model parameters; the individual sample sizes N_h need not be large. The normal approximations also hold if all the N_hs are large regardless of the size of n.

In Section 11.3 we found the least squares estimates for linear regression models. Although we did not explicitly give the likelihood function for regression models with normally distributed data, we mentioned that the least squares estimates were also maximum likelihood estimates. Unfortunately, for logistic regression there are no closed form solutions for the estimates and standard errors like those presented for measurement data in Chapter 11. For logistic regression, different computer programs may give *slightly* different results because the computations are more complex.

Maximum likelihood theory also provides a (generalized) likelihood ratio (LR) test for a full model versus a reduced model. Suppose the full model is

$$\log\left(\frac{p_{Fh}}{1-p_{Fh}}\right) = m_F(x_h).$$

Fitting the model leads to estimated probabilities $\hat{p}_{Fh}$. The reduced model must be a special case of the full model, say,

$$\log\left(\frac{p_{Rh}}{1-p_{Rh}}\right) = m_R(x_h),$$

with fitted probabilities $\hat{p}_{Rh}$. The commonly used form of the likelihood ratio test statistic is,

$$\begin{aligned} G^2 &= -2\log\left(\frac{L(\hat{p}_{R1}, \ldots, \hat{p}_{Rn})}{L(\hat{p}_{F1}, \ldots, \hat{p}_{Fn})}\right) \\ &= 2\sum_{h=1}^{n}\left[N_h y_h \log(\hat{p}_{Fh}/\hat{p}_{Rh}) + (N_h - N_h y_h)\log((1-\hat{p}_{Fh})/(1-\hat{p}_{Rh})\right], \end{aligned}$$

where the second equality is based on equation (4). An alternative to the LR test statistic is the Pearson test statistic which is

$$X^2 = \sum_{h=1}^{n} \frac{(N_h\hat{p}_{Fh} - N_h\hat{p}_{Rn})^2}{N_h\hat{p}_{Rn}(1-\hat{p}_{Rn})} = \sum_{h=1}^{n}\left[\frac{\hat{p}_{Fh} - \hat{p}_{Rn}}{\sqrt{\hat{p}_{Rn}(1-\hat{p}_{Rn})/N_h}}\right]^2.$$

We make minimal use of X^2 in our discussions.

If the reduced model is true and the sample size n is large relative to the number of parameters in the full model, G^2 and X^2 have asymptotic χ^2 distributions where the degrees of freedom is the difference in the number of (functionally distinct) parameters between the two models. The same χ^2 distribution holds even if n is not large when the N_hs are all large.

Many computer programs for fitting a model report the value of the log-likelihood,

$$\ell(\hat{p}_1, \ldots, \hat{p}_n) \equiv \log\left[L(\hat{p}_1, \ldots, \hat{p}_n)\right].$$

To compare a full and reduced model, G^2 is twice the absolute value of the difference between these values. When using logistic regression programs (as opposed to generalized linear model programs) this is how one needs to compute G^2.

The smallest interesting logistic model that we can fit to the data is the *intercept only model*

$$\log\left(\frac{p_h}{1-p_h}\right) = \beta_0. \tag{21.4.5}$$

The largest logistic model that we can fit to the data is the *saturated model* that has a separate parameter for each case,

$$\log\left(\frac{p_h}{1-p_h}\right) = \gamma_h. \tag{21.4.6}$$

Interesting models tend to be somewhere between these two. Many computer programs automatically report the results of testing the fitted model against both of these.

Testing a fitted model $m(\cdot)$ against the saturated model (21.4.6) is called a *goodness-of-fit test*. The fitted probabilities under model (21.4.6) are just the observed proportions for each case, the y_hs. The *deviance* for a fitted model is defined as G^2 for testing the fitted model against the saturated model (21.4.6),

$$\begin{aligned} G^2 &= -2\log\left(\frac{L(\hat{p}_1,\ldots,\hat{p}_n)}{L(y_1,\ldots,y_n)}\right) \\ &= 2\sum_{h=1}^{n}\left[N_h y_h \log(y_h/\hat{p}_h) + (N_h - N_h y_h)\log((1-y_h)/(1-\hat{p}_h))\right]. \end{aligned}$$

In this formula, if $y_h = 0$, then $y_h \log(y_h)$ is taken as zero. The degrees of freedom for the deviance are n (the number of parameters in model (21.4.6)) minus the number of (functionally distinct) parameters in the fitted model.

The problem with the goodness-of-fit test is that the number of parameters in model (21.4.6) is the sample size n, so the only way for G^2 to have an asymptotic χ^2 distribution is if all the N_hs are large. For the rat death data, the N_hs are all 10, which is probably fine, but for a great many data sets, all the N_hs are 1, so a χ^2 test of the goodness-of-fit statistic is not appropriate. A similar conclusion holds for the Pearson statistic.

As also discussed in the next section, in an effort to increase the size of the N_hs, many logistic regression computer programs pool together any cases for which $x_h = x_i$. Thus, instead of having two cases with $N_h y_h \sim \text{Bin}(N_h, p_h)$ and $N_i y_i \sim \text{Bin}(N_i, p_i)$, the two cases get pooled into a single case with $(N_h y_h + N_i y_i) \sim \text{Bin}(N_h + N_i, p_h)$. Note that if $x_h = x_i$, it follows that $p_h = p_i$ and the new proportion would be $(N_h y_h + N_i y_i)/(N_h + N_i)$. I have never encountered regression data with so few distinct x_h values that this pooling procedure actually accomplished its purpose of making all the group sizes reasonably large. Although if the mice data were presented as 120 mice that either died or not along with their dose, such pooling would work fine.

Ideally, the deviance G^2 could be used analogously to the *SSE* in normal theory and the degrees of freedom for the deviance would be analogous to the dfE. To compare a full and reduced model you just subtract their deviances (rather than their *SSE*s) and compare the test statistic to a χ^2 with degrees of freedom equal to the difference in the deviance degrees of freedom (rather than differencing the dfEs). This procedure works just fine when fitting the models using programs for fitting generalized linear models. The invidious thing about the pooling procedure of the previous paragraph is that when you change the model from reduced to full, you often change the predictor vector x_h in such a way that it changes which cases have $x_h = x_i$. When comparing a full and a reduced model, the models may well have different cases pooled together, which means that the difference in deviances no longer provide the appropriate G^2 for testing the models. In such cases G^2 needs to be computed directly from the log-likelihood.

After discussing the commonly reported goodness-of-fit statistics in the next section, we will no longer discuss any deviance values that are obtained by pooling. After Subsection 21.5.1, the deviances we discuss may not be those reported by a logistic regression program but they should be those obtained by a generalized linear models program.

Table 21.3: *O-ring Failure Data*

Case	Flight	Failure	Temperature	Case	Flight	Failure	Temperature
1	14	1	53	13	2	1	70
2	9	1	57	14	11	1	70
3	23	1	58	15	6	0	72
4	10	1	63	16	7	0	73
5	1	0	66	17	16	0	75
6	5	0	67	18	21	1	75
7	13	0	67	19	19	0	76
8	15	0	67	20	22	0	76
9	4	0	68	21	12	0	78
10	3	0	69	22	20	0	79
11	8	0	70	23	18	0	81
12	17	0	70				

21.5 Binary data

Logistic regression is often used when the binomial sample sizes are all 1. The resulting binary data consist entirely of 0s and 1s.

EXAMPLE 21.5.2. *O-ring Data.*
Table 21.3 presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-*Challenger* space shuttle launches. *Challenger* was the shuttle that blew up on take-off. Atmospheric temperature is the predictor variable. The *Challenger* explosion occurred during a takeoff at 31 degrees Fahrenheit. Each flight is viewed as an independent trial. The result of a trial is 1 if any field O-rings failed on the flight and 0 if all the O-rings functioned properly. A simple logistic regression uses temperature to model the probability that any O-ring failed. Such a model allows us to predict O-ring failure from temperature.

Let p_i be the probability that any O-ring fails in case i. The simple linear logistic regression model is

$$\text{logit}(p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i,$$

where x_i is the known temperature and β_0 and β_1 are unknown intercept and slope parameters (coefficients).

Maximum likelihood theory gives the coefficient estimates, standard errors, and t values as

Table of Coefficients: O-rings

Predictor	*Est*	SE	t	P
Constant	15.0429	7.37862	2.04	0.041
Temperature	−0.232163	0.108236	−2.14	0.032

The t values are the estimate divided by the standard error. For testing $H_0 : \beta_1 = 0$, the value $t = -2.14$ yields a P value that is approximately .03, so there is evidence that temperature does help predict O-ring failure. Alternatively, a model based test of $\beta_1 = 0$ compares the simple linear logistic model to an intercept only model and gives $G^2 = 7.952$ with $df = 1$ and $P = 0.005$. These tests should be reasonably valid because $n = 23$ is reasonably large relative to the 2 parameters in the fitted model. The log-likelihood is $\ell = -10.158$.

Figure 21.2 gives a plot of the estimated probabilities as a function of temperature,

$$\hat{p}(x) = \frac{e^{15.0429-0.232163x}}{1+e^{15.0429-0.232163x}}.$$

The *Challenger* was launched at $x = 31$ degrees, so the predicted log odds are $15.04 - .2321(31) = 7.8449$ and the predicted probability of an O-ring failure is $e^{7.8449}/(1+e^{7.8449}) = .9996$. Actually,

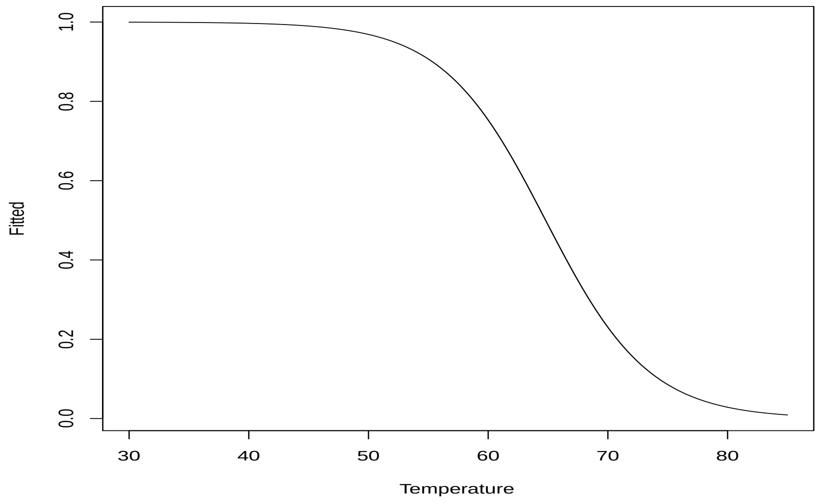

Figure 21.2: *O-ring Failure Probabilities.*

there are problems with this prediction because we are predicting very far from the observed data. The lowest temperature at which a shuttle had previously been launched was 53 degrees, very far from 31 degrees. According to the fitted model, a launch at 53 degrees has probability .939 of O-ring failure, so even with the caveat about predicting beyond the range of the data, the model indicates an overwhelming probability of failure.

21.5.1 Goodness-of-Fit Tests

Many specialized logistic regression computer programs report the following goodness-of-fit statistics for the O-ring data.

Goodness-of-Fit Tests

Method	Chi-Square	df	P
Pearson	11.1303	14	0.676
Deviance	11.9974	14	0.607
Hosmer-Lemeshow	9.7119	8	0.286

For 0-1 data, these are all useless. The Hosmer-Lemeshow statistic does not have a χ^2 distribution. For computing the Pearson and Deviance statistics the 23 original cases have been pooled into $\tilde{n} = 16$ new cases based on duplicate temperatures. This gives binomial sample sizes of $\tilde{N}_6 = 3$, $\tilde{N}_9 = 4$,$\tilde{N}_{12} = \tilde{N}_{13} = 2$, and $\tilde{N}_h = 1$ for all other cases. With two parameters in the fitted model, the reported degrees of freedom are $14 = 16 - 2$. To have a valid $\chi^2(14)$ test, all the $\tilde{N}_h$s would need to be large, but none of them are. Pooling does not give a valid χ^2 test and it also eliminates the deviance as a useful tool in model testing.

Henceforward, we only report deviances that are not obtained by pooling. These are the likelihood ratio test statistics for the fitted model against the saturated model with the corresponding degrees of freedom. Test statistics for any full and reduced models can then be obtained by subtracting the corresponding deviances from each other just as the degrees of freedom for the test can be obtained by subtraction. These deviances can generally be found by fitting logistic models as special cases in programs for fitting generalized linear models. When using specialized logistic regression software, great care must be taken and the safest bet is to always use log-likelihoods to obtain test statistics.

EXAMPLE 21.5.1 CONTINUED. For the simple linear logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i. \tag{21.5.1}$$

Without pooling the deviance is $G^2 = 20.315$ with $21 = 23-2 = n-2$ degrees of freedom. For the intercept only model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 \tag{21.5.2}$$

the deviance is $G^2 = 28.267$ with $22 = 23-1 = n-1$ degrees of freedom. Since $N_i = 1$ for all i, neither of these G^2's is compared directly to a chi-squared distribution. However, the model based test for $H_0 : \beta_1 = 0$ has $G^2 = 28.267 - 20.315 = 7.952$ on $df = 22-21 = 1$ which agrees with the test reported earlier even though the deviance for model (21.5.1) is different from that reported earlier. Comparing $G^2 = 7.952$ to a $\chi^2(1)$ distribution, the P value for the test is approximately .005. It is considerably smaller than the P value for the t test of H_0. □

It can be difficult to get even generalized linear model programs to fit the intercept only model but the deviance G^2 can be obtained from the formula in Section 4. Given the estimate $\hat{\beta}_0$ for model (21.5.2), we get $\hat{p}_i = e^{\hat{\beta}_0}/(1+e^{\hat{\beta}_0})$ for all i, and apply the formula. In general, for the intercept only model $\hat{p}_i = \sum_{i=1}^n N_i y_i / \sum_{i=1}^n N_i$ which, for binary data, reduces to $\hat{p}_i = \sum_{i=1}^n y_i/n$. The degrees of freedom are the number of cases minus the number of fitted parameters, $n-1$.

21.5.2 Case Diagnostics

The residuals and leverages in Table 21.4 have also been computed using pooling which is why some values are missing. Cases 6, 7, 8, cases 11, 12, 13, 14, cases 17, 18, and cases 19, 20 all have duplicated temperatures with residuals and leverages reported only for the first case. Without pooling the reported leverage for case 6, 0.22, would otherwise be distributed as 0.22/3 for each of cases 6, 7, and 8.

21.5.3 Assessing Predictive Ability

$$R^2 = .346$$

```
Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs        Number  Percent  Summary Measures
Concordant       85     75.9  Somers' D              0.56
Discordant       22     19.6  Goodman-Kruskal Gamma  0.59
Ties              5      4.5  Kendall's Tau-a        0.25
Total           112    100.0
```

21.5.4 Computer Commands

The data are in a file TAB21-3.DAT that contains six columns similar to Table 21.3. ID indicates the case, flt is the flight, f indicates failure of any O-rings, and temp is temperature. The fourth column s is one minus f. The last column in the file contains the actual number of O-rings that failed on a flight.

The following commands are for the generalized linear model procedure in SAS, genmod.

Table 21.4: *Diagnostics for* Challenger *data.*

Case	y_h	$\hat{p}_h$	r_h	$\tilde{r}_h$	Leverage	Cook
1	1	0.939248	0.25433	0.27874	0.167481	
2	1	0.859317	0.40462	0.45459	0.207773	
3	1	0.828845	0.45442	0.51093	0.208982	
4	1	0.602681	0.81194	0.87704	0.142942	
5	0	0.430493	-0.86943	-0.90961	0.086393	
6	0	0.374724	-1.34085	-1.52147	0.223334	
7	0	0.374724	*	*	*	
8	0	0.374724	*	*	*	
9	0	0.322094	-0.68930	-0.71359	0.066923	
10	0	0.273621	-0.61375	-0.63417	0.063342	
11	0	0.229968	1.28338	1.48289	0.250981	
12	0	0.229968	*	*	*	
13	1	0.229968	*	*	*	
14	1	0.229968	*	*	*	
15	0	0.158049	-0.43326	-0.44831	0.065987	
16	0	0.129546	-0.38578	-0.39958	0.067861	
17	0	0.085544	2.09565	2.25756	0.138293	
18	1	0.085544	*	*	*	
19	0	0.069044	-0.38514	-0.41441	0.136286	
20	0	0.069044	*	*	*	
21	0	0.044541	-0.21591	-0.22306	0.063106	
22	0	0.035641	-0.19225	-0.19823	0.059440	
23	0	0.022703	-0.15242	-0.15645	0.050869	

```
options ps=60 ls=72 nodate;
data oring;
   infile ’oring.dat’;
   input ID flt f s temp no;
   n = 1;
proc genmod data = oring;
   model f/n = temp  / link=logit  dist=binomial;
run;
```

The first line controls printing of the output. The next four lines define the data. The variable “n” is used to specify that there is only one trial in each of the 23 binomials. PROC GENMOD needs the data specified: “data = oring”. GENMOD also needs information on the model. “link = logit” and “dist = binomial” both are needed to specify that a logistic regression is being fitted. “model f/n = temp” indicates that we are modeling the number of failures in “f” out of “n” trials using the predictor “temp” (and implicitly an intercept).

The SAS program for logistic regression, PROC LOGISTIC, provides more specialized output.

21.6 Multiple Logistic Regression

This section examines regression models for the log-odds of a two-category response variable in which we use more than one predictor variable. The discussion is centered around an example.

EXAMPLE 21.6.1. *Chapman Data.*
Dixon and Massey (1983) and Christensen (1997) present data on 200 men taken from the Los Angeles Heart Study conducted under the supervision of John M. Chapman, UCLA. The data consist of seven variables:

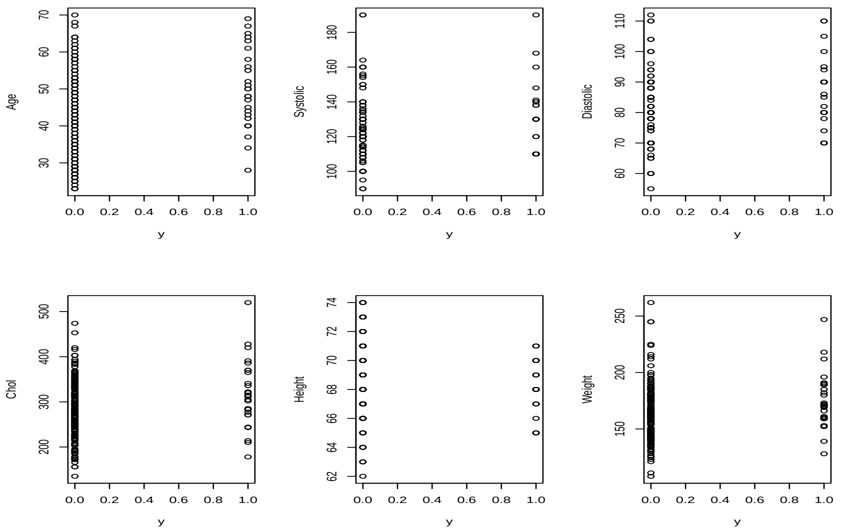

Figure 21.3: *Coronary incident scatterplot matrix.*

Abbreviation	Variable	Units
Ag	Age:	in years
S	Systolic Blood Pressure:	millimeters of mercury
D	Diastolic Blood Pressure:	millimeters of mercury
Ch	Cholesterol:	milligrams per DL
H	Height:	inches
W	Weight:	pounds
CN	Coronary incident:	1 if an incident had occurred in the previous ten years; 0 otherwise

Of the 200 cases, 26 had coronary incidents. The data are available on the internet via STATLIB as well as through the webpage:

`http://stat.unm.edu/~fletcher`

The data are part of the data that go along with the book *Log-Linear Models and Logistic Regression*. Figure 21.3 plots each variable against $y = \text{CN}$. Figures 21.4 through 21.7 provide a scatterplot matrix of the predictor variables.

Let p_i be the probability of a coronary incident for the ith man. We begin with the logistic regression model

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_3 D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i \qquad (21.6.1)$$

$i = 1, \ldots, 200$. The maximum likelihood fit of this model is given in Table 21.5 The deviance df is the number of cases, 200, minus the number of fitted parameters, 7. Based on the t values, none of the variables really stand out. There are suggestions of age, cholesterol, and weight effects. The (unpooled) deviance G^2 would look good except that, as discussed earlier, with $N_i = 1$ for all i there is no basis for comparing it to a $\chi^2(193)$ distribution.

Prediction follows as usual,

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \hat{\beta}_0 + \hat{\beta}_1 Ag_i + \hat{\beta}_2 S_i + \hat{\beta}_3 D_i + \hat{\beta}_4 Ch_i + \hat{\beta}_5 H_i + \hat{\beta}_6 W_i\,.$$

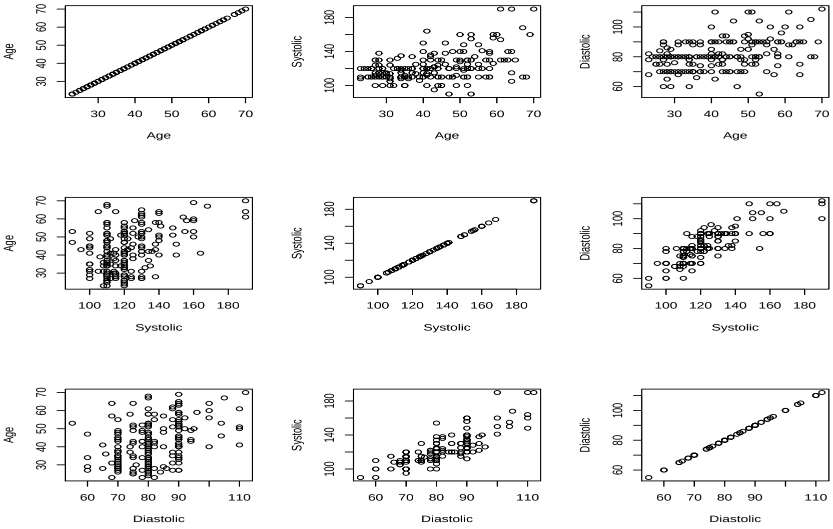

Figure 21.4: *Coronary incident scatterplot matrix.*

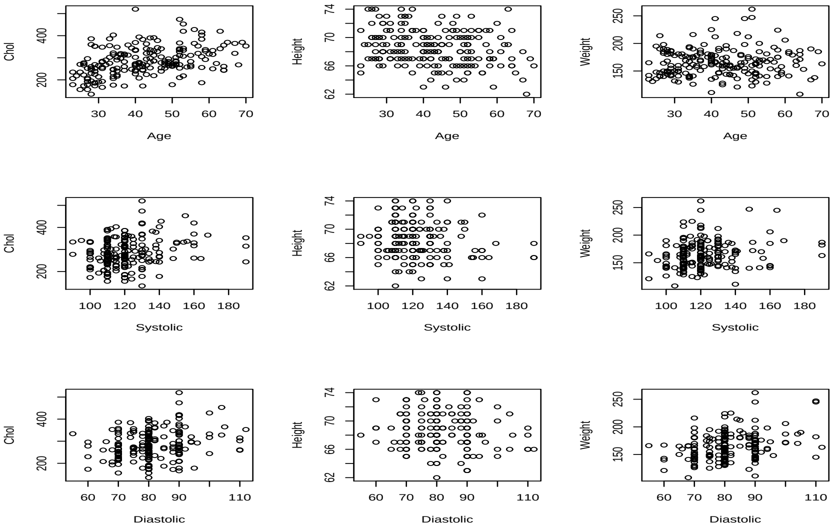

Figure 21.5: *Coronary incident scatterplot matrix.*

Table 21.5: *Table of Coefficients: Model (21.6.1)*

Predictor	*Est*	SE	t
Constant	−4.5173	7.481	−0.60
Ag	0.04590	0.02354	1.95
S	0.00686	0.02020	0.34
D	−0.00694	0.03835	−0.18
Ch	0.00631	0.00363	1.74
H	−0.07400	0.10622	−0.70
W	0.02014	0.00987	2.04
Deviance: $G^2 = 134.9$ $df = 193$			

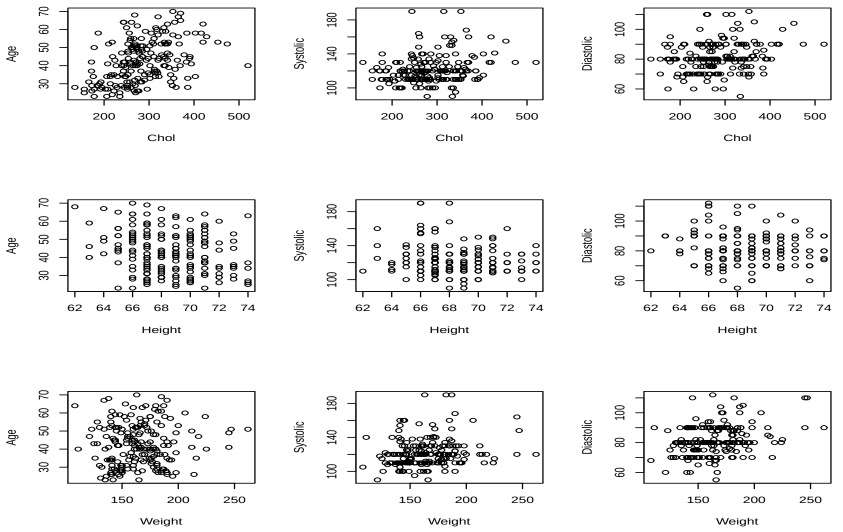

Figure 21.6: *Coronary incident scatterplot matrix.*

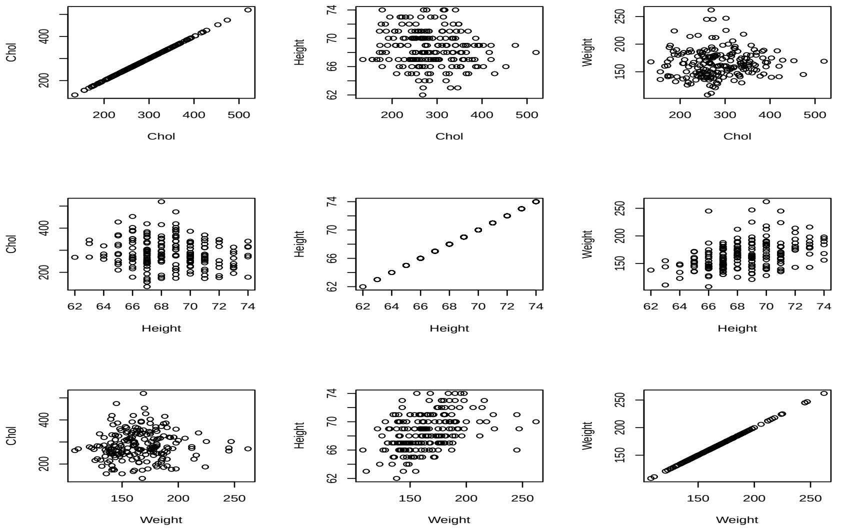

Figure 21.7: *Coronary incident scatterplot matrix.*

For a 60 year old man with blood pressure of 140 over 90, a cholesterol reading of 200, who is 69 inches tall and weighs 200 pounds, the estimated log odds of a coronary incident are

$$\log[\hat{p}/(1-\hat{p})] = -4.5173 + .04590(60) + .00686(140) - .00694(90) \\ + .00631(200) - 0.07400(69) + 0.02014(200) = -1.2435.$$

The probability of a coronary incident is estimated as

$$\hat{p} = \frac{e^{-1.2435}}{1+e^{-1.2435}} = .224\,.$$

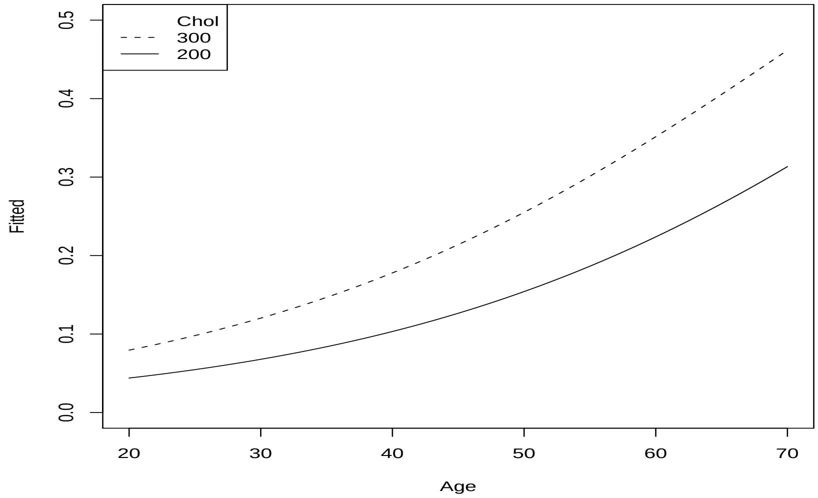

Figure 21.8 *Coronary incident probabilities as a function of age for* $S = 140$*,* $D = 90$*,* $H = 69$*,* $W = 200$*. Solid* $Ch = 200$*, dashed* $Ch = 300$*.*

Figure 21.8 plots the estimated probability of a coronary incident as a function of age for people with $S = 140$, $D = 90$, $H = 69$, $W = 200$ and either $Ch = 200$ (solid line) or $Ch = 300$ (dashed line).

Diagnostic quantities for the cases with the largest Cook's distances are given in Table 21.6. They include 19 of the 26 cases that had coronary incidents. The large residuals are for people who had low probabilities for a coronary incident but had one nonetheless. High leverages correspond to unusual data. For example, case 41 has the highest cholesterol. Case 108 is the heaviest man in the data.

We now consider fitting some reduced models. Simple linear logistic regressions were fitted for each of the variables with high t values, i.e., Ag, Ch, and W. Regressions with variables that seem naturally paired were also fitted, i.e., S,D and H,W. Table 21.7 contains the models along with df, G^2, $A-q$, and A^*. The first two of these are the deviance degrees of freedom and the deviance. No P values are given because the asymptotic χ^2 approximation does not hold. Also given are two analogues of Mallow's C_p statistic, $A-q$ and A^*. $A-q \equiv G^2 - 2(df)$ is the *Akaike information criterion (AIC)* less twice the number of trials ($q \equiv 2n$). A^* is a version of the Akaike information criterion defined for comparing model (21.6.1) to various submodels. It gives numerical values similar to the C_p statistic and is defined by

$$A^* = (G^2 - 134.9) - (7 - 2p).$$

Here 134.9 is the deviance G^2 for the full model (21.6.1), 7 comes from the degrees of freedom for the full model (6 explanatory variables plus an intercept), and p comes from the degrees of freedom for the submodel ($p = 1 +$ number of explanatory variables). The information in $A-q$ and A^* is identical, $A^* = 258.1 + (A-q)$. (The value $258.1 = 2n - G^2[\text{full model}] - p[\text{full model}] = 400 - 134.9 - 7$ does not depend on the reduced model.) A^* is listed because it is a little easier to look at since it takes values similar to C_p. Computer programs rarely report $A-q$ or A^*. (The glm procedure in the R language provides a version of the AIC.) $A-q$ is very easy to compute from the deviance and its degrees of freedom.

Of the models listed in Table 21.7

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i \tag{21.6.2}$$

Table 21.6: *Diagnostics for Chapman data.*

Case	y_h	$\hat{p}_h$	Leverage	r_h	$\tilde{r}_h$	Cook
5	1	0.36	0.13	1.32	1.42	.043
19	1	0.46	0.15	1.08	1.17	.036
21	1	0.08	0.02	3.34	3.37	.028
27	1	0.21	0.03	1.97	1.99	.016
29	1	0.11	0.01	2.73	2.75	.016
39	1	0.16	0.03	2.33	2.36	.022
41	1	0.31	0.15	1.46	1.59	.065
42	1	0.12	0.03	2.60	2.63	.027
44	1	0.41	0.09	1.19	1.24	.021
48	1	0.18	0.06	2.14	2.21	.045
51	1	0.34	0.06	1.39	1.44	.019
54	1	0.19	0.03	2.07	2.09	.017
55	1	0.52	0.08	0.96	1.00	.012
81	1	0.32	0.06	1.44	1.49	.021
84	0	0.36	0.20	-0.74	-0.83	.026
86	1	0.03	0.01	5.95	5.98	.052
108	0	0.45	0.17	-0.91	-1.00	.029
111	1	0.56	0.11	0.89	0.95	.015
113	0	0.37	0.21	-0.76	-0.85	.027
114	0	0.46	0.14	-0.93	-1.00	.024
116	0	0.41	0.10	-0.84	-0.89	.013
123	1	0.36	0.07	1.35	1.40	.022
124	1	0.12	0.02	2.70	2.72	.019
126	1	0.13	0.04	2.64	2.70	.047

Table 21.7: *Models for Chapman data.*

Variables	df	G^2	$A-q$	A^*
Ag,S,D,Ch,H,W	193	134.9	−251.1	7
Ag	198	142.7	−253.3	4.8
W	198	150.1	−245.9	12.2
H,W	197	146.8	−247.2	10.9
Ch	198	146.9	−249.1	9.0
S,D	197	147.9	−246.1	12.0
Intercept	199	154.6	−243.4	14.7

is the only model that is better than the full model based on the information criterion, i.e., A^* is 4.8 for this model, less than the 7 for model (21.6.1).

Asymptotically valid tests of submodels against model (21.6.1) are available. These are performed in the usual way, i.e., the differences in deviance degrees of freedom and deviance G^2s give the appropriate values for the tests. For example, the test of model (21.6.2) versus model (21.6.1) has $G^2 = 142.7 - 134.9 = 7.8$ with $df = 198 - 193 = 5$. This and other tests are given below.

Tests against Model (21.6.1)

Model	df	G^2
Ag	5	7.8
W	5	15.2**
H,W	4	11.9*
Ch	5	12.0*
S,D	4	13.0*
Intercept	6	19.7**

All of the test statistics are significant at the .05 level, except for that associated with model (21.6.2). This indicates that none of the models other than (2) is an adequate substitute for the full model

Table 21.8: *Chapman data models that include Age.*

Variables	df	G^2	A^*
Ag,S,D,Ch,H,W	193	134.9	7.0
Ag,S,D	196	141.4	7.5
Ag,S,Ch	196	139.3	5.4
Ag,S,H	196	141.9	8.0
Ag,S,W	196	138.4	4.5
Ag,D,Ch	196	139.0	5.1
Ag,D,H	196	141.4	7.5
Ag,D,W	196	138.5	4.6
Ag,Ch,H	196	139.9	6.0
Ag,Ch,W	196	135.5	1.6
Ag,H,W	196	138.1	4.2
Ag,S	197	141.9	6.0
Ag,D	197	141.4	5.5
Ag,Ch	197	139.9	4.0
Ag,H	197	142.7	6.8
Ag,W	197	138.8	2.9
Ag	198	142.7	4.8

(21.6.1). In the table above, one asterisk indicates significance at the .05 level and two asterisks denotes significance at the .01 level.

Our next step is to investigate models that include Ag and some other variables. If we can find one or two variables that account for most of the value $G^2 = 7.8$, we may have an improvement over model (21.6.2). If it takes three or more variables to explain the 7.8, model (21.6.2) will continue to be the best-looking model. (Note that $\chi^2(.95, 3) = 7.81$, so a model with three more variables than model (21.6.2) and the same G^2 fit as model (21.6.1) would still not demonstrate a significant lack of fit in model (21.6.2).)

Fits for all models that involve Ag and either one or two other explanatory variables are given in Table 21.8. Based on the A^* values, two models stand out

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i + \gamma_2 W_i \tag{21.6.3}$$

with $A^* = 2.9$ and

$$\log[p_i/(1-p_i)] = \eta_0 + \eta_1 Ag_i + \eta_2 W_i + \eta_3 Ch_i \tag{21.6.4}$$

with $A^* = 1.6$.

The estimated parameters and standard errors for model (21.6.3) are

Table of Coefficients: Model (21.6.3).

Variable	Parameter	Est	SE
Intercept	γ_0	−7.513	1.706
Ag	γ_1	0.06358	0.01963
W	γ_2	0.01600	0.00794

For model (21.6.4), these are

Table of Coefficients: Model (21.6.4).

Variable	Parameter	Est	SE
Intercept	η_0	−9.255	2.061
Ag	η_1	0.05300	0.02074
W	η_2	0.01754	0.003575
Ch	η_3	0.006517	0.008243

The coefficients for Ag and W are quite stable in the two models. The coefficients of Ag, W, and Ch are all positive, so that a small increase in age, weight, or cholesterol is associated with a small increase in the odds of having a coronary incident. Note that we are establishing association, not causation. The data tell us that higher cholesterol is related to higher probabilities, not that it causes higher probabilities.

As in standard regression, interpreting regression coefficients can be very tricky. The fact that the regression coefficients are all positive conforms with the conventional wisdom that high values for any of these factors is associated with increased chance of heart trouble. However, as in standard regression analysis, correlations between predictor variables can make interpretations of individual regression coefficients almost impossible.

It is interesting to note that from fitting model (21.6.1) the estimated regression coefficient for D, diastolic blood pressure, is negative, cf. Table 21.5. A naive interpretation would be that as diastolic blood pressure goes up, the probability of a coronary incident goes down. (If the log odds go down, the probability goes down.) This is contrary to common opinion about how these things work. Actually, this is really just an example of the fallacy of trying to interpret regression coefficients. The regression coefficients have been determined so that the fitted model explains these particular data as well as possible. As mentioned, correlations between the predictor variables can have a huge effect on the estimated regression coefficients. The sample correlation between S and D is .802, so estimated regression coefficients for these variables are unreliable. Moreover, it is not even enough just to check pairwise correlations between variables; any large partial correlations will also adversely affect coefficient interpretations. Fortunately, such correlations should not normally have an adverse affect on the predictive ability of the model; they only adversely affect attempts to interpret regression coefficients. Another excuse for the D coefficient $\hat{\beta}_3$ being negative is that, from the t value, β_3 is not significantly different from 0.

The estimated blood pressure coefficients from model (21.6.1) also suggest an interesting hypothesis. (The hypothesis would be more interesting if the individual coefficients were significant, but we wish to demonstrate a modeling technique.) The coefficient for D is -0.00694, which is approximately the negative of the coefficient for S, 0.00686. This suggests that perhaps $\beta_3 = -\beta_2$ in model (21.6.1). If we incorporate this hypothesis into model (21.6.1) we get

$$\begin{aligned}
&\log[p_i/(1-p_i)] \\
&= \beta_0 + \beta_1 Ag_i + \beta_2 S_i + (-\beta_2) D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i \qquad (21.6.5) \\
&= \beta_0 + \beta_1 Ag_i + \beta_2 (S_i - D_i) + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i,
\end{aligned}$$

which gives deviance $G^2 = 134.9$ on $df = 194$. This model is a reduced model relative to model (21.6.1), so from Table 21.8 a test of it against model (21.6.1) has

$$G^2 = 134.9 - 134.9 = 0.0,$$

with $df = 194 - 193 = 1$. The G^2 is essentially 0, so the data are consistent with the reduced model. Of course this reduced model was suggested by the fitted full model, so any formal test would be biased — but then one does not accept null hypotheses anyway, and the whole point of choosing this reduced model was that it seemed likely to give a G^2 close to that of model (21.6.1). We note that the new variable $S - D$ is still not significant in model (21.6.5); it only has a t value of $.006834/.01877 = .36$.

If we wanted to test something like $\beta_3 = -0.005$, the reduced model is

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + (-0.005) D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i$$

and involves a known term $(-0.005)D_i$ in the linear predictor. This known term is called an *offset*. To fit a model with an offset, most computer programs require that the offset be specified separately and that the model be specified without it, i.e., as

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i.$$

The use of an offset is illustrated in Section 22.6.

We learned earlier that, relative to model (21.6.1), either model (21.6.3) or (21.6.4) does an adequate job of explaining the data. This conclusion was based on looking at A^* values, but would also be obtained by doing formal tests of models.

Christensen (1997, Section 4.4) discusses how to perform best subset selection, similar to Section 10.2, for logistic regression. His preferred method requires access to a standard best subset selection program that allows weighted regression. He does not recommend the score test procedure used by SAS in PROC LOGISTIC.

21.6.1 Computer Commands

Below are SAS commands for obtaining a logistic regression. The data are in a file 'chapman.dat' with eight columns: the case index, Ag, S, D, Ch, H, W, and CN. The file looks like this.

```
  1 44 124  80 254 70 190 0
  2 35 110  70 240 73 216 0
  3 41 114  80 279 68 178 0
  4 31 100  80 284 68 149 0
       data continue
199 50 128  92 264 70 176 0
200 31 105  68 193 67 141 0
```

A simple way to fit the logistic regression model (21.6.4) in SAS is to use PROC GENMOD. The first line controls printing. The next four lines involve defining and reading the data and creating a variable "n" that gives the total number of possible successes for each case. The remaining lines specify the model and that a logistic regression is to be performed.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W CN;
   n = 1;
proc genmod data=chapman ;
   model CN/n = Ag Ch W / link=logit dist=binomial;
run;
```

A more powerful program for logistic regression (but one that pools cases for goodness-of-fit tests and diagnostics) is PROC LOGISTIC.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W CN;
proc logistic data=chapman descending;
   model CN=Ag Ch W ;
run;
```

On the line with "proc logistic", one specifies the data being used and the command "descending". The command "descending" is used so that the program models the probabilities of events coded as 1 rather than events coded as 0. In other words, it makes the program model the probability of a coronary incident rather than the probability of no coronary incident.

21.7 ANOVA Type Logit Models

In this section analysis of variance–type models for the log odds of a two-category response variable are discussed. For ANOVA models, binary data can often be pooled to obtain reasonably large group sizes. More often, the data come presented in groups. We begin with a standard example.

Table 21.9: *Muscle tension change data.*

			Drug (k)	
Tension (h)	Weight (i)	Muscle (j)	Drug 1	Drug 2
High	High	Type 1	3	21
		Type 2	23	11
	Low	Type 1	22	32
		Type 2	4	12
Low	High	Type 1	3	10
		Type 2	41	21
	Low	Type 1	45	23
		Type 2	6	22

EXAMPLE 21.7.1. A study on mice examined the relationship between two drugs and muscle tension. Each mouse had a muscle identified and its tension measured. A randomly selected drug was administered to the mouse and the change in muscle tension was evaluated. Muscles of two types were used. The weight of the muscle was also measured. Factors and levels are as follow.

Factor	Abbreviation	Levels
Change in muscle tension	T	High, Low
Weight of muscle	W	High, Low
Muscle type	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

The data in Table 21.9 are counts (rather than proportions) for every combination of the factors. Probabilities p_{hijk} can be defined for every factor combination.

Change in muscle tension is a response factor. Weight, muscle type, and drug are all predictor variables. We model the log odds of having a high change in muscle tension (given the levels of the explanatory factors), so the observed proportion of, say, high change for Weight = Low, Muscle = 2, Drug = 2 is, from Table 21.9, $4/(4+6)$. The most general ANOVA model (saturated model) includes all main effects and all interactions between the explanatory factors, i.e.,

$$\begin{aligned}\log(p_{1ijk}/p_{2ijk}) &= G+W_i+M_j+D_k \\ &\quad +(WM)_{ij}+(WD)_{ik}+(MD)_{jk} \\ &\quad +(WMD)_{ijk}\,. \end{aligned} \tag{21.7.1}$$

As usual, this is equivalent to a model with just the highest order effects,

$$\log(p_{1ijk}/p_{2ijk}) = (WMD)_{ijk}\,.$$

As introduced in earlier chapters, we denote this model [WMD] with similar notations for other models that focus on the highest order effects.

Models can be fitted by maximum likelihood. Reduced models can be tested. Estimates and asymptotic standard errors can be obtained. The analysis of model (21.7.1) is similar to that of an unbalanced three-factor ANOVA model as illustrated in Chapter 16.

Table 21.10 gives a list of ANOVA type logit models, deviance df, deviance G^2, P values for testing the fitted model against model (21.7.1), and $A-q$ values. Clearly, the best fitting logit models are the models [MD][W] and [WM][MD]. Both involve the muscle type—drug interaction and a main effect for weight. One of the models includes the muscle type—weight interaction. Note that P values associated with saturated model goodness-of-fit tests are appropriate here because we are not dealing with 0-1 data. (The smallest group size is $3+3=6$.)

The estimated odds for a high tension change using [MD][W] are given in Table 21.11. The estimated odds are 1.22 times greater for high weight muscles than for low-weight muscles. For

Table 21.10: *Statistics for Logit Models*

Logit Model	df	G^2	P	$A-q$
$[WM][WD][MD]$	1	0.111	0.7389	-1.889
$[WM][WD]$	2	2.810	0.2440	-1.190
$[WM][MD]$	2	0.1195	0.9417	-3.8805
$[WD][MD]$	2	1.059	0.5948	-2.941
$[WM][D]$	3	4.669	0.1966	-1.331
$[WD][M]$	3	3.726	0.2919	-2.274
$[MD][W]$	3	1.060	0.7898	-4.940
$[W][M][D]$	4	5.311	0.2559	-2.689
$[W][M]$	5	11.35	0.0443	1.35
$[W][D]$	5	12.29	0.0307	2.29
$[M][D]$	5	7.698	0.1727	-2.302

Table 21.11: *Estimated Odds of High Tension Change for [MD][W]*

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.625	1.827
	Type 2	.590	.592
Low	Type 1	.512	1.496
	Type 2	.483	.485

example, in Table 21.11, $.625/.512 = 1.22$ but also $1.22 = .590/.483 = 1.827/1.495 = .592/.485$. This corresponds to the main effect for weight in the logit model. The odds also involve a muscle type—drug interaction. To establish the nature of this interaction, consider the four estimated odds for high weights with various muscles and drugs. These are the four values at the top of Table 21.11, e.g., for muscle type 1, drug 1 this is .625. In every muscle type—drug combination other than type 1, drug 2, the estimated odds of having a high tension change are about .6. The estimated probability of having a high tension change is about $.6/(1+.6) = .375$. However, for type 1, drug 2, the estimated odds are 1.827 and the estimated probability of a high tension change is $1.827/(1+1.827) = .646$. The chance of having a high tension change is much greater for the combination muscle type 1, drug 2 than for any other muscle type—drug combination. A similar analysis holds for the low weight odds $\hat{p}_{12jk}/(1-\hat{p}_{12jk})$ but the actual values of the odds are smaller by a multiplicative factor of 1.22 because of the main effect for weight.

The other logit model that fits quite well is [WM][MD]. Tables 21.12 and 21.13 both contain the estimated odds of high tension change for this model. The difference between Tables 21.12 and 21.13 is that the rows of Table 21.12 have been rearranged in Table 21.13. This sounds like a trivial change, but examination of the tables shows that Table 21.13 is easier to interpret. The reason for changing from Table 21.12 to Table 21.13 is the nature of the logit model. The model [WM][MD] has M in both terms, so it is easiest to interpret the fitted model when fixing the level of M. For a fixed level of M, the effects of W and D are additive in the log odds, although the size of those effects change with the level of M.

Table 21.12: *Estimated Odds for [WM][MD]*

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.809	2.202
	Type 2	.569	.512
Low	Type 1	.499	1.358
	Type 2	.619	.557

Table 21.13: *Estimated Odds for [WM][MD]*

		Drug	
Muscle	Weight	Drug 1	Drug 2
Type 1	High	.809	2.202
	Low	.499	1.358
Type 2	High	.569	.512
	Low	.619	.557

Looking at the type 2 muscles in Table 21.13, the high weight odds are .919 times the low weight odds. Also, the drug 1 odds are 1.111 times the drug 2 odds. Neither of these are really very striking differences. For muscle type 2, the odds of a high tension change are about the same regardless of weight and drug. Contrary to our previous model, they do not seem to depend much on weight and to the extent that they do depend on weight, the odds go down rather than up for higher weights.

Looking at the type 1 muscles, we see the dominant features of the previous model reproduced. The odds of high tension change are 1.622 times greater for high weights than for low weights. The odds of high tension change are 2.722 times greater for drug 2 than for drug 1.

Both models indicate that for type 1 muscles, high weight increases the odds and drug 2 increases the odds. Both models indicate that for type 2 muscles, drug 2 does not substantially change the odds. The difference between the models [MD][W] and [WM][MD] is that [MD][W] indicates that for type 2 muscles, high weight should increase the odds, but [WM][MD] indicates little change for high weight and, in fact, what change there is indicates a decrease in the odds.

21.7.1 Computer Commands

The muscle tension data are listed in the file ‘tenslr.dat’ with one column for the number of high tension scores, one column for the low tension scores, and three columns of indices that specify the level of weight (high is 1), muscle type, and drug, respectively.

```
 3  3 1 1 1
21 10 1 1 2
23 41 1 2 1
11 21 1 2 2
22 45 2 1 1
32 23 2 1 2
 4  6 2 2 1
12 22 2 2 2
```

The following commands fit the model [WM][WD][MD] using SAS PROC GENMOD. Note that the variable “n” is the total number of individuals with each level of weight, muscle type, and drug. The “class” command is used to distinguish ANOVA type factors from regression predictors.

```
options ps=60 ls=72 nodate;
data tension;
   infile 'TAB21-9.DAT';
   input H L W M D;
   n = H+L;
proc genmod data=tension;
   class W M D;
   model H/n = W*M W*D M*D / link=logit dist=binomial;
run;
proc print data=chdiag;
run;
```

Table 21.14: *Abortion Opinion Data*

			AGE					
RACE	SEX	OPINION	18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5

21.8 Ordered Categories

In dealing with ANOVA models, when one or more factors had quantitative levels, it was useful to model effects with polynomials. Similar results apply to logit models.

EXAMPLE 21.8.1. Consider data in which there are four factors defining a $2 \times 2 \times 2 \times 6$ table. The factors are

Factor	Abbreviation	Levels
Race (h)	R	White, Nonwhite
Sex (i)	S	Male, Female

Opinion (j)	O	Yes = Supports Legalized Abortion
		No = Opposed to Legalized Abortion

Age (k)	A	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

Opinion is the response factor. Age has ordered categories. The data are given in Table 21.14. The probability of a Yes opinion for Race h, Sex i, Age k is $p_{hik} \equiv p_{hi1k}$. The corresponding No probability has $1 - p_{hik} \equiv p_{hi2k}$.

As in the previous section, we could fit a three-factor ANOVA type logit model to these data. From the deviances and $A - q$ in Table 21.15 a good fitting logit model is

$$\log[p_{hik}/(1-p_{hik})] = (RS)_{hi} + A_k. \tag{21.8.1}$$

Fitting this model gives the estimated odds of supporting relative to opposing legalized abortion that follow.

Odds of Support versus Opposed: Model (21.8.1)

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	2.52	2.14	2.09	1.60	1.38	1.28
	Female	3.18	2.70	2.64	2.01	1.75	1.62
Nonwhite	Male	2.48	2.11	2.06	1.57	1.36	1.26
	Female	5.08	4.31	4.22	3.22	2.79	2.58

The deviance G^2 is 9.104 with 15 df. The G^2 for fitting [R][S][A] is 11.77 on 16 df. The difference in G^2's is not large, so the reduced logit model $\log[p_{hik}/(1-p_{hik})] = R_{(h)} + S_{(i)} + A_{(k)}$ may fit adequately but we continue to examine model (21.8.1).

Table 21.15: *Logit Models for the Abortion Opinion Data*

Model	df	G^2	$A-q$
[RS][RA][SA]	5	4.161	−5.839
[RS][RA]	10	4.434	−15.566
[RS][SA]	10	8.903	−11.097
[RA][SA]	6	7.443	−4.557
[RS][A]	15	9.104	−20.896
[RA][S]	11	7.707	−14.23
[SA][R]	11	11.564	−10.436
[R][S][A]	16	11.772	−20.228
[R][S]	21	40.521	−1.479
[R][A]	17	21.605	−12.395
[S][A]	17	14.084	−19.916
[R]	22	49.856	5.856
[S]	22	43.451	−0.549
[A]	18	23.799	−12.201
[]	23	52.636	6.636

The odds suggest two things: 1) odds decrease as age increases and 2) the odds for males are about the same, regardless of race. We fit models that incorporate these suggestions. Of course, because the data are suggesting the models, formal tests of significance will be even less appropriate than usual but G^2s still give a reasonable measure of the quality of model fit.

To model odds that are decreasing with age we incorporate a linear trend in ages. In the absence of specific ages to associate with the age categories we simply use the scores $k = 1,2,\ldots,6$. These quantitative levels suggest fitting the ACOVA model

$$\log[p_{hik}/(1-p_{hik})] = (RS)_{hi} + \gamma k. \tag{21.8.2}$$

The deviance G^2 is 10.18 on 19 df, so the linear trend in coded ages fits very well. Recall that model (21.8.1) has $G^2 = 9.104$ on 15 df, so a test of model (21.8.2) versus model (21.8.1) has $G^2 = 10.18 - 9.104 = 1.08$ on $19 - 15 = 4$ df.

To incorporate the idea that males have the same odds of support, recode the indices for races and sexes. The indices for the $(RS)_{hi}$ terms are $(h,i) = (1,1),(1,2),(2,1),(2,2)$. We recode with new indexes (f,e) having the correspondence

(h,i)	(1,1)	(1,2)	(2,1)	(2,2)
(f,e)	(1,1)	(2,1)	(1,2)	(3,1)

The model

$$\log[p_{fek}/(1-p_{fek})] = (RS)_{fe} + A_k$$

gives exactly the same fit as model (21.8.1). Together, the subscripts f, e and k still distinguish all of the cases in the data. The point of this recoding is that the single subscript f distinguishes between males and the two female groups but does not distinguish between white and nonwhite males, so now if we fit the model

$$\log[p_{fek}/(1-p_{fek})] = (RS)_{f} + A_k, \tag{21.8.3}$$

we have a model that treats the two male groups the same. To fit this, you generally do not need to define the index e in your data file, even though it will implicitly exist in the model.

Of course, model (21.8.3) is a reduced model relative to model (21.8.1). Model (21.8.3) has deviance $G^2 = 9.110$ on 16 df, so the comparison between models has $G^2 = 9.110 - 9.104 = .006$ on $16 - 15 = 1$ df. We have lost almost nothing by going from model (21.8.1) to model (21.8.3).

Finally, we can write a model that incorporates both the trend in ages and the equality for males

$$\log[p_{fek}/(1-p_{fek})] = (RS)_{f} + \gamma k. \tag{21.8.4}$$

This has $G^2 = 10.19$ on 20 df. Thus relative to model (21.8.1), we have dropped 5 df from the model, yet only increased the G^2 by $10.19 - 9.10 = 1.09$. Rather than fitting model (21.8.4), we fit the equivalent model that includes an intercept (grand mean) μ. The estimates and standard errors for this model, using the side condition $(RS)_1 = 0$, are

Table of Coefficients: Model related to (21.8.4)

Parameter	*Est*	SE	*t*
μ	1.071	0.1126	9.51
$(RS)_1$	0	—	—
$(RS)_2$	0.2344	0.09265	2.53
$(RS)_3$	0.6998	0.2166	3.23
γ	−0.1410	0.02674	−5.27

All of the terms seem important. With this side condition, $(\widehat{RS})_2$ is actually an estimate of $(RS)_2 - (RS)_1$, so the t score 2.53 is an indication that white females have an effect on the odds of support that is different from males. Similarly, $(\widehat{RS})_3$ is an estimate of the difference in effect between nonwhite females and males.

The estimated odds of support are

Odds of Support: Model (21.8.4).

	Age					
Race-Sex	18-25	26-35	36-45	46-55	56-65	65+
Male	2.535	2.201	1.912	1.661	1.442	1.253
White female	3.204	2.783	2.417	2.099	1.823	1.583
Nonwhite female	5.103	4.432	3.850	3.343	2.904	2.522

The odds can be transformed into probabilities of support. To most people, probabilities are easier to interpret than odds. The estimated probability that a white female between 46 and 55 years of age supports legalized abortion is $2.099/(1+2.099) = .677$. The odds are about 2, so the probability is about twice as great that such a person will support legalized abortion rather than oppose it.

21.8.1 Computer Commands

To fit the data in Table 21.14 they need to be manipulated. In Minitab, the data are read into columns c1 through c5. The following commands allow for fitting model (21.8.1). Here "EPRO2" gives the fitted probabilities from which the odds can be computed and placed into a separate column.

```
MTB > names c1 "h" c2 "i" c3 "k" c4 "j" c5 "count"
sort the data so that all the yes counts are together
and all the no counts are together
MTB > sort c4 c1 c2 c3 c5 c11 c12 c13 c14 c15
Now c11=j, c12=h, c13=i, c14=k, and c15=count
separate the yes data from the no data into a new worksheet
MTB > Unstack (C12 C13 C14 C15);
SUBC>   Subscripts C11;
SUBC>   NewWS;
SUBC>   VarNames.
Now c4 has the yes counts and c8 has the no counts with
c1=c5=h, c2=c6=i, c3=c7=k
MTB > let c10=c4+c8
MTB > Name c13 "EPRO2"
MTB > Blogistic C4 C10 = c1|c2  c3;
SUBC>   ST;
SUBC>   Factors c1 c2 c3;
SUBC>   Logit;
```

Table 21.16: *French convictions*

Year	Convictions	Accusations
1825	4594	7234
1826	4348	6988
1827	4236	6929
1828	4551	7396
1829	4475	7373
1830	4130	6962

Table 21.17: *Heights and chest circumferences*

	Heights					
Chest	64–65	66–67	68–69	70–71	71–73	Total
39	142	442	341	117	20	1062
40	118	337	436	153	38	1082
Total	260	779	777	270	58	2144

```
SUBC>   Eprobability 'EPRO2';
SUBC>   Brief 2.
MTB > Name c14 "Odds"
MTB > let c14 = c13/(1-c13)
```

21.9 Exercises

EXERCISE 21.9.1. Fit a logistic model to the data of Table 21.16 that relates probability of conviction to year. Is there evidence of a trend in the conviction rates over time? Is there evidence for a lack of fit?

EXERCISE 21.9.2. Stigler (1986, p. 208) reports data from the *Edinburgh Medical and Surgical Journal* (1817) on the relationship between heights and chest circumferences for Scottish militia men. Measurements were made in inches. We concern ourselves with two groups of men, those with 39 inch chests and those with 40 inch chests. The data are given in Table 21.17. Test whether the distribution of heights is the same for these two groups, cf. Chapter 5.

Is it reasonable to fit a logistic regression to the data of Table 21.17 Why or why not? Explain what such a model would be doing. Whether reasonable or not, fitting such a model can be done. Fit a logistic model and discuss the results. Is there evidence for a lack of fit?

EXERCISE 21.3. Chapman, Masinda, and Strong (1995) give the data in Table 21.18. These are the number out of 150 popcorn kernels that fail to pop when microwaved for given amount of time. There are three replicates. Fit a logistic regression with time as the predictor.

EXERCISE 21.1. Reanalyze the chloracetic acid data using the log of the dose as a predictor variable. Which model gives a smaller deviance?

Table 21.18: *Unpopped Kernels*

	Trials		
Time	1	2	3
30	144	145	141
45	125	125	118
60	101	138	119
120	197	112	92
150	109	101	61
165	64	54	78
180	34	23	50
210	25	31	36
225	25	27	8
240	11	12	27
255	3	0	2