Model selection.

↙ ↘

grouped ↓ contingency table.

ungrouped data.

Marijuana use

race	gender	Yes	No
White	F	420	620
White	M	483	579
Others	F	25	55
Others	M	32	62

Y - Marijuana use (Y, X)

X - gender

Z - race.

Test independence for X and Y

μ_{ij} - expected frequencies $\mu_{ij} = n\pi_{ij}$

$$\widehat{\mu_{ij}} = n\,\widehat{\pi}_{ij}$$

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

G^2 - test Pearson χ^2 - test.

$$\text{logit}(P(Y=1)) = \alpha + \underset{\substack{\downarrow\\ \text{gender}}}{\beta_1 X} + \underset{\substack{\downarrow\\ \text{race}}}{\beta_2 Z}.$$

X, Y are independent

H_0: $\beta_1 = 0$ Wald test

H_1: $\beta_1 \neq 0$ LRT- test

Goodness of fit test.

↙ comparison with a saturated model

↓

LRT → G^2 test

$$G^2 = 2 \sum \text{observed} \log \frac{\text{observed}}{\text{fitted}}.$$

score test. Pearson χ^2 test

$$X^2 = \frac{\sum(\text{observed} - \text{fitted})^2}{\text{fitted}} \sim \chi^2_{df}.$$

↘ Comparison with a null model.

↘

$$\left\{ \begin{array}{l} \text{logit}(P(Y=1)) = \alpha + \beta_1 x + \beta_2 z \\ \text{logit}(P(Y=1)) = \alpha \end{array} \right.$$

H_0: $\beta_1 = \beta_2 = 0$

H_1: at least one is not zero.

Residuals.

Pearson residuals $e_i = \dfrac{\overset{= y_i}{\text{observed}} - \text{predicted}}{\sqrt{Var(\hat{y}_i)}}$

score $X^2 = \sum_{i=1}^{n} e_i^2$

Stand. res. $= \dfrac{y_i - n\widehat{\pi_i}}{SE}$ $SE = \sqrt{Var(\hat{y}_i)(1-h_i)}$

St. res. $\sim$ $N(0,1)$ if the fit is good.

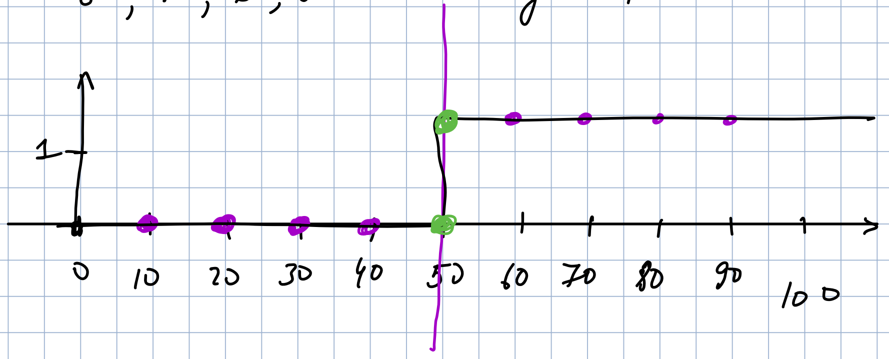
$|\text{st res}| < 2$.

< 3

Sparse data.

$x =$ 10, 20, 30, 40 $y = 0$

$x =$ 60, 70, 80, 90 $y = 1$



$\beta = \infty$ perfect separation.

quasi complete

$x =$ 10, 20, 30, 40, 50 $y = 0$

$x =$ 50, 60, 70, 80, 90 $y = 1$

Summary on model selection.

Model selection → backward step algorithm

→ search all models and select the best one using AIC (BIC) criteria

→ penalized regression (Lasso, Ridge;

Goodness of fit.

Method of goodness of fit or comparison of models.	Grouped data	Ungrouped data
1. Comparison with a saturated model.	Yes G^2-test LRT Pearson χ^2-test.	NO chi-sq. approx doesnot work.
2. Comparison of two nested model.	Yes LRT	Yes. LRT
3. Comparison with the null model	Yes LRT	Yes LRT
4. Residual analysis.	Yes	Yes.
5 Correlation	No	Yes.
6. ROC AUC	NO	Yes.

Alternative link functions.

$Y \nearrow 1$, $\searrow 0$

$$\text{logit}\,(P(Y=1)) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p .$$

$$'' \quad \log\left(\frac{\pi}{1-\pi}\right) \leftarrow \text{logit link.}$$

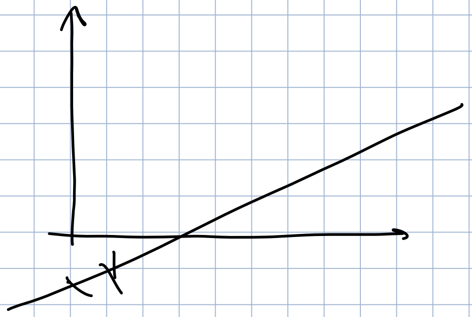
Identity link function,

- $\pi = \alpha + \beta_1 x_1 + --- + \beta_p x_p .$

$P(Y=1) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p .$

Problem ?

$\pi(x) = \alpha + \beta x$

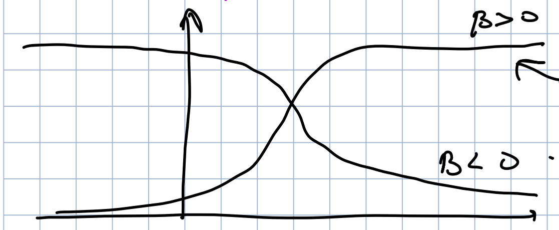


$\hat{\pi}$ can be < 0.

$\hat{\pi} > 1$.

lack of convergance and give an error message.

Probit model and norma latent variable model.



$\text{logit}\,(\pi(x)) = \alpha + \beta x$

cdf of $\quad \beta > 0$. S-Shaped.

some distribution (logistic distr.)

$Y \sim N(0,1)$ $\quad \Phi(y)$ - cdf of a st. normal distr.

$$\pi(x) = \Phi(\alpha + \beta x)$$

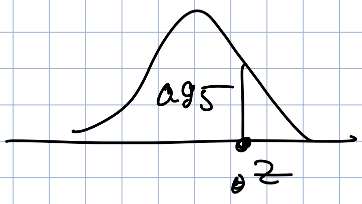
$$\underbrace{\Phi^{-1}(\pi(x))} = \alpha + \beta x$$

$$\text{probit}(\pi(x)) = \alpha + \beta x$$

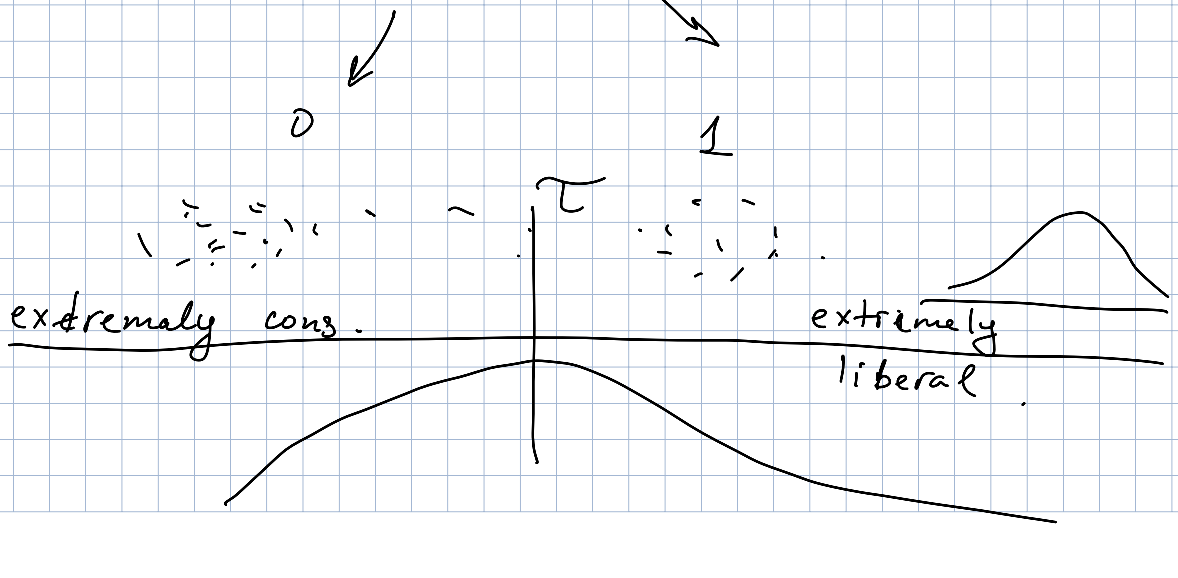
Probit link transforms $P(Y=1)$ to a z-score.

probit $(0.5) = 0$

probit $(0.95) = 1.645$



Interpretation of the parameters in probit model is simplest when we can relate the model to a corresponding normal linear model.



$$y^* \quad \text{if} \quad \begin{array}{ll} y^* < \tau & y = 0 \\ y^* > \tau & y = 1 \end{array}$$

$$y^* = \alpha + \beta_1 x_1 + \dots + \beta_p x_p + \boxed{\varepsilon}$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$\text{probit}(P(Y=1)) = \alpha + \beta x_1 + \dots + \beta_p x_p.$$

We can interpret $\hat{\beta}_j$ from probit model fit as representing the estimated changes in $E\, y^*$ for 1 unit increase in x_j adjusting for other explanatory variables

For arbitrary value for $\text{var}(\varepsilon)$ $\hat{\beta}_j$ is the estimated # of st. dev. that the distr. y^* shifts.

y^* - latent (unobserved variable)