Chapter 11

Logistic Regression

Logistic regression analysis is used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function. Logistic regression is frequently used to refer to the problem in which the dependent variable is binary — that is, the number of available categories is two — and problems with more than two categories are referred to as multinomial logistic regression or, if the multiple categories are ordered, as ordered logistic regression.

Logistic regression measures the relationship between a categorical dependent variable and usually (but not necessarily) one or more continuous independent variables, by converting the dependent variable to probability scores. As such it treats the same set of problems as does probit regression using similar techniques.

11.1 Generalized linear model variance and link families

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have other than a normal distribution. The GLM generalizes linear regression by allowing the linear
model to be related to the response variable via a link function and by allowing
the magnitude of the variance of each measurement to be a function of its
predicted value.

In R, the basic tool for fitting generalized linear models is the \texttt{glm()} function,
which has the following general structure:

\begin{verbatim}
  glm(formula, family, data, weights, subset, ...)
\end{verbatim}

where “...” stands for additional options. The key parameter here is \texttt{family},
which is a simple way of specifying a choice of variance and link functions. Some
choices of family are listed in the table. As can be seen, each of the first five
choices has an associated variance function (for binomial the binomial variance
$\mu(1 - \mu)$), and one or more choices of link functions (for binomial the logit,
probit, or complementary log-log).

<table>
<thead>
<tr>
<th>Family</th>
<th>Variance</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>gaussian</td>
<td>gaussian</td>
<td>identity</td>
</tr>
<tr>
<td>binomial</td>
<td>binomial</td>
<td>logit, probit, or cloglog</td>
</tr>
<tr>
<td>poisson</td>
<td>poisson</td>
<td>log, identity, or sqrt</td>
</tr>
<tr>
<td>Gamma</td>
<td>Gamma</td>
<td>inverse, identity, or log</td>
</tr>
<tr>
<td>inverse.gaussian</td>
<td>inverse.gaussian</td>
<td>$1/\mu^2$</td>
</tr>
<tr>
<td>quasi</td>
<td>user-defined</td>
<td>user-defined</td>
</tr>
</tbody>
</table>

As long as you want the default link, all you have to specify is the family
name. If you want an alternative link, you must add a link argument. For
example to do probits you use:

\begin{verbatim}
  glm(formula, family = binomial(link = probit))
\end{verbatim}

The last family on the list, quasi, is there to allow fitting user-defined models
by maximum quasi-likelihood.

The rest of this chapter concerns logistic regression with a binary response
variable.
11.2 Example: Age of Menarche in Warsaw

The data\(^1\) below are from a study conducted by Milicer and Szczotka on pre-teen and teenage girls in Warsaw, Poland in 1965. The subjects were classified into 25 age categories. The number of girls in each group (Total) and the number that reached menarche (Menarche) at the time of the study were recorded. The age for a group corresponds to the midpoint for the age interval.

```r
### Example: Menarche
# menarche dataset is available in MASS package
# (remove previous instance if it exists, important if rerunning code)
rm(menarche)
library(MASS)
# these frequencies look better in the table as integers
menarche$Total <- as.integer(menarche$Total)
menarche$Menarche <- as.integer(menarche$Menarche)
str(menarche)
## 'data.frame': 25 obs. of 3 variables:
## $ Age  : num 9.21 10.21 10.58 10.83 11.08 ... 
## $ Total: int 376 200 93 120 90 88 105 111 100 93 ... 
## $ Menarche: int 0 0 0 2 2 5 10 17 16 29 ... 

# create estimated proportion of girls reaching menarche for each age group
menarche$p.hat <- menarche$Menarche / menarche$Total
```

The researchers were curious about how the proportion of girls that reached menarche ($\hat{p} = \text{Menarche}/\text{Total}$) varied with age. One could perform a test of homogeneity (Multinomial goodness-of-fit test) by arranging the data as a 2-by-25 contingency table with columns indexed by age and two rows: ROW1 = Menarche, and ROW2 = number that have not reached menarche = (Total − Menarche). A more powerful approach treats these as regression data, using the proportion of girls reaching menarche as the response and age as a predictor.

A plot of the observed proportion $\hat{p}$ of girls that have reached menarche shows that the proportion increases as age increases, but that the relationship is nonlinear. The observed proportions, which are bounded between zero and one, have a lazy $S$-shape (a **sigmoidal function**) when plotted against age. The change in the observed proportions for a given change in age is much smaller when the proportion is near 0 or 1 than when the proportion is near 1/2. This phenomenon is common with regression data where the response is a proportion.

The trend is nonlinear so linear regression is inappropriate. A sensible alternative might be to transform the response or the predictor to achieve near linearity. A common transformation of response proportions following a sig-
moidal curve is to the **logit scale** \( \hat{\mu} = \log_e \{\hat{p} / (1 - \hat{p})\} \). This transformation is the basis for the **logistic regression model**. The natural logarithm (base e) is traditionally used in logistic regression.

The logit transformation is undefined when \( \hat{p} = 0 \) or \( \hat{p} = 1 \). To overcome this problem, researchers use the **empirical logits**, defined by \( \log \{ (\hat{p} + 0.5/n) / (1 - \hat{p} + 0.5/n) \} \), where \( n \) is the sample size or the number of observations on which \( \hat{p} \) is based.

A plot of the empirical logits against age is roughly linear, which supports a logistic transformation for the response.

```r
library(ggplot2)
p <- ggplot(menarche, aes(x = Age, y = p.hat))
p <- p + geom_point()
p <- p + labs(title = paste("Observed probability of girls reaching menarche,\n" "Warsaw, Poland in 1965", sep=""))
print(p)

# emperical logits
menarche$emp.logit <- log(( menarche$p.hat + 0.5/menarche$Total) /
                         (1 - menarche$p.hat + 0.5/menarche$Total))
library(ggplot2)
p <- ggplot(menarche, aes(x = Age, y = emp.logit))
p <- p + geom_point()
p <- p + labs(title = "Empirical logits")
print(p)
```

---

**Observed probability of girls reaching menarche,**
**Warsaw, Poland in 1965**

**Empirical logits**
11.3 Simple logistic regression model

The simple logistic regression model expresses the population proportion $p$ of individuals with a given attribute (called the probability of success) as a function of a single predictor variable $X$. The model assumes that $p$ is related to $X$ through

$$\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X$$

or, equivalently, as

$$p = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

The logistic regression model is a binary response model, where the response for each case falls into one of two exclusive and exhaustive categories, success (cases with the attribute of interest) and failure (cases without the attribute of interest).

The odds of success are $p/(1 - p)$. For example, when $p = 1/2$ the odds of success are 1 (or 1 to 1). When $p = 0.9$ the odds of success are 9 (or 9 to 1). The logistic model assumes that the log-odds of success is linearly related to $X$. Graphs of the logistic model relating $p$ to $X$ are given below. The sign of the slope refers to the sign of $\beta_1$.

I should write $p = p(X)$ to emphasize that $p$ is the proportion of all individuals with score $X$ that have the attribute of interest. In the menarche data, $p = p(X)$ is the population proportion of girls at age $X$ that have reached menarche.
The data in a logistic regression problem are often given in summarized or **aggregate** form:

\[
\begin{array}{ccc}
X_i & n_i & y_i \\
X_1 & n_1 & y_1 \\
X_2 & n_2 & y_2 \\
\vdots & \vdots & \vdots \\
X_m & n_m & y_m \\
\end{array}
\]

where \( y_i \) is the number of individuals with the attribute of interest among \( n_i \) randomly selected or representative individuals with predictor variable value \( X_i \). For **raw data** on individual cases, \( y_i = 1 \) or \( 0 \), depending on whether the case at \( X_i \) is a success or failure, and the sample size column \( n \) is omitted with raw data.

For logistic regression, a plot of the sample proportions \( \hat{p}_i = y_i/n_i \) against \( X_i \) should be roughly sigmoidal, and a plot of the empirical logits against \( X_i \) should be roughly linear. If not, then some other model is probably appropriate. I find the second plot easier to calibrate, but neither plot is very informative when the sample sizes are small, say 1 or 2. (Why?).

There are a variety of other binary response models that are used in practice. The **probit** regression model or the **complementary log-log** regression...
model might be appropriate when the logistic model does fit the data.

The following section describes the standard MLE strategy for estimating the logistic regression parameters.

### 11.3.1 Estimating Regression Parameters via LS of empirical logits

*(This is a naive method; we will discuss a better way in the next section.)*

There are two unknown population parameters in the logistic regression model

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

A simple way to estimate $\beta_0$ and $\beta_1$ is by least squares (LS), using the empirical logits as responses and the $X_i$s as the predictor values.

Below we use standard regression to calculate the LS fit between the empirical logits and age.

```r
lm.menarche.e.a <- lm(emp.logit ~ Age, data = menarche)
# LS coefficients
coef(lm.menarche.e.a)
## (Intercept) Age
## -22.027933 1.676395
```

The LS estimates for the menarche data are $b_0 = -22.03$ and $b_1 = 1.68$, which gives the fitted relationship

$$\log \left( \frac{\hat{p}}{1-\hat{p}} \right) = -22.03 + 1.68 \text{ Age}$$

or

$$\hat{p} = \frac{\exp(-22.03 + 1.68 \text{ Age})}{1 + \exp(-22.03 + 1.68 \text{ Age})},$$

where $\hat{p}$ is the predicted proportion (under the model) of girls having reached menarche at the given age. I used $\hat{p}$ to identify a predicted probability, in contrast to $\hat{p}$ which is the observed proportion at a given age.

The power of the logistic model versus the contingency table analysis discussed earlier is that the model gives estimates for the population proportion
reaching menarche at all ages within the observed age range. The observed proportions allow you to estimate only the population proportions at the observed ages.

11.3.2 Maximum Likelihood Estimation for Logistic Regression Model

There are better ways to fit the logistic regression model than LS which assumes that the responses are normally distributed with constant variance. A deficiency of the LS fit to the logistic model is that the observed counts $y_i$ have a Binomial distribution under random sampling. The Binomial distribution is a discrete probability model associated with counting the number of successes in a fixed size sample, and other equivalent experiments such as counting the number of heads in repeated flips of a coin. The distribution of the empirical logits depend on the $y_i$'s so they are not normal (but are approximately normal in large samples), and are extremely skewed when the sample sizes $n_i$ are small. The response variability depends on the population proportions, and is not roughly constant when the observed proportions or the sample sizes vary appreciably across groups.

The differences in variability among the empirical logits can be accounted for using weighted least squares (WLS) when the sample sizes are large. An alternative approach called maximum likelihood uses the exact Binomial distribution of the responses $y_i$ to generate optimal estimates of the regression coefficients. Software for maximum likelihood estimation is widely available, so LS and WLS methods are not really needed.

In maximum likelihood estimation (MLE), the regression coefficients are estimated iteratively by minimizing the deviance function (also called the likelihood ratio chi-squared statistic)

$$D = 2 \sum_{i=1}^{m} \left\{ y_i \log \left( \frac{y_i}{n_i p_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - n_i p_i} \right) \right\}$$
over all possible values of $\beta_0$ and $\beta_1$, where the $p_i$s satisfy the logistic model

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

The ML method also gives standard errors and significance tests for the regression estimates.

The deviance is an analog of the residual sums of squares in linear regression. The choices for $\beta_0$ and $\beta_1$ that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a “likelihood sense”.

Suppose that $b_0$ and $b_1$ are the MLEs of $\beta_0$ and $\beta_1$. The deviance evaluated at the MLEs,

$$D = 2 \sum_{i=1}^{m} \left\{ y_i \log \left( \frac{y_i}{n_i \hat{p}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - n_i \hat{p}_i} \right) \right\},$$

where the fitted probabilities $\hat{p}_i$ satisfy

$$\log \left( \frac{\hat{p}_i}{1 - \hat{p}_i} \right) = b_0 + b_1 X_i,$$

is used to test the adequacy of the model. The deviance is small when the data fits the model, that is, when the observed and fitted proportions are close together. Large values of $D$ occur when one or more of the observed and fitted proportions are far apart, which suggests that the model is inappropriate.

If the logistic model holds, then $D$ has a chi-squared distribution with $m - r$ degrees of freedom, where $m$ is the the number of groups and $r$ (here 2) is the number of estimated regression parameters. A p-value for the deviance is given by the area under the chi-squared curve to the right of $D$. A small p-value indicates that the data does not fit the model.

Alternatively, the fit of the model can be evaluated using the chi-squared approximation to the Pearson $X^2$ statistic:

$$X^2 = \sum_{i=1}^{m} \left\{ \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i} + \frac{((n_i - y_i) - n_i(1 - \hat{p}_i))^2}{n_i(1 - \hat{p}_i)} \right\} = \sum_{i=1}^{m} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i(1 - \hat{p}_i)}. $$
11.3.3 Fitting the Logistic Model by Maximum Likelihood, Menarche

```r
# For our summarized data (with frequencies and totals for each age)
# The left-hand side of our formula binds two columns together with cbind():
# the columns are the number of "successes" and "failures".
# For logistic regression with logit link we specify family = binomial,
# where logit is the default link function for the binomial family.
glm.m.a <- glm(cbind(Menarche, Total - Menarche) ~ Age, family = binomial, menarche)
```

The `glm()` statement creates an object which we can use to create the fitted probabilities and 95% CIs for the population proportions at the ages in `menarche`. The fitted probabilities and the limits are stored in columns labeled `fitted.values`, `fit.lower`, and `fit.upper`, respectively.

```r
# put the fitted values in the data.frame
menarche$fitted.values <- glm.m.a$fitted.values
pred <- predict(glm.m.a, data.frame(Age = menarche$Age), type = "link", se.fit = TRUE)
menarche$fit <- pred$fit
menarche$se.fit <- pred$se.fit
# CI for fitted values
menarche <- within(menarche, {
  fit.lower = exp(fit - 1.96 * se.fit) / (1 + exp(fit - 1.96 * se.fit))
  fit.upper = exp(fit + 1.96 * se.fit) / (1 + exp(fit + 1.96 * se.fit))
})
#round(menarche, 3)
```

This printed summary information is easily interpreted. For example, the estimated population proportion of girls aged 15.08 (more precisely, among girls in the age interval with midpoint 15.08) that have reached menarche is 0.967. You are 95% confident that the population proportion is between 0.958 and 0.975. A variety of other summaries and diagnostics can be produced.

<table>
<thead>
<tr>
<th>Age</th>
<th>Total</th>
<th>Menarche</th>
<th>p.hat</th>
<th>emp.logit</th>
<th>fitted.values</th>
<th>fit</th>
<th>se.fit</th>
<th>fit.upper</th>
<th>fit.lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>15.08</td>
<td>112.00</td>
<td>0.96</td>
<td>3.06</td>
<td>0.97</td>
<td>3.38</td>
<td>0.14</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

UNM, Stat 428/528 ADA2
The summary table gives MLEs and standard errors for the regression parameters. The z-value column is the parameter estimate divided by its standard error. The p-values are used to test whether the corresponding parameters of the logistic model are zero.

```
summary(glm.m.a)
##
## Call:
## glm(formula = cbind(Menarche, Total - Menarche) ~ Age, family = binomial, 
## data = menarche)
##
## Deviance Residuals:
##          Min     1Q   Median     3Q    Max
## -2.0363 -0.9953 -0.4900  0.7780  1.3675
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.22639    0.77068  -27.54  <2e-16 ***
## Age          1.63197    0.05895   27.68  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3693.884 on 24 degrees of freedom
## Residual deviance: 26.703 on 23 degrees of freedom
## AIC: 114.76
```
If the model is correct and when sample sizes are large, the residual deviance $D$ has an approximate chi-square distribution,

$$\text{residual } D = \chi^2_{\text{residual df}}.$$

If $D$ is too large, or the $p$-value is too small, then the model does not capture all the features in the data.

The deviance statistic is $D = 26.70$ on $25 - 2 = 23$ df. The large p-value for $D$ suggests no gross deficiencies with the logistic model. The observed and fitted proportions ($\hat{p}$ and fitted.values in the output table above are reasonably close at each observed age. Also, emp.logit and fit are close. This is consistent with $D$ being fairly small. The data fits the logistic regression model reasonably well.

```r
# Test residual deviance for lack-of-fit (if > 0.10, little-to-no lack-of-fit)
glm.m.a$deviance
## [1] 26.70345

glm.m.a$df.residual
## [1] 23

dev.p.val <- 1 - pchisq(glm.m.a$deviance, glm.m.a$df.residual)
dev.p.val
## [1] 0.2687953
```

The MLEs $b_0 = -21.23$ and $b_1 = 1.63$ for the intercept and slope are close to the LS estimates of $b_{LS0} = -22.03$ and $b_{LS1} = 1.68$, respectively from page 298. The two estimation methods give similar predicted probabilities here. The MLE of the predicted probabilities satisfy

$$\log \left( \frac{\tilde{p}}{1 - \tilde{p}} \right) = -21.23 + 1.63 \text{ Age}$$

or

$$\tilde{p} = \frac{\exp(-21.23 + 1.63 \text{ Age})}{1 + \exp(-21.23 + 1.63 \text{ Age})}.$$
library(ggplot2)
p <- ggplot(menarche, aes(x = Age, y = p.hat))
# predicted curve and point-wise 95% CI
p <- p + geom_ribbon(aes(x = Age, ymin = fit.lower, ymax = fit.upper), alpha = 0.2)
p <- p + geom_line(aes(x = Age, y = fitted.values), color = "red")  
# fitted values
p <- p + geom_point(aes(y = fitted.values), color = "red", size=2)
# observed values
p <- p + geom_point(size=2)
p <- p + labs(title = paste("Observed and predicted probability of girls reaching menarche,\nWarsaw, Poland in 1965", sep="))
print(p)

If the model holds, then a slope of \( \beta_1 = 0 \) implies that \( p \) does not depend on AGE, i.e., the proportion of girls that have reached menarche is identical across age groups. The Wald p-value for the slope is < 0.0001, which leads to rejecting \( H_0 : \beta_1 = 0 \) at any of the usual test levels. The proportion of girls that have reached menarche is not constant across age groups. Again, the power of the model is that it gives you a simple way to quantify the effect of age on the proportion reaching menarche. This is more appealing than testing
homogeneity across age groups followed by multiple comparisons.

Wald tests can be performed to test the global null hypothesis, that all non-intercept $\beta$s are equal to zero. This is the logistic regression analog of the overall model F-test in ANOVA and regression. The only predictor is AGE, so the implied test is that the slope of the regression line is zero. The Wald test statistic and p-value reported here are identical to the Wald test and p-value for the AGE effect given in the parameter estimates table. The Wald test can also be used to test specific contrasts between parameters.

```r
# Testing Global Null Hypothesis
library(aod)
coef(glm.m.a)
## (Intercept) Age
## -21.226395 1.631968
# specify which coefficients to test = 0 (Terms = 2:4 would be terms 2, 3, and 4)
wald.test(b = coef(glm.m.a), Sigma = vcov(glm.m.a), Terms = 2:2)
## Wald test:
##  
## Chi-squared test:
## X2 = 766.3, df = 1, P(> X2) = 0.0
```

### 11.4 Example: Leukemia white blood cell types

Feigl and Zelen\(^2\) reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG−, indicating the presence or absence of a certain morphological characteristic in the white cells. Four variables are given in the data set: WBC, a binary factor or **indicator variable** AG (1 for AG+, 0 for AG−), NTOTAL (the number of patients with

\(^2\)Feigl, P., Zelen, M. (1965) Estimation of exponential survival probabilities with concomitant information. Biometrics 21, 826-838. Survival times are given for 33 patients who died from acute myelogenous leukaemia. Also measured was the patient’s white blood cell count at the time of diagnosis. The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at the time of diagnosis.

UNM, Stat 428/528 ADA2
the given combination of AG and WBC), and NRES (the number of NTOTAL that survived at least one year from the time of diagnosis).

The researchers are interested in modelling the probability $p$ of surviving at least one year as a function of WBC and AG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values.

### Example: Leukemia

#### Leukemia white blood cell types example
- `ntotal` = number of patients with IAG and WBC combination
- `nres` = number surviving at least one year
- `ag` = 1 for AG+, 0 for AG-
- `wbc` = white cell blood count
- `lwbc` = log white cell blood count
- `p.hat` = Empirical Probability

```r
leuk <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch11_leuk.dat", header = TRUE)
leuk$ag <- factor(leuk$ag)
leuk$lwbc <- log(leuk$wbc)
leuk$p.hat <- leuk$nres / leuk$ntotal
str(leuk)
```

```r
## 'data.frame': 30 obs. of 6 variables:
##  $ ntotal: int 1 1 1 1 1 1 1 1 3 1 ...
##  $ nres : int 1 1 1 1 1 1 1 1 1 1 ...
##  $ ag   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...
##  $ wbc  : int 75 230 260 430 700 940 1000 1050 10000 300 ...
##  $ lwbc : num 4.32 5.44 5.56 6.06 6.55 ...
##  $ p.hat: num 1 1 1 1 1 1 ...
```
As an initial step in the analysis, consider the following model:

$$\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \, \text{LWBC} + \beta_2 \, \text{AG},$$

where \( \text{LWBC} = \log(\text{WBC}) \). The model is best understood by separating the \( \text{AG}+ \) and \( \text{AG}− \) cases. For \( \text{AG}− \) individuals, \( \text{AG}=0 \) so the model reduces to

$$\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \, \text{LWBC} + \beta_2 \times 0 = \beta_0 + \beta_1 \, \text{LWBC}.$$
For AG+ individuals, AG=1 and the model implies

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 \times 1 = (\beta_0 + \beta_2) + \beta_1 \text{LWBC}.
\]

The model without AG (i.e., \(\beta_2 = 0\)) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with \(\beta_2 = 0\) implies that there is no effect of the AG level on the survival probability once LWBC has been taken into account.

Including the binary predictor AG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. A natural extension would be to include a product or interaction effect, a point that I will return to momentarily.

The parameters are easily interpreted: \(\beta_0\) and \(\beta_0 + \beta_2\) are intercepts for the population logistic regression lines for AG− and AG+, respectively. The lines have a common slope, \(\beta_1\). The \(\beta_2\) coefficient for the AG indicator is the difference between intercepts for the AG+ and AG− regression lines. A picture of the assumed relationship is given below for \(\beta_1 < 0\). The population regression lines are parallel on the logit scale only, but the order between AG groups is preserved on the probability scale.
Before looking at output for the equal slopes model, note that the data set has 30 distinct AG and LWBC combinations, or 30 “groups” or samples. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving ≥ 1 year/group sample size) are 0 (i.e., 0/1) or 1 (i.e., 1/1). This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (because the estimated proportions are almost all 0 or 1). Although significance tests on the regression coefficients do not require large group sizes, the chi-squared approximation to the deviance statistic is suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-value for $D$ literally. Instead, they would use the p-values to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

```r
glm.i.l <- glm(cbind(nres, ntotal - nres) ~ ag + lwbc, family = binomial, leuk)
# Test residual deviance for lack-of-fit (if > 0.10, little-to-no lack-of-fit)
dev.p.val <- 1 - pchisq(glm.i.l$deviance, glm.i.l$df.residual)
dev.p.val
```
The large p-value for $D$ indicates that there are no gross deficiencies with the model. Recall that the Testing Global Null Hypothesis gives p-values for testing the hypothesis that the regression coefficients are zero for each predictor in the model. The two predictors are LWBC and AG, so the small p-values indicate that LWBC or AG, or both, are important predictors of survival. The p-values in the estimates table suggest that LWBC and AG are both important. If either predictor was insignificant, I would consider refitting the model omitting the least significant effect, as in regression.

```r
# Testing Global Null Hypothesis
library(aod)
coef(glm.i.l)
## (Intercept)  ag1  lwbc
## 5.543349  2.519562 -1.108759
# specify which coefficients to test = 0  (Terms = 2:3 is for terms 2 and 3)
wald.test(b = coef(glm.i.l), Sigma = vcov(glm.i.l), Terms = 2:3)
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 8.2, df = 2, P(> X2) = 0.017
```

Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. This test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model).

```r
summary(glm.i.l)
##
## Call:
## glm(formula = cbind(nres, ntotal - nres) ~ ag + lwbc, family = binomial,
##     data = leuk)
##
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6599 -0.6595 -0.2776  0.6438  1.7131
##
## Coefficients:
A plot of the predicted survival probabilities as a function of LWBC, using AG as the plotting symbol, indicates that the probability of surviving at least one year from the time of diagnosis is a decreasing function of LWBC. For a given LWBC the survival probability is greater for AG+ patients than for AG− patients. This tendency is consistent with the observed proportions, which show little information about the exact form of the trend.

```r
# put the fitted values in the data.frame
leuk$fitted.values <- glm.i.l$fitted.values
pred <- predict(glm.i.l, data.frame(lwbc = leuk$lwbc, ag = leuk$ag), type = "link", se.fit = TRUE)
leuk$fit <- pred$fit
leuk$se.fit <- pred$se.fit

# CI for fitted values
leuk <- within(leuk, {
  fit.lower = exp(fit - 1.96 * se.fit) / (1 + exp(fit - 1.96 * se.fit))
  fit.upper = exp(fit + 1.96 * se.fit) / (1 + exp(fit + 1.96 * se.fit))
})

#round(leuk, 3)

library(ggplot2)
p <- ggplot(leuk, aes(x = lwbc, y = p.hat, colour = ag, fill = ag))

# predicted curve and point-wise 95% CI
p <- p + geom_ribbon(aes(x = lwbc, ymin = fit.lower, ymax = fit.upper), alpha = 0.2)
p <- p + geom_line(aes(x = lwbc, y = fitted.values))

# fitted values
p <- p + geom_point(aes(y = fitted.values), size=2)

# observed values
p <- p + geom_point(size = 2, alpha = 0.5)
p <- p + labs(title = "Observed and predicted probability of 1+ year survival")
print(p)
```
### Observed and predicted probability of 1+ year survival

<table>
<thead>
<tr>
<th></th>
<th>ntotal</th>
<th>nres</th>
<th>ag</th>
<th>wbc</th>
<th>lwbc</th>
<th>p.hat</th>
<th>fitted.values</th>
<th>fit</th>
<th>se.fit</th>
<th>fit.upper</th>
<th>fit.lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>75</td>
<td>4.32</td>
<td>1.00</td>
<td>0.96</td>
<td>3.28</td>
<td>1.44</td>
<td>1.00</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>230</td>
<td>5.44</td>
<td>1.00</td>
<td>0.88</td>
<td>2.03</td>
<td>0.99</td>
<td>0.98</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>260</td>
<td>5.56</td>
<td>1.00</td>
<td>0.87</td>
<td>1.90</td>
<td>0.94</td>
<td>0.98</td>
<td>0.51</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>430</td>
<td>6.06</td>
<td>1.00</td>
<td>0.79</td>
<td>1.34</td>
<td>0.78</td>
<td>0.95</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>700</td>
<td>6.55</td>
<td>1.00</td>
<td>0.69</td>
<td>0.80</td>
<td>0.66</td>
<td>0.89</td>
<td>0.38</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>940</td>
<td>6.85</td>
<td>1.00</td>
<td>0.62</td>
<td>0.47</td>
<td>0.61</td>
<td>0.84</td>
<td>0.33</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1000</td>
<td>6.91</td>
<td>1.00</td>
<td>0.60</td>
<td>0.40</td>
<td>0.61</td>
<td>0.83</td>
<td>0.31</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1050</td>
<td>6.96</td>
<td>1.00</td>
<td>0.59</td>
<td>0.35</td>
<td>0.60</td>
<td>0.82</td>
<td>0.30</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>10000</td>
<td>9.21</td>
<td>0.33</td>
<td>0.10</td>
<td>-2.15</td>
<td>1.12</td>
<td>0.51</td>
<td>0.01</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>300</td>
<td>5.70</td>
<td>1.00</td>
<td>0.31</td>
<td>-0.78</td>
<td>0.87</td>
<td>0.72</td>
<td>0.08</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>440</td>
<td>6.09</td>
<td>1.00</td>
<td>0.23</td>
<td>-1.21</td>
<td>0.83</td>
<td>0.61</td>
<td>0.06</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>540</td>
<td>6.29</td>
<td>0.00</td>
<td>0.75</td>
<td>1.09</td>
<td>0.72</td>
<td>0.92</td>
<td>0.42</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>600</td>
<td>6.40</td>
<td>0.00</td>
<td>0.73</td>
<td>0.97</td>
<td>0.69</td>
<td>0.91</td>
<td>0.41</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1700</td>
<td>7.44</td>
<td>0.00</td>
<td>0.45</td>
<td>-0.18</td>
<td>0.61</td>
<td>0.73</td>
<td>0.20</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3200</td>
<td>8.07</td>
<td>0.00</td>
<td>0.29</td>
<td>-0.89</td>
<td>0.73</td>
<td>0.63</td>
<td>0.09</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3500</td>
<td>8.16</td>
<td>0.00</td>
<td>0.27</td>
<td>-0.99</td>
<td>0.75</td>
<td>0.62</td>
<td>0.08</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5200</td>
<td>8.56</td>
<td>0.00</td>
<td>0.19</td>
<td>-1.42</td>
<td>0.88</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>150</td>
<td>5.01</td>
<td>0.00</td>
<td>0.50</td>
<td>0.01</td>
<td>1.02</td>
<td>0.88</td>
<td>0.12</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>400</td>
<td>5.99</td>
<td>0.00</td>
<td>0.25</td>
<td>-1.10</td>
<td>0.84</td>
<td>0.63</td>
<td>0.06</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>530</td>
<td>6.27</td>
<td>0.00</td>
<td>0.20</td>
<td>-1.41</td>
<td>0.83</td>
<td>0.55</td>
<td>0.05</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>900</td>
<td>6.80</td>
<td>0.00</td>
<td>0.12</td>
<td>-2.00</td>
<td>0.86</td>
<td>0.42</td>
<td>0.02</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>6.91</td>
<td>0.00</td>
<td>0.11</td>
<td>-2.12</td>
<td>0.87</td>
<td>0.40</td>
<td>0.02</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1900</td>
<td>7.55</td>
<td>0.00</td>
<td>0.06</td>
<td>-2.83</td>
<td>1.01</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2100</td>
<td>7.65</td>
<td>0.00</td>
<td>0.05</td>
<td>-2.94</td>
<td>1.03</td>
<td>0.29</td>
<td>0.01</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2600</td>
<td>7.86</td>
<td>0.00</td>
<td>0.04</td>
<td>-3.18</td>
<td>1.09</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2700</td>
<td>7.90</td>
<td>0.00</td>
<td>0.04</td>
<td>-3.22</td>
<td>1.11</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2800</td>
<td>7.94</td>
<td>0.00</td>
<td>0.04</td>
<td>-3.26</td>
<td>1.12</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3100</td>
<td>8.04</td>
<td>0.00</td>
<td>0.03</td>
<td>-3.37</td>
<td>1.15</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7900</td>
<td>8.97</td>
<td>0.00</td>
<td>0.01</td>
<td>-4.41</td>
<td>1.48</td>
<td>0.18</td>
<td>0.00</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10000</td>
<td>9.21</td>
<td>0.00</td>
<td>0.01</td>
<td>-4.67</td>
<td>1.57</td>
<td>0.17</td>
<td>0.00</td>
</tr>
</tbody>
</table>
The estimated survival probabilities satisfy
\[
\log \left( \frac{\hat{p}}{1 - \hat{p}} \right) = 5.54 - 1.11 \text{ LWBC} + 2.52 \text{ AG}.
\]

For AG− individuals with AG=0, this reduces to
\[
\log \left( \frac{\hat{p}}{1 - \hat{p}} \right) = 5.54 - 1.11 \text{ LWBC},
\]
or equivalently,
\[
\hat{p} = \frac{\exp(5.54 - 1.11 \text{ LWBC})}{1 + \exp(5.54 - 1.11 \text{ LWBC})}.
\]

For AG+ individuals with AG=1,
\[
\log \left( \frac{\hat{p}}{1 - \hat{p}} \right) = 5.54 - 1.11 \text{ LWBC} + 2.52(1) = 8.06 - 1.11 \text{ LWBC},
\]
or
\[
\hat{p} = \frac{\exp(8.06 - 1.11 \text{ LWBC})}{1 + \exp(8.06 - 1.11 \text{ LWBC})}.
\]

Using the logit scale, the difference between AG+ and AG− individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated AG regression coefficient
\[
(8.06 - 1.11 \text{ LWBC}) - (5.54 - 1.11 \text{ LWBC}) = 2.52.
\]

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is \(\exp(2.52) = 12.42\) times larger than the odds that an AG− patient lives at least one year, regardless of LWBC.

This summary, and a CI for the AG odds ratio, is given in the Odds Ratio table. Similarly, the estimated odds ratio of 0.33 for LWBC implies that the odds of surviving at least one year is reduced by a factor of 3 for each unit increase of LWBC.

We can use the confint() function to obtain confidence intervals for the coefficient estimates. Note that for logistic models, confidence intervals are based on the profiled log-likelihood function.
## CIs using profiled log-likelihood
```r
confint(glm.i.l)
# Waiting for profiling to be done...
## 2.5 % 97.5 %
## (Intercept) 0.1596372 12.4524409
## ag1 0.5993391 5.0149271
## lwbc -2.2072275 -0.3319512
```

We can also get CIs based on just the standard errors by using the default method.

## CIs using standard errors
```r
confint.default(glm.i.l)
# 2.5 % 97.5 %
## (Intercept) -0.3804137 11.467112
## ag1 0.3818885 4.657236
## lwbc -2.0121879 -0.205330
```

You can also exponentiate the coefficients and confidence interval bounds and interpret them as odds-ratios.

## coefficients and 95% CI
```r
cbind(OR = coef(glm.i.l), confint(glm.i.l))
# Waiting for profiling to be done...
## OR 2.5 % 97.5 %
## (Intercept) 5.543349 0.1596372 12.4524409
## ag1 2.519562 0.5993391 5.0149271
## lwbc -1.108759 -2.2072275 -0.3319512
```

## odds ratios and 95% CI
```r
exp(cbind(OR = coef(glm.i.l), confint(glm.i.l)))
# Waiting for profiling to be done...
## OR 2.5 % 97.5 %
## (Intercept) 255.5323676 1.1730851 2.558741e+05
## ag1 12.4231582 1.8209149 1.506452e+02
## lwbc 0.3299682 0.1100052 7.175224e-01
```

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an interaction, or product term, AG \times LWBC to the model. The logistic model with an AG effect and the AG \times LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.
Interpreting odds ratios in logistic regression

Let’s begin with probability\(^3\). Let’s say that the probability of success is 0.8, thus \( p = 0.8 \). Then the probability of failure is \( q = 1 - p = 0.2 \). The odds of success are defined as \( \text{odds(success)} = \frac{p}{q} = \frac{0.8}{0.2} = 4 \), that is, the odds of success are 4 to 1. The odds of failure would be \( \text{odds(failure)} = \frac{q}{p} = \frac{0.2}{0.8} = 0.25 \), that is, the odds of failure are 1 to 4. Next, let’s compute the odds ratio by \( \text{OR} = \frac{\text{odds(success)}}{\text{odds(failure)}} = \frac{4}{0.25} = 16 \). The interpretation of this odds ratio would be that the odds of success are 16 times greater than for failure. Now if we had formed the odds ratio the other way around with odds of failure in the numerator, we would have gotten something like this, \( \text{OR} = \frac{\text{odds(failure)}}{\text{odds(success)}} = \frac{0.25}{4} = 0.0625 \).

Another example

This example is adapted from Pedhazur (1997). Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted. The probabilities for admitting a male are, \( p = \frac{7}{10} = 0.7 \) and \( q = 1 - 0.7 = 0.3 \). Here are the same probabilities for females, \( p = \frac{3}{10} = 0.3 \) and \( q = 1 - 0.3 = 0.7 \). Now we can use the probabilities to compute the admission odds for both males and females, \( \text{odds(male)} = \frac{0.7}{0.3} = 2.33333 \) and \( \text{odds(female)} = \frac{0.3}{0.7} = 0.42857 \). Next, we compute the odds ratio for admission, \( \text{OR} = \frac{2.33333}{0.42857} = 5.44 \). Thus, the odds of a male being admitted are 5.44 times greater than for a female.

Leukemia example

In the example above, the OR of surviving at least one year increases 12.43 times for AG+ vs AG−, and increases 0.33 times (that’s a decrease) for every unit increase in lwbc.

Example: Mortality of confused flour beetles

This example illustrates a quadratic logistic model.

---

\(^3\)Borrowed graciously from UCLA Academic Technology Services at http://www.ats.ucla.edu/stat/sas/faq/oratio.htm
The aim of an experiment originally reported by Strand (1930) and quoted by Bliss (1935) was to assess the response of the confused flour beetle, *Tribolium confusum*, to gaseous carbon disulphide (CS₂). In the experiment, prescribed volumes of liquid carbon disulphide were added to flasks in which a tubular cloth cage containing a batch of about thirty beetles was suspended. Duplicate batches of beetles were used for each concentration of CS₂. At the end of a five-hour period, the proportion killed was recorded and the actual concentration of gaseous CS₂ in the flask, measured in mg/l, was determined by a volumetric analysis. The mortality data are given in the table below.

### Example: Beetles

## Beetles data set

<table>
<thead>
<tr>
<th>Conc</th>
<th>Y</th>
<th>N</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.06</td>
<td>2</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>52.99</td>
<td>7</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>56.91</td>
<td>9</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>60.84</td>
<td>14</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>64.76</td>
<td>23</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>68.69</td>
<td>29</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>72.61</td>
<td>29</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>76.54</td>
<td>29</td>
<td>29</td>
<td>1</td>
</tr>
</tbody>
</table>

# Conc = CS₂ concentration  
# Y = number of beetles killed  
# N = number of beetles exposed  
# Rep = Replicate number (1 or 2)

```
beetles <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch11_beetles.dat", header = TRUE)
beetles$rep <- factor(beetles$rep)

<table>
<thead>
<tr>
<th>Conc</th>
<th>Y</th>
<th>N</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.06</td>
<td>2</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>52.99</td>
<td>7</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>56.91</td>
<td>9</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>60.84</td>
<td>14</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>64.76</td>
<td>23</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>68.69</td>
<td>29</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>72.61</td>
<td>32</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>76.54</td>
<td>31</td>
<td>31</td>
<td>2</td>
</tr>
</tbody>
</table>
```

```
beetles$conc2 <- beetles$conc^2  # for quadratic term (making coding a little easier)
beetles$p.hat <- beetles$y / beetles$n  # observed proportion of successes

# empirical logits
beetles$emp.logit <- log((beetles$p.hat + 0.5/beetles$n) / (1 - beetles$p.hat + 0.5/beetles$n))

```

Plot the observed probability of mortality and the empirical logits with linear and quadratic LS fits (which are not the same as the logistic MLE fits).

```
library(ggplot2)
p <- ggplot(beetles, aes(x = conc, y = p.hat, shape = rep))  
# observed values
p <- p + geom_point(color = "black", size = 3, alpha = 0.5)
p <- p + labs(title = "Observed mortality, probability scale")
print(p)
```
In a number of articles that refer to these data, the responses from the first two concentrations are omitted because of apparent non-linearity. Bliss himself remarks that

...in comparison with the remaining observations, the two lowest concentrations gave an exceptionally high kill. Over the remaining concentrations, the plotted values seemed to form a moderately straight line, so that the data were handled as two separate sets, only the results at 56.91 mg of CS$_2$ per litre being included in both sets.

However, there does not appear to be any biological motivation for this and so here they are retained in the data set.

Combining the data from the two replicates and plotting the empirical logit of the observed proportions against concentration gives a relationship that is
better fit by a quadratic than a linear relationship,

\[
\log \left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 X + \beta_2 X^2.
\]

The right plot below shows the linear and quadratic model fits to the observed values with point-wise 95% confidence bands on the logit scale, and on the left is the same on the proportion scale.

```r
# fit logistic regression to create lines on plots below
# linear
glm.beetles1 <- glm(cbind(y, n - y) ~ conc, family = binomial, beetles)
# quadratic
glm.beetles2 <- glm(cbind(y, n - y) ~ conc + conc2, family = binomial, beetles)

## put model fits for two models together
beetles1 <- beetles
# put the fitted values in the data.frame
beetles1$fitted.values <- glm.beetles1$fitted.values
pred <- predict(glm.beetles1, data.frame(conc = beetles1$conc), type = "link", se.fit = TRUE)
beetles1$fit <- pred$fit
beetles1$se.fit <- pred$se.fit
# CI for fitted values
beetles1 <- within(beetles1,
{  
  fit.lower = exp(fit - 1.96 * se.fit) / (1 + exp(fit - 1.96 * se.fit))
  fit.upper = exp(fit + 1.96 * se.fit) / (1 + exp(fit + 1.96 * se.fit))
})
beetles1$modelorder <- "linear"

beetles2 <- beetles
# put the fitted values in the data.frame
beetles2$fitted.values <- glm.beetles2$fitted.values
pred <- predict(glm.beetles2, data.frame(conc = beetles2$conc, conc2 = beetles2$conc2), type = "link", se.fit = TRUE)
beetles2$fit <- pred$fit
beetles2$se.fit <- pred$se.fit
# CI for fitted values
beetles2 <- within(beetles2,
{  
  fit.lower = exp(fit - 1.96 * se.fit) / (1 + exp(fit - 1.96 * se.fit))
  fit.upper = exp(fit + 1.96 * se.fit) / (1 + exp(fit + 1.96 * se.fit))
})
beetles2$modelorder <- "quadratic"

beetles.all <- rbind(beetles1, beetles2)
beetles.all$modelorder <- factor(beetles.all$modelorder)

# plot on logit and probability scales
library(ggplot2)
```
11.4: Example: Leukemia white blood cell types

```r
p <- ggplot(beetles.all, aes(x = conc, y = p.hat, shape = rep, colour = modelorder, fill = modelorder))
# predicted curve and point-wise 95% CI
p <- p + geom_ribbon(aes(x = conc, ymin = fit.lower, ymax = fit.upper), linetype = 0, alpha = 0.1)
# fitted values
p <- p + geom_line(aes(x = conc, y = fitted.values, linetype = modelorder), size = 1)
# observed values
p <- p + geom_point(aes(y = fitted.values), size=2)
p <- p + geom_point(color = "black", size = 3, alpha = 0.5)
p <- p + labs(title = "Observed and predicted mortality, probability scale")
print(p)

library(ggplot2)
p <- ggplot(beetles.all, aes(x = conc, y = emp.logit, shape = rep, colour = modelorder, fill = modelorder))
# predicted curve and point-wise 95% CI
p <- p + geom_ribbon(aes(x = conc, ymin = fit - 1.96 * se.fit, ymax = fit + 1.96 * se.fit), linetype = 0, alpha = 0.1)
# fitted values
p <- p + geom_line(aes(x = conc, y = fit, linetype = modelorder), size = 1)
# observed values
p <- p + geom_point(aes(y = fit), size=2)
p <- p + geom_point(color = "black", size = 3, alpha = 0.5)
p <- p + labs(title = "Observed and predicted mortality, logit scale")
print(p)
```

**Observed and predicted mortality, probability scale**

**Observed and predicted mortality, logit scale**
11.5 Example: The UNM Trauma Data

The data to be analyzed here were collected on 3132 patients admitted to The University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, the attending physician recorded their age, their revised trauma score (RTS), their injury severity score (ISS), whether their injuries were blunt (i.e., the result of a car crash: BP=0) or penetrating (i.e., gunshot/knife wounds: BP=1), and whether they eventually survived their injuries (SURV=0 if not, SURV=1 if survived). Approximately 10% of patients admitted to the UNM Trauma Center eventually die from their injuries.

The ISS is an overall index of a patient’s injuries, based on the approximately 1300 injuries cataloged in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with 3 or more life threatening injuries. The ISS is the standard injury index used by trauma centers throughout the U.S. The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient’s systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS can take on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.

Champion et al. (1981) proposed a logistic regression model to estimate the probability of a patient’s survival as a function of RTS, the injury severity score ISS, and the patient’s age, which is used as a surrogate for physiologic reserve. Subsequent survival models included the binary effect BP as a means to differentiate between blunt and penetrating injuries.

We will develop a logistic model for predicting survival from ISS, AGE, BP, and RTS, and nine body regions. Data on the number of severe injuries in each of the nine body regions is also included in the database, so we will also assess whether these features have any predictive power. The following labels were used to identify the number of severe injuries in the nine regions: AS = head, BS = face, CS = neck, DS = thorax, ES = abdomen, FS = spine, GS = upper extremities, HS = lower extremities, and JS = skin.
### Example: UNM Trauma Data

```r
trauma <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch11_trauma.dat", header = TRUE)
```

#### Variables

- **surv**: survival (1 if survived, 0 if died)
- **rts**: revised trauma score (range: 0 no vital signs to 7.84 normal vital signs)
- **iss**: injury severity score (0 no injuries to 75 for 3 or more life threatening injuries)
- **bp**: blunt or penetrating injuries (e.g., car crash BP=0 vs gunshot/knife wounds BP=1)

Severe injuries: add the severe injuries 3--6 to make summary variables

```r
trauma <- within(trauma, {
  as = a3 + a4 + a5 + a6 # as = head
  bs = b3 + b4 + b5 + b6 # bs = face
  cs = c3 + c4 + c5 + c6 # cs = neck
  ds = d3 + d4 + d5 + d6 # ds = thorax
  es = e3 + e4 + e5 + e6 # es = abdomen
  fs = f3 + f4 + f5 + f6 # fs = spine
  gs = g3 + g4 + g5 + g6 # gs = upper extremities
  hs = h3 + h4 + h5 + h6 # hs = lower extremities
  js = j3 + j4 + j5 + j6 # js = skin
})
```

# keep only columns of interest

```r
names(trauma)
```

```r
## [1] "id"   "surv"  "a1"  "a2"  "a3"  "a4"  "a5"  "a6"  "b1"  "b2"
## [11] "b3"  "b4"  "b5"  "b6"  "c1"  "c2"  "c3"  "c4"  "c5"  "c6"
## [21] "d1"  "d2"  "d3"  "d4"  "d5"  "d6"  "e1"  "e2"  "e3"  "e4"
## [31] "e5"  "e6"  "f1"  "f2"  "f3"  "f4"  "f5"  "f6"  "g1"  "g2"
## [41] "g3"  "g4"  "g5"  "g6"  "h1"  "h2"  "h3"  "h4"  "h5"  "h6"
## [51] "j1"  "j2"  "j3"  "j4"  "j5"  "j6"  "iss"  "iciss" "bp"  "rts"
## [61] "age"  "prob"  "js"  "hs"  "gs"  "fs"  "es"  "ds"  "cs"  "bs"
## [71] "as"
```

```r
trauma <- subset(trauma, select = c(id, surv, as:js, iss:prob))
```

```r
head(trauma)
```

```r
## id surv as bs cs ds es fs gs hs js iss iciss bp rts age prob
## 1 1238385 1 0 0 0 1 0 0 0 0 13 0.8612883 0 7.8408 13 0.9909890
## 2 1238393 1 0 0 0 0 0 0 0 0 5 0.9421876 0 7.8408 23 0.9947165
## 3 1238898 1 0 0 0 0 2 0 0 0 13 0.7251130 0 7.8408 43 0.9947165
## 4 1239516 1 1 0 0 0 0 0 0 0 16 1.0000000 0 5.9672 17 0.9615540
## 5 1239961 1 1 0 0 0 0 0 0 0 9 0.9346634 0 4.8040 20 0.9338096
## 6 1240266 1 0 0 0 0 0 0 1 0 13 0.9004691 0 7.8408 32 0.9947165
```

```r
str(trauma)
```

I made side-by-side boxplots of the distributions of ISS, AGE, RTS, and AS through JS for the survivors and non-survivors. In several body regions the number of injuries is limited, so these boxplots are not very enlightening. Survivors tend to have lower ISS scores, tend to be slightly younger, tend to
have higher RTS scores, and tend to have fewer severe head (AS) and abdomen injuries (ES) than non-survivors. The importance of the effects individually towards predicting survival is directly related to the separation between the survivors and non-survivors scores.

```r
# Create boxplots for each variable by survival
library(reshape2)
trauma.long <- melt(trauma, id.vars = c("id", "surv", "prob"))

# Plot the data using ggplot
library(ggplot2)
p <- ggplot(trauma.long, aes(x = factor(surv), y = value))
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.1)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                        alpha = 0.75, colour = "red")
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, alpha = 0.8)
p <- p + facet_wrap(~ variable, scales = "free_y", ncol = 4)
p <- p + labs(title = "Boxplots of variables by survival")
print(p)
```
11.5.1 Selecting Predictors in the Trauma Data

The same automated methods for model building are available for the `glm()` procedure, including backward elimination, forward selection, and stepwise methods, among others. Below we perform a stepwise selection using AIC starting at the full model, starting with a full model having 13 effects: ISS, BP, RTS,
AGE, and AS–JS. Revisit Chapter 10 for more information.

In our previous logistic regression analyses, the cases in the data set were pre-aggregated into groups of observations having identical levels of the predictor variables. The numbers of cases in the success category and the group sample sizes were specified in the model statement, along with the names of the predictors. The trauma data set, which is not reproduced here, is raw data consisting of one record per patient (i.e., 3132 lines). The logistic model is fitted to data on individual cases by specifying the binary response variable (SURV) with successes and $1 - \text{SURV}$ failures with the predictors on the right-hand side of the formula. Keep in mind that we are defining the logistic model to model the success category, so we are modeling the probability of surviving.

As an aside, there are two easy ways to model the probability of dying (which we don’t do below). The first is to swap the order the response is specified in the formula: \texttt{cbind}(1 - surv, surv). The second is to convert a model for the log-odds of surviving to a model for the log-odds of dying by simply changing the sign of each regression coefficient in the model.

I only included the summary table from the backward elimination, and information on the fit of the selected model.

```r
glm.tr <- glm(cbind(surv, 1 - surv) ~ as + bs + cs + ds + es + fs + gs + hs + js + iss + rts + age + bp, family = binomial, trauma)
```

```r
# option: trace = 0 doesn't show each step of the automated selection
glm.tr.red.AIC <- step(glm.tr, direction="both", trace = 0)
```

```r
# the anova object provides a summary of the selection steps in order
glm.tr.red.AIC$anova
```

<table>
<thead>
<tr>
<th>Step</th>
<th>Df</th>
<th>Deviance Resid. Df</th>
<th>Resid. Dev</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>869.4309</td>
<td>897.4309</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>as</td>
<td>0.04911674</td>
<td>3119</td>
<td>869.4800</td>
</tr>
<tr>
<td>3</td>
<td>bs</td>
<td>0.12242766</td>
<td>3120</td>
<td>869.6024</td>
</tr>
<tr>
<td>4</td>
<td>fs</td>
<td>0.15243093</td>
<td>3121</td>
<td>869.7549</td>
</tr>
<tr>
<td>5</td>
<td>cs</td>
<td>0.78703127</td>
<td>3122</td>
<td>870.5419</td>
</tr>
<tr>
<td>6</td>
<td>ds</td>
<td>0.81690443</td>
<td>3123</td>
<td>871.3588</td>
</tr>
<tr>
<td>7</td>
<td>gs</td>
<td>1.18272567</td>
<td>3124</td>
<td>872.5415</td>
</tr>
<tr>
<td>8</td>
<td>hs</td>
<td>1.17941462</td>
<td>3125</td>
<td>873.7209</td>
</tr>
<tr>
<td>9</td>
<td>js</td>
<td>1.71204029</td>
<td>3126</td>
<td>875.4330</td>
</tr>
</tbody>
</table>
```

The final model includes effects for ES (number of severe abdominal in-
juries), ISS, RTS, AGE, and BP. All of the effects in the selected model are significant at the 5% level.

```r
summary(glm.tr.red.AIC)
```

## Call:
## glm(formula = cbind(surv, 1 - surv) ~ es + iss + rts + age +
##     bp, family = binomial, data = trauma)
##
## Deviance Residuals:
##   Min       1Q   Median       3Q      Max
## -3.1546   0.0992   0.1432   0.2316   3.4454
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.355845    0.442943  0.8030   0.4218
## es            -0.461317    0.110098 -4.1900  2.79e-05 ***
## iss           -0.056920    0.007411 -7.6800  1.59e-14 ***
## rts            0.843143    0.055339 15.2360  < 2e-16 ***
## age           -0.049706    0.005291 -9.3940  < 2e-16 ***
## bp            -0.635137    0.249597 -2.5450   0.0109 *
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1825.37 on 3131 degrees of freedom
## Residual deviance: 875.43 on 3126 degrees of freedom
## AIC: 887.43
##
## Number of Fisher Scoring iterations: 7
```

The p-value for $D$ is large, indicating no gross deficiencies with the selected model.

```r
# Test residual deviance for lack-of-fit (if > 0.10, little-to-no lack-of-fit)
dev.p.val <- 1 - pchisq(glm.tr.red.AIC$deviance, glm.tr.red.AIC$df.residual)
```

`dev.p.val`  # [1] 1

Letting $p$ be the probability of survival, the estimated survival probability is given by

$$
\log \left( \frac{\hat{p}}{1 - \hat{p}} \right) = 0.3558 - 0.4613 \text{ES} - 0.6351 \text{BP} - 0.0569 \text{ISS} + 0.8431 \text{RTS} - 0.0497 \text{AGE}.
$$

Let us interpret the sign of the coefficients, and the odds ratios, in terms of the
impact that individual predictors have on the survival probability.

```r
## coefficients and 95% CI
cbind(OR = coef(glm.tr.red.AIC), confint(glm.tr.red.AIC))
## Waiting for profiling to be done...
## OR  2.5 %     97.5 %
## (Intercept)  0.35584499 -0.51977300  1.21869015
## es          -0.46131679 -0.67603693 -0.24307991
## iss         -0.05691973 -0.07159539 -0.04249502
## rts         0.84314317  0.73817886  0.95531089
## age         -0.04970641 -0.06020882 -0.03943822
## bp          -0.63513735 -1.12051508 -0.14005288

## odds ratios and 95% CI
exp(cbind(OR = coef(glm.tr.red.AIC), confint(glm.tr.red.AIC)))
## Waiting for profiling to be done...
## OR  2.5 %     97.5 %
## (Intercept)  1.4273863  0.5946555  3.3827539
## es          0.6304529  0.5086287  0.7842088
## iss         0.9446699  0.9309075  0.9583952
## rts         2.3236592  2.0921220  2.5994786
## age         0.9515087  0.9415679  0.9613293
## bp          0.5298627  0.3261118  0.8693123
```

### 11.5.2 Checking Predictions for the Trauma Model

To assess the ability of the selected model to accurately predict survival, assume that the two types of errors (a false positive prediction of survival, and a false negative prediction) have equal costs. Under this assumption, the optimal classification rule is to predict survival for an individual with an estimated survival probability of 0.50 or larger.

In the below script, a table is generated that gives classifications for cutoff probabilities `thresh = 0.0, 0.1, \ldots, 1.0` based on the selected model. While I do not do this below, it is common to use the `jackknife` method for assessing classification accuracy. To implement the jackknife method, each observation is temporarily held out and the selected model is fitted to the remaining (3131) cases. This leads to an estimated survival probability for the case, which is compared to the actual result. The results are summarized in terms of total number of correct and incorrect classifications for each possible outcome. In the table below, the columns labeled `Event` (a survival) and `Non-Event` (a death)
refer to the classification of observations, and not to the actual outcomes. The columns labeled **Correct** and **Incorrect** identify whether the classifications are accurate.

```r
# thresholds for classification given model proportion predictions for each observation
thresh <- seq(0,1,by=0.1)

# predicted probabilities
Yhat <- fitted(glm.tr.red.AIC)

# Name: lower (0) = NonEvent, higher (1) = Event
YObs <- cut(trauma$surv, breaks = c(-Inf, mean(trauma$surv), Inf)
  , labels = c("NonEvent", "Event"))

classify.table <- data.frame(Thresh = rep(NA, length(thresh))
  , Cor.Event = rep(NA, length(thresh))
  , Cor.NonEv = rep(NA, length(thresh))
  , Inc.Event = rep(NA, length(thresh))
  , Inc.NonEv = rep(NA, length(thresh))
  , Cor.All = rep(NA, length(thresh))
  , Sens = rep(NA, length(thresh))
  , Spec = rep(NA, length(thresh))
  , Fal.P = rep(NA, length(thresh))
  , Fal.N = rep(NA, length(thresh)))

for (i.thresh in 1:length(thresh)) {
  # choose a threshold for dichotomizing according to predicted probability
  YhatPred <- cut(Yhat, breaks = c(-Inf, thresh[i.thresh], Inf)
    , labels = c("NonEvent", "Event"))

  # contingency table and marginal sums
  cTab <- table(YhatPred, YObs)
  addmargins(cTab)

  # Classification Table
  classify.table$Thresh [i.thresh] <- thresh[i.thresh]
  classify.table$Cor.Event[i.thresh] <- cTab[2,2]
  classify.table$Cor.NonEv[i.thresh] <- cTab[1,1]
  classify.table$Inc.Event[i.thresh] <- cTab[2,1]
  classify.table$Inc.NonEv[i.thresh] <- cTab[1,2]
  classify.table$Cor.All [i.thresh] <- 100 * sum(diag(cTab)) / sum(cTab)
  classify.table$Sens [i.thresh] <- 100 * cTab[2,2] / sum(cTab[,2])
  classify.table$Spec [i.thresh] <- 100 * cTab[1,1] / sum(cTab[,1])
  classify.table$Fal.P [i.thresh] <- 100 * cTab[2,1] / sum(cTab[2,])
  classify.table$Fal.N [i.thresh] <- 100 * cTab[1,2] / sum(cTab[1,])
}

round(classify.table, 1)
```

```r
## Thresh Cor.Event Cor.NonEv Inc.Event Inc.NonEv Cor.All Sens Spec Fal.P Fal.N
```
The data set has 2865 survivors and 267 people that died. Using a 0.50 cut-off, 2825 of the survivors would be correctly identified, and 40 misclassified. Similarly, 139 of the patients that died would be correctly classified and 128 would not. The overall percentage of cases correctly classified is \((2825 + 139)/3132 = 94.6\%\). The sensitivity is the percentage of survivors that are correctly classified, \(2825/(2825 + 40) = 98.6\%\). The specificity is the percentage of patients that died that are correctly classified, \(139/(139 + 128) = 52.1\%\). The false positive rate, which is the % of those predicted to survive that did not, is \(128/(128 + 2825) = 4.3\%\). The false negative rate, which is the % of those predicted to die that did not is \(40/(40 + 139) = 22.5\%\).

The misclassification rate seems small, but you should remember that approximately 10% of patients admitted to UNM eventually die from their injuries. Given this historical information only, you could achieve a 10% misclassification rate by completely ignoring the data and classifying each admitted patient as a
survivor. Using the data reduces the misclassification rate by about 50% (from 10% down to 4.4%), which is an important reduction in this problem.

11.6 Historical Example: O-Ring Data

The table below presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-Challenger space shuttle launches. Temperature at lift-off and O-ring joint pressure are predictors. The binary version of the data views each flight as an independent trial. The result of a trial \( y \) is a 1 if at least one field O-ring failure occurred on the flight and a 0 if all six O-rings functioned properly. A regression for the binary data would model the probability that any O-ring failed as a function of temperature and pressure. The binomial version of these data is also presented. It views the six field O-rings on each flight as independent trials. A regression for the binomial data would model the probability that a particular O-ring on a given flight failed as a function of temperature and pressure. The two regressions model different probabilities. The Challenger explosion occurred during a takeoff at 31 degrees Fahrenheit.

### Example: Shuttle O-ring data

```r
case flight y six temp pressure
1 1 14 1 2 53 50
2 2 9 1 1 57 50
3 3 23 1 1 58 200
4 4 10 1 1 63 50
5 5 1 0 0 66 200
6 6 5 0 0 67 50
7 7 13 0 0 67 200
8 8 15 0 0 67 50
9 9 4 0 0 68 200
10 10 3 0 0 69 200
11 11 8 0 0 70 50
12 12 17 0 0 70 200
13 13 2 1 1 70 200
14 14 11 1 1 70 200
15 15 6 0 0 72 200
16 16 7 0 0 73 200
17 17 16 0 0 75 100
18 18 21 1 2 75 200
19 19 19 0 0 76 200
20 20 22 0 0 76 200
21 21 12 0 0 78 200
22 22 20 0 0 79 200
23 23 18 0 0 81 200
```

UNM, Stat 428/528 ADA2
The regression model for the binomial data (i.e., six trials per launch) is
suspect on substantive grounds. The binomial model makes no allowance for
an effect due to having the six O-rings on the same flight. If these data were
measurements rather than counts, they would almost certainly be treated as
dependent repeated measures. If interest is focused on whether one or more
O-rings failed, the simplest, most direct data are the binary data.

Consider fitting a logistic regression model using temperature and pressure
as predictors. Let \( p_i \) be the probability that any O-ring fails in case \( i \) and model
this as

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 \text{Pressure}_i.
\]

Logistic histogram plots of the data show a clear marginal relationship of
failure with temp but not with pressure. We still need to assess the model with
both variables together.

We fit the logistic model below using \( Y = 1 \) if at least one O-ring failed,
and 0 otherwise. We are modelling the chance of one or more O-ring failures
as a function of temperature and pressure.

The \( D \) goodness-of-fit statistic suggest no gross deviations from the model.
Furthermore, the test of $H_0 : \beta_1 = \beta_2 = 0$ (no regression effects) based on the Wald test has a p-value of 0.1, which suggests that neither temperature or pressure, or both, are useful predictors of the probability of O-ring failure. The z-test test p-values for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually are 0.037 and 0.576, respectively, which indicates pressure is not important (when added last to the model), but that temperature is important. This conclusion might be anticipated by looking at data plots above.

```r
glm.sh <- glm(cbind(y, 1 - y) ~ temp + pressure, family = binomial, shuttle)
# Test residual deviance for lack-of-fit (if > 0.10, little-to-no lack-of-fit)
dev.p.val <- 1 - pchisq(glm.sh$deviance, glm.sh$df.residual)
dev.p.val
## [1] 0.4589415

# Testing Global Null Hypothesis
library(aod)
coef(glm.sh)
## (Intercept) temp pressure
## 16.385319489 -0.263404073 0.005177602

# specify which coefficients to test = 0 (Terms = 2:3 is for terms 2 and 3)
wald.test(b = coef(glm.sh), Sigma = vcov(glm.sh), Terms = 2:3)
## Wald test:
## ********
## # Chi-squared test:
## # X2 = 4.6, df = 2, P(> X2) = 0.1

# Model summary
summary(glm.sh)
##
## Call:
## glm(formula = cbind(y, 1 - y) ~ temp + pressure, family = binomial,
##     data = shuttle)
##
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
##    -1.1928 -0.7879 -0.3789  0.4172  2.2031
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 16.385319  8.027474  2.041 0.0412 *
## temp     -0.263404  0.126371 -2.084 0.0371 *
## pressure  0.005178  0.009257  0.559 0.5760
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
A reasonable next step would be to refit the model, after omitting pressure as a predictor. The target model is now

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Temp}_i.$$
results for two predictor model. The model estimates the log-odds of (at least one) O-ring failure to be

$$\log \left( \frac{\tilde{p}}{1 - \tilde{p}} \right) = 15.043 - 0.2322 \text{ Temp}. $$

The estimated probability of (at least one) O-ring failure is

$$\tilde{p} = \frac{\exp(15.043 - 0.2322 \text{ Temp})}{1 + \exp(15.043 - 0.2322 \text{ Temp})}. $$

This is an decreasing function of temperature.

The *Challenger* was set to launch on a morning where the predicted temperature at lift-off was 31 degrees. Given that temperature appears to affect the probability of O-ring failure (a point that NASA missed), what can/should we say about the potential for O-ring failure?

Clearly, the launch temperature is outside the region for which data were available. Thus, we really have no prior information about what is likely to occur. If we assume the logistic model holds, and we can extrapolate back to 31 degrees, what is the estimated probability of O-ring failure?

The following gives the answer this question. I augmented the original data set to obtain predicted probabilities for temperatures not observed in the data set. Note that the fitted model to data with missing observations gives the same model fit because `glm()` excludes observations with missing values. Predictions are then made for all temperatures in the dataset and the resulting table and plot are reported.

```r
# append values to dataset for which we wish to make predictions
shuttle.pred <- data.frame(  case = rep(NA, 5),
, flight = rep(NA, 5),
, y = rep(NA, 5),
, six = rep(NA, 5),
, temp = c(31, 35, 40, 45, 50) # temp values to predict
, pressure = rep(NA, 5)
)
shuttle <- rbind(shuttle.pred, shuttle)
# fit model
glm.sh <- glm(cbind(y, 1 - y) ~ temp, family = binomial, shuttle)
# Note: same model fit as before since glm() does not use missing values
```

UNM, Stat 428/528 ADA2
```
round(summary(glm.sh)$coefficients, 3)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   15.043     7.379  2.039  0.041
## temp          -0.232     0.108 -2.145  0.032

# put the fitted values in the data.frame
shuttle$fitted.values <- c(rep(NA, 5), glm.sh$fitted.values)
# predict() uses all the temp values in dataset, including appended values
pred <- predict(glm.sh, data.frame(temp = shuttle$temp), type = "link", se.fit = TRUE)
shuttle$fit <- pred$fit
shuttle$se.fit <- pred$se.fit
# CI for fitted values
shuttle <- within(shuttle,
{  
    # added "fitted" to make predictions at appended temp values
    fitted = exp(fit) / (1 + exp(fit))
    fit.lower = exp(fit - 1.96 * se.fit) / (1 + exp(fit - 1.96 * se.fit))
    fit.upper = exp(fit + 1.96 * se.fit) / (1 + exp(fit + 1.96 * se.fit))
})
```
p <- p + geom_point(aes(y = fitted.values), size=2, colour="red")
# observed values
p <- p + geom_point(size = 2)

p <- p + ylab("Probability")
p <- p + labs(title = "Observed events and predicted probability of 1+ O-ring failures")
print(p)

## Warning: Removed 5 rows containing missing values (geom_point).
## Warning: Removed 5 rows containing missing values (geom_point).