MONASH
BUSINESS
SCHOOL

ETC3580: Advanced Statistical Modelling

Week 3: Binary responses

Outline

Logistic regression

Suppose response variable Y_i takes values 0 or 1 with probability $P(Y_i = 1) = p_i$. (i.e., a Bernoulli distribution)

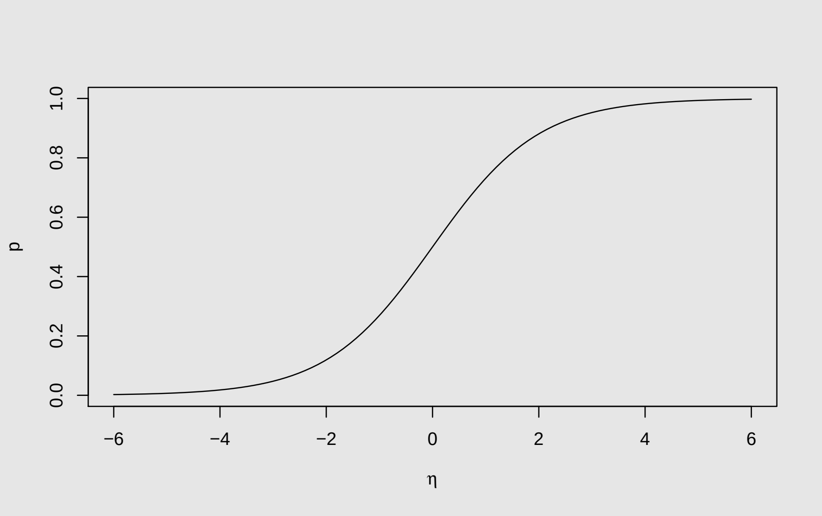
We relate p_i to the predictors:

$$p_i = e^{\eta_i}/(1 + e^{\eta_i}) = P(Y_i = 1)$$

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

- $g(p) = \log(p/(1 - p))$ is the "logit" function. It maps $(0, 1) \to \mathbb{R}$.
- The inverse logit is $g^{-1}(\eta) = e^{\eta}/(1 + e^{\eta})$ which maps $\mathbb{R} \to (0, 1)$.
- g is called the "link" function.

Inverse logit function



- If $p_i \approx 0.5$, logistic and linear regression similar.

Log-likelihood

Likelihood

$$L = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\begin{aligned}
\log L(\beta) &= \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \\
&= \sum_{i=1}^{n} y_i[\eta_i - \log(1 + e^{\eta_i})] - (1 - y_i)\log(1 + e^{\eta_i}) \\
&= \sum_{i=1}^{n} [y_i\eta_i - \log(1 + e^{\eta_i})]
\end{aligned}$$

Log-likelihood

Likelihood

$$L = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\begin{aligned}
\log L(\beta) &= \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \\
&= \sum_{i=1}^{n} y_i[\eta_i - \log(1 + e^{\eta_i})] - (1 - y_i)\log(1 + e^{\eta_i}) \\
&= \sum_{i=1}^{n} [y_i\eta_i - \log(1 + e^{\eta_i})]
\end{aligned}$$

- MLE: maximize $\log L$ to obtain $\hat{\beta}$
- Equivalent to iterative weighted least squares
- No closed form expressions

Logistic regression in R

```r
fit <- glm(y ~ x1 + x2, family=binomial,
                data=df)
```

- Bernoulli is equivalent to Binomial with only two levels.
- First (alphabetical) level is set to 0, other to 1. Use `relevel` if you want to change it.
- `glm` uses MLE with a logit link function (when `family=binomial`).

Interpreting a logistic regression

Odds = $p/(1-p)$.

- Example: 3-1 odds means $p = 3/4$.
- Example: 5-2 odds means $p = 5/7$.
- A horse at "15-1 against" means $p = 1/16$.
- Odds are unbounded.
- $\log(\text{odds}) = \log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- A unit increase in x_1 with x_2 held fixed increases log-odds of success by β_1.
- A unit increase in x_1 with x_2 held fixed increases odds of success by a factor of e^{β_1}.

Outline

Fitted values

Link: on $(-\infty, \infty)$

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_q x_{i,q}$$

Fitted values

Link: on $(-\infty, \infty)$

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_q x_{i,q}$$

Response: on $(0, 1)$ = probability of "success"

$$e^{\hat{\eta}_i} / (1 + e^{\hat{\eta}_i})$$

Fitted values

Link: on $(-\infty, \infty)$

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_q x_{i,q}$$

Response: on $(0, 1)$ = probability of "success"

$$e^{\hat{\eta}_i}/(1 + e^{\hat{\eta}_i})$$

```
fit <- glm(y ~ x1 + x2, family='binomial',
           data=df)
library(broom)
augment(fit, type.predict='link') # default
augment(fit, type.predict='response')
```

OLS Residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{y}_i$$

OLS Residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{y}_i$$

Pearson residuals: Standardized

$$r_i = e_i / \hat{\sigma}$$

OLS Residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{y}_i$$

Pearson residuals: Standardized

$$r_i = e_i / \hat{\sigma}$$

- Mean 0, variance 1.

OLS Residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{y}_i$$

Pearson residuals: Standardized

$$r_i = e_i / \hat{\sigma}$$

- Mean 0, variance 1.

Deviance residuals: Signed root contribution to $-2 \log L$.

$$-2 \log L = c + \frac{1}{\hat{\sigma}^2} \sum e_i^2 = c + \sum d_i^2$$

$$d_i = e_i / \hat{\sigma}$$

Logistic regression residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{p}_i$$

Logistic regression residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{p}_i$$

Pearson residuals: Standardized

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

Logistic regression residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{p}_i$$

Pearson residuals: Standardized

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

- Mean 0, variance 1.

Logistic regression residuals

Response residuals: Observation – estimate

$$e_i = y_i - \hat{p}_i$$

Pearson residuals: Standardized

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

- Mean 0, variance 1.

Deviance residuals: Signed root contribution to

$$-2 \log L = -2 \sum [y_i \eta_i - \log(1 + e^{\eta_i})] = \sum d_i^2$$

$$d_i = \text{sign}(y_i - \hat{p}_i) \sqrt{2 \left[-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) \right]}$$

Logistic regression residuals

```
fit <- glm(y ~ x1 + x2, family='binomial',
           data=df)
library(broom)
augment(fit, type.resid='response')
augment(fit, type.resid='pearson')
augment(fit, type.resid='deviance') # default
```

- Residual plots can be hard to interpret
- Don’t expect residuals to be normally distributed

Partial residuals

- Let $e_i = y_i - \hat{p}_i$ be a response residual. Then

$$e_i^* = \frac{e_i}{\hat{p}_i(1-\hat{p}_i)}$$

  is the "logit residual" (compare Pearson residuals).

- The "logit partial residual" for the jth variable is

$$e_i^* + \hat{\beta}_j x_{i,j}$$

- We can plot these against $x_{i,j}$ to identify potential nonlinearity.
- `visreg` plots $d_i + \hat{\beta}_j x_{i,j}$ instead.

Outline

Deviance

Generalization of sum of squared residuals based on likelihood ratio:

$$D = \sum d_i^2 = -2 \log L + c$$

where L is the likelihood of the model and c is constant that depends on data but not model.

- Difference between deviances equivalent to a likelihood ratio test.
- $D_1 - D_2 \sim \chi^2_{q_2 - q_1}$ where q_i is df for model i assuming
  1. smaller model is correct
  2. models are nested
  3. distributional assumptions true
- Null deviance is for model with only an intercept.

Deviance test

In R:

```
fit <- glm(y ~ x1 + x2, family='binomial',
           data=df)
anova(fit, test="Chisq")
drop1(fit, test="Chisq")
anova(fit1, fit2, test="Chisq")
```

- NOT equivalent to t-tests on coefficients
- Deviance tests preferred

Confidence intervals for coefficients

- Standard intervals based on normal distribution are poor approximations.
- Better to use “profile likelihood” confidence intervals
- Let $L_p(\theta)$ be the profile likelihood (the likelihood without the nuisance parameters). LR test for $H_0 : \theta = \theta_0$ is

  $$LR = 2\left[\log L_p(\hat{\theta}) - \log L_p(\theta_0)\right]$$

  Confidence interval consists of those values θ_0 for which test is not significant. (A contour region of the likelihood.)
- Implemented in R using `confint`

Model selection

Akaike's Information Criterion

$$\text{AIC} = -2\log L + 2q = c + D + 2q$$

- Select model with smallest AIC
- Beware of hypothesis tests after variable selection
- `step()` works for `glm` objects in the same way as for `lm` objects, and minimizes the AIC.

Outline

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is "standard logistic", then $F(w) = 1/[1 + e^{-w}]$.
- So $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

That is, we can think of logistic regression as an ordinary regression with logistic noise, and we observe only if it is above or below 0.

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is “standard normal”, then $F(w) = \Phi(w)$.
- So $\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is “standard normal”, then $F(w) = \Phi(w)$.
- So $\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.
- Here Φ^{-1} is the link function.

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is "standard normal", then $F(w) = \Phi(w)$.
- So $\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.
- Here Φ^{-1} is the link function.

That is, probit regression is an ordinary regression with normal noise, and we observe only if it is above or below 0.

Latent variable interetation

General binary model

$$p_i = g(\eta_i) = P(Y_i = 1)$$

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

where g maps $\mathbb{R} \to (0, 1)$.

- $g(\eta) = e^\eta/(1 + e^\eta)$: logit link, logistic regression
- $g(\eta) = \Phi(\eta)$: normal cdf link, probit regression
- $g(\eta) = 1 - \exp(-\exp(\eta))$: log-log link

Latent variable interetation

General binary model

$$p_i = g(\eta_i) = P(Y_i = 1)$$

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

where g maps $\mathbb{R} \to (0, 1)$.

- $g(\eta) = e^\eta/(1 + e^\eta)$: logit link, logistic regression
- $g(\eta) = \Phi(\eta)$: normal cdf link, probit regression
- $g(\eta) = 1 - \exp(-\exp(\eta))$: log-log link

```
fit <- glm(y ~ x1 + x2,
  family=binomial(link=probit), data=df)
```

Why prefer logit over probit?

- odds ratio interpretation of coefficients
- non-biased with disproportionate stratified sampling. Only link function with this property.
- non-biased with clustered observations. Only link function with this property.
- logit link is “canonical”: ensures $\sum x_{ij} y_i$ ($j = 1, \ldots, q$) are sufficient for estimation. So p-values are exact.