An Introduction to Simple Logistic Regression

In statistics, logistic regression is used to model the probability of dichotomous or binary events such as whether a person with the Coronavirus displays symptoms. (An infected person who does not develop symptoms is said to be asymptomatic, and may still be able to transmit the virus to others unwittingly.)

In logistic regression, the independent variables are of the same types allowable in multiple regression, i.e., either quantitative or categorical, where categorical variables may be represented by dummy variables. Logistic regression may be used to predict the risk of developing a disease, such as diabetes, cancer, or coronary heart disease, based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.). Another example might be to predict whether an independent voter will vote Republican in a presidential election based on their age, income, sex, race, state of residence, etc.

Odds and Log-Odds

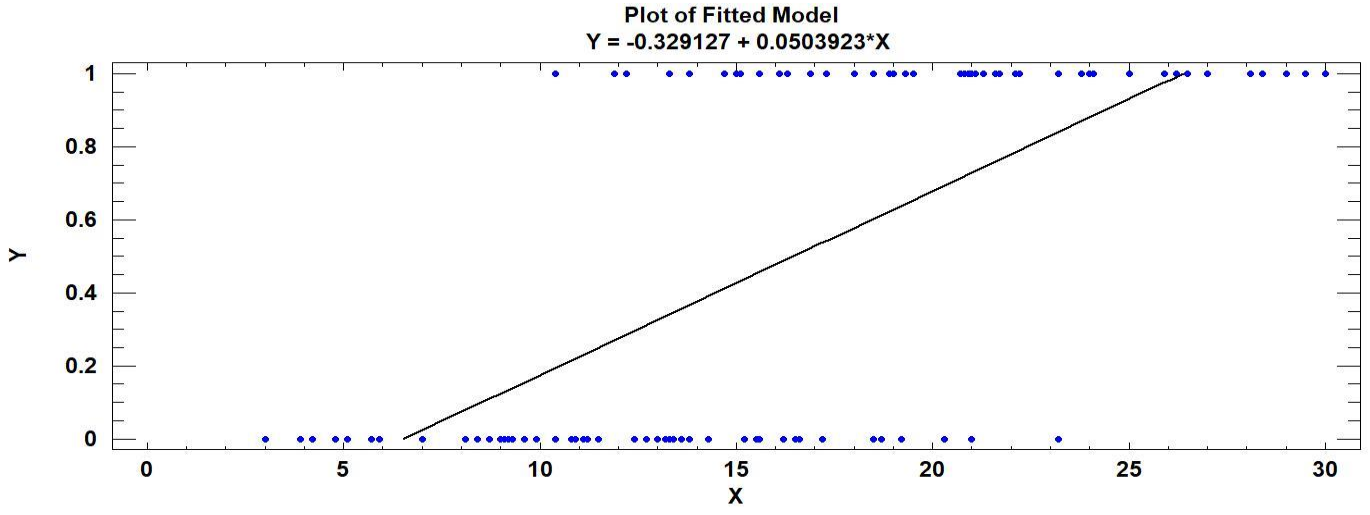
You may have seen odds discussed in a previous class or you may be familiar with the concept from sports or gambling. The odds (in favor) of an event is the ratio of the probability that the event will happen to the probability that the event will not happen. Mathematically, this is a Bernoulli trial, as it has exactly two outcomes (the event either occurs or it doesn't), typically represented as 1 (if the event occurs) and 0 (if the event does not occur). If p is the probability that event E occurs, then the odds that E occurs is defined as $\frac{p}{1-p}$.

Similarly, the odds that E does *not* occur is $\frac{1-p}{p}$.

Example: For a well shuffled deck, the odds of drawing a heart are $\frac{P(\text{Heart})}{P(\text{non Heart})} = \frac{13/52}{39/52} = \frac{1}{3}$.

The Need for Logistic Regression

To demonstrate the need for a new form of regression when the response is binary, consider the result of regressing Y on X for the data in the file *Simple Logistic Regression*. Recall that a regression line estimates the mean value of $\boldsymbol{Y}$ at X. The expected value of a Bernoulli random variable is just the probability it equals 1, so the regression line below represents a probability. However, the plot of the fitted model below predicts negative probabilities when X is less than 7 and predicts probabilities greater than 1 when X is greater than 27.



Nonlinear models have been developed for a fitting binary response variable, and the most popular of these uses a logistic curve (rather than a line) for the fit. (Hence the term logistic regression.) You may be familiar with logistic functions from another course, such as Differential Equations, where they are used to model population growth. They take the form $P(t) = \dfrac{Ce^{kt}}{1 + Be^{kt}}$. (**Note:** for $k > 0$, as $t \to \infty$, $P(t) \to C / B$, and as $t \to -\infty$, $P(t) \to 0$.) The graphs of logistic functions have a characteristic *S*-shape to them for $k > 0$, and are reverse-*S* for $k < 0$. The logistic model for a single predictor (independent) variable is given below.

Let $(X_i, \boldsymbol{Y}_i)$ be the i^{th} observation (out of *n*) on the joint probability distribution of ($\boldsymbol{X}$, $\boldsymbol{Y}$), where X_i is fixed, then

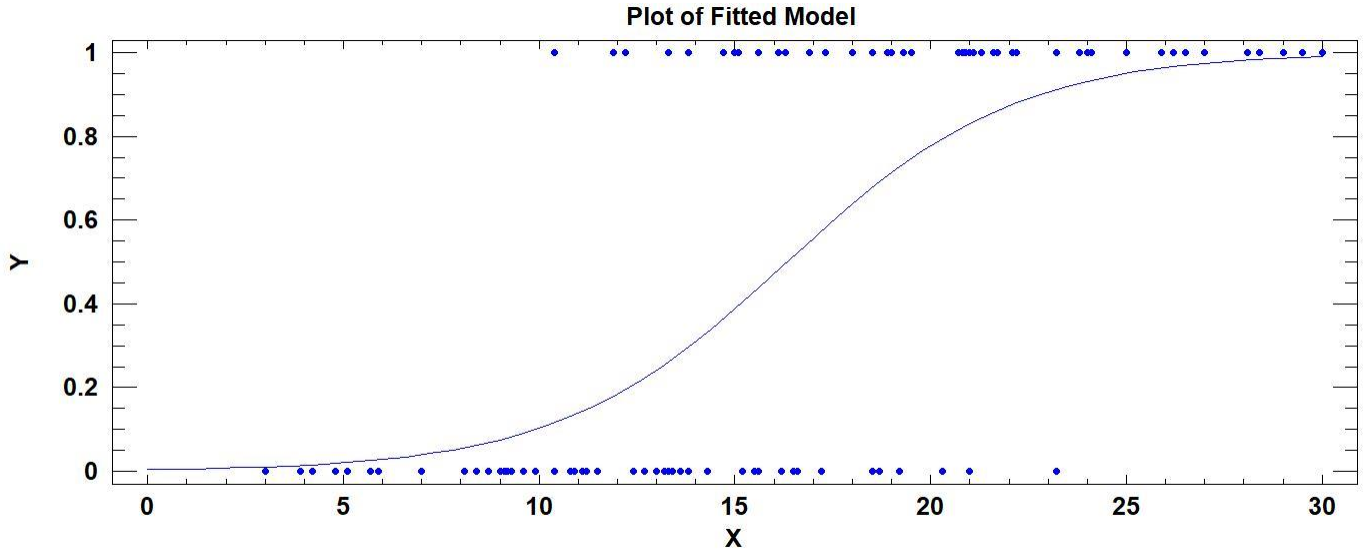
$$E(\boldsymbol{Y}_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

Note: the linear expression $\beta_0 + \beta_1 X_i$ from the simple linear regression model is preserved in the exponents.

Logistic Regression in Statgraphics

Logistic regression is accessed in Statgraphics by following *Relate* → *Attribute Data* → *Logistic Regression*. You'll have the option to enter independent variables as quantitative or categorical factors. For the *Simple Logistic Regression* data, the predictor is quantitative. As you may have guessed, I am using this dataset to demonstrate logistic regression with a single independent variable, comparable to simple linear regression.

Example: For the data set *Simple Logistic Regression* in the previous example, the fitted logistic curve appears below.



You may have noticed that the logistic model $E(\boldsymbol{Y}_i) = \dfrac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} = f(X_i)$ is not of the form $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$, seen in linear regression . There are several reasons for this.

For any value X_i, the response variable $\boldsymbol{Y}_i$ is Bernoulli, i.e., $\boldsymbol{Y}_i \sim Bernoulli(p_i)$, and from a property of Bernoulli distributions $E(\boldsymbol{Y}_i) = p_i$. Thus, the value $f(X_i)$ fitted by the regression function f estimates the probability that the response variable equals 1. Put another way, $f(X_i)$ estimates $P(\boldsymbol{Y} = 1 | \boldsymbol{X} = X_i)$.

The previous paragraph implies that the error $\boldsymbol{\varepsilon}_i = \boldsymbol{Y}_i - f(X_i)$ cannot be normally distributed with equal variance for *any* choice of regression function $f(X)$ because,

- $\boldsymbol{\varepsilon}_i$ has only *two* possible values: $\boldsymbol{\varepsilon}_i = 1 - f(X_i)$ and $\boldsymbol{\varepsilon}_i = -f(X_i)$ because $\boldsymbol{Y}_i$ has only two possible values, 1 and 0. Thus, the error is definitely *not* normally distributed.
- The variance of the error, $\sigma^2_{\boldsymbol{\varepsilon}_i} = p_i(1 - p_i) = f(X_i)(1 - f(X_i))$, is a function of X_i, i.e., not constant.

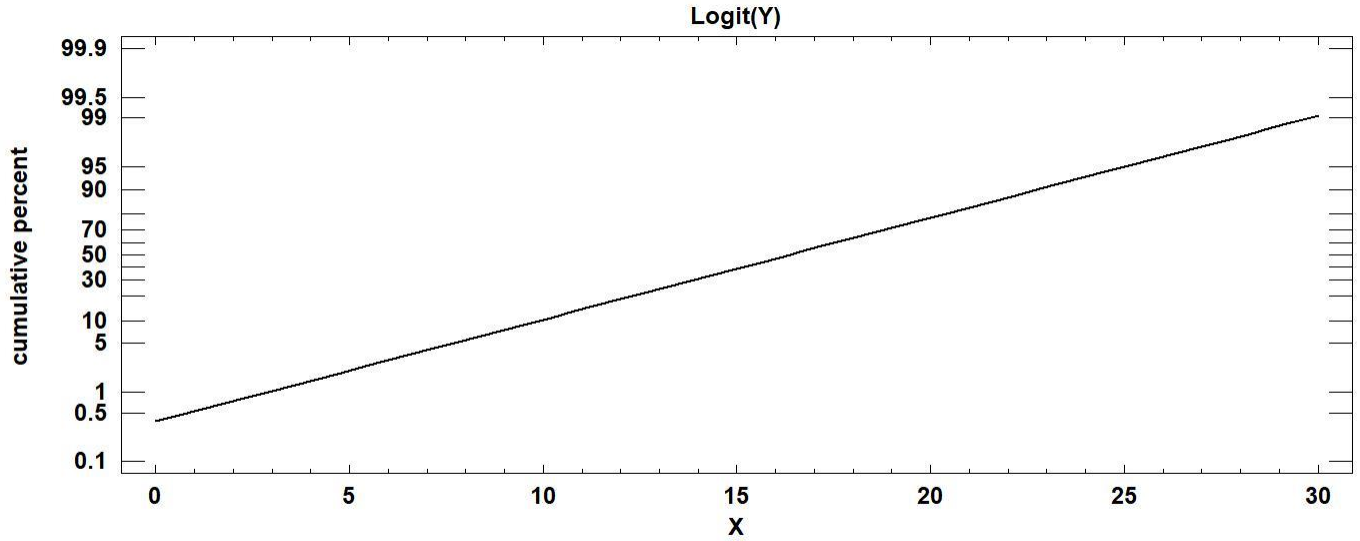
In addition, the regression function $f(X)$ is not of the form $\beta_0 + \beta_1 X$, which classifies it as nonlinear. However, as we'll see next, $E(\boldsymbol{Y}_i) = f(X_i)$ is *linked* to the linear predictor $\beta_0 + \beta_1 X_i$ through a function called, appropriately, a Link Function. This makes simple logistic regression an example of a Generalized Linear Model, abbreviated GLM.

The Logit (Link) Function

Beginning with the logistic model $E(\boldsymbol{Y}) = p = \dfrac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$, the following are connected,

- $P(\boldsymbol{Y} = 1 | X) = p = \dfrac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$, which is the (true) mean value of the response in the logistic model.
- Odds that $\boldsymbol{Y} = 1$: $\dfrac{p}{1 - p} = e^{\beta_0 + \beta_1 X}$, which explains why I began by defining the odds for an event.
- Log-Odds that $\boldsymbol{Y} = 1$: $ln\left(\dfrac{p}{1 - p}\right) = \beta_0 + \beta_1 X$, which is linear in the regression coefficients.

The Logit Function: In logistic regression we define the Logit Function $ln\left(\dfrac{p}{1 - p}\right) = \beta_0 + \beta_1 X$, which is the log-odds that $\boldsymbol{Y} = 1$ given X. Below is Statgraphics' plot of the (estimated) logit for our running example.



Estimating β_0 and β_1

The procedure we used to estimate β_0 and β_1 in simple linear regression, called ordinary least squares (OLS), cannot be applied because of the lack of constant variance. One of two methods is usually used to estimate the regression coefficients, weighted least squares (WLS) or maximum likelihood estimation (MLE). Since exploring these two methods is beyond the scope of this course, we will let software do the work for us. Statgraphics defaults to the maximum likelihood estimates of the coefficients, which are the most common.

Estimating the Mean Response and the Logit Function

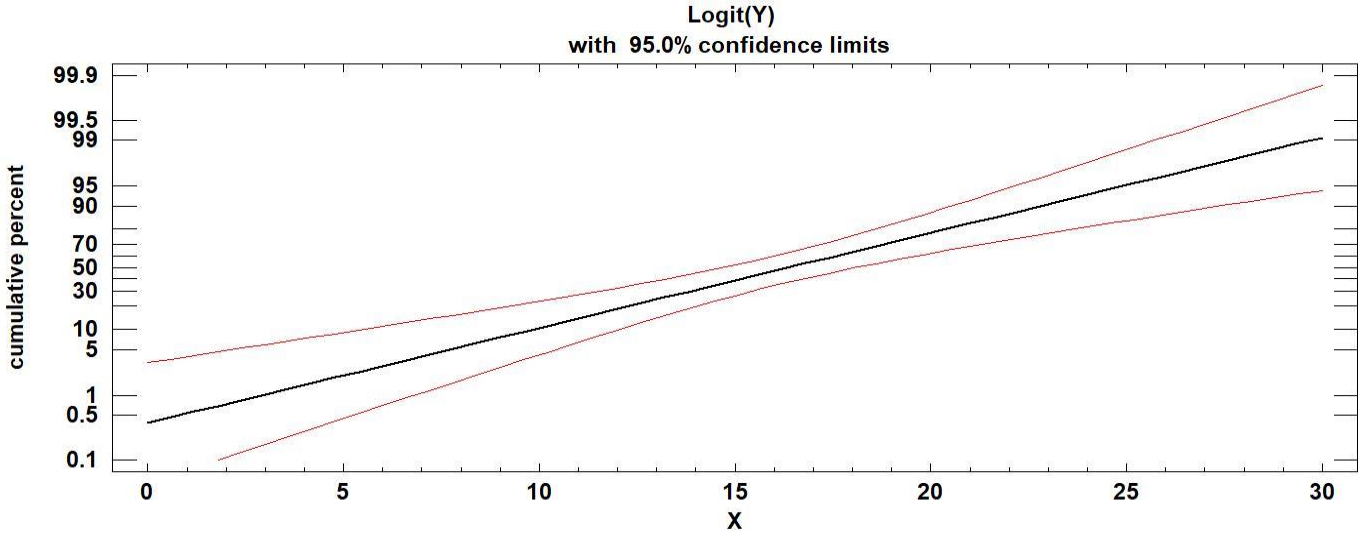
Once β_0 and β_1 have been estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$, the mean of $\mathbf{Y}_i$, $E(\mathbf{Y}_i) = p_i$, is estimated by

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}}$$

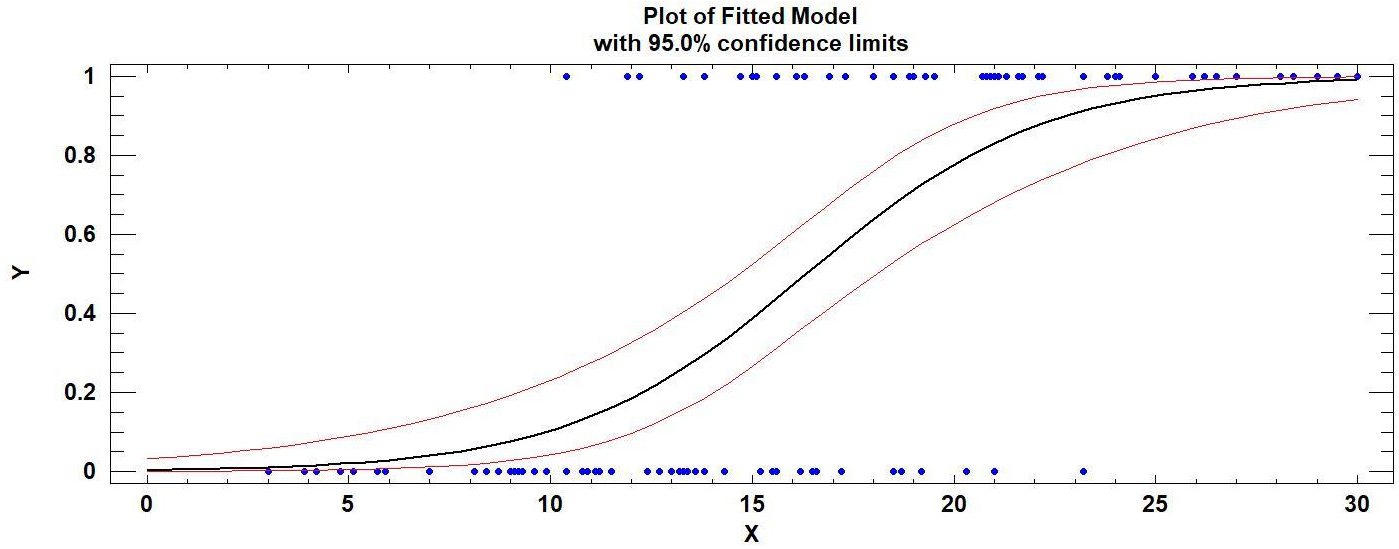
Confidence Intervals and Confidence Bands

One of the advantages of working with log-odds is that, unlike $\mathbf{Y}_i$ which is binary (either 0 or 1), and $E(\mathbf{Y}_i) = p_i$ which is between 0 and 1, the log-odds ranges from $-\infty$ to ∞ which can be useful for fitting error. If we hypothesize the model $\hat{\pi} = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X$, then maximum likelihood confidence intervals can be created for $E(\hat{\pi}) = \pi$ given any appropriate value of *X*. Confidence bands for π can also be constructed over the range of *X* appearing in the data. Finally, confidence intervals and bands for π can be transformed into intervals and bands for $p = E(\mathbf{Y})$ over the range of *X* by applying the logistic function to the lower and upper limits for π. The results for our running example are displayed below, starting with 95% confidence bands for π and followed by 95% confidence bands for $E(\mathbf{Y})$.

Discussion: Unlike confidence intervals created in least squares, confidence intervals created using maximum likelihood don't place restrictive assumptions on the form of the error.



95% Confidence Bands for $\pi = ln\left(\frac{E(\mathbf{Y})}{1-E(\mathbf{Y})}\right) = \beta_0 + \beta_1 X$



95% Confidence Bands for $E(\boldsymbol{Y})$

Discussion: While the original confidence bands around $\hat{\pi} = \hat{\beta}_0 + \hat{\beta}_1 X$ are symmetric about the line, the transformed confidence bands are not symmetric about the estimated logistic curve $\hat{p} = \dfrac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$ because the logistic transformation of the original bands is nonlinear.

The lower and upper limits of confidence intervals for the probabilities $p_i = \dfrac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$ can be saved to the data spreadsheet by selecting the *Save Results* icon.

Lower and upper limits of confidence intervals for the probability that $\boldsymbol{Y} = 1$ for other values of X can be found by appending the additional values of X to the end of the data spreadsheet and then scrolling to the bottom of the *Predictions* table. (I'll wait until a later example to demonstrate this in Statgraphics, but the method is similar to the way Statgraphics handles prediction in multiple regression.) The dataset in our running example has 100 additional values of X already added. I suspect they were originally meant to be used to Cross-Validate the model (a way to check if a model created using only part of the data is useful in predicting the remaining data). Cross-validation is useful in model building, but we have not covered it in this course.

Interpreting $\hat{\beta}_1$

The interpretation of $\hat{\beta}_1$ is based on the odds-ratio (I know, this is where heads begin to explode). Let $\dfrac{\hat{q}}{1-\hat{q}} = e^{\hat{\beta}_0 + \hat{\beta}_1 (X+1)}$ and $\dfrac{\hat{p}}{1-\hat{p}} = e^{\hat{\beta}_0 + \hat{\beta}_1 X}$ be the estimated odds that $\boldsymbol{Y} = 1$ for $X+1$ and X, respectively. Then, the odds-ratio $\dfrac{\dfrac{\hat{q}}{1-\hat{q}}}{\dfrac{\hat{p}}{1-\hat{p}}} = \dfrac{e^{\hat{\beta}_0 + \hat{\beta}_1 (X+1)}}{e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = e^{\hat{\beta}_1}$, and $\dfrac{\hat{q}}{1-\hat{q}} = \left(\dfrac{\hat{p}}{1-\hat{p}}\right) e^{\hat{\beta}_1}$. So, the odds that $\boldsymbol{Y} = 1$ change by a factor of $e^{\hat{\beta}_1}$ for each additional 1 unit increase in X.

If $\hat{\beta}_1 > 0$, the odds that $\boldsymbol{Y} = 1$ increase as X increases, while the odds that $\boldsymbol{Y} = 1$ decrease as X increases if $\hat{\beta}_1 < 0$.

Example: in our running example, $\hat{\beta}_1 = 0.341469$ is the estimated regression coefficient of X in the table below.

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-5.58023	1.1063	
X	0.341469	0.0665176	1.40701

So, the odds that $\boldsymbol{Y} = 1$ increase by a factor of $e^{0.341469} = 1.407$ (see the column labeled *Estimated Odds Ratio*), or about 40.7%, for each additional 1 unit increase in X. (The interpretation as the *percentage* change in the odds $\boldsymbol{Y} = 1$ per unit increase in X is convenient, and is the interpretation I will use in class.)

The Analysis of Deviance

Deviance serves the same role in logistic regression as variance does in regression. I'm not planning on covering deviance or its interpretation except to say that it's a more appropriate measure than variance when using maximum likelihood estimation. The table below, which appears in the *Analysis Summary* window in Statgraphics, is similar to the ANOVA table in regression and analysis of variance, but you'll notice that there is no column for mean squares and no mean square error estimate of variance.

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	55.0288	1	0.0000
Residual	83.5606	98	0.8505
Total (corr.)	138.589	99	

Nevertheless, the deviances do have properties reminiscent of the sums of squares in ordinary least squares, i.e. linear, regression.

- The sum of the deviances for the model and residuals equals the total deviance.
- The degrees of freedom for the model and residuals sum to the total degrees of freedom. (The degrees of freedom in logistic regression with k predictors even equal those in a k variable linear regression.)
  - $df_{\text{model}} = k$, where k is the number of independent variables in the model (k = 1 here)
  - $df_{\text{residuals}} = n - k - 1$
  - $df_{\text{total}} = n - 1$

In simple logistic regression, the P-value for the model in the table (P-value = 0.0000 in our example), is the P-value for the hypothesis test

- $\mathbf{H}_0: \beta_1 = 0$
- $\mathbf{H}_A: \beta_1 \neq 0$

In multiple logistic regression with k independent variables, the hypotheses tested are

- $\mathbf{H}_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- $\mathbf{H}_A$: At least one $\beta_i \neq 0$, for $i = 1, 2, \cdots, k$

As always, we are led to reject the null hypothesis for small P-values.

The test of the statistical significance of coupon value to the probability the coupon is redeemed is also shown in the table of likelihood ratio tests found in the *Analysis Summary* window (see below). Comparing this to the *Analysis of Deviance* table, you'll notice that the same test, a chi-square likelihood ratio test, is conducted in both cases. For multiple logistic regression, however, a likelihood ratio test is conducted for each independent variable to determine its significance.

Likelihood Ratio Tests

Factor	*Chi-Square*	*Df*	*P-Value*
X	55.0288	1	0.0000

The *P*-value for residuals in the table (*P*-value = 0.8505 in our example), is the *P*-value for a hypothesis test in which the null hypothesis is that the model cannot be significantly improved. Large *P*-values suggest that the model cannot be improved (at least not without introducing new independent variables). For those of you in Stat 115, Statgraphics conducts a right-tailed likelihood ratio test using a χ^2 distribution with $n-k-1$ degrees of freedom. In our example, $P\left(\chi^2_{98} \geq 83.5606/98\right) = 0.8505$.

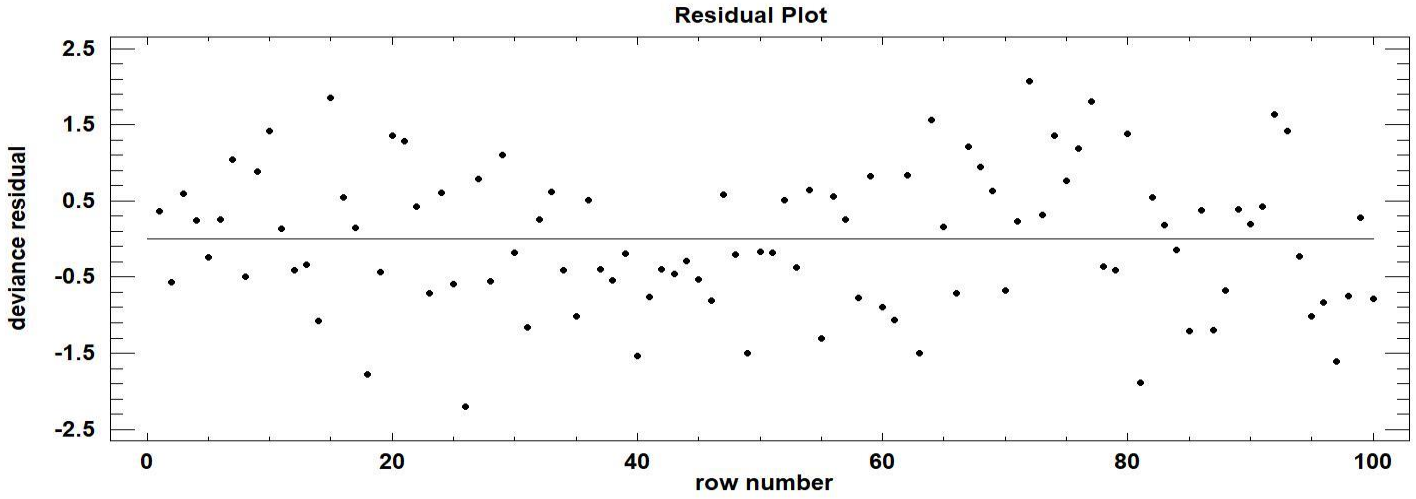
A Measure of Fit

There are even statistics similar to R^2 and R^2_{Adj}, with similar interpretations, that can serve as measures of fit. Below the Analysis of Deviance table, you'll find the following:

Percentage of deviance explained by model = 39.7063
Adjusted percentage = 36.8201

Additional Diagnostics

Logistic regression diagnostics are more difficult than those for ordinary least squares (OLS) regression covered earlier in the course. In addition to the *P*-value for a test of the appropriateness of the model discussed previously, a plot of deviance residuals may be used to locate potential outliers. Below is plot of the deviance residuals versus row number. None of the deviance residuals are unusually large (in absolute value), so there is no obvious indication of outliers.



Predicting Whether ***Y*** will be 0 or 1

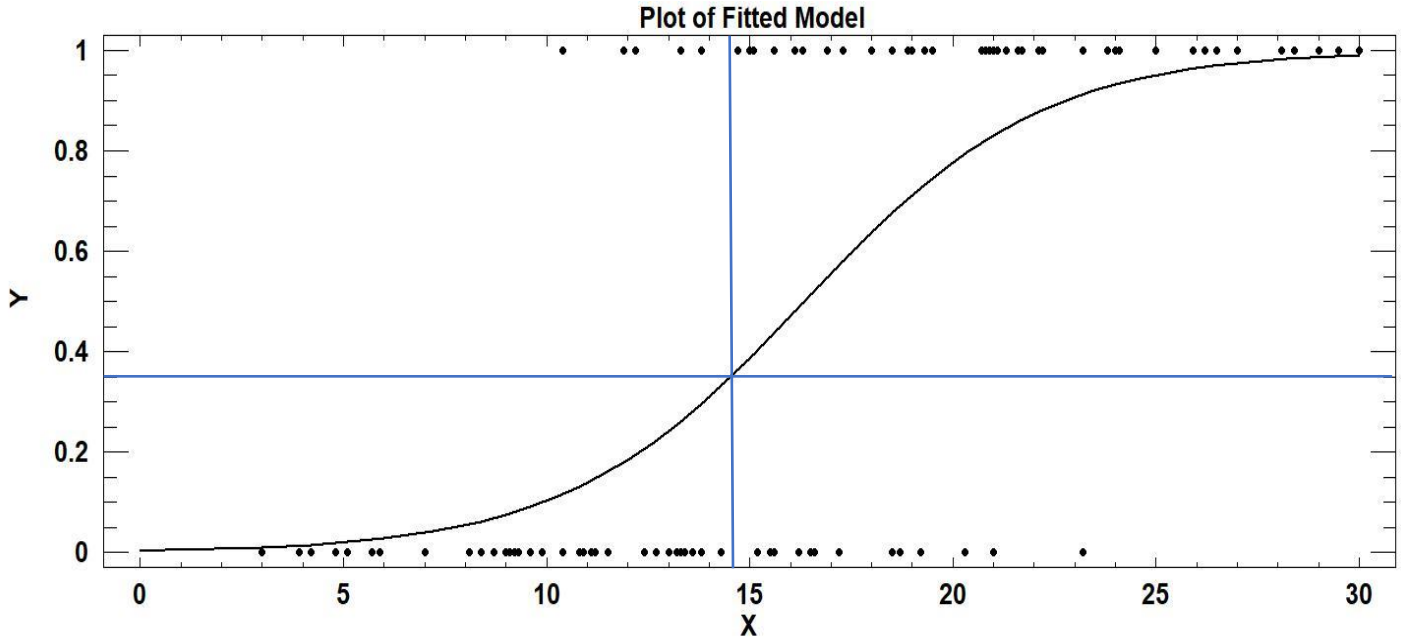
We've already discussed point estimates and confidence intervals for the mean value of ***Y***, $E(\boldsymbol{Y}) = p$, but suppose our interest is in predicting whether the next observation on ***Y*** for some value of *X* is 0 or 1. Unlike ordinary regression, *prediction intervals* don't make sense for binary outcomes. ***Y*** can only assume one of two values, 0 or 1. In logistic regression, predictions are obtained by picking a threshold, or cutoff, value for $p = E(\boldsymbol{Y})$ above which we predict that ***Y*** = 1. An obvious choice for a cutoff value is $p = 0.5$, but other choices may be more appropriate depending on the relative consequences of predicting false positives and false negatives.

The first table in the *Predictions* window displays the percentages (among the observations in the dataset) correctly predicted to be 1 or 0, respectively, for different choices of cutoff. For example, suppose the following decision rule is used: predict ***Y*** = 1 if $p > 0.35$, and predict ***Y*** = 0 otherwise (*TRUE* corresponds to ***Y*** = 1). The rule would have correctly predicted 89.80 percent of the ones observed (44 out of 49), and correctly predicted 70.59 percent of the zeros observed (36 out of 51) in the data. Overall, 80% of all observations in the data would have been correctly predicted by the rule ((44 + 36) out of 100). Statgraphics has flagged the cutoff of $p = 0.35$ because it maximizes the total percent of correct predictions, but the optimal choice for a particular application would depend on the relative consequences of predicting false positives and false negatives.

Prediction Performance - Percent Correct

Cutoff	*TRUE*	*FALSE*	*Total*
0.0	100.00	0.00	49.00
0.05	100.00	17.65	58.00
0.1	100.00	37.25	68.00
0.15	97.96	47.06	72.00
0.2	93.88	49.02	71.00
0.25	93.88	58.82	76.00
0.3	89.80	68.63	79.00
0.35	89.80	70.59	80.00
0.4	79.59	70.59	75.00
0.45	77.55	78.43	78.00
0.5	73.47	80.39	77.00
0.55	71.43	84.31	78.00
0.6	69.39	86.27	78.00
0.65	67.35	86.27	77.00
0.7	65.31	92.16	79.00
0.75	57.14	94.12	76.00
0.8	57.14	96.08	77.00
0.85	44.90	98.04	72.00
0.9	34.69	98.04	67.00
0.95	24.49	100.00	63.00
1.0	0.00	100.00	51.00

To visualize the use of the cutoff $p = 0.35$, the plot of the fitted logistic regression model is shown below with horizontal and vertical blue lines added. The horizontal line intersects the vertical axis at $p = 0.35$, and the vertical line intersect the *x*-axis at approximately 14.5. The decision rule in the previous paragraph predicts $Y = 1$ when the logistic curve lies above the horizontal blue line, i.e., where $p > 0.35$. Equivalently, the decision rule predicts $Y = 1$ to the right of the vertical line, where $X > 14.5$. This rule missed the leftmost 5 observations where $Y = 1$, but correctly predicted the other 44.



Inverse Predictions

An inverse prediction is exactly what it sounds like, predicting the value of X at which $P(\boldsymbol{Y}=1|\,X)=p$ assumes a particular value. Inverse predictions can be found in the *Inverse Predictions* table in Statgraphics. Part of the table provided in Statgraphics is reproduced below. For example, the value of X corresponding to a predicted 35% success rate, i.e., $\hat{p}=0.35$, is 14.529. This agrees with the horizontal and vertical blue lines in the plot of the fitted model above, which provided a cruder estimate of ≈ 14.5.

Table of Inverse Predictions for X

		Lower 95.0%	*Upper 95.0%*
Percent	*X*	*Conf. Limit*	*Conf. Limit*
0.1	-3.88477	-16.3975	1.80123
0.5	0.840259	-8.80225	5.26865
1.0	2.88491	-5.52566	6.77919
2.0	4.94454	-2.23571	8.31143
3.0	6.162	-0.298426	9.22455
4.0	7.03483	1.08574	9.88394
5.0	7.71897	2.16713	10.4043
6.0	8.2839	3.05721	10.8369
7.0	8.76665	3.81537	11.209
8.0	9.18936	4.47707	11.537
9.0	9.5663	5.06516	11.8314
10.0	9.90721	5.59523	12.0995
15.0	11.262	7.67995	13.1868
20.0	12.282	9.21701	14.0379
25.0	13.1245	10.4548	14.7726
30.0	13.8605	11.5033	15.4473
35.0	**14.529**	12.4206	16.095
40.0	15.1544	13.2418	16.7383
45.0	15.7542	13.9899	17.3943
50.0	16.3418	14.6827	18.0774