ETH zürich

Applied Generalized Linear Models (FS 20)

Binary outcomes: logistic regression and probit models

Viviana Amati

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02 due to 19.03
25.02	Introduction to R and review of the linear regression mode	
03.03	The general linear model: ANOVA and ANCOVA	
10.03	Practical: ANOVA and ANCOVA	
17.03	**Binary outcomes: logistic regression and probit models**	Ass. 2 released on 17.03 due to 23.04
24.03	Practical: logistic regression and probit models	
31.03	Nominal outcomes: multinomial logistic regression	
07.04	Practical: multinomial logistic regression	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 21.04 due to 21.05
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models (lecture+practical)	
26.05	Regular lecture: panel data model	

A bit of (re-)organization

- Live streaming of the lecture
- Every Tuesday (but 14.04) at 17.15 via zoom
- No break during the lecture
- Join URL: https://ethz.zoom.us/j/765282314
  Meeting ID: 765 282 314
- For questions: emails and skype/zoom meetings
- Exam: I will let you know how we will proceed once we have all the available options

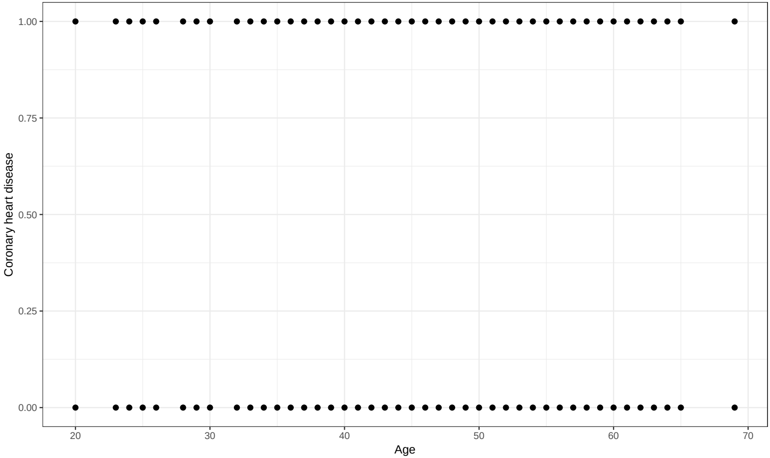
Today's agenda

- Models for binary outcomes
- Logistic regression model
- Probit models
- Latent variable models

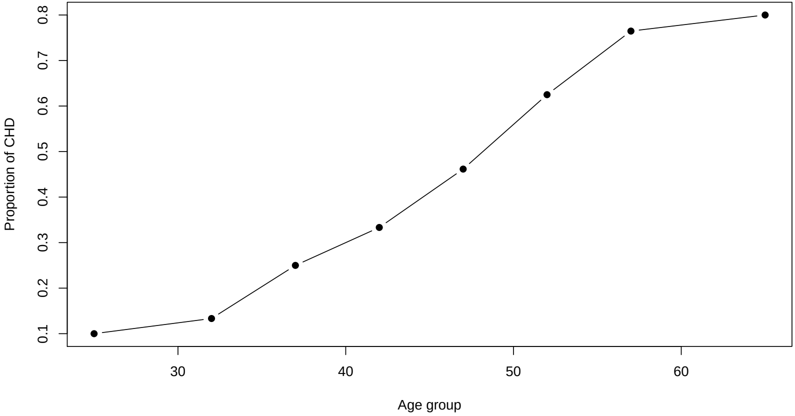
An example: CHD and age

- Hypothetical study of risk factors for heart disease (Hosmer et al., 2013)
- Y: presence (1) absence (0) of a coronary heart disease CHD
- X: age in years
- Developing a coronary heart disease as a function of age

An example: CHD and age



An example: CHD and age



Models for binary outcomes

- Binary dependent variable

$$Y = \begin{cases} 1 & \text{if an event occurs (success)} \\ 0 & \text{otherwise (failure)} \end{cases} ,$$

- Y has a Bernoulli distribution

$$P(Y=1) = \pi \qquad P(Y=0) = 1 - \pi$$

- Investigate the relation between Y and $X_1, \ldots, X_p$

- Information on n entities independently sampled from a population

Linear probability model (LPM)

- Model formulation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

- If $E[\varepsilon_i] = 0$

$$E[Y \,|\, \mathbf{X}] = \pi(\mathbf{x}) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

with

$$E[Y \,|\, \mathbf{X}] = P(Y = 1 | \, \mathbf{X}) \ .$$

- LPM is an LRM for the probability $\pi(\mathbf{x})$

- β_j: expected change in $\pi(\mathbf{x})$ for a unit increase of X_j, controlling for all the other variables

Linear probability model

The LPM has three "structural defects":

1. Range
   $\pi(\mathbf{x})$ takes values in the unit interval $[0, 1]$
   The LPM can predict values of $\pi(\mathbf{x})$ greater than 1 or less than 0

2. Linearity
   Typically, the relation between X_j and $\pi(\mathbf{x})$ is described by an s-shaped curve

Linear probability model - issues

3. Normality and homoschnedasticity of the error term are not met

   – Given
   $$Y_i = E[Y \,|\, \mathbf{X}] + \varepsilon_i = \pi(\mathbf{x}_\mathrm{i}) + \varepsilon_i \ ,$$
   the error term can only take the two values
   $$\varepsilon_i = \begin{cases} 1 - \pi(\mathbf{x}_\mathrm{i}) & \text{if } y_i = 1 \\ -\pi(\mathbf{x}_\mathrm{i}) & \text{if } y_i = 0 \end{cases}$$

   – the variance of ε_i is not constant across the observations
   $$\mathrm{E}[\varepsilon_i] = 0 \qquad \text{and} \qquad \mathrm{Var}[\varepsilon_i] = \pi(\mathbf{x}_\mathrm{i})(1 - \pi(\mathbf{x}_\mathrm{i})) \quad .$$

Linear probability model - issues

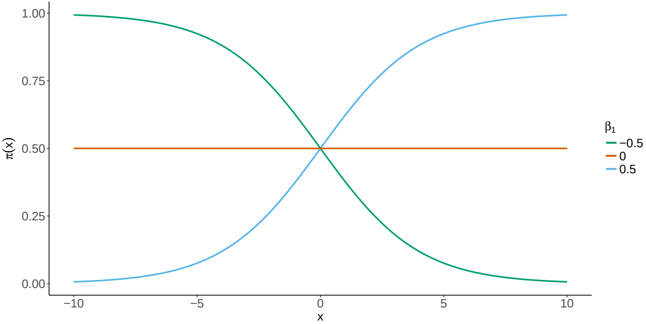
- The simplest model: one explanatory variable

$$\text{logit}[\pi(\mathbf{x})] = \log \Bigg[\underbrace{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}}_{\text{odds}} \Bigg] = \beta_0 + \beta_1 X_1$$

- The logistic distribution

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

The logistic distribution



More than one explanatory variable

- The simplest model: one explanatory variable

$$\text{logit}[\pi(\mathbf{x})] = \log\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

- The logistic distribution

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

Binary logistic regression model as a GLM

1. The random component $Y \,|\, \mathbf{X}$ has a binomial distribution

2. The systematic component is $\sum_{j=1}^{p} \beta_j X_j$

3. The link function is the *logit* function defined as the logarithm of the odds of a success conditional on $\mathbf{x}$

$$\text{logit}[\pi(\mathbf{x})] = \log \left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right] .$$

Parameter estimation

- Maximum likelihood estimation

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \max_{\beta} L(\boldsymbol{\beta}, \mathbf{y}) \\
&= \prod_{i=1}^{n} P(Y_i = y_i | \, \mathbf{x}) \\
&= \prod_{i=1}^{n} \pi(\mathbf{x}_{\mathrm{i}})^{y_i} [1 - \pi(\mathbf{x}_{\mathrm{i}})]^{(1-y_i)}
\end{aligned}
$$

Maximum likelihood estimation

- Log-likelihood

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \max_{\beta} \ell(\boldsymbol{\beta}, \mathbf{y}) \\ &= \sum_{i=1}^{n} \Big[y_i \log \pi(\mathbf{x}_\mathrm{i}) + (1 - y_i) \log[1 - \pi(\mathbf{x}_\mathrm{i})] \Big]\end{aligned}$$

- System non-linear in the parameters

$$\begin{cases} \sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_\mathrm{i})] = 0 & \text{derivative w.r.t. } \beta_0 \\ \sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_\mathrm{i})] x_{ij} = 0 & \text{derivative w.r.t. } \beta_j \end{cases}$$

- Approximation using iterative methods

Maximum likelihood estimator

- Asymptotic distribution

$$\mathbf{B} \sim N\big(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta})\big)$$

with $\mathbf{I}^{-1}(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \mathbf{y})}{\partial \beta_i, \partial \beta_j}\right]$ is the Fisher information matrix

- $\mathbf{B}$ is (asymptotically) the minimum variance unbiased estimator (MVUE)

Hypotheses testing: Wald test

One single parameter β_j

- Hypotheses
$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$
H_0: no association between Y and X_j

- Test statistic
$$W = \frac{B_j}{s.e.(B_j)} \sim Z$$
with $Z \sim N(0, 1)$

- Rejection region
$$\left| \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \right| \geq z_{1-\alpha/2}$$
with $z_{1-\alpha/2}$ the quantile of the standard normal that leaves on its left a probability of $1 - \alpha/2$.

Hypotheses testing: Likelihood ratio test

All β_j

- Hypotheses

$$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one } \beta_j \neq 0,\ j = 1, \ldots, p$$

  H_0: the model does not explain the variability of Y

- Reduced and full model

$$\begin{aligned} \text{logit}[\pi(\mathbf{x})] &= \beta_0 && \text{(under } H_0\text{)} \\ \text{logit}[\pi(\mathbf{x})] &= \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p && \text{(under } H_1\text{)} \end{aligned}$$

- Test statistic

$$G = -2 \log \left[\frac{L\left(\hat{\beta}, \mathbf{y} | H_0\right)}{L\left(\hat{\beta}, \mathbf{y} | H_1\right)} \right] = -2\ell\left(\hat{\beta}, \mathbf{y} | H_0\right) + 2\ell\left(\hat{\beta}, \mathbf{y} | H_1\right)$$

- Rejection region

$$G > \chi^2_{p,1-\alpha}$$

  $\chi^2_{p,1-\alpha}$ is the quantile of a χ^2_p leaving on its left a probability of $1 - \alpha$.

Parameter interpretation

- Odds ratio

$$OR = \frac{\pi(\mathbf{x}+\delta)/[1-\pi(\mathbf{x}+\delta)]}{\pi(\mathbf{x})/[1-\pi(\mathbf{x})]} = e^{\delta\beta_j}$$

  with

  - $\mathbf{x} = (x_1, \dots, x_j, \dots, x_p)$
  - $\mathbf{x} + \delta = (x_1, \dots, x_j + \delta, \dots, x_p)$

- Percentage changes in the odds:

$$OR_{\%} = 100 \cdot \left[\frac{\text{odds}(\mathbf{x}+1) - \text{odds}(\mathbf{x})}{\text{odds}(\mathbf{x})}\right] = 100 \cdot [e^{\beta_j} - 1]$$

Example

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.309	1.134	-4.683	0.000
age	0.111	0.024	4.610	0.000

- Significant association between age and CHD
- $OR = e^{0.111}$=1.117:
  For a one year increase in age, the odds of a CHD are increased by a factor of 1.117, holding all other variables constant
- $OR_{\%} = 100 \cdot [e^{0.111} - 1] = 11.7\%$:
  For each additional year in age, the odds of being admitted are increased 11.7%, holding all other variables constant.constant

Probit model

A GLM

1. The random component $Y \mid \mathbf{X}$ has a binomial distribution

2. The systematic component is $\sum_{j=1}^{p} \beta_j X_j$

3. The link function is the inverse of the cumulative distribution function of a standard normal distribution:

$$\phi^{-1}[\pi(\mathbf{x})] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

   with ϕ denotes the cumulative distribution function of a standard normal distribution

Probit vs. logit model

- Usually lead to the same results
- Logit transformation is preferred to the probit because
  - the logit allows to write $\pi(\mathbf{x})$ in a closed form
  - the logit can be easily interpreted using the OR
  - the probit model is more difficult to estimate

Latent variable model

- Y binary variable
- Y^* latent continuous variable ranging from $-\infty$ to ∞
- Y^* is assumed to be linearly related to X through the model

$$y_i^* = \beta_0 + \beta_1 X + \varepsilon_i$$

- The variable y_i^* is linked to the observed binary variable y_i by the equation

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq \tau \\ 0 & \text{if } y_i^* < \tau \end{cases}$$

  where τ is a threshold

Latent variable model

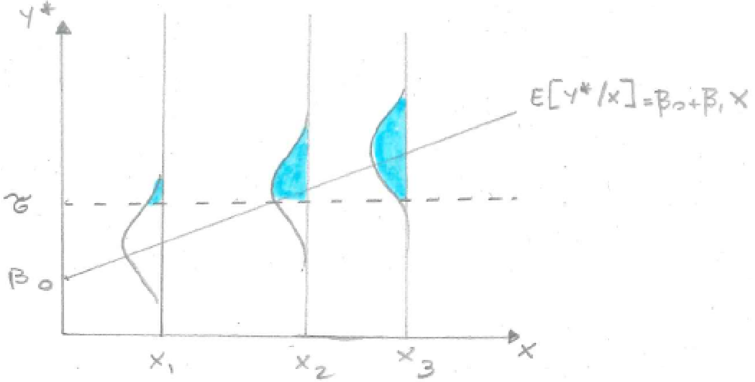
- To estimate the parameters of

$$y_i^* = \beta_0 + \beta_1 X + \varepsilon_i$$

  we use the MLE

- Assumptions on the distribution of the error terms
  - ε_i has the logistic distribution $\rightarrow$ binary logistic regression model
  - ε_i has the normal distribution $\rightarrow$ probit model

Latent variable model



References

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.