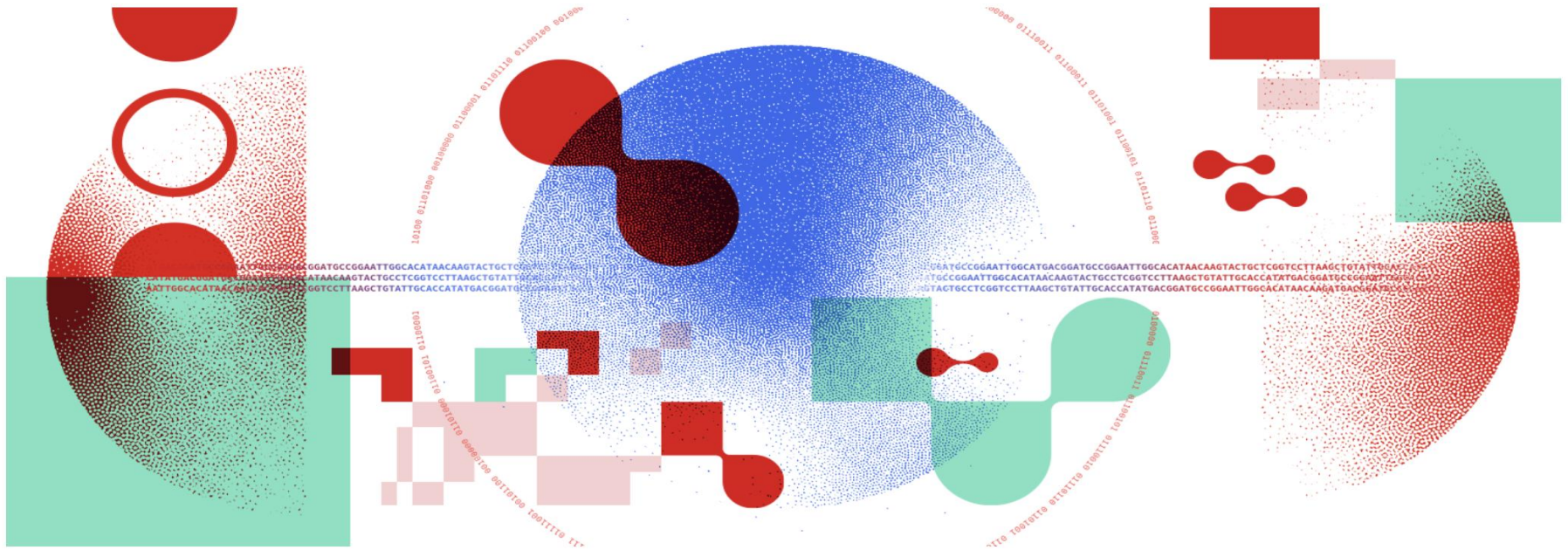




Advanced Statistical Modelling

Lausanne, September 2026

Paolo Angelino and Rachel Marcone

Logistic Regression and GLM



Statistical Models

The models we have seen until now looked like

$$Y = f(X) + \text{Error}$$

We have seen

Splines, Polynomial regression, Step functions, Local regression, Simple and Multiple regression

Statistical Models

Are used for explanation and prediction

Statistical models predicts the mean Y for any combination of predictors.

General form: $g(Y) = f(X) (+ \text{Error})$

with a stochastic process ($\Leftrightarrow$ errors)

Y : dependent variable
(response variable,
observed outcome)

X : independent/
explanatory variable(s)
(grouping variable,
predictor)

Types of response and predictors variables

- binary (2 groups)

(e.g. yes/no, passed/failed, male/female, ill/healthy, responder/NR, ...)

- categorical (k groups)

(e.g. phenotype, genotype, degree of physical activity, ...)

- continuous (i.e. pot. infinite number of groups)

(e.g. weight, blood pressure, gene expression value, ...)

- Censored data (e.g. survival data for patients, living beings, technical devices,...)

Types of variables

Response variable's type determines the suitable regression method(s) :

continuous response -> Linear regression

binary response -> **Logistic regression**

count response -> Poisson regression

Logistic regression

What is Logistic Regression?

Form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.

Discrete ~ continuous / discrete

Example: Responder status ~ dosis

Binary Logistic Regression Model

Y = Binary **response**, ex. **Responder status (1 / 0)**

X = Quantitative **predictor**, ex. Dose, genotype for gene G, ...

π = Proportion / **Probability** of »event 1« at any X

Given π we assume that a (always identical) stochastic process “determines” the event(s) observed outcome.

For a group with the same configurations of X , the same π :
there is a **binomial** distribution $B(n,p)$ of »event 1«
 n = number of observations at this X ,
 p = prob. of event 1 and

Proportion of “success”

In linear regression the model predicts the mean Y for any combination of prediction (the $E[Y | X]$) resp. $E[P(Y=1) | X]$.

What’s the mean of a 0/1 indicator variable?

*The **Proportion** of “cases 1” among n observations.*

$$\pi = \bar{y} = \frac{\sum y_i}{n}$$

*Goal of logistic regression: Predict the “**true**” probability of success, π , at any value of the predictor(s).*

Binary Logistic Regression Model

In a simple regression we have

$$Y = f(X) + \text{error}$$

Now with 0 and 1s we need an intermediate quantity that makes the computation possible

This probability is exactly that: an **intermediate quantity** that makes the statistical modelling possible.

And for each patient gives us an interpretable result, if you try to predict a possibility to get cancer (1) or be healthy (0) according to the age at the end you get a $[0, 1]$ probability given the age to be a 1.

But for computing is it really enough? Almost we are still stucked with values between $[0, 1]$ so we do not want the linear modelling to predict a probability of -0.5

Relation probability – odds

$$\mathbf{odds} = \frac{\pi}{1 - \pi} \Leftrightarrow \pi = \frac{\mathbf{odds}}{1 + \mathbf{odds}}$$

π in $[0, 1]$, odds in $(0, +\infty)$,

$\pi = 0.5$ odds = 1

$\pi = 0.9$ odds = 9

$\pi = 0.1$ odds = $1/9 = 0.111$

not symmetric

Logistic curve

Logit is the **logarithm of the odds**
($\log = \ln = \log_e$)

$$\log\left(\frac{\pi}{1-\pi}\right)$$

Probability of success

Probability of failure

$$\pi = 0.50 \Leftrightarrow \text{logit} = 0$$

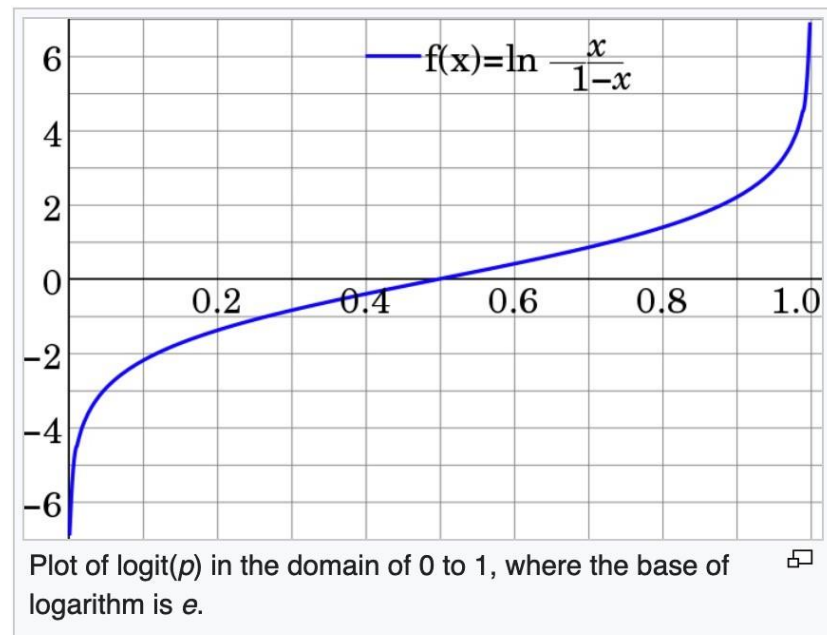
$$\pi = 0.70 \Leftrightarrow \text{logit} = 0.84$$

$$\pi = 0.30 \Leftrightarrow \text{logit} = -0.84$$

$$\pi \rightarrow 1 \Leftrightarrow \text{logit} \rightarrow \text{inf}$$

$$\pi \rightarrow 0 \Leftrightarrow \text{logit} \rightarrow -\text{inf}$$

Symmetric, range $(-\text{inf}, +\text{inf})$



Binary Logistic Regression Model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

The predictors acts at the level of the log odds

$X \rightarrow \text{logit} \rightarrow \underset{\text{prob. } [0,1]}{E(Y)} \rightarrow \text{individual values } Y_i$

The logit is called a link function, links the level of the observed events (response level) to the level at which the predictors effects are acting (link level)

GLM LOGISTIC

Predictors $X \Rightarrow E(\text{logit}(\pi) \mid X) \Rightarrow$ observations Y_i

**$\Rightarrow$ The assumed
model of effects**

**$\Rightarrow$ The assumed underlying
stochastic process
(generating the data)**

**$\Rightarrow$ ex. Logit
is linear in β 's**

$\Rightarrow$ ex. Binomial distribution

Binary Logistic Regression Model

$$\textit{odds} = \frac{\pi}{1 - \pi} \Leftrightarrow \pi = \frac{\textit{odds}}{1 + \textit{odds}} = \frac{1}{1 + 1/\textit{odds}}$$

Logit form

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

Link - Level

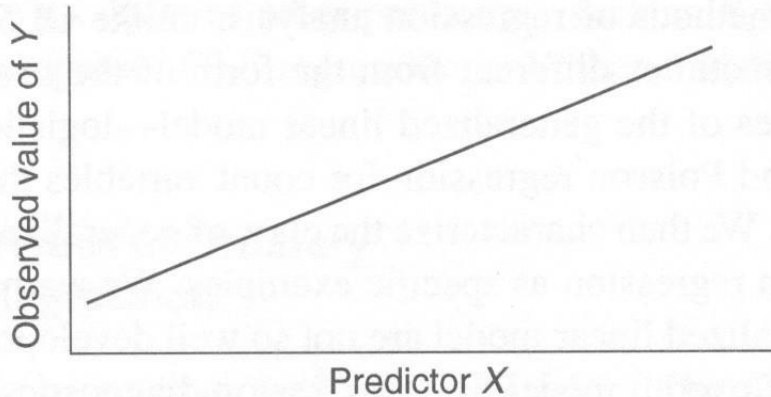
Probability form

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

Response - Level

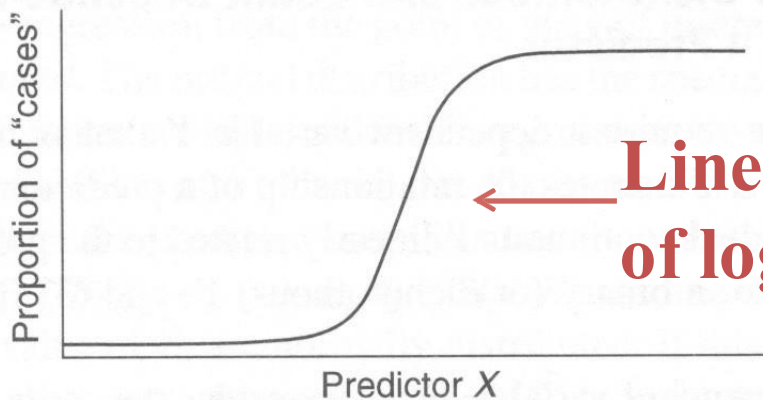
The logistic function

(A) For a continuous outcome variable Y , the numerical value of Y at each value of X .



Change in probability is not constant (linear) with constant changes in X

(B) For a binary outcome variable, the proportion of individuals who are “cases” (exhibit a particular outcome property) at each value of X .



**Linear part
of logistic fit**

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Odds for X:

$$odds = e^{\beta_0 + \beta_1 X}$$

Odds for X+1:

$$odds = e^{\beta_0 + \beta_1 (X+1)}$$

Odds ratio (odds for X+1 / odds for X):

$$\frac{e^{\beta_0 + \beta_1 (X+1)}}{e^{\beta_0 + \beta_1 X}} = e^{\beta_0 + \beta_1 (X+1) - (\beta_0 + \beta_1 X)} = e^{\beta_1}$$

We increase X_1 by one unit (**+1, additive**)

The log odds is increased by β_1 (additive)

The odds is increased by a factor $\exp(\beta_1)$ (**multiplicative**)

The probability is increased by ?(**question!**)

Assumptions

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

The logistic model assumes a linear relationship between the *predictors* and the *log(odds)*.

$$odds = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X}$$

Logistic regression

is a special case of a

Generalized Linear Model
GLM

Ordinary Least Squares regression provides linear models of continuous variables. However, much data of interest to statisticians and researchers are not continuous and so other methods must be used to create useful predictive models.

GLM LOGISTIC

Predictors $X \Rightarrow E(\text{logit}(\pi) \mid X) \Rightarrow$ observations Y_i

**$\Rightarrow$ The assumed
model of effects**

**$\Rightarrow$ The assumed underlying
stochastic process
(generating the data)**

**$\Rightarrow$ ex. Logit
is linear in β 's**

$\Rightarrow$ ex. Binomial distribution

GLM Poisson

Predictors $X \Rightarrow E(\text{link}(Y) | X) \Rightarrow$ observations Y_i

**$\Rightarrow$ The assumed
model of effects**

$\Rightarrow$ stochastic process

Log (λ_i) linear in β 's

Data $Y_i \sim$ Poisson distribution

Poi (mean $\lambda = E[Y]$)

Stdev = sqrt (λ)

The dispersion is the one expected for a »pure random sampling« that is without any factor of variability increasing the dispersion.

Stddev = sqrt(mean)

ML-estimation, deviance, LRT,
Wald test on coefficients etc:
Like Logistic Regression

Poisson regression

- Basic standard model used for Count data
- Distribution: Poisson, (Restriction: mean = variance : $E(Y)=V(Y)=\lambda$)
- Default Link Function: log link:

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

$$\Rightarrow \lambda(X_1, \dots, X_k) = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}$$

Tests are conducted as in Logistic regression

Poisson regression - assumptions

- Poisson at the Response Level : the response variable is a count per unit of time or space, described by a Poisson distribution.
- Linearity at Link Level : the log of the mean rate, $\log(\lambda)$, is be a linear function of the predictor x .
- Independence: the observations are independent of one another.
- Mean=Variance: the mean of a Poisson random variable is equal to its variance.

Generalized Linear Models

Generalized linear models are fit using the `glm()` function. The form of the `glm` function is

`glm(formula, family=familytype(link=linkfunction), data=)`

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

How to find the »best fit«

Standard method: **maximum likelihood estimation MLE**

Probability of observations

(Likelihood of the model given the data) = maximum

The **MLE** is the preferred method for statisticians in many situations. It has a series of good properties (best method given some criteria).

The t-test for example is the maximum likelihood-based test to compare the mean of two normal distributions.

The MLE estimate for lm models leads to the same solution like the least squares (under given assumptions).

How to find the »best fit«

Solution:

Generally there is no closed solution (formula) for the parameters in function of the data

The point estimates are determined by a multi-step iterative algorithms

GLM diagnostic 1: Hat and Cook

- **Detection of influential observations and poor fitting**

- 1) **Hat values h_i** In analogy to LM there is a definition of a hat matrix for logistic regression fits and large diagonal values suggest a potential high influence (**leverage**) of a point on the obtained fit.
Limit $\sim 2p / n$ or $3p / n$.
- 2) **Cook's distance Cd_i** is a measure of a change in estimated coefficients when the observation i is ignored. Large values ($> \sim 4/n$) suggest a large influence, pointing to observations one might want to "investigate".
- 3) (The square of the) individual **deviance residuals** or **studentized residuals** can also indicate single observation points with potential high influence or outlying character.

Plots: `influencePlot(model);`
`residualPlot(model2, type = "response")`
`residualPlot(model2, type = "pearson")`
`residualPlot(model2, type = "deviance")`

GLM diagnostic 2: analysis of residuals

- Deviance residuals vs fitted values
- **Missing patterns: Deviance residuals vs each of the available covariates**
- **Dispersion check: Quantile Residuals**

Many more checking procedures are known, but **interpretation** and recommended actions rarely straightforward

GLM diagnostic 3: analysis of residuals

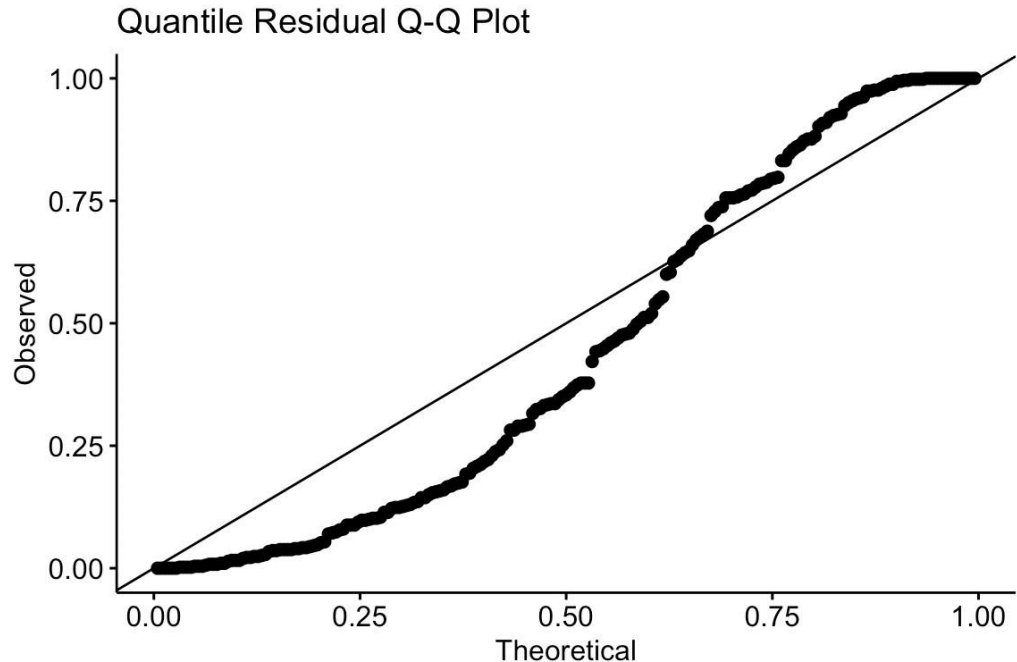
Quantile Residuals QR_i

Example: (Poisson GLM QRs)

Several excessively extreme (larger and smaller than expected) QRs in this checking

suggests overdispersion of data

might suggest the use of a **quasi-Poisson** or a **Negativ Binomial** approach instead of the Poisson, as these have higher variances



GLM diagnostic 4: analysis of residuals

Quantile Residuals QR_i

One can generate faked simulated data from the fitted model, like distribution of predicted new values and compare to the observed values (for each data point, each $\mathbf{x}_i$).

The nb of fakes $<$ obs, y_i is called quantile residual QR_i .

If the data are distributed as specified by the model these follow a uniform $U[0,1]$ distribution.

A Q-Q plot of calculated vs. expected quantile residuals can detect significant departures and suggest modifications to the model.

GLM diagnostic 5: analysis of residuals

“Raw Residuals” $RR_i = Y_i - \text{fitted } E[Y | X]$, where $Y_i = 0$ or 1

Pearson Residuals PR_i :

are adjusted for expected variance (given X) and are expected to follow approximately a normal distribution at each X_i (under assumptions).

Can reveal potential outliers.

Large residuals (in absolute value) are “somewhat strange” compared to their “neighbour points”, but not necessarily to be considered outliers (in general some large residuals have to be expected).

A (linear) trend in a plot of PR_i against covariates might identify predictors **that have been** omitted in the model but should maybe be included.

Trend: add a loess to the graph

A curved trend might indicate that adding a higher order term of the covariate could be useful (ex. x^2).

GLM diagnostic 6: autocorrelation

A quick test if there is any obvious evidence of **non-independence** of the observations:

Check for «autocorrelation», function `acf()`

Challenge 1

Using the babies dataset

- Fit a logistic regression to find parameters explaining the probability of prematurity ?
- What is the effect of birth weight on the probability of prematurity ?
- What about parity ?

Solution

```
> model2 <- glm(prem ~ bwt, family=binomial)
> summary(model2)
```

Call:

```
glm(formula = prem ~ bwt, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2879	-0.3985	-0.2784	-0.1810	3.0710

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.017338	0.717952	6.988	2.78e-12 ***
bwt	-0.067061	0.006808	-9.851	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 664.66 on 1173 degrees of freedom
Residual deviance: 545.31 on 1172 degrees of freedom
AIC: 549.31

Number of Fisher Scoring iterations: 6

How to test »significance« and determine CI ?

Given standard errors SE of β :

test-statistics = estimate / SE = z

approx. Normal (under the null hypothesis)

called a **Wald-test**

CI width = approx. 1.96 * SE

CI symmetric for β and the log odds scale

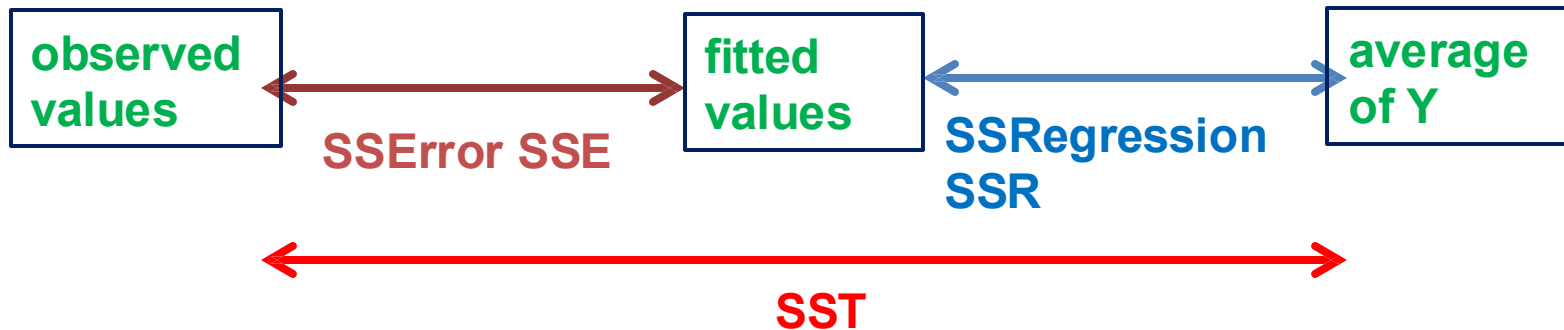
⇒ not symmetric for the multiplicative effect $\exp(\beta)$ on the odds scale

⇒ not symmetric for the effect on the probability π

$$SST = SSR + SSE$$

Total sum of squares = regression SS + residual SS

HEURISTIC REPRESENTATION



$$R^2 = SSR / SST = 1 - (SSE / SST)$$

MLE likelihood and deviance

Highest (log) likelihood possible, predictors best adapted to each Y_i

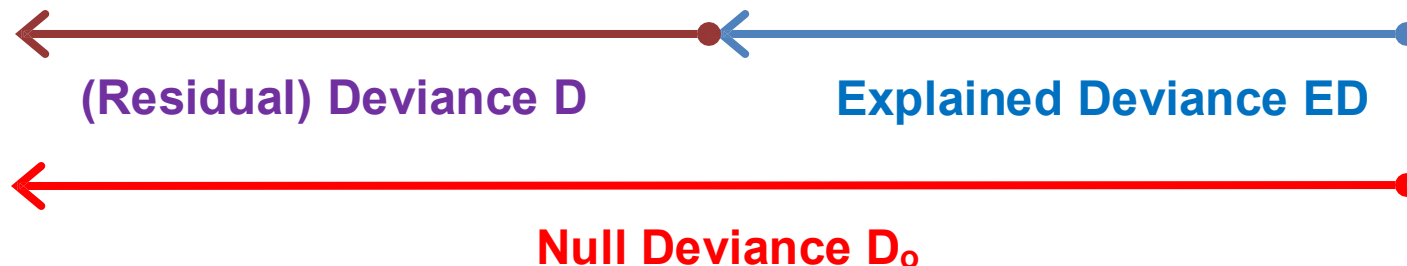
*Deviance : difference in $2 * \text{Log Lik}$*

(log) likelihood if data from a fixed distribution with no individual observation-predictors

observed values,
Saturated model

fitted , current
proposed, Model

“Null model” ,
only 1 parameter



$$D_o = ED + D; \quad ED = D_o - D$$

$$R^2 \Rightarrow ED / D_o = (D_o - D) / D_o = 1 - (D / D_o) = \text{called a pseudo- } R^2$$

Notes: $\log \text{Lik} \leq 0$; good Log Lik is close to 0;

Deviance > 0 : a measure of “lack of fitting”,

good is small positive close to 0

Parameter Optimization: Maximal (Log) Likelihood $\sim$ Minimal Deviance

Deviance

- ❑ **Null deviance:** how well (or bad) the response variable is predicted by a model that includes only the intercept (overall mean, logistic: binomial with fixed p) compared to the best possible model
- ❑ **Residual deviance:** how much deviance is missing compared to the best model after including the proposed set of independent variables (residual lack of fit)

Solution

```
> model2 <- glm(prem ~ bwt, family=binomial)
> summary(model2)
```

Call:

```
glm(formula = prem ~ bwt, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2879	-0.3985	-0.2784	-0.1810	3.0710

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.017338	0.717952	6.988	2.78e-12 ***
bwt	-0.067061	0.006808	-9.851	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 664.66 on 1173 degrees of freedom
Residual deviance: 545.31 on 1172 degrees of freedom
AIC: 549.31

Number of Fisher Scoring iterations: 6

Multiple Logistic Regression

Extension to more than one predictor variable (either numeric or dummy variables).

With k predictors, the model is written:

$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

Solution

```
> model4 <- glm(prem ~ bwt*smoke+parity, family=binomial)
> summary(model4)
```

Call:

```
glm(formula = prem ~ bwt * smoke + parity, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4798	-0.3998	-0.2784	-0.1682	2.9571

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	4.839354	0.978834	4.944	7.65e-07	***
bwt	-0.062082	0.008741	-7.103	1.22e-12	***
smokesmoker	2.247047	1.609071	1.396	0.1626	
paritynot first	-0.470085	0.283836	-1.656	0.0977	.
bwt:smokesmoker	-0.028043	0.015781	-1.777	0.0756	.

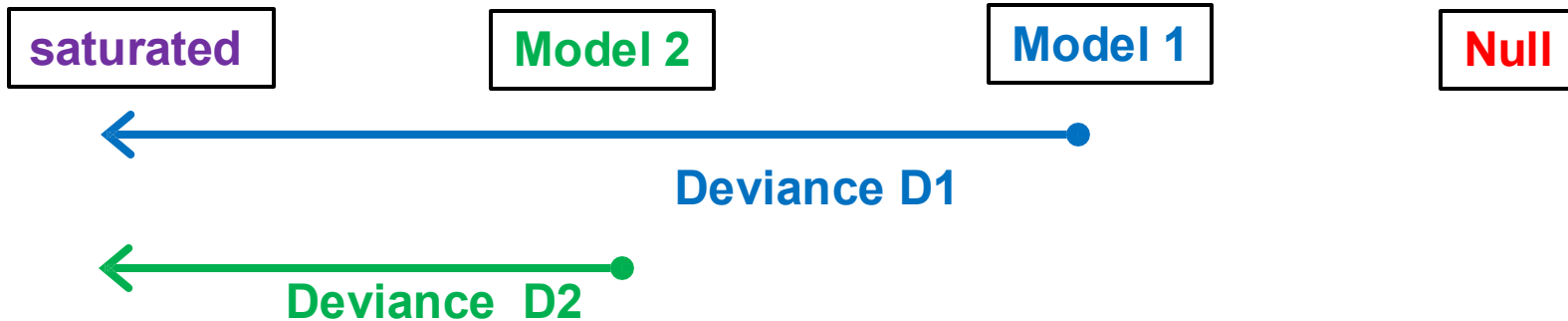
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 664.66 on 1173 degrees of freedom
Residual deviance: 532.93 on 1169 degrees of freedom
AIC: 542.93

Number of Fisher Scoring iterations: 6

model comparison tests



Likelihood Ratio Test LRT

Nested models

- the smaller model 1 is a special case of the larger model 2;
- larger model 2 has all predictors of model 1 and some additional predictors

Comparison of **nested** models:

- we can test if the improvement of model 2 over model 1 is statistically significant with a **likelihood ratio test (LRT) = deviance test = Wilks test**

model comparison tests

Likelihood Ratio Test LRT

Test statistic = $2 \times \text{Log Lik Ratio} = \text{Deviance D1} - \text{Deviance D2}$

~ **chi2 distribution** with degrees of freedom = df for smaller model (higher df)
- df for larger model

Example R code :

```
anova(model1, model2, test = "Chisq")
```

The same function encodes the analogous model comparison test for LM models

Example for the class data :

```
anova(model.0, model.3)
```

Model 1: Height ~ Age

Model 2: Height ~ Age + Weight

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	17	1042.71				
2	16	524.94	1	517.77	15.781	0.0011 **

Model quality indices

R squared

Analogous of R^2 in LM

$$R^2 = \text{ED} / \text{D}_o = (\text{D}_o - \text{D}) / \text{D}_o = 1 - (\text{D} / \text{D}_o)$$

called a pseudo- R^2

Many different R-Squared and adjusted R-Squared have been proposed for GLM
Some are fairly widely used but generally model selection is best done with LRT

Akaike Information Criterion (AIC)

- allows to assess the quality of a model through comparison of related models
- based on the Deviance, but penalizes for the number of parameters (like adjusted R-squared, it's intent is to correct for irrelevant predictors)

Solution

```
> model4 <- glm(prem ~ bwt*smoke+parity, family=binomial)
> summary(model4)
```

Call:

```
glm(formula = prem ~ bwt * smoke + parity, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4798	-0.3998	-0.2784	-0.1682	2.9571

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	4.839354	0.978834	4.944	7.65e-07	***
bwt	-0.062082	0.008741	-7.103	1.22e-12	***
smokesmoker	2.247047	1.609071	1.396	0.1626	
paritynot first	-0.470085	0.283836	-1.656	0.0977	.
bwt:smokesmoker	-0.028043	0.015781	-1.777	0.0756	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 664.66 on 1173 degrees of freedom
Residual deviance: 532.93 on 1169 degrees of freedom
AIC: 542.93

Number of Fisher Scoring iterations: 6

Model selection

Nested Models: LRT !!!!

Otherwise: complicated
Nothing simple works reliably.

Multiple data methods (separate datasets learning-testing,
cross-validation, bootstraps)

Some methods incorporate cross-validation for optimization:
(ex. **penalized regression**, f.ex. package glmnet)

Other Questions 1

Does another **link function** give a better fit ?

(example: binomial family regression: logit or complementary log-log which can better fit cases asymmetric about 0.5 , ...)

For the Binomial Model

- There is the Probit link function, but very close to the Logit

- There is an asymmetric link function:

Complementary Log-Log transformation (cloglog)

$\log \{-\log [1 - \pi(x)]\}$ linear in X , $=X\beta$

$$\pi(x) = 1 - \exp(-\exp(X\beta))$$

Results are frequently close to the Logit results

Other Questions 2

Is the model appropriate ?

Does another model type («**family**») give a better fit ?
(example: binomial vs. Poisson vs. Quasi Poisson)