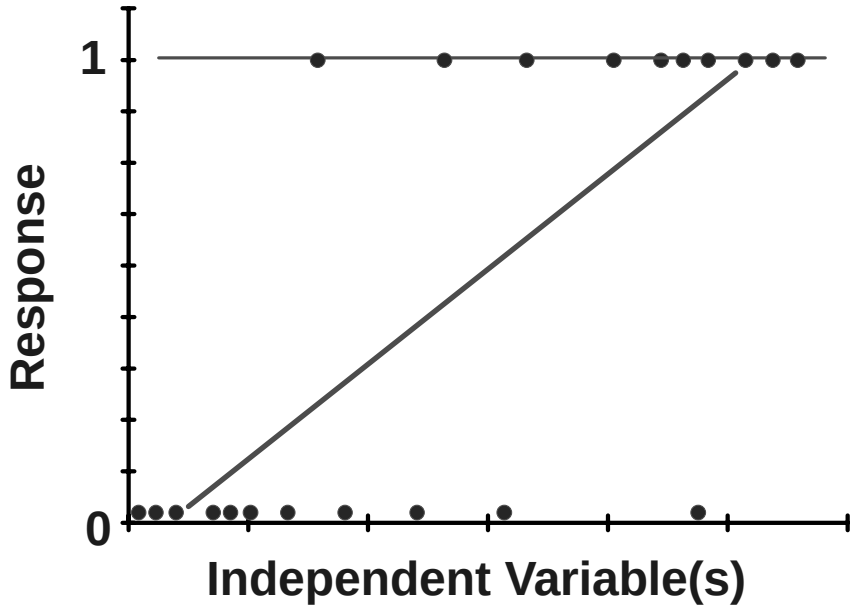
REGRESSION ON AN INDICATOR VARIABLE –

In this technique, the dependent variable (Y) is an indicator, and takes a value of either 0 or 1.

This is called a binary response variable



Examples – any two categories, any binomial or binary variable

a) Success-failure, Gender (Male-female), mortality, presence-absence, pass-fail, etc.

The results of a simple linear regression is a slope and intercept which will produce a predicted value which ranges from 0 to 1 over most of the range of X

This b_1 can be interpreted as a probability of obtaining a 1 per unit of X, and the predicted value is the probability of obtaining a 1 at some particular value of X.

Problems with regression on indicator variables

1) Nonnormal errors : given that $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$, then

When $Y_i = 1$, then $\epsilon_i = 1 - \beta_0 - \beta_1 X_i$

When $Y_i = 0$, then $\epsilon_i = -\beta_0 - \beta_1 X_i$

2) Nonconstant errors

Let $P(Y_i=1) = \pi_i$ and $P(Y_i=0) = 1 - \pi_i$

then $E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = \beta_0 + \beta_1 X_i$

and $\sigma^2_{Y_i} = E[Y_i - E(Y_i)]^2 = (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i)$

$= \pi_i (1 - \pi_i) = E(Y_i)(1 - E(Y_i))$

finally, $Var(\epsilon_i) = Var(Y_i)$, since $\epsilon_i = Y_i - \pi_i$ and π_i is a constant

so $\sigma^2_{\epsilon_i} = \pi_i (1 - \pi_i)$

$= E(Y_i)(1 - E(Y_i)) = (\beta_0 - \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$

and the variance is a function of X_i

3) Constraints on the response function

If the function is fitted with a line, at some point the predicted value will be <0 or >1. As a probability, the true value must be between 0 and 1, so we must place some restraint on the predicted value.

So we would like to find a function which solves some of these problems, we might also expect a curve instead of a simple linear and we would like a curve that can go from 0 to 1 (asymptotically)

Several sigmoid possibilities have been considered, especially

a) Logistic (symmetric)

b) cumulative normal distribution (Probit analysis)

This version of the logistic has several advantages,

$$E(Y) = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$$

particularly that it can be readily linearized by the transformation

$$\pi' = \log_e \left[\frac{\pi}{1-\pi}\right]$$

This is called a LOGIT transformation, and π' is called a logit mean response.

We can then fit

$$\pi_i' = b_0 + b_1 X_i$$

and we should closely approximate the logistic.

The logistic can also be fitted directly with nonlinear techniques.

A similar, but more difficult and less flexible, transformation exists for the **cumulative normal distribution**, and is called a PROBIT transformation

Weighting to improve variance : the logit only linearizes the logistic function, it does not cure the nonhomogeneous variance problem

The logit,

$$\pi' = \log_e \left[\frac{\pi}{1-\pi}\right]$$

is estimated by

$$p_i' = \log_e \left[\frac{p_i}{1-p_i}\right]$$

The variance of p_i is

$$\text{Var}(p_i') = \frac{1}{n_i \pi_i (1-\pi_i)}$$

which is estimated by,

$$s_{p_i'} = \frac{1}{n_i p_i (1-p_i)}$$

we could therefore weight by

$$w_i = n_i p_i (1 - p_i)$$

in order to homogenize the variance.

Notes:

1) logits are readily extendible to multiple regression.

2) Logistic regression has many applications. One common application in the biological sciences is the calculation of the dose needed to cause mortality. However, small doses cause small mortalities and large doses cause large mortalities. We therefore calculate an LD_{50}, which is the "lethal dose for 50% mortality".

for example, given the equation below

$$\hat{\pi}'_i = b_0 + b_1 X_i = -2.64 + 0.673*\text{dose}$$

the LD_{50} is given by

$$\hat{\pi}'_{50} = \log_e \left[\frac{50}{1-50}\right] = 0$$

$$0 = -2.64 + 0.673*\text{dose}_{50}$$

$$\text{dose}_{50} = \frac{2.64}{0.673} = 3.923, \text{ or a dose of about 4}$$