Logit Regression.

Categorical variables.

-> why linear Reg. Not possible

Response Not on a ratio scale.

Error terms not Normally distributed.

Linear Reg. can generate any real value

but Categorical var can only take Limited nos of discrete values.

-> Fundamental difference of Logit with L.R.

L.R. -> Exp. value of the dep. var = Linear combination of indep vars. and their corresp. Parameters.

Generalized linear models. equate the linear combination of indep vars. to some function of the Probability of a given outcome on the dep. var.

Logit. Regr.] -> That function is Logit ~~trans~~ form. Natural Log. of the odds that some event will occur.

Parameter estimation in Linear Reg.

-> Minimizing sum of squares of dev. of ~~Predicted~~ values from observed values.

-> Solution involves solving system of N. equations.

The Model → Binomial Logit Regression.

Consider ith Row → has distinct Combo. of Values of indep. variables.

N → total Nos of populations

$\begin{bmatrix} Y_i \end{bmatrix}$ N → Nos of successes of Z for population i

$\begin{bmatrix} y_i \end{bmatrix}$ → Observed counts of the Nos of Success For Each population.

$\begin{bmatrix} \pi_i \end{bmatrix}$ → Prob of success. For any given Obs in ith population

The Logit Model.] Equates the logit transform (Log odds) of the Prob. of success.

$$\operatorname{Log}\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{ik}\beta_k , \quad i = 1,2,\cdots N$$

Getting an expression For $\pi_i = P(x)$.

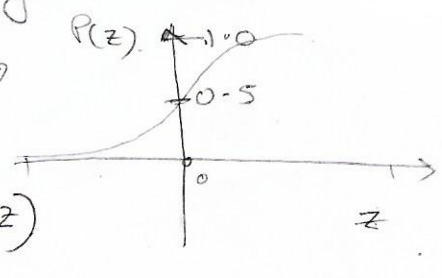
$$\frac{P(x)}{1-P(x)} = \exp\left(\sum_{j=0}^{k} x_k \beta_k\right) = \prod_{j=0}^{k} \exp(x_k\beta_k)$$

Let $Z = \prod_{j=0}^{k} x_k \beta_k$

Then $\dfrac{P(x)}{1-P(x)} = \exp(Z)$

or $P(x) = \dfrac{\exp(Z)}{1+\exp(Z)} = P(Z)$ Sigmoid Function. Maps a Real Line to (0,1)

Thus $P(Z) = \dfrac{\exp(Z)}{1+\exp(Z)} = \dfrac{1}{1+\exp(-Z)}$



$P'(z) = P(z)(1-P(z)) \rightarrow$ Root. $P(z) = \frac{1}{1-e^{-z}} = (1-e^{-z})^{-1}$

$\therefore P'(z) = -1(1-e^{-z})^{-2}(e^{-z})$

$= \frac{-e^{-z}}{(1-e^{-z})(1-e^{-z})}$

now $P(z) = \frac{1}{1-e^{-z}}$

and $1 - P(z) = 1 - \frac{1}{1-e^{-z}}$

$= \frac{(1-e^{-z}) - (1)}{(1-e^{-z})} = \frac{-e^{-z}}{(1-e^{-z})}$

Thus. $\boxed{P'(z) = P(z)(1-P(z))}$

Note $P'(\beta) = P(z)(1-P(z)).(z')$ where z' is gradient taken wrt β.

Maximum Likelihood Estimation.

$L(X|P) =$ ~~ΠP~~ Product of Predicted Probabilities of the N individual observations.

$L(X|P) = \prod_{\substack{i=1 \\ y_i=1}}^{n} P(x_i) \prod_{\substack{y_i=0 \\ i=1}}^{n} (1-P(x_i))$

$X_{(K+1,\, N)}$ Each column corresponds to an observation. First row is 1.

y_N dim vector of Responses.

$(X, y) \Rightarrow$ is a set of observations.]

Log likelihood.

$$\mathcal{L}(X|p) = \sum_{i=1, y_i=1}^{N} \text{Log}\, p(x_i) + \sum_{i=0, y_i=0}^{N} \text{Log}\,(1-p(x_i))$$

we want to maximize Log likelihood.

$$\nabla_b \mathcal{L} = \sum_{\substack{i=0 \\ y_i=1}}^{N} \frac{p_i'}{p_i} x_i - \sum_{\substack{i=0 \\ y_i=0}}^{N} \frac{p_i'}{(1-p_i)} x_i$$

Now $p_i' \equiv p(1-p)$.

Thus we actually have.

$$\nabla_b \mathcal{L} = \sum_{\substack{i=1 \\ y_i=1}}^{N} \frac{p_i(1-p_i)}{p_i} x_i - \sum_{\substack{i=1 \\ y_i=0}}^{N} \frac{p_i(1-p_i)}{1-p_i} x_i$$

Thus

$$\nabla_b \mathcal{L} = \sum_{\substack{i=1 \\ y_i=1}}^{N} (1-p_i)x_i - \sum_{\substack{i=1 \\ y_i=0}} p_i x_i$$

$$= \sum_{i=1}^{N} \left[y_i(1-p_i) - (1-y_i)p_i \right] x_i$$

after cancelling terms.

$$= \sum_{i=1}^{N} \left[y_i x_i - p_i x_i \right] = 0$$

Notes:

Equations to be solved are in terms of probabilities (p). Not in terms of b.

$\therefore$ Logit Models are coordinate free.

[Solving for coeff.]

Say for a vector valued func. $f(b)_{opt} = 0$

Taylor expansion around initial guess.

$f(b_0+\Delta) \approx f(b_0) + f'(b_0)\Delta$ $\quad f' \rightarrow$ Jacobean of first derivatives of f wrt. b.

Now setting LHS to zero, we have.

$$\Delta_0 = \left[f'(b_0)\right]^{-1} f(b_0)$$

Now update estimate for b: $b_1 = b_0 + \Delta_0$. iterate till convergence.

Now $H = \frac{\partial}{\partial \vec{b}} \nabla_{\vec{b}} \mathcal{L}$

Multi variable Taylor's Series.

$$f(x,y) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + \frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(y-b)^2\right] + \dots$$

Now we already have:

$$\nabla_b \mathcal{L} = \sum_{i=1}^{N}\left[y_i x_i - P_i x_i\right]$$

Now.

$$H = \frac{\partial}{\partial b}\nabla_b \mathcal{L} = -\sum_{i=1}^{N} x_i \nabla_b P_i = -\sum_{i=1}^{N} x_i P_i (1-P_i) x_i^T$$

or $H = XWX^T$

Thus we have the following.

$$\Delta_k = (X W_k X^T)^{-1} X (y - P_k)$$

where we have the following.

-> W is a diagonal matrix of the derivatives P_i'
-> The ith. col of X corresponds to ith observation.

Compare with LINEAR Regression.

$$y = x^T b$$

$$Xy = XX^T b$$

$$b = (XX^T)^{-1} X y$$

1) Comparing the two we find. that at each. iteration Δ. is the solution of a weighted Least squares.

Notes:.

Coeff. tends to infinity → Sign that a input is perfectly correlated.

Large. Coeff. Magnitudes: → indication of Correlated inputs.

[The AIC Value to Compare Models.] Akaike Information (AIC) Criterion

$$AIC = 2k - 2\ln(\hat{L}).$$

→ Maximized value of the likelihood function

k: Nos of Estimated Parameters

→ Lower the AIC Value better the model

Notes:]

1] it Supports better Fit

2] Penalizes using more Parameters.

Classifying using Linear Regression.

→ Train Model to Predict [0, 1]

→ $>0.5 \rightarrow 1$ $<0.5 = 0$

Cons] → Raw output from LR → can Predict value outside this Range

→ LR → Designed to Min Sq Err. against line of best fit.

↓

→ Regression Line shifts ⟶ Thus decision boundary highly sensitive to influential observations.

Logit Function

Sigmoid $f(x) = \frac{1}{1+e^{-x}}$ output $[0,1]$ always.

Deviance Computation

↳ Used in lieu of Sum of Squares Computation.

$$D = -2 \ln \frac{\text{likelihood of fitted model}}{\text{likelihood of Saturated model}}$$

'D' approx follows Chi Square distribution.

When Saturate model not avail.

$$D = -2 \ln \text{likelihood of fitted model})$$

Null Deviance → Model with just intercept vs saturated.

$$D_{null} = -2 \ln \left(\frac{\text{likelihood of Null model}}{\text{likelihood of Satu. model}} \right).$$

Model Deviance

$$D_{fitted} = -2 \ln \left(\frac{\text{likelihood of fitted model}}{\text{likelihood of SaturatedM}} \right.$$

Good Fit

For Model to have good fit.
Model Deviance Should be Significantly Smaller than NULL Deviance.

Likelihood Function. For Logestics Regression

Notes. → For Each training data Point we have a vector of featurs x_i, and an observed class y_i

Prob of that Class was p if $y_i = 1$
or $1-p$ if $y_i = 0$

Thus one can write

the likelihood as.

$$L(\beta_0, \beta) = \prod p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

The log likelihood is then.

$$l(\beta_0, \beta) = \sum_{i=1}^{n} y_i \log p(x_i) + (1-y_i) \log (1 - p(x_i))$$

[Goodness of Fit test]

$$\text{Pseudo } R^2 = \frac{D_{NULL} - D_{fitted}}{D_{NULL}}$$

$$= 1 - \frac{D_{fit}}{D_{NULL}}$$

Logit Regression
(Log Likelihood Max)

$$P(z) = \frac{e^z}{1+e^z} = (e^z)(1+e^z)^{-1}$$

Note.

$$1 - P(z) = 1 - \frac{e^z}{1+e^z} = \frac{1}{1+e^z}$$

$$P'(z) = (e^z)(1+e^z)^{-1} + (-1)(e^z)(1+e^z)^{-2} e^z$$

$$= \frac{e^z}{1+e^z} - \frac{(e^z)^2}{(1+e^z)^2} = \frac{e^z(1+e^z)}{(1+e^z)^2} - \frac{(e^z)^2}{(1+e^z)^2}$$

$$= \frac{e^z}{(1+e^z)} \cdot \frac{1}{(1+e^z)}$$

Thus $P'(z) = P(z)(1-P(z))$

Log likelihood]> $\mathcal{L}(X/P) = \sum_{i=1, y_i=1}^{N} \text{Log } P(x_i) + \sum_{i=0, y_i=0}^{N} \text{Log } (1-P(x_i))$

$$\nabla_b \mathcal{L} = \sum_{\substack{i=0 \\ y_i=1}}^{N} \frac{P_i'}{P_i} x_i - \sum_{\substack{i=0 \\ y_i=0}}^{N} \frac{P_i'}{1-P_i} \cdot x_i$$

$$\Rightarrow \sum_{\substack{i=1 \\ y_i=1}} (1-P_i) x_i - \sum_{\substack{i=1 \\ y_i=0}} P_i x_i$$

$$= \sum_{i=1}^{N} [y_i (1-P_i) - (1-y_i) P_i] x_i$$

Thus we arrive at the set of Simultaneous equations that are true at the optimum.

$$\sum_{i=1}^{N} y_i x_i - P_i x_i = 0$$

How to solve for coefficients: ?

Solving for coeff :→ [Logit R]

Suppose you have a vector valued function.
$f: y = f(b)$ you want the value b_{opt}.
Such that $f(b)_{opt} = 0$, Assume we start with initial guess (b_0)

$$\underbrace{f(b_0+\Delta)}_{0} \simeq f(b_0) + f'(b_0)\Delta$$

$$\Delta_0 = -\frac{f(b_0)}{f'(b_0)}$$

$$b_1 = b_0 + \Delta_0$$

In our case. $f = \nabla_b L = 0$
$H = \frac{\partial}{\partial b} \nabla_b L.$

Now $\nabla_b L = \sum y_i x_i - P_i x_i = 0$

$$H = \frac{\partial}{\partial b}(\nabla_b L) = -\sum_{i=1}^{N} x_i (\nabla_b P_i)$$

$$= -\sum_{i=1}^{N} x_i P_i (1-P_i) x_i^T$$

$$= XWX^T$$

thus $\Delta_k = (XW_kX^T)^{-1} X(y - P_k)$

Compare with Linear Regression.

$y = x^T b$

$Xy = XX^T b$

$b = (XX^T)^{-1} Xy$

L.R →
Like a weighted Least squares problem