Simple linear regression on an indicator variable – a precursor to logistic regression

Basically it is a simple linear regression where the dependent variable has a value of either 0 or 1.

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad \text{where } Y_i = 0, 1$$

This is called a binary response, and the interpretation of $E(Y_i)$ is different from the usual response variable. Given that $E(\varepsilon_i) = 0$, then $E(Y_i) = \beta_0 + \beta_1 X_i$. If Y_i is a Bernoulli random variable then the probability distribution is

when $Y_i = 1$, $P(Y_i = 1) = \pi_i$ and when $Y_i = 0$, $P(Y_i = 1) = 1 - \pi_i$. The expected value of the distribution is then given by $E(Y_i) = \beta_0 + \beta_1 X_i = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$.

Issues when the response variable is binary.

1) The residuals are not normally distributed. The residuals, defined as $Y_i - (\beta_0 + \beta_1 X_i)$ will only take on two values. When $Y_i = 1$ the values is $1 - \beta_0 - \beta_1 X_i$ and when $Y_i = 0$ the value is $-\beta_0 - \beta_1 X_i$.

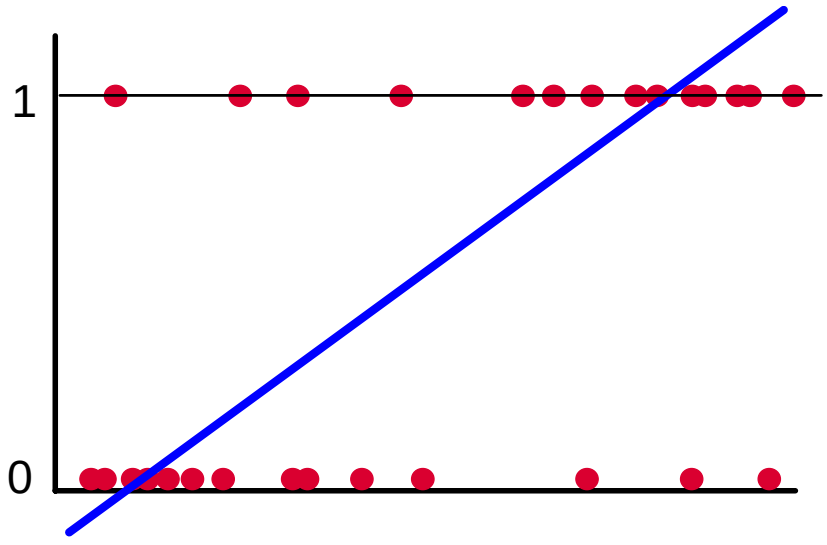
2) The residuals are not homogeneous. The variance is given by $\sigma^2_{Y_i} = E\{(Y_i - E(Y_i))^2\} =$ $(1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i) = \pi_i(1-\pi_i) = (E\{Y_i\})(1-E\{Y_i\})$.

Finally the variance for residual is the same as for Y_i because $\varepsilon_i = Y_i - \pi_i$, and π_i is a constant. $\sigma^2_{Y_i} = E\{(Y_i - E(Y_i))^2\}$ $\sigma^2_{Y_i} = E\{(Y_i - E(Y_i))^2\}$

3) The last issue is that since the response ranges between 0 and 1 there should be a constraint on the response such that $0 \le E\{Y_i\} = \pi \le 1$.

Any solution should address these issues.

First model, simple linear regression on an indicator variable.

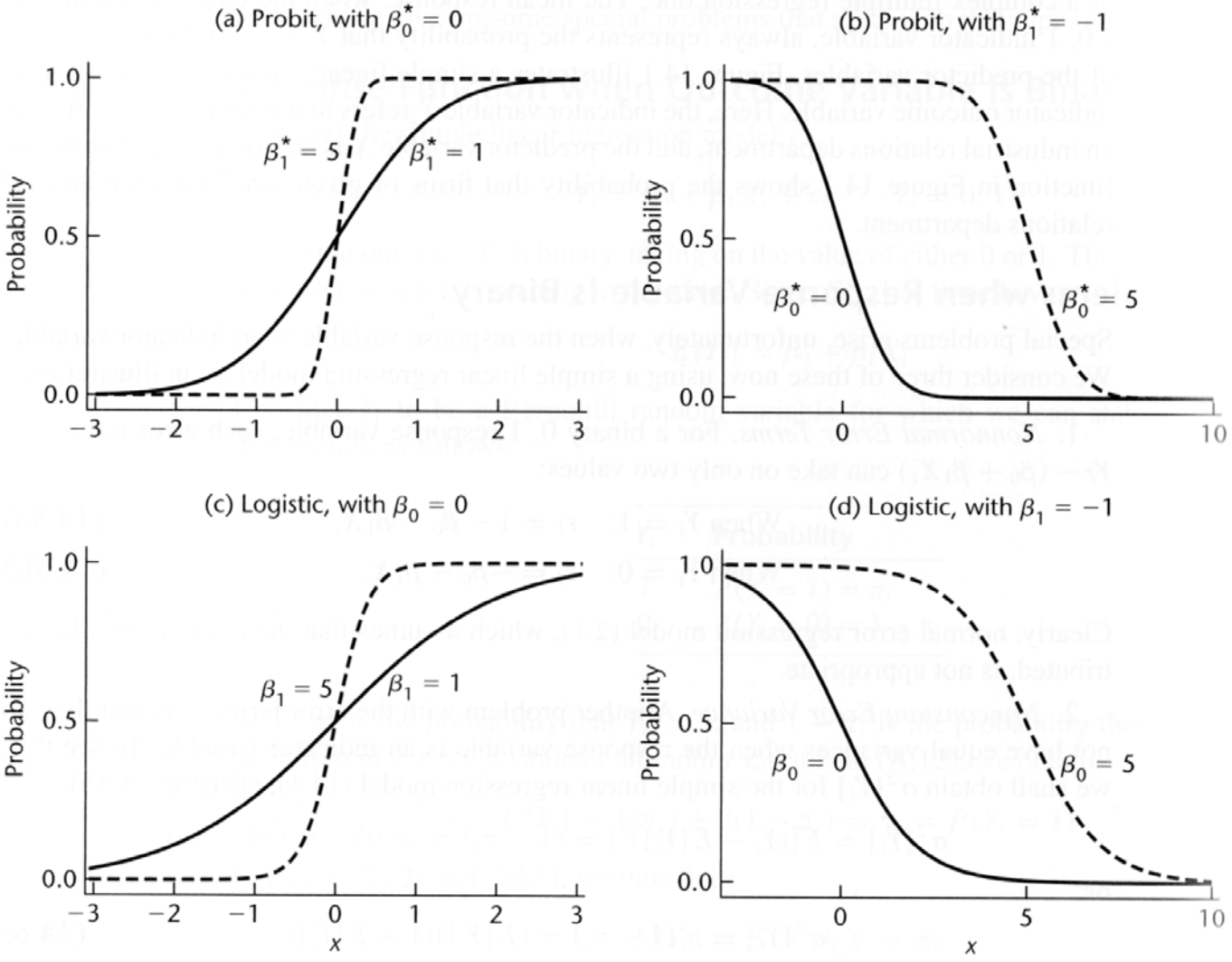


This is a "primitive" version of regression on an indicator variable. The predicted value ($\hat{Y}$) is interpreted as “the probability of getting a 1”. However, this fitted line does not address any of the issues stated above. It does nothing to address the lack of normality, the problem with homogeneity of variance or to keep the line from going below zero and above 1. This solution has not particularly desirable properties.

See SAS example – SLR on p and on indicator variable

Second model, a sigmoid response variable.

FIGURE 14.2 Examples of Probit and Logistic Mean Response Functions.



The probit analysis, based on a standard normal cumulative distribution, will be discussed later. However, it is similar to the logistic function. The probit distribution has a normal density function and the logistic function is very similar.

The full version of the Logistic Model was discussed in the section on nonlinear models as a common growth model. This three-parameter model is not linear and cannot be fitted with PROC REG or PROC GLM or even PROC Logistic. $E(Y_i) = \dfrac{\beta_2}{1+\left(\frac{\beta_2-\beta_0}{\beta_2}\right)e^{-\beta_1 X_i}} = \dfrac{\beta_2}{1+e^{-\beta_0-\beta_1 X_i}}$

The model fitted by PROC LOGISTIC is much simplified since the upper bound (β_2) is known to be 1.

The logistic mean response function is $E\{Y_i\} = \dfrac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} = [1+\exp(-\beta_0-\beta_1 X_i)]^{-1}$. The logistic can be derived as the logit transformation of the probability π_i which is $\log_e\left(\dfrac{\pi_i}{1-\pi_i}\right)$.

The final logistic model is then $\log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$. The ratio $\frac{\pi_i}{1-\pi_i}$ is called the "odds" and the natural log of this is called the *logit* response function.

What are odds?

Odds are an expression of the likelihood of some event happens compared to the likelihood that it does not happen. If the odds on a horse in a race are 30 to 1, is that horse likely to win? Or lose. If the odds of an event happening are 50:50, what does that mean? What if the odds are 1:1? How is that different?

The odds is simply the ratio of the probability of the occurance of an event to the probability of that event not occurring. The values of $^{50}/_{50}$ or $^{1}/_{1}$ both produce odds equal to "1". They have the same odds.

If the odds ~~ratio~~ is 1, then the likelihood of something happening is equal to the likelihood that it will not happen.

To simplify our concepts we will think of the odds as the ratio of two probabilities. The probability that some event happens (success) will be equal to p. The probability of failure will be 1–p. The odds is given by p/(1–p).

What if the odds ~~ratio~~ is 2? This means that p is twice as large as (1–p), so success is twice as likely as failure. If the odds ratio is 10 the probability of success is 10 times more likely than failure.

Odds are also commonly expressed as percents, so an odds of 1.5 means success is 50% greater than the probability of failure. For odds of 2 the probability of success is 100% greater than the probability of success.

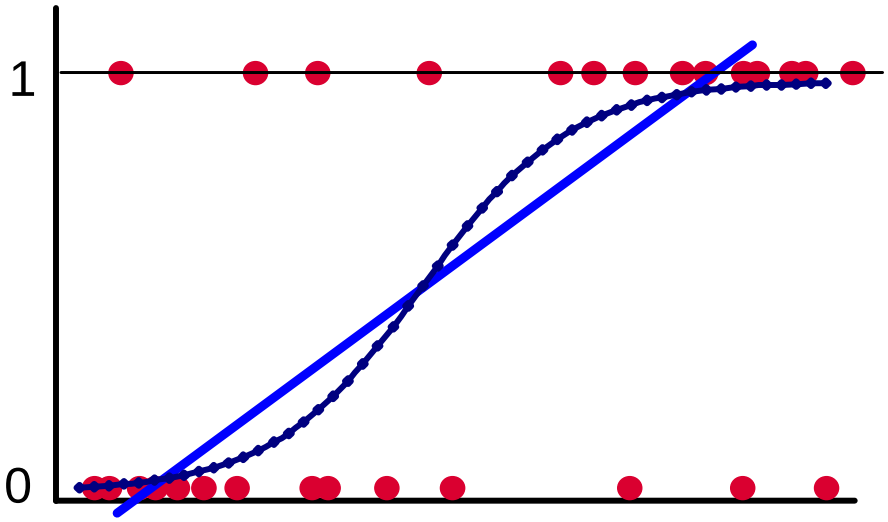
If the odds are 0.5, then the probability of failure is twice as likely. An odds of 0.1 means the probability of failure is ten times more likely than success.

Detransforming odds – The logistic analysis produces "log odds" as predicted values of the dependent variable. Odds are obtained by taking the antilog [$\exp(YHat_i) = odds_i$). The probability can be obtained by calculating $p_i = odds_i / (1 + odds_i)$.

Odds ratios

Although the odds are a ratio they are usually referred to as just "odds". The "odds ratio" is a different value. In Analysis of Variance the tests of interest are often difference in means, and in regression the tests of interest often involve the change in Y per unit X, which is the difference between the mean of Y at X_i and the mean of Y at X_i+1. When odds are used as the dependent variable the difference in "means" is the difference in the estimated odds. For analysis of variance the difference would be $\log_e(\pi_i/(1-\pi_i)) - \log_e(\pi_j/(1-\pi_j)) = \log_e\left(\pi_i/(1-\pi_i) \Big/ \pi_j/(1-\pi_j)\right)$. This term would be referred to as the "odds ratio". Likewise for regression, where the slope is the change in Y per unit X the slope would be given as an odds ratio, $\log_e\left(\pi_{X+1}/(1-\pi_{X+1}) \Big/ \pi_X/(1-\pi_X)\right)$.

Recall the three issues previously mentioned for working with 0, 1 indicator variables. The use of logistic or probit analyses addresses the third issue mentioned above, the constraint that the response ranges between 0 and 1. These sigmoid curves can be limited to this range. Also note that the odds are not restricted to the 0, 1 range. The probit function is based on a Z distribution transformation, and has a standard normal density function. The logistic density function is very similar, but not quite normal (having slightly heavier tails).



To address the last issue of lack of homogeneity of the residuals the analysis is sometimes weighted. The weights apply when there are repeat observations at several levels of the response variable in which case the outcomes are binomial distributed and a p_i can be calculated. If the transformed variables are designated $\pi_j' = \log_e\left[\frac{\pi_j}{1-\pi_j}\right]$ for the logit, which is estimated as $p_j' = \log_e\left[\frac{p_j}{1-p_j}\right]$, the variance is $Var(\pi_j') = \frac{1}{n_j\pi_j(1-\pi_j)}$ and the estimated variance is $s^2_{p_j'} = \frac{1}{n_j p_j(1-p_j)}$. To address the lack of homogeneity of variance weighting is done by the inverse of the variance, or $n_j\pi_j(1-\pi_j)$.

See examples – weighted SLR on logit and NLIN on p

Logistic regression in SAS

The procedure is fitted using maximum likelihood. Several common fit statistics are provided including –2 residual log likelihood (–2 Log L) the Akaike Information Criterion (AIC), and the Schwartz criterion (SC). The last two are penalized log likelihood based estimates.

The procedure supports weight and frequency statements, and the CLASS statement. The model statement looks similar to regression, but can be set up in one of two ways.

1) MODEL Indicator = independent variables

2) MODEL Success / TotalTrials = independent variables

The analysis provides tests of the model. The Wald test is an application of large sample statistics and is based on the Z distribution. Wald can also be used to place confidence intervals on the estimates. A likelihood ratio test is also available.

Model fit statistics

Model Fit Statistics	Intercept Only	Intercept and Covariates
AIC	2063.911	1684.291
SC	2069.225	1694.917
-2 Log L	2061.911	1680.291

1) Akaike Information Criterion $\text{AIC} = -2\log(L) + 2p$

where Log(L) is the log likelihood and p is the number of parameters

2) Schwarz Criterion $\text{SC} = -2\log(L) + p\log(\sum_j f_j)$

3) -2log L $-2\sum_{i-1}^{n}\left[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)\right]$

This is analogous to the SSE in regression and is given in SAS as the "-2 Log L".

Two models (full and reduced) can be compared by calculating the difference in "-2 Log L" for both models. This difference follows a chi square distribution with a d.f. equal to the difference in d.f. for the two models.

4) Generalized R^2 $1-\left(\frac{L(0)}{L(\theta)}\right)^{\frac{2}{n}}$, where L(0) is the intercept only model.

Since this value reaches its maximum of less than 1 for discrete models an adjustment has been proposed. This is called the Max-rescaled Rsquare in SAS. $\frac{R^2}{R^2_{\max}}$

Global tests

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	381.6204	1	<.0001
Score	352.7372	1	<.0001
Wald	296.9818	1	<.0001

Wald : used to test individual parameter estimates and to place confidence intervals. It is based on a large sample assumption of asymptotic normality. The Chi-square test is given by $\beta_i^2 / Var(\beta_i) = [\beta_i / Stderr(\beta_i)]^2$ and the confidence interval is $P\left(e^{(\hat{\beta}_i - 1.96 S_{\hat{\beta}_i})} \le \beta i \le e^{(\hat{\beta}_i - 1.96 S_{\hat{\beta}_i})}\right) = 0.95$.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6435	0.1561	286.7841	<.0001
X	1	0.6740	0.0391	296.9818	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
X	1.962	1.817	2.118