Logistic Regression Continued

Logistic regression with a proportional response variable and a single continuous explanatory variable

Proportional data is clearly bounded between 0 and 1, the errors are not normally distributed and the variance is not constant, therefore linear regression is unsuitable for modeling this type of data. We can use logistic regression, once again, we have just one continuous explanatory variable x_1 so the logistic regression model will be of the form:

$$\log\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 x_1$$

where p_i represents the proportion of observations observed to respond in a given way.

Example 2

Let's model the Challenger data again, however this time instead of modelling whether O-ring damage had occurred to at least one O-ring or not (a binary response variable), we will model the proportion of damaged O-rings.

Once again, the explanatory variable, x_i, is the temperature in ^oF at the time of launch. Table 1 below shows the data from the 23 shuttle missions prior to the Challenger launch.

Table 1

Temperature ^oF (x_i)	Number of damaged O-rings	Proportion of damaged O-rings (y_i)
53	5	$\frac{5}{6}$
57	1	$\frac{1}{6}$
58	1	$\frac{1}{6}$
63	1	$\frac{1}{6}$
66	0	0
67	0	0
67	0	0
67	0	0
68	0	0
69	0	0
70	1	$\frac{1}{6}$
70	0	0
70	1	$\frac{1}{6}$
70	0	0
72	0	0
73	0	0
75	0	0
75	1	$\frac{1}{6}$
76	0	0
76	0	0
79	0	0
79	0	0
81	0	0

Figure 1 shows that O-ring damage is more likely to occur at low temperatures.

Fitting a logistic regression model to the data gives the following output:

Table 2

	Estimate	Std Error	z value	$Pr\,(>z)$
Intercept (β_0)	11.55	1.35	8.87	< 0.001
Temperature (β_1)	-0.22	0.02	-9.96	< 0.001

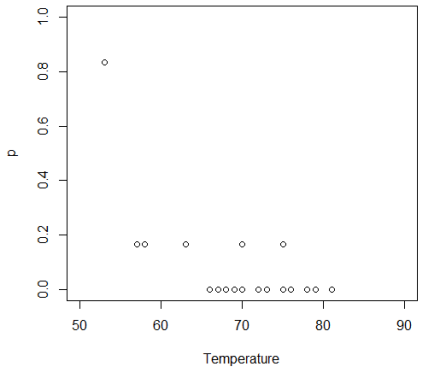

Figure 1: Figure 1

Using the estimates for β_0 and $\beta_{1,}$ the model is:

$$\begin{aligned} \log\left(\frac{p_i}{(1-p_i)}\right) &= 11.55 - 0.22x \\ \frac{p_i}{(1-p_i)} &= e^{11.55-0.22x} \end{aligned}$$

A plot of the fitted model is shown in Figure 2 below.

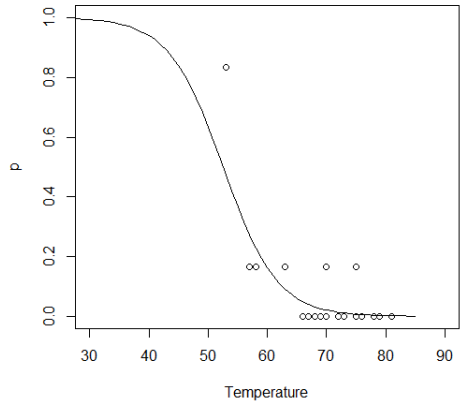

Figure 2

Example 3

More logistic regression with a proportional response variable and a single continuous explanatory variable

We wish to find out whether the sex ratio in a particular species of insects is density dependent. We define the response variable y_i to be the proportion of males and the explanatory variable, x_i, is the population density at the time of reproduction. The proportion of males is then $p_i = \frac{\text{males}}{n_i}$ where n_i is the population density of observation i. The data for the experiment is shown in Table 3 below.

Table 3

Population Density (x_i)	Females	Males	p_i
1	1	0	0
4	3	1	0.25
10	7	3	0.3
22	18	4	0.81
55	22	33	0.6
121	41	80	0.66
210	52	158	0.75
444	79	365	0.82

The data from Table 3 is shown in Fig. 3 below.

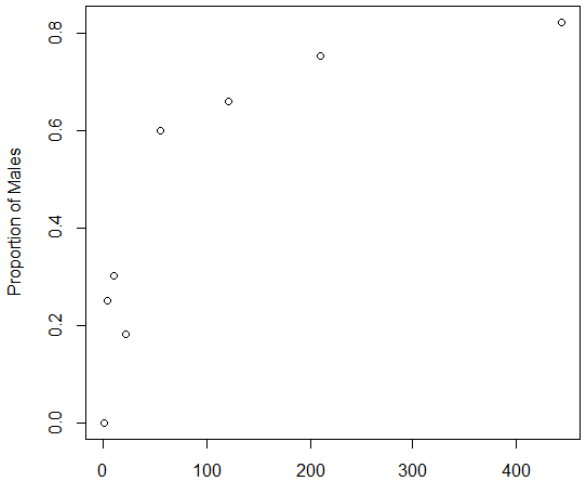

Figure 3

It seems that increasing population density leads to an increase in the proportion of males in the population. If we try to fit a linear model to this data (i.e. a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$), we run into the problems mentioned previously.
Instead, we will use the logistic model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x$ which allows for non constant variance and does not predict outcomes outside the interval $[0, 1]$. However as we see in Figure 4, the model does not fit the data well.

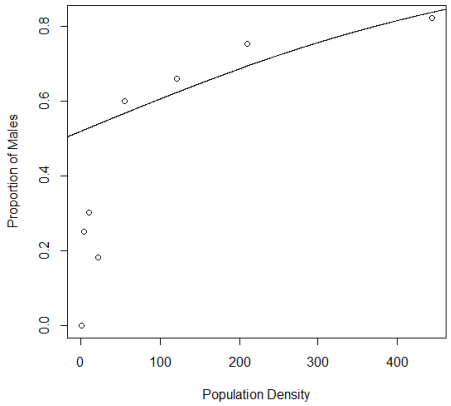

Figure 4

Figure 4 shows that the logistic model relating the proportion of males to the population density is not a good fit. The model does not fit the data with low population densities well. If we look at the scale of the explanatory variable we see that the small values are close together and the large values spread apart. A log transformation to the explanatory variable improves the fit of this model see Figure 5.

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

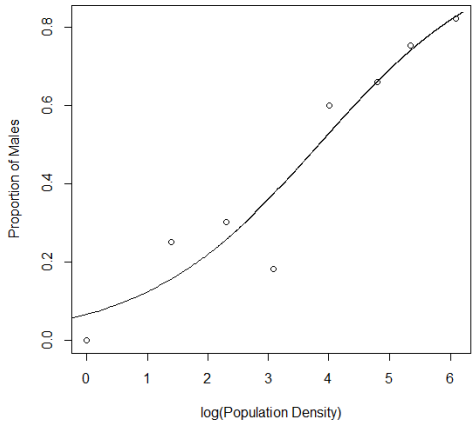

Figure 5

Figure 5 shows that the logistic model relating the proportion of males to the log of the population density is a good fit. There is a positive relationship between the proportion of males and the population density. The summary of the model fitted in R is shown in Table 4 below.

Table 4

	Estimate	Std Error	z value	$Pr\,(>z)$
Intercept (β_0)	-2.66	0.49	-5.45	5×10^{-8}
Temperature (β_1)	0.69	0.09	7.66	2×10^{-14}

Using the estimates for β_0 and $\beta_{1,}$ the model is:

$$\log\left(\frac{p_i}{(1-p_i)}\right) = -2.66 + 0.69\log(x)$$

If there was no relationship between the proportion of males and the log of the population density then we would expect the coefficient of the x predictor (β_1) to be 0.

How to interpret the model The positive value of the coefficient β_1 indicates that the proportion of males was greater at high population densities. A unit increase in the log of the population density will **increase** the **log odds** of the proportion of males by 0.69. Alternatively we can say that a unit increase in the log of the population density will **increase** the **odds** of the proportion of males by a factor of $e^{(0.69)}$.

Model Checking

As in the case for linear models we can evaluate the fit of the model by examining the residuals. To evaluate the fit of a logistic model we examine the **deviance residuals** which are defined as:

$$d_i = \pm\sqrt{2y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)}$$

where y_i represents the observed values (the number of "successes"), $\widehat{y}_i$ represents the fitted values and n_i represents the number of trials.
The sign is: $+$ if $y_i > \widehat{y}_1$ and is $-$ if $y_i < \widehat{y}_i$.
Figure 6 shows a plot of the residual values (vs) fitted values for Example.2. Observations with a deviance residual greater than two may indicate lack of fit.

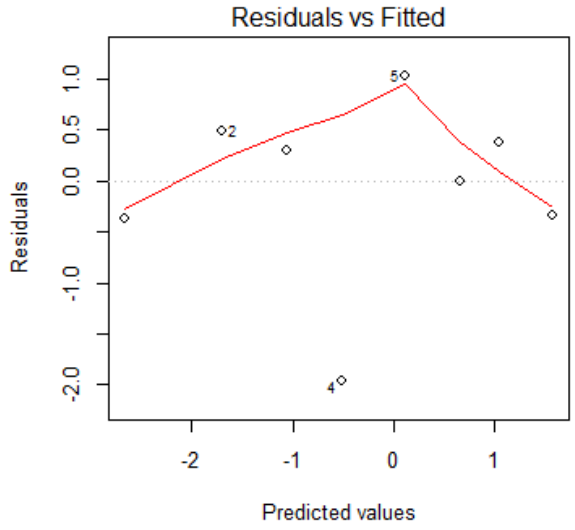

Figure 6

Goodness of fit

In linear regression models, goodness of fit is measured using R^2. In logistic regression analysis, deviance (D) can sometimes be used as a measure of goodness of fit. Squaring the deviance residuals, d_i and summing over all observations yields the deviance statistic:

$$D = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)$$

This quantity is calculated by comparing a given model with the saturated model – a model with a theoretically perfect fit. Smaller values of D indicate a better fit as the fitted model deviates less from the saturated model. If n_i is sufficiently large ($n_i > 5$ has been suggested) then the residual deviance, D, is approximately chi-square distributed. When the residual deviance is tested using a chi-square distribution, nonsignificant chi-square values indicate very little unexplained variance and thus, good model fit. Conversely, a significant chi-square value indicates that a significant amount of the variance is unexplained.
For the Challenger example, since y_i was a binary response variable (either 0, or 1) this meant that $n_i = 1$ and so the residual deviance did not follow a chi-squared distribution therefore the residual deviance D is not suitable for testing goodness of fit.

For the insect sex ratio example, the full R output was:

Deviance Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
-1.9697	*-0.3411*	*0.1499*	*0.4019*	*1.0372*

Coefficients:

	Estimate	*Std.Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-2.65927*	*0.48758*	*-5.454*	*4.92e-08 ****
log(sexratio$density)	*0.69410*	*0.09056*	*7.665*	*1.80e-14 ****

—

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 71.1593 on 7 degrees of freedom
Residual deviance: 5.6739 on 6 degrees of freedom

Null deviance is the deviance of a model that contains only the intercept, that is, describes a fixed probability, it is often denoted by D_0.
Residual deviance is the deviance statistic D which corresponds to the residual sum of squares in ordinary regression analysis.

To test the goodness of fit for this model we can test the residual deviance, D. For the insect sex ratio example, we expect $D \tilde{} \chi^2_{N-k}$. Since there were 8 observations $N = 8$ and our model estimates just two parameters (β_0 and β_1), $k = 2$, therefore, we expect $D \tilde{} \chi^2_6$.

Using R, to calculate the p value from the χ^2_6 distribution, the probability of obtaining $D = 5.6739$ with 6 degrees of freedom is 0.46. The p-value of 0.46 indicates that the model does not differ significantly from the saturated (perfect) model and therefore the model with just two parameters is an adequate fit.

The null deviance represents the difference between a model with only the intercept i.e. no explanatory variables and the saturated model. In this respect, the null model provides a baseline upon which to compare models. Comparison of the model deviance to the null deviance can be used for model selection. This comparison asks whether the model including the explanatory variables fits significantly better than a model with just an intercept (i.e., a null model). The test statistic is the difference between the residual deviance for the model with explanatory variables and the null model. The test statistic is has a chi-square distribution with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of explanatory variables in the model). It is important to note that this comparison is valid for data where n_i is small so can be used for data where the response variable has binary outcomes.
For the insect sex ratio example, the null deviance was 71.1593 and the model deviance was 5.6739, therefore to test whether the fitted model explained a significant amount of the variation in the data, we test $71.1593 - 5.6739 = 65.4854$ using the χ^2_1 distribution. The p-value obtained from R was of 5×10^{-16} which indicates that the fitted model explained a significant amount of the variation in the data compared to the null model.

An alternative approach to assessing goodness of fit is based on prediction errors. Suppose we were to use the fitted model to predict 'success' if the fitted probability exceeds 0.5 and 'failure' otherwise. We could then crosstabulate the observed and predicted responses, and calculate the proportion of cases predicted correctly. One problem with this approach is that a model that fits the data may not necessarily predict well, since this depends on how predictable the outcome is. If prediction was the main objective of the analysis, however, the proportion classified correctly would a suitable criterion for model comparison.

Overdispersion

In logistic regression analysis, it is assumed that the response variable follows a binomial distribution, i.e. $y_i \tilde{} Bin(n_i, p_i)$ where $E[y_i] = n_i p_i$ and $Var[y_i] = n_i p_i (1 - p_i)$. If the variance observed is larger than $n_i p_i (1 - p_i)$ then we have **overdispersion**. If the residual deviance D is greater than the residual degrees of freedom we may have overdispersion,

however the large residual deviance could also be due to other factors such as a misspecified model.