Generalized Linear Models

Lecture 6. Models with binary response

Model with binary response

Response has two possible values: yes/no, success/failure, exists/does not exist

Values are usually coded as 1/0, so that

$$\mathbf{P}(Y = 1) = \pi; \ \mathbf{P}(Y = 0) = 1 - \pi$$

Question of interest

How is the occurrence probability π related to the arguments?

Distribution of a binary random variable

Bernoulli distribution (Swiss mathem. Jacob Bernoulli, 1654–1705)

Random variable Y has Bernoulli distribution, $Y \sim B(1, \pi)$ if its pmf is

$$p(y; \pi) = \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\}$$

If $Y \sim B(1, \pi)$ then $\mathbf{E}Y = \pi$ and $\mathbf{D}Y = \pi(1 - \pi)$

If $Y_1, \ldots, Y_n$, $Y_i \sim B(1, \pi)$ are mutually independent then $\sum_{i=1}^n Y_i \sim B(n, \pi)$, π is often interpreted as the probability of 'success'

Binomial distribution

Random variable Y has binomial distribution, $Y \sim B(n, \pi)$ if its pmf is

$$p(y; n, \pi) = C_n^y \pi^y (1 - \pi)^{n-y}, \; C_n^y = \frac{n!}{y!(n-y)!}$$

If $Y \sim B(n, \pi)$ then $\mathbf{E}Y = n\pi$ and $\mathbf{D}Y = n\pi(1 - \pi)$

If $n \to \infty$ then $B(n, \pi)$ converges to normal $N(n\pi, \; n\pi(1 - \pi))$

Grouped and ungrouped data

Grouped data: group sizes $n_1, \ldots, n_n$ (n – number of groups)
Observations can be treated as proportions:

$$\frac{y_1}{n_1}, \ldots, \frac{y_n}{n_n},$$

y_i – number of successes in n_i trials

If the observations are independent and the probability of success is constant for each element in a group, then the response has binomial distribution

For ungrouped data $n_1 = \ldots = n_n = 1$

In case of grouped data, response has *binomial distribution*
In case of ungrouped data, response has *Bernoulli distribution*

Example. Ungrouped data

Vasoconstriction data, Finney (1947)

Reaction at fingertips while breathing in deeply (narrowing of blood vessels)
$y = 1$ (reaction); $y = 0$ (no reaction)
Arguments: volume of inhaled air and rate of inhalation (both continuous)

```
  y    volume     rate
  1     3.70     0.285
  1     3.50     1.090
 ....
  0     0.60     0.750
  0     1.10     1.700
```

Example. Grouped data

Data about seed sprouting

n – cultivated seeds, r – sprouted seeds
cult $= 0/1$ – two different cultures, soil $= 0/1$ – two different soil conditions

n	r	cult	soil
16	8	0	0
51	26	0	0
81	23	1	0
30	10	1	0
....			
51	32	0	1
72	55	0	1
74	53	1	1
56	12	1	1

Bernoulli distribution and exponential family

Let us start with Bernoulli pmf

$$p(y_i; \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

and rewrite it as

$$p(y_i; \pi_i) = \exp[y_i \ln \frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i)]$$

Thus we have

- $\theta_i = \ln \frac{\pi_i}{1-\pi_i}$
- $b(\theta_i) = -\ln(1-\pi_i) = \ln(1+e^{\theta_i})$
- $\varphi_i = 1$

We can also derive mean $\mu_i = b^{'}(\theta_i) = \pi_i$ and variance $\varphi_i b^{''}(\theta_i) = \pi_i(1-\pi_i)$
Prove it!

Canonical link is **Logit**: $g(\mu_i) = g(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}$

Binomial distribution and exponential family, 1

Let us start with binomial pmf (assume n_i is known)

$$p(y_i; n_i, \pi_i) = C_{n_i}^{y_i}\ \pi^{y_i}(1-\pi_i)^{n_i - y_i}$$

and rewrite it as

$$\begin{aligned} p(y_i; n_i, \pi_i) &= \exp[y_i \ln(\pi_i) + n_i \ln(1-\pi_i) - y_i \ln(1-\pi_i) + \ln C_{n_i}^{y_i}] \\ &= \exp[y_i \ln \frac{\pi_i}{1-\pi_i} + n_i \ln(1-\pi_i) + \ln C_{n_i}^{y_i}] \end{aligned}$$

Thus we have

- $\theta_i = \ln \frac{\pi_i}{1-\pi_i}$
- $b(\theta_i) = -n_i \ln(1-\pi_i) = n_i \ln(1+e^{\theta_i})$
- $\varphi_i = 1$

We can also derive mean $\mu_i = b^{'}(\theta_i) = n_i\pi_i$ and variance $\varphi b^{''}(\theta_i) = n_i\pi_i(1-\pi_i)$
Prove it!

Canonical link is $g(\mu_i) = \ln \frac{\pi_i}{1-\pi_i}$

Binomial distribution and exponential family, 2

Consider a GLM setup with y_i being a realization from $B(n_i, \pi_i)$

Question(s)

Are y_i-s comparable? Do we actually want to model y_i-s against arguments?

Binomial distribution and exponential family, 2

Consider a GLM setup with y_i being a realization from $B(n_i, \pi_i)$

Question(s)

Are y_i-s comparable? Do we actually want to model y_i-s against arguments?

Answer

Not really, it would be more informative to consider $\frac{y_i}{n_i}$ instead.

In other words, instead of $Y_i \sim B(n_i, p_i)$ we consider $Y_i^* = Y_i/n_i$
For Y_i^* we have

- $\theta_i = \ln \frac{\pi_i}{1-\pi_i}$
- $b(\theta_i) = -\ln(1-\pi_i) = \ln(1+e^{\theta_i})$
- $\varphi = 1$, $a_i = \frac{1}{n_i}$, $\varphi_i = \frac{1}{n_i}$
- $\mu_i = b^{'}(\theta_i) = \pi_i$
- $\varphi b^{''}(\theta_i) = \pi_i(1-\pi_i)$

Prove it!

Canonical link is *Logit*: $g(\mu_i) = \ln \frac{\pi_i}{1-\pi_i} = \ln \frac{\mu_i}{1-\mu_i}$

Goodness of fit. Grouped data

Let us denote $y_i^* = y_i/n_i$, $\boldsymbol{y}^* = (y_i^*, \dots, y_n^*)^T$
Deviance for (scaled) binomial model:

$$D = -2[l(\hat{\boldsymbol{\pi}}) - l(\boldsymbol{y}^*)] = 2[l(\boldsymbol{y}^*) - l(\hat{\boldsymbol{\pi}})]$$
$$= 2\sum \left(n_i[y_i^* \ln \frac{y_i^*}{\hat{\pi}_i} + (1 - y_i^*) \ln \frac{1 - y_i^*}{1 - \hat{\pi}_i}] \right) = 2\sum o \ln \frac{o}{e}$$

Pearson χ^2-statistic

$$\chi_P^2 = \sum \frac{n_i(y_i^* - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)} = \sum \frac{(o - e)^2}{e},$$

o – observed e – expected

If H_0 holds (NB! in case of grouped data the asymptotic means $n_i \to \infty$)

$$D \overset{a}{\sim} \chi^2_{n-p}, \quad \chi^2_P \overset{a}{\sim} \chi^2_{n-p}$$

Goodness of fit. Ungrouped data

Hosmer-Lemeshow' test (1989): used in case of ungrouped data

Idea: subjects will be decided into classes (e.g. by estimated deciles, 10 classes). Pearson's χ^2-statistic is used to measure the agreement between observed and expected values

Pros: simple to use, implemented in most statistical packages

Cons:
– conservative test, often has too low power
– arbitrary to choice of bins and method of computing quantiles
– in case of small number of classes (less than 5) almost always shows fit

Hallett (1999) *Goodness of fit tests in logistic regression*

Example. Goodness of fit

Vaso data:

```
> v1=glm("y~volume+rate",family="binomial",data=vaso)
> library(ResourceSelection)
> hoslem.test(x=vaso$y,y=predict(v1,type="response"))
Hosmer and Lemeshow goodness of fit (GOF) test

data:  vaso$y, predict(v1, type = "response")
X-squared = 17.812, df = 8, p-value = 0.02268
```

Conclusion?

Measures for goodness of fit

- Akaike information criterion $AIC = -2\log L + 2p$
- Schwarz (Bayes) criterion $BIC = -2\log L + p\ln n$
- Generalized coefficients of determination R^2

  Cox & Snell generalized coefficient of determination (1989):

$$R^2_{CS} = 1 - \left\{ \frac{L(0)}{L(\hat{\beta})} \right\}^{2/n}$$

  $L(0)$ – likelihood of constant model $L(\hat{\beta})$ – likelihood of current model

  $R^2_{CS} < 1$ since $R^2_{CS_{max}} = \{L(0)\}^{2/n}$, n – sample size

  Others:

  - Nagelkerke (1991) max-rescaled R^2: $\tilde{R}^2 = \frac{R^2_{CS}}{R^2_{CS_{max}}}$
  - McFadden's R^2: $R^2_{McF} = 1 - \frac{l(M)}{l(0)}$
  - Deviance R^2: $R^2_D = \frac{l(M)-l(0)}{l(S)-l(0)}$

Remarks about generalized R^2

- $R^2 \in (0, 1)$, the bigger, the better model
- values are relatively small
  empirical estimates: $R^2_{MF} \in (0.2, 0.4)$ is considered satisfactory
- are used to compare models with same number of arguments
- does not have nice reasonable explanation

NB! Definitely can not be interpreted as a measure describing the response's variability!

```
> library(DescTools)
> PseudoR2(v1,which="all")
      McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson VeallZimmermann          Effron
     0.4490675       0.3380383       0.4632616       0.6178176       0.1590559       0.5366965       0.5344613
McKelveyZavoina            Tjur             AIC             BIC          logLik         logLik0              G2
     0.7326604       0.5198015      35.7723045      40.7629894     -14.8861522     -27.0199181      24.2675318
```

Remark about coding (ordering) the response

Usual assumption for binary response is that we estimate the probability of "success" (i.e. value 1)

$$Logit(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}; \qquad \pi_i = \mathbf{P}(Y_i = 1)$$

$$Logit(\pi_i) = -Logit(1-\pi_i)$$

$\Rightarrow$ estimating model for $Y_i = 0$ means the change of signs for the coefficients

R: for grouped data (Seeds example):

```
s1 = glm(cbind(r,n-r)~cult+soil, family="binomial", data=seeds)
```

vs

```
s2 = glm(cbind(n-r,r)~cult+soil, family="binomial", data=seeds)
```

Choices of link function

GLM with binary/binomial response:

- Model: $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \eta_i = g(\pi_i), \quad \pi_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta})$
- As $\mu_i = \pi_i$ is a probability, it is restricted to $[0, 1]$
- Linear predictor η_i can take any values on real line
- Canonical link: Logit-function

In general, any one-to-one continuous and differentiable transformation that maps probabilities into real line could be used to produce a GLM

Now, consider some cdf F such that

$$\pi_i = F(\eta_i), \ -\infty < \eta_i < \infty$$

Then the inverse $\eta_i = F^{-1}(\pi_i)$ can be considered as a link function
Popular choices:

- normal distribution
- logistic distribution (gives canonical link)
- extreme value distribution

A latent variable formulation

Y_i – binary r.v., manifest response
Y_i^* – continuous r.v., latent (unobservable) such that

$$Y_i = \begin{cases} 1, & \text{iff } Y_i^* \geq \theta \\ 0, & \text{iff } Y_i^* < \theta, \end{cases}$$

where θ is some threshold
The latent variable defining the binary process is also called **tolerance**

$\Rightarrow$ probability of success is given by

$$\pi_i = \mathbf{P}(Y_i = 1) = \mathbf{P}(Y_i^* > \theta)$$

Now, as location and scale Y_i^* are arbitrary, we take $\theta = 0$ and standardize Y_i^* to identify the model:

$$Y_i^* = \boldsymbol{x}_i^T \beta + U_i, \qquad U_i \sim F$$

$$\pi_i = \mathbf{P}(Y_i^* > 0) = \mathbf{P}(U_i > -\eta_i) = 1 - F(-\eta_i)$$

- symmetric F: $1 - F(-\eta_i) = F(\eta_i), \quad \eta_i = g(\pi_i) = F^{-1}(\pi_i)$
- general F: $\eta_i = g(\pi_i) = -F^{-1}(1 - \pi_i)$

Latent variable and manifest response

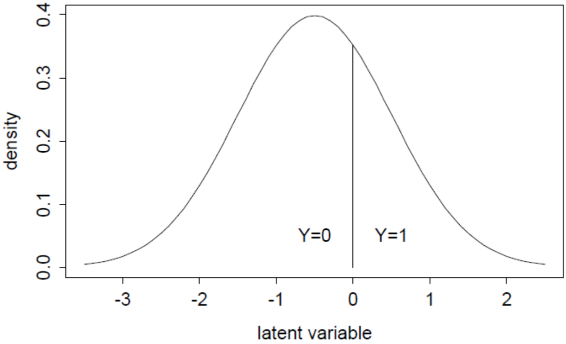


FIGURE 3.6: Latent Variable and Manifest Response

Distribution of tolerance

- $U_i \sim N(0,1)$ – **Probit** link

$$\pi_i = F(\eta_i) = \Phi(\eta_i), \quad \eta_i = \Phi^{-1}(\pi_i)$$

- U_i (standard) logistic – **Logit** link

$$\pi_i = F(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \qquad \eta_i = \ln\frac{\pi_i}{1-\pi_i}$$

- U_i extreme min distributed (Gompertz dist.) – **complementary log-log** (also CLL, Gombit, Gompit)

$$\pi_i = F(\eta_i) = 1 - \exp(-\exp(\eta_i)), \quad \eta_i = \log(-\log(1-\pi_i))$$

- U_i extreme max distributed (Gumbel dist.) – log-log link

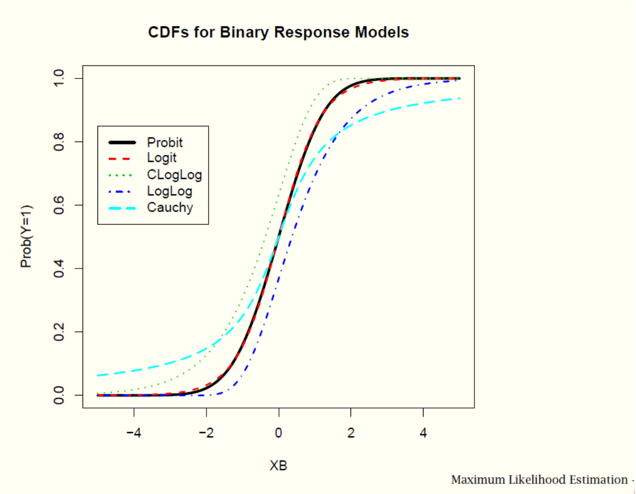
$$\pi_i = F(\eta_i) = \exp(-\exp(\eta_i)), \quad \eta_i = \log(-\log(\pi_i))$$

- U_i Cauchy distributed (also called Cauchy-Lorentz)

$$\pi_i = F(\eta_i) = \pi^{-1}\arctan(\eta_i) + \frac{1}{2}, \quad \eta_i = \tan[\pi(\pi_i - \frac{1}{2})], \quad \pi = 3.1415\ldots$$

Note: Cauchy distribution has heavy tails, moments do not exist

Distributions of tolerance



Less known link functions

- U_i exponentially distributed – complementary log link

$$\pi_i = 1 - \exp(\eta_i), \quad \eta_i = -\log(1 - \pi_i)$$

or log-link, if $1 - \pi_i$ is chosen instead of π_i:

$$\pi_i = \exp(\eta_i), \quad \eta_i = \log(\pi_i)$$

- id-model, identity link $\pi_i = \eta_i$ (seldom used, only if the range of arguments is restricted)

Nagler (1994): *Scobit* link (Skewed Logit) – an alternative to *Logit* and *Probit*

Scobit is not symmetric w.r.t. 0.5 but 0.5α: if $\alpha = 1$, it reduces to *Logit* model

Probit model

Distribution of tolerance is normal, $U_i \sim N(0,1)$

$$\pi_i = F(\eta_i) = \Phi(\eta_i), \quad \eta_i = \Phi^{-1}(\pi_i)$$

In general, one can take $U_i \sim N(0,\sigma^2)$
Using normal U_i, the latent variable model $Y_i^* = \boldsymbol{x}_i^T \boldsymbol{\beta} + U_i$ gives

$$\pi_i = \mathbf{P}(Y^* > 0) = \mathbf{P}(U_i > -\eta_i) = \mathbf{P}(\frac{U_i}{\sigma} > \frac{-\eta_i}{\sigma}) = 1 - \Phi(-\frac{\eta_i}{\sigma}) = \Phi(\frac{\eta_i}{\sigma})$$

$\Rightarrow$ we can not separately estimate $\boldsymbol{\beta}$ and σ, i.e. scale of the latent variable is not uniquely defined

Choosing $\sigma = 1$ means that we interpret parameters $\boldsymbol{\beta}$ in units of std. dev. of the latent variable

Pros: good numerical solution methods

Cons: no analytic form, hard to interpret

Logit model

Distribution of tolerance is logistic distribution

$$\pi_i = F(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \qquad \eta_i = \ln\frac{\pi_i}{1-\pi_i}$$

Standard logistic distribution is symmetric: $\mu = 0, \quad \sigma^2 = \frac{\pi^2}{3} \approx \frac{3.14^2}{3} \approx 3.29$

Alternative to *Probit* model, shape similar to normal, tails a bit heavier, simple analytic form, easy to interpret, Logit-link is canonical for binary response.

Logit vs *Probit*
– both symmetric w.r.t.0.5
– similar results if $\pi \in (0.1, 0.9)$
Note that only the comparison of ratio $\boldsymbol{\beta}/\sigma$ makes sense. **Why?**

In case of *Probit* model $\sigma = 1$, in case of *Logit* model $\sigma = \frac{\pi}{\sqrt{3}} \approx \frac{3.14}{\sqrt{3}} \approx 1.814$

$\Rightarrow$ it is reasonable to compare $\hat{\boldsymbol{\beta}}_{Logit}$ with $1.81\hat{\boldsymbol{\beta}}_{Probit}$ or, equivalently, $\hat{\boldsymbol{\beta}}_{Probit}$ with $0.55\hat{\boldsymbol{\beta}}_{Logit}$

Clog-log model, log-log model

Clog-log model: distribution of tolerance is extreme min distribution (Gompertz):

$$F(\eta_i) = 1 - \exp(-\exp(\eta_i)), \quad \eta_i = \log(-\log(1-\pi_i))$$

Log-log model: distribution of tolerance is extreme max distribution (Gumbel)

$$F(\eta_i) = \exp(-\exp(\eta_i)), \quad \eta_i = \log(-\log(\pi_i))$$

NB! Gompertz and Gumbel distributions are not symmetric

- If U_i is Gompertz-distributed then $1 - U_i$ is Gumbel-distributed
- modelling successes with Gompertz (for tolerance) is equivalent to modelling failures with Gumbel
- Standard Gumbel distribution: $\mu = \gamma$, $\sigma^2 = \frac{\pi^2}{6}$; $\gamma = 0.5772156649$ is Euler constant, transcendental number (number that is not algebraic)

1735 Euler, 1790 Mascheroni calculated 16 digits, 1999 Gourdon, Demichel 108 mln digits

Link functions. Conclusion

- *Logit* link is canonical
- *Logit* link is preferred due to its good interpretability
- *Logit* and *Probit* are symmetric w.r.t. $\pi_i = 0.5$ and are fairly similar unless some π_i-s are very big or very small
- *Logit*, *Probit* and *Clog-log* are similar in case of small probabilities
- *Cauchy* is not sensitive to big probabilities, fits if the probabilities are > 0.9.

Interpretation of logistic model

Odds

Odds of an event is defined as

$$\Pi_i = \frac{\pi_i}{1 - \pi_i}$$

$\Rightarrow$ *Logit* function is log-odds: $Logit(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}$

Odds ratio

Odds ratio is defined as

$$\frac{\Pi_k}{\Pi_i} = \frac{\frac{\pi_k}{1-\pi_k}}{\frac{\pi_i}{1-\pi_i}}$$

Change in argument value x_{ij} by c units corresponds to $e^{c\hat{\beta}_j}$ times change in odds (if other conditions remain the same), i.e. the odds ratio is $e^{c\hat{\beta}_j}$

- In practice, odds ratio is often used as it provides nice interpretation for the model (especially if $c = 1$)
- Also, if (a_j, b_j) is confidence interval for parameter β_j, (e^{a_j}, e^{b_j}) is CI corresponding to odds ratio

Example. Interpretation of logit model

Seeds example (grouped data)

```
> s1=glm(cbind(r,n-r)~cult+soil,family="binomial",data=seeds)
> library(MASS)
> exp(cbind(coef(s1), confint(s1)))
Waiting for profiling to be done...
                          2.5 %    97.5 %
(Intercept) 0.7092206 0.5536289 0.9063096
cult        0.6971842 0.5303025 0.9152451
soil        2.4750647 1.8925019 3.2466015
```

Interpretation?

Interpretation of Probit model

We predict the probability and assume *Probit* link, $\eta_i = \Phi^{-1}(\pi_i)$

$$\hat{\pi}_i = \Phi(\eta_i) = \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta})$$

Positive coefficients increase the probability of an event and negative coefficients decrease

Interpretation of the intercept: Calculating $\Phi(\hat{\beta}_0)$ gives the probability of the event if all arguments are 0

Interpretation of a coefficient β_j: change in argument value (x_{ij}) influences response through the change in the argument of standard normal cdf

NB! The relation is not linear, result depends on the values of other arguments as well as the starting value x_{ij}

To interpret a model, a **base level** of other arguments is chosen (e.g. mean)

Example. Interpretation of Probit model

Question: how does the admission (binary variable) depend on GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution ($n = 400$)

Probit model: $\pi = \Phi(-2.09 + 0.00140 gre + 0.464 gpa - 0.332 rank)$

Let us interpret the dependence from school rank (1–2–3)

(1) taking other arguments to zero:

$$\Phi(-2.09 - 0.332 \cdot 1) = 0.0077$$

$$\Phi(-2.09 - 0.332 \cdot 2) = 0.0029$$

$$\Phi(-2.09 - 0.332 \cdot 3) = 0.0010$$

(2) taking other arguments equal to their means:

$$\Phi(-2.09 + 0.0014 \cdot 587.7 + 0.464 \cdot 3.4 - 0.332 \cdot 1) = 0.491$$

$$\Phi(-2.09 + 0.0014 \cdot 587.7 + 0.464 \cdot 3.4 - 0.332 \cdot 2) = 0.362$$

$$\Phi(-2.09 + 0.0014 \cdot 587.7 + 0.464 \cdot 3.4 - 0.332 \cdot 3) = 0.246$$

Source: `https://stats.idre.ucla.edu/r/dae/probit-regression/`

Interpretation of Clog-log model

Probability of the event and the linear predictor are connected through

$$\pi_i = 1 - \exp(-\exp(\eta_i)), \quad \eta_i = \log(-\log(1-\pi_i))$$

Obviously the link is not linear w.r.t. probability

The effect of change in argument value can be analyzed similarly to Probit model: fix the remaining arguments to their mean level and compare the change of value of function $1 - \exp(-\exp(\eta_i))$

NB! It is actually possible to derive the effect of the change without assuming anything about the remaining arguments. **How?**

Estimation of parameters

Model $\eta_i = \mathbf{x}_i^T \beta, \quad \eta_i = g(\pi_i), \quad \pi_i = h(\mathbf{x}_i^T \boldsymbol{\beta})$
Estimation of parameters:

- Start from sample log-likelihood
- Take the derivatives, solve the score equations $s(\boldsymbol{\beta}) = 0$

  Two main options: Newton-Raphson or Fisher method of scoring (equivalent to weighted least squares)
  Both methods give estimates on the form (r – iteration step)

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}_{r-1} + \tilde{\mathbf{F}}_{r-1}^{-1}(\hat{\boldsymbol{\beta}}_{r-1}) s_{r-1}(\hat{\boldsymbol{\beta}}_{r-1})$$

  Estimated covariance matrices are slightly different

Confidence intervals for parameters, 1

Two approaches:

1. CI based on profile likelihood
   (iterative algorithm, based on asymptotic chi-square distribution of log-likelihood test)
2. CI based on asymptotic normality (simply using SEs), also called Wald's CI:
   $100(1-\alpha)\%$ CI for $\boldsymbol{\beta}_j$ is
   $$\hat{\boldsymbol{\beta}}_j \pm z_{1-\frac{\alpha}{2}}\hat{\sigma}_j$$
   $z_{1-\frac{\alpha}{2}}$ – standard normal quantile
   $\hat{\boldsymbol{\beta}}_j$ – MLE estimate for $\boldsymbol{\beta}_j$
   $\hat{\sigma}_j$ – SE for $\hat{\boldsymbol{\beta}}_j$

The asymptotics for Wald's method works if sample size is big and probabilities lie within (0.1, 0.9)

Confidence intervals for parameters, 2

In R:

- for Wald's CI: function
  - `confint` in package `stats`
  - `confint.default` if package `MASS` is loaded
- for profile likelihood CI: `confint` (or `confint.glm`) in package `MASS`

Overdispersion and underdispersion

For a properly chosen model

$$\frac{\chi^2_P}{n-p} \approx 1 \qquad \frac{D}{n-p} \approx 1$$

If the ratio is > 1, there is **overdispersion**, if the ratio is < 1, there is **underdispersion**

In other words: the variability estimated from data does not match the theoretical

Causes:

- data has an error or an outlier
- too big or too small probabilities of the observed event ('success')
- bad link function choice
- missing covariate, wrong scale of a covariate or some covariate has different effect on subjects
- correlated observations

In case of ungrouped data, over/underdispersion issues are rare

The essence of over/underdispersion (Tutz, 2012)

Let $Y_{ij} \sim B(1, \pi_i)$ and $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, $Y_i \sim B(n_i, \pi_i)$

A usual assumption is that observations are independent. If this is violated, i.e. if $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ are correlated, we have

$$\mathbf{D}Y_i = \mathbf{D}(\sum_{j=1}^{n_i} Y_{ij}) = \sum_{j=1}^{n_i} \mathbf{D}Y_{ij} + \sum_{r \neq s} cov(Y_{ir}, Y_{is})$$

Taking into account that $\mathbf{D}Y_{ij} = \pi_i(1 - \pi_i)$ and $cov(Y_{ir}, Y_{is}) = \rho\sqrt{\mathbf{D}Y_{ir}\mathbf{D}Y_{is}}$, we get

$$\mathbf{D}Y_i = n_i\pi_i(1 - \pi_i)[(1 + (n_i - 1)\rho] = n_i\pi_i(1 - \pi_i)\varphi_i,$$

where $\varphi_i = 1 + (n_i - 1)\rho$ and ρ is the coefficient of correlation

Thus

- if $n_i = 1$ (ungrouped data), overdispersion is not present
- if $\rho > 0$ (pos. correlation between observations) ⇒ overdispersion
- if $\rho < 0$ (neg. correlation between observations) ⇒ underdispersion

Taking over/underdispersion into account

Taking over- or underdispersion into account means adjusting the (variability) parameter estimates

1. Use quasi-likelihood instead on likelihood: change the covariance matrix of parameter estimates based on estimated scale $\hat{\varphi}$
   (a) group sizes n_i are almost equal: estimate $\hat{\varphi}$ using Pearson χ^2-statistic or deviance

   $$\hat{\varphi} = \frac{\chi^2_P}{n-p}, \qquad \hat{\varphi} = \frac{D}{n-p}$$

   In R: use option `family="quasibinomial"`

   (b) group sizes n_i are different: estimate $\hat{\varphi}$ using Williams (1982) method that proposes iterative algorithm for ρ and then φ_i takes into account the group size $\hat{\varphi}_i = 1 + (n_i - 1)\rho$

   In R: use function `glm.binomial.disp` from package `dispmod`
2. Use link function that stabilizes the variance $\eta_i = \arcsin\sqrt{\pi_i}$ or Cauchy link $\eta_i = \tan[\pi(\pi_i - \frac{1}{2})], \pi = 3.14$
3. Use another distribution...

Using quasi-likelihood

Quasi-likelihood function has similar properties to likelihood function but does not correspond to any probability distribution

Some remarks:

- We assume that the means $\mu_i = h(\mathbf{x}_i^T \beta)$ are specified correctly but the variance differs from theoretical
- Estimates are based on quasi-score function and are found solving the GEE (*Generalized Estimating Equations*)
- The estimate for parameter β does not depend on scale $\varphi \Rightarrow$ parameter estimates are the same as for regular likelihood, but the covariance matrix is multiplied by $\hat{\varphi}$, i.e. standard errors are multiplied by $\sqrt{\hat{\varphi}}$

Example. Overdispersion, 1

Let us look again the seeds data:

```
> s1=glm(cbind(r,n-r)~cult+soil,family="binomial",data=seeds)
> summary(s1)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3436     0.1256  -2.735  0.00624 **
cult         -0.3607     0.1392  -2.592  0.00954 **
soil          0.9063     0.1376   6.585 4.55e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.195  on 19  degrees of freedom
Residual deviance:  68.544  on 17  degrees of freedom
AIC: 154.73

Number of Fisher Scoring iterations: 4
```

Example. Overdispersion, 2

```
> s1q=glm(cbind(r,n-r)~cult+soil,family="quasibinomial",data=seeds)
> summary(s1q)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3436     0.2491  -1.379  0.18562
cult         -0.3607     0.2759  -1.307  0.20850
soil          0.9063     0.2729   3.321  0.00404 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 3.931169)

    Null deviance: 118.195  on 19  degrees of freedom
Residual deviance:  68.544  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Example. Overdispersion, 3

```
> library(dispmod)
> s1w=glm.binomial.disp(s1)
Binomial overdispersed logit model fitting...
...
Estimated dispersion parameter: 0.06958442
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3785     0.2454  -1.542  0.12296
cult         -0.2907     0.2825  -1.029  0.30342
soil          0.8059     0.2822   2.856  0.00429 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.677  on 19  degrees of freedom
Residual deviance: 17.485  on 17  degrees of freedom
AIC: 44.888

Number of Fisher Scoring iterations: 3
```

Model diagnostics. Residuals

- Pearson residuals:

$$r_{Pi} = \frac{y_i^* - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}}$$

- Deviance residuals (for scaled binomial model):

$$r_{Di} = sign(y_i^* - \hat{\pi}_i)\sqrt{2n_i[y_i^* \ln\frac{y_i^*}{\hat{\pi}_i} + (1-y_i^*)\ln\frac{1-y_i^*}{1-\hat{\pi}_i}]}$$

- Deviance residuals for $n_i = 1$ (binary model):

$$r_{Di} = sign(y_i - \hat{\pi}_i)\sqrt{-2\ln(1-|y_i-\hat{\pi}_i|)}$$

Remarks:

- $\sum r_{Pi}^2 = \chi_{Pi}^2$
- for small n_i, r_{Pi}^2 are rather skewed, transformation to Anscombe residuals can be considered as an alternative
- $\sum r_{Di}^2 = D$
- standardization: divide the residuals by $\sqrt{1-h_{ii}}$
- Rule of thumb: standardized residuals > 3 are too large

Model diagnostics. Leverage. Influential observations

- **Leverage**: measures how far the argument values of an observation are from those of the other observations. Leverage of observation i is the corresponding diagonal element of the generalized hat matrix, $h_{ii} = (\mathbf{H})_{ii}$. Elements that are $>$2 times larger than the average are considered large Recall the generalized hat matrix:

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2},$$

  where the (diagonal) weight matrix $\mathbf{W}$ depends on parameter estimates, $\mathbf{W} = \mathbf{W}(\hat{\beta})$
- **Influential observations**: have big influence to the model (parameters). Influence of ith observation is measured by comparing the models with and without the ith observation.
  - Cook's distance – influence of i-th observation to the model (response)
  - difference of betas $\texttt{dfbetas}_{ij}$ – influence of i-th observation to $\hat{\beta}_j$

Models with binary response. Summary

1. Choosing the model
   – Choosing arguments and scale
   – Choosing the link function
2. Parameter estimation
   – Does an estimate exist?
3. Model fit
   – Is there overdispersion?
4. Model diagnostics
5. Choosing the best model
6. Interpreting the model

Existence of estimates

Lemma (Claudia Czado, München, 2004)

The log-likelihood $\ln L(\beta)$ in logistic regression is strict concave in β if $rank(\mathbf{X}) = p$.

This implies that the score equations can have at most one solution

$\Rightarrow$ if a ML estimate of β exists, it is unique and it is a solution to score equations

Existence of estimates, R

```
> sep1=glm(y~x1+x2,data=separ,family="binomial")
Warning message:
glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

Existence of estimates, R

```
> sep1=glm(y~x1+x2,data=separ,family="binomial")
Warning message:
glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

What's going on?

Existence of estimates, R

```
> sep1=glm(y~x1+x2,data=separ,family="binomial")
Warning message:
glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

What's going on?

What is the cause?

Existence of estimates, R

```
> sep1=glm(y~x1+x2,data=separ,family="binomial")
Warning message:
glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

What's going on?

What is the cause?

How to proceed?

Infinite parameter estimates (1)

Parameter estimates $\hat{\beta}$ are found using ML method

Estimates for parameters exist $\Leftrightarrow$ iteration converges

The existence of a MLE depends on points in the observation space, i.e. data (Albert & Anderson, 1984)

ML estimates exist if there exists no hyperplane separating the values of the response

Three possible scenarios

- complete separation
- quasi-complete separation
- overlap

Separation – a covariate or a set of covariates determine the response ($y_i = 0$ or $y_i = 1$)

Large standard errors of parameters are an indication of possible separation issues

Infinite parameter estimates (2)

Complete separation

Arguments can divide the response values to exact groups (prediction in each group exactly 1 or 0)

ML estimates do not exist, log-likelihood tends to zero when the number of iterations increases

Quasi-complete separation

For at least one subject, it is not exactly fixed to which response group it belongs

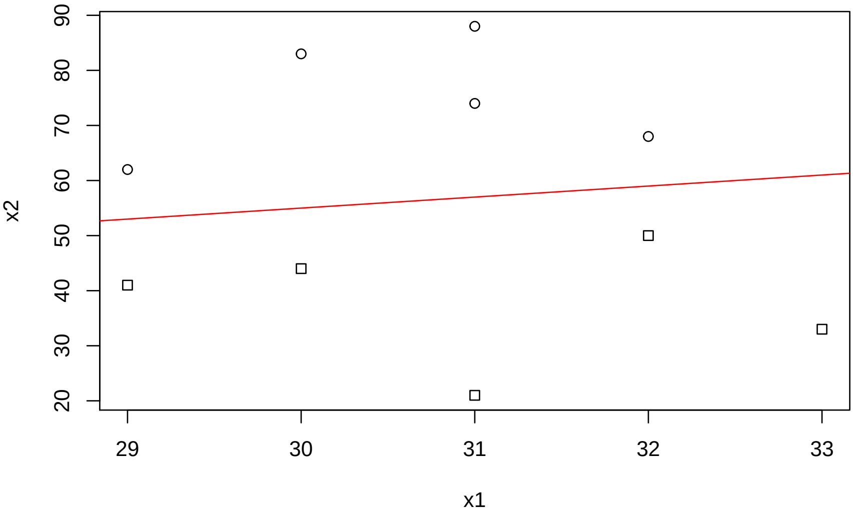
ML estimates do not exist, log-likelihood does not tend to zero, but the information matrix is unbounded and the inverse does not exist

Overlap

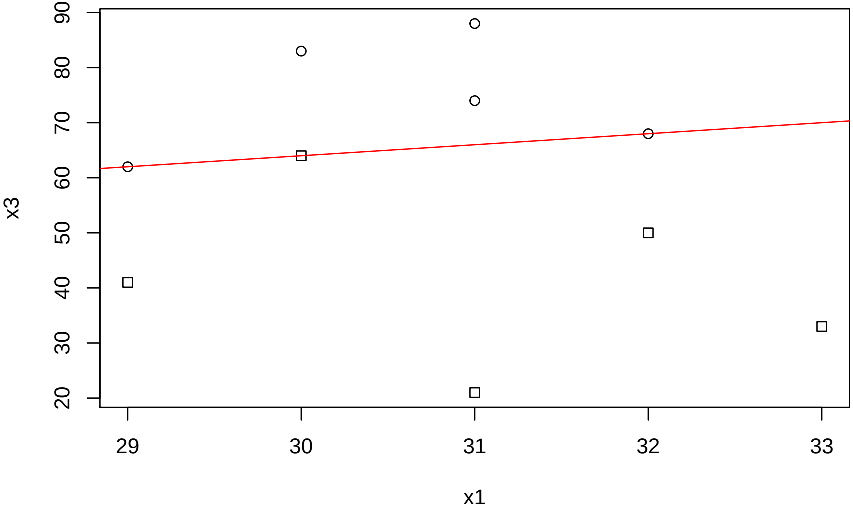
If there is no separation, then there is overlap

ML estimate $\hat{\beta}$ exists and is unique

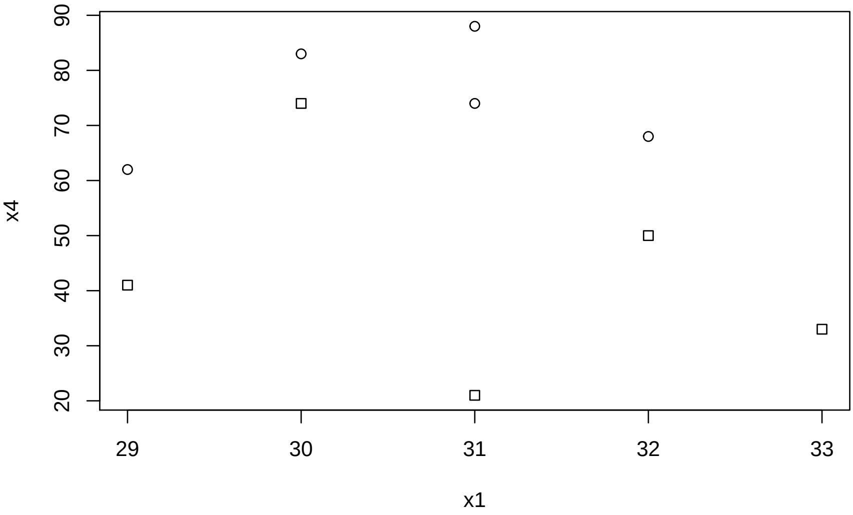
One can describe overlap as a number observations that can be removed to reach partial separation, i.e. situation when the parameters can no longer be estimated.



Complete separation



Quasi-complete separation



Overlap

How to solve the separation issue?

How to proceed?

- Find out which arguments cause the problem, leave some arguments (or observations) out, or re-code
- *Penalized likelihood* (Firth, 1993) – add an adjustment term (which results in skewed estimates), used for continuous arguments
- *Exact logistic regression* – used when number of parameters is small, samples are small and arguments are discrete

Firth's method (Firth, 1993)

Idea: add an adjustment (penalty) term to the log-likelihood and maximize the penalized log-likelihood. Information matrix remains unchanged

Method is asymptotically consistent, i.e. the estimate converges to ML estimate

Idea is similar to *ridge regression*, which is used in case of multicollinearity

In R: function `logistf` (package `logistf`)

```
> logistf(y~x1+x2,data=separ)
logistf(formula = y ~ x1 + x2, data = separ)
Model fitted by Penalized ML
Confidence intervals and p-values by Profile Likelihood

                  coef    se(coef)   lower 0.95 upper 0.95       Chisq
(Intercept) -1.7984331 22.19477577 -52.60524094 65.3741588 0.006847239
x1          -0.1642127  0.72706874  -2.78341758  1.2547411 0.053029128
x2           0.1216285  0.06997688   0.02404056  0.3756169 7.380218208
                      p
(Intercept) 0.934051880
x1          0.817873779
x2          0.006594517
```

Classification problem

A GLM (eventually) predicts the probabilities of an event (or nonevent)

If we need to classify the results, how to do that?

Classification problem

A GLM (eventually) predicts the probabilities of an event (or nonevent)

If we need to classify the results, how to do that?

Simplest way is to say that $pi_i \leq 0.5$ means 0 and $\pi_i > 0.5$ means 1, but is it actually the best option? Maybe another cut-off point is better? Which one? How to compare the classification ability of models using different cut-offs?

Classification problem

A GLM (eventually) predicts the probabilities of an event (or nonevent)

If we need to classify the results, how to do that?

Simplest way is to say that $pi_i \leq 0.5$ means 0 and $\pi_i > 0.5$ means 1, but is it actually the best option? Maybe another cut-off point is better? Which one? How to compare the classification ability of models using different cut-offs?

Confusion matrix:

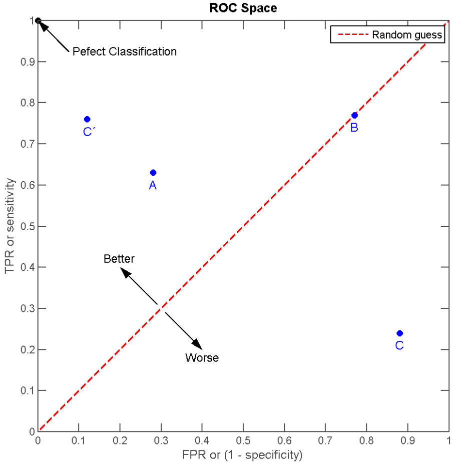
		Actual (true) value	
		1	0
Predicted value	1	True positives (TP)	False positives (FP)
	0	False negatives (FN)	True negatives (TN)

Now

- $TP + FN = P$ – total actual positives
- $TN + FP = N$ – total actual negatives
- $\frac{TP}{P}$ – true positive rate (also sensitivity)
- $\frac{TN}{N}$ – true negative rate (also specificity)
- $\frac{TP+TN}{P+N}$ – accuracy of the model

Receiver operating characteristic

ROC curve allows us to compare models based on their classification ability
The bigger area under ROC curve (AUC), the better



ROC Space. Source: Wikipedia

Decisions

Which cut-off point (probability) to choose?

Possible options:

- the point closest to perfect classification
- the point farthest from the random guess line
- the point that maximizes accuracy
- the point that minimizes the cost (individual costs are specified by a cost matrix)
- the point where sensitivity = specificity

In R: library `ROCR` or `pROC`

A nice example by Arthur Charpentier:
`https://freakonometrics.hypotheses.org/48285`