



Logistic Regression

Psy 524
Ainsworth

What is Logistic Regression?

- Form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.
- Addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors (the predictors do not have to be normally distributed, linearly related or have equal variance in each group)

What is Logistic Regression?

- Logistic regression is often used because the relationship between the DV (a discrete variable) and a predictor is non-linear
- Example from the text: the probability of heart disease changes very little with a ten-point difference among people with low-blood pressure, but a ten point change can mean a drastic change in the probability of heart disease in people with high blood-pressure.

Questions

- Can the categories be correctly predicted given a set of predictors?
 - Usually once this is established the predictors are manipulated to see if the equation can be simplified.
 - Can the solution generalize to predicting new cases?
 - Comparison of equation with predictors plus intercept to a model with just the intercept

Questions

- What is the relative importance of each predictor?
 - How does each variable affect the outcome?
 - Does a predictor make the solution better or worse or have no effect?

Questions

- Are there interactions among predictors?
 - Does adding interactions among predictors (continuous or categorical) improve the model?
 - Continuous predictors should be centered before interactions made in order to avoid multicollinearity.
- Can parameters be accurately predicted?
- How good is the model at classifying cases for which the outcome is known ?

Questions

- What is the prediction equation in the presence of covariates?
- Can prediction models be tested for relative fit to the data?
 - So called “goodness of fit” statistics
- What is the strength of association between the outcome variable and a set of predictors?
 - Often in model comparison you want non-significant differences so strength of association is reported for even non-significant effects.

Assumptions

- The only “real” limitation on logistic regression is that the outcome must be discrete.

Assumptions

- If the distributional assumptions are met than discriminant function analysis may be more powerful, although it has been shown to overestimate the association using discrete predictors.
- If the outcome is continuous then multiple regression is more powerful given that the assumptions are met

Assumptions

- Ratio of cases to variables – using discrete variables requires that there are enough responses in every given category
 - If there are too many cells with no responses parameter estimates and standard errors will likely blow up
 - Also can make groups perfectly separable (e.g. multicollinear) which will make maximum likelihood estimation impossible.

Assumptions

- Linearity in the logit – the regression equation should have a linear relationship with the logit form of the DV. There is no assumption about the predictors being linearly related to each other.

Assumptions

- Absence of multicollinearity
- No outliers
- Independence of errors – assumes a between subjects design. There are other forms if the design is within subjects.

Background

- Odds – like probability. Odds are usually written as “5 to 1 odds” which is equivalent to 1 out of five or .20 probability or 20% chance, etc.
 - The problem with probabilities is that they are non-linear
 - Going from .10 to .20 doubles the probability, but going from .80 to .90 barely increases the probability.

Background

- Odds ratio – the ratio of the odds over 1 – the odds. The probability of winning over the probability of losing. 5 to 1 odds equates to an odds ratio of $.20/.80 = .25$.

Background

- Logit – this is the natural log of an odds ratio; often called a log odds even though it really is a log odds ratio. The logit scale is linear and functions much like a z-score scale.

Background

LOGITS ARE CONTINUOUS, LIKE Z SCORES

$p = 0.50$, then logit = 0

$p = 0.70$, then logit = 0.84

$p = 0.30$, then logit = -0.84

Plain old regression

- $Y =$ A BINARY RESPONSE (DV)
 - 1 POSITIVE RESPONSE (Success) $\rightarrow P$
 - 0 NEGATIVE RESPONSE (failure) $\rightarrow Q = (1-P)$
- $MEAN(Y) = P$, observed proportion of successes
- $VAR(Y) = PQ$, maximized when $P = .50$, variance depends on mean (P)
- $X_j =$ ANY TYPE OF PREDICTOR \rightarrow
Continuous, Dichotomous, Polytomous

Plain old regression

$$Y|X = B_0 + B_1X_1 + \varepsilon$$

- and it is assumed that errors are normally distributed, with mean=0 and constant variance (i.e., homogeneity of variance)

Plain old regression

$$E(\hat{Y} | X) = B_0 + B_1 X_1$$

- an expected value is a mean, so

$$(\hat{Y} = \hat{\pi}) = P_{Y=1} | X$$

• The predicted value equals the proportion of observations for which $Y|X = 1$; P is the probability of $Y = 1$ (A SUCCESS) given X , and $Q = 1 - P$ (A FAILURE) given X .

Plain old regression

- For any value of X , only two errors ($Y - \hat{Y}$) are possible, $1 - \hat{\pi}$ AND $0 - \hat{\pi}$. Which occur at rates $P|X$ AND $Q|X$ and with variance $(P|X)(Q|X)$

Plain old regression

- Every respondent is given a probability of success and failure which leads to every person having drastically different variances (because they depend on the mean in discrete cases) causing a violation of the homoskedasticity assumption.

Plain old regression

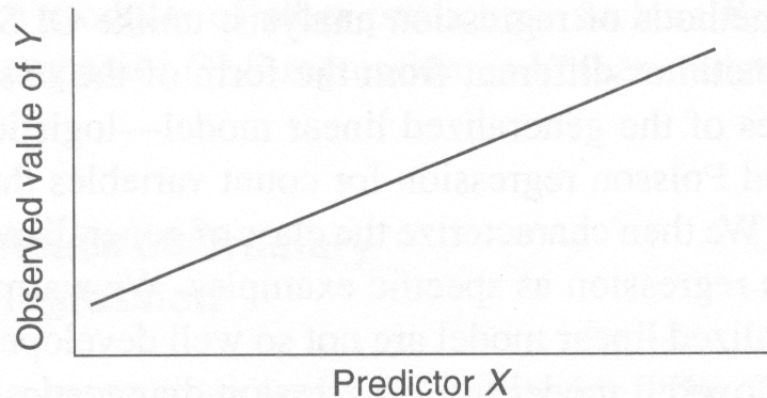
- Long story short – you can't use regular old regression when you have discrete outcomes because you don't meet homoskedasticity.

An alternative – the ogive function

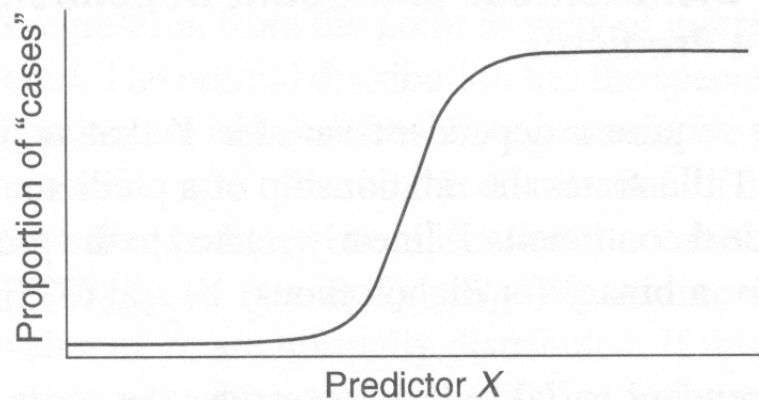
- An ogive function is a curved s-shaped function and the most common is the logistic function which looks like:

The logistic function

(A) For a continuous outcome variable Y , the numerical value of Y at each value of X .



(B) For a binary outcome variable, the proportion of individuals who are “cases” (exhibit a particular outcome property) at each value of X .



The logistic function

$$\hat{Y}_i = \frac{e^u}{1 + e^u}$$

- Where \hat{Y} is the estimated probability that the i th case is in a category and u is the regular linear regression equation:

$$u = A + B_1X_1 + B_2X_2 + \cdots + B_KX_K$$

The logistic function

$$\hat{\pi}_i = \frac{e^{b_0 + b_1 X_1}}{1 + e^{b_0 + b_1 X_1}}$$

The logistic function

- Change in probability is not constant (linear) with constant changes in X
- This means that the probability of a success ($Y = 1$) given the predictor variable (X) is a non-linear function, specifically a logistic function

The logistic function

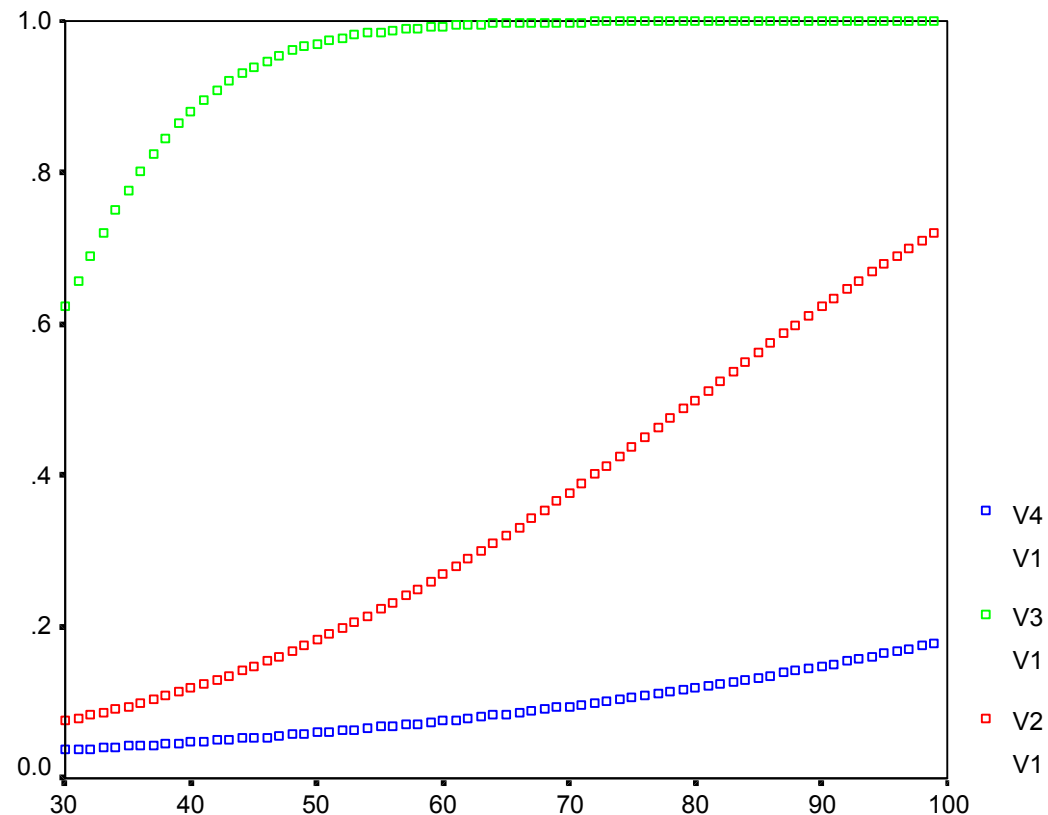
- It is not obvious how the regression coefficients for X are related to changes in the dependent variable (Y) when the model is written this way
- Change in Y (in probability units)| X depends on value of X . Look at S-shaped function

The logistic function

- The values in the regression equation b_0 and b_1 take on slightly different meanings.
- b_0 ← The regression constant (moves curve left and right)
- b_1 ← The regression slope (steepness of curve)
- $\frac{-b_0}{b_1}$ ← The threshold, where probability of success = .50

Logistic Function

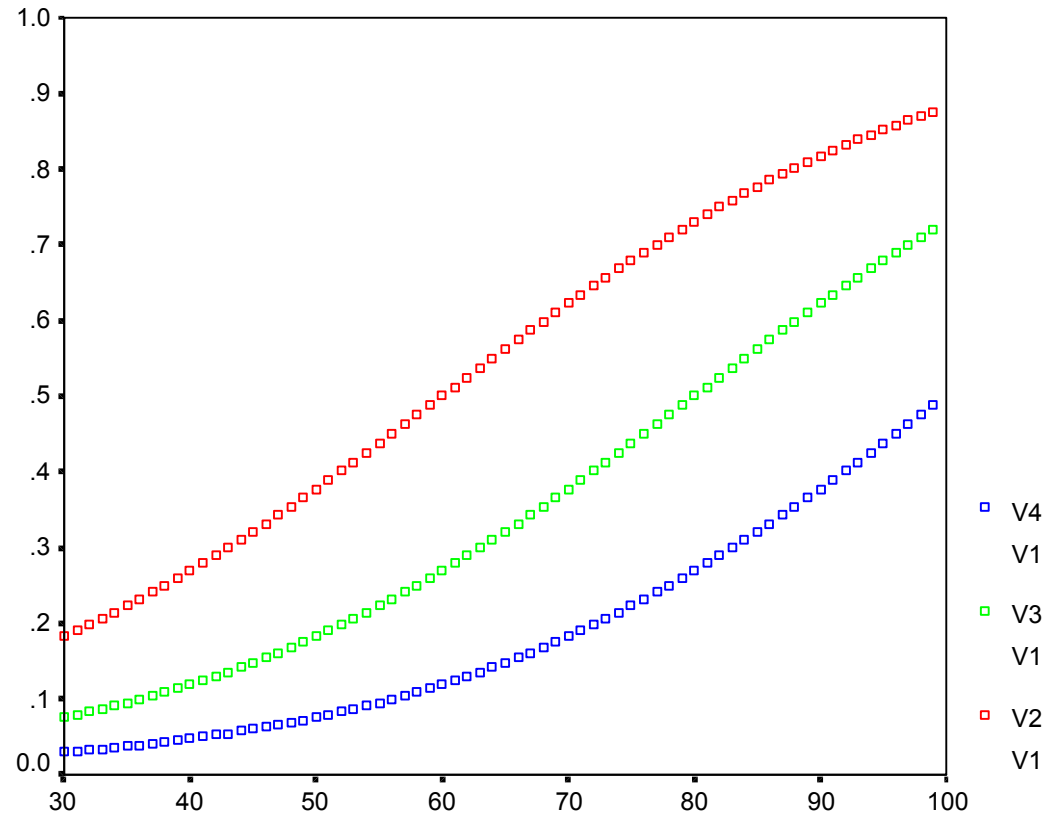
- Constant regression constant different slopes
 - v2: $b_0 = -4.00$
 $b_1 = 0.05$ (middle)
 - v3: $b_0 = -4.00$
 $b_1 = 0.15$ (top)
 - v4: $b_0 = -4.00$
 $b_1 = 0.025$ (bottom)



Logistic Function

- Constant slopes with different regression constants

- v2: $b_0 = -3.00$
 $b_1 = 0.05$ (top)
- v3: $b_0 = -4.00$
 $b_1 = 0.05$ (middle)
- v4: $b_0 = -5.00$
 $b_1 = 0.05$ (bottom)



The Logit

- By algebraic manipulation, the logistic regression equation can be written in terms of an odds ratio for success:

$$\left[\frac{P(Y=1 | X_i)}{(1 - P(Y=1 | X_i))} \right] = \left[\frac{\hat{\pi}}{(1 - \hat{\pi})} \right] = \exp(b_0 + b_1 X_{1i})$$

The Logit

- Odds ratios range from 0 to positive infinity
- Odds ratio: P/Q is an odds ratio; less than 1 = less than .50 probability, greater than 1 means greater than .50 probability

The Logit

- Finally, taking the natural log of both sides, we can write the equation in terms of logits (log-odds):

$$\ln \left[\frac{P(Y=1|X)}{(1-P(Y=1|X))} \right] = \ln \left[\frac{\hat{\pi}}{(1-\hat{\pi})} \right] = b_0 + b_1 X_1$$

For a single predictor

The Logit

$$\ln \left[\frac{\hat{\pi}}{(1-\hat{\pi})} \right] = b_0 + b_1 X_1 + b_2 X_2 \dots + b_k X_k$$

- For multiple predictors

The Logit

- Log-odds are a linear function of the predictors
- The regression coefficients go back to their old interpretation (kind of)
 - The expected value of the logit (log-odds) when $X = 0$
 - Called a 'logit difference'; The amount the logit (log-odds) changes, with a one unit change in X ; the amount the logit changes in going from X to $X + 1$

Conversion

- $\text{EXP}(\text{logit}) \text{ or } = \text{odds ratio}$
- $\text{Probability} = \text{odd ratio} / (1 + \text{odd ratio})$