• ESTIMATION

data $(y_1, \dots, y_n)$ from $Y_i \sim Ber(\pi_i)$ with $logit(\pi_i) = \tilde{x}_i^T\beta$ $\quad i = 1, \dots, n$

distribution

$$p(y_1, \dots, y_n) = \prod_{i=1}^n p(y_i) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$
$$= \prod_{i=1}^n \left(\frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\tilde{x}_i^T\beta}}\right)^{1-y_i}$$

likelihood

$$L(\underline{\beta}) \propto \prod_{i=1}^n p(y_i) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad \text{where} \quad \pi_i = \frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}}$$

log likelihood

$$\ell(\underline{\beta}) = \sum_{i=1}^n \{y_i \log\pi_i + (1-y_i)\log(1-\pi_i)\}$$

$$\log\pi_i = \log\frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}} = \tilde{x}_i^T\beta - \log(1+e^{\tilde{x}_i^T\beta})$$

$$\log(1-\pi_i) = \log\frac{1}{1+e^{\tilde{x}_i^T\beta}} = -\log(1+e^{\tilde{x}_i^T\beta})$$

$$\ell(\underline{\beta}) = \sum_{i=1}^n \left\{y_i\tilde{x}_i^T\beta - \cancel{y_i\log(1+e^{\tilde{x}_i^T\beta})} - \log(1+e^{\tilde{x}_i^T\beta}) + \cancel{y_i\log(1+e^{\tilde{x}_i^T\beta})}\right\}$$
$$= \sum_{i=1}^n \left\{y_i\tilde{x}_i^T\beta - \log(1+e^{\tilde{x}_i^T\beta})\right\}$$

score function

$$\ell_*(\beta) = \left\{\frac{\partial}{\partial\beta_r}\ell(\beta)\right\}_{r=1,\dots,p} \quad \text{where} \quad \frac{\partial\ell(\beta)}{\partial\beta_r} = \sum_{i=1}^n\left\{y_i x_{ir} - \frac{1}{1+e^{\tilde{x}_i^T\beta}}\cdot e^{\tilde{x}_i^T\beta}\cdot x_{ir}\right\}$$

$$\ell_*(\underline{\beta}) = \sum_{i=1}^n \tilde{\underline{x}}_i\left(y_i - \frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}}\right) = \sum_{i=1}^n \tilde{\underline{x}}_i(y_i - \pi_i) = X^T(\underline{y} - \underline{\pi})$$

likelihood equation: $\ell_*(\underline{\beta}) = 0 \Rightarrow X^T(\underline{y} - \underline{\pi}) = 0$ again they resemble the normal equations but they are not linear in β

Similarly to the Poisson case, we need to solve the equation numerically and we do not have a closed-form expression of the MLE $\hat{\underline{\beta}}$.

Finally, the 2nd derivative is

$$\frac{\partial^2\ell(\beta)}{\partial\beta_r\partial\beta_s} = \frac{\partial}{\partial\beta_s}\left(\sum_{i=1}^n\left\{\cancel{y_i x_{ir}} - \frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}}x_{ir}\right\}\right)$$
$$= -\sum_{i=1}^n \frac{e^{\tilde{x}_i^T\beta}\cdot x_{is}\cdot x_{ir}(1+e^{\tilde{x}_i^T\beta}) - e^{\tilde{x}_i^T\beta}x_{ir}\cdot e^{\tilde{x}_i^T\beta}x_{is}}{(1+e^{\tilde{x}_i^T\beta})^2}$$
$$= -\sum_{i=1}^n \frac{e^{\tilde{x}_i^T\beta}x_{ir}x_{is} + \cancel{e^{2\tilde{x}_i^T\beta}x_{ir}x_{is}} - \cancel{e^{2\tilde{x}_i^T\beta}x_{ir}x_{is}}}{(1+e^{\tilde{x}_i^T\beta})^2}$$
$$= -\sum_{i=1}^n \frac{e^{\tilde{x}_i^T\beta}x_{is}x_{ir}}{(1+e^{\tilde{x}_i^T\beta})^2} = -\sum_{i=1}^n x_{ir}x_{is}\pi_i(1-\pi_i)$$

$$\Rightarrow \frac{\partial\ell(\beta)}{\partial\beta\partial\beta^T} = -X^TUX \quad \text{with} \quad U = diag\{\pi_1(1-\pi_1), \dots, \pi_n(1-\pi_n)\} = U(\beta)$$

observed information

$$j(\underline{\beta}) = -\ell_{**}(\beta) = X^TUX \quad \text{and} \quad j(\hat{\underline{\beta}}) = X^TU(\hat{\underline{\beta}})X$$

• INFERENCE

Inference is analogous to the Poisson glm.

• DISTRIBUTION of the MAXIMUM LIKELIHOOD ESTIMATOR of the REGRESSION PARAMETERS

$$\hat{\underline{\beta}} \mathrel{\dot\sim} N_p(\underline{\beta}, j(\hat{\underline{\beta}})^{-1})$$

the marginal distribution for the j-th element is $\hat{\beta}_j \mathrel{\dot\sim} N(\beta_j, [j(\hat{\beta})^{-1}]_{jj})$ $\quad j = 1, \dots, p$

• CONFIDENCE INTERVAL FOR β_j

A pivotal quantity is

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} \mathrel{\dot\sim} N(0,1)$$

a confidence interval with level $(1-\alpha)$ for β_j $(j = 1, \dots, p)$ can be obtained as

$$\mathbb{P}\left(z_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

with the data:

$$\hat{\beta}_j - \sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}\cdot z_{1-\frac{\alpha}{2}} < \beta_j < \hat{\beta}_j + \sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}\cdot z_{1-\frac{\alpha}{2}}$$

$$\Rightarrow \beta_j \in \hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}$$

• TEST ABOUT β_j

consider the test $\begin{cases} H_0: \beta_j = b_j \\ H_1: \beta_j \neq b_j \end{cases}$

We can use the test statistic

$$z_j = \frac{\hat{\beta}_j - b_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} \mathrel{\dot\sim} N(0,1) \quad \text{under } H_0$$

the observed value of the test is z_j^{obs}

- if we use a fixed significance level α
  $\alpha = \mathbb{P}_{H_0}(|z_j| > z_{1-\frac{\alpha}{2}})$
  $\rightarrow$ reject region is $R = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, +\infty)$

- if we use the observed significance level
  the p-value is $\alpha^{obs} = \mathbb{P}_{H_0}(|z_j| \geq |z_j^{obs}|) = 2(1 - \Phi(|z_j^{obs}|))$

• TEST for COMPARING NESTED MODELS
(test about a subset of the parameters)

The proposed "full" model is
$Y_i \sim Bern(\pi_i)$ indep for $i = 1, \dots, n$
with $\pi_i = \frac{e^{\tilde{x}_i^T\beta}}{1+e^{\tilde{x}_i^T\beta}}$

and specifically $\tilde{x}_i^T\beta = \beta_1 + \beta_2 x_{i2} + \dots + \beta_{p_0}x_{ip_0} + \beta_{p_0+1}x_{i\,p_0+1} + \dots + \beta_p x_{ip}$

we want to test
$\begin{cases} H_0: \beta_{p_0+1} = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$

as usual, we partition $\underline{\beta} = \begin{bmatrix} \underline{\beta}^{(0)} \\ \underline{\beta}^{(1)} \end{bmatrix}$ $\quad \underline{\beta}^{(0)} \in \mathbb{R}^{p_0}$, $\underline{\beta}^{(1)} \in \mathbb{R}^{p-p_0}$

so the test can be reformulated as
$\begin{cases} H_0: \underline{\beta}^{(1)} = \underline{0} \\ H_1: \underline{\beta}^{(1)} \neq \underline{0} \end{cases}$

Similarly to what we have seen with the Poisson regression, to perform this test we use the LIKELIHOOD RATIO TEST:

$$W = 2\log\frac{\hat{L}(\text{model})}{\tilde{L}(\text{restricted})} = 2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted})\} \mathrel{\dot\sim} \chi^2_{p-p_0} \quad \text{under } H_0$$

$p-p_0$: number of parameters we are testing

We estimate the full model (H_1), obtaining $\hat{\underline{\beta}} = (\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)})$
$$\hat{\ell}(\text{model}) = \ell(\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)}) = \sum_{i=1}^n y_i\log\hat{\pi}_i + \sum_{i=1}^n(1-y_i)\log(1-\hat{\pi}_i) \quad \text{with} \quad \hat{\pi}_i = \frac{e^{\tilde{x}_i^T\hat{\beta}}}{1+e^{\tilde{x}_i^T\hat{\beta}}}$$

We estimate the restricted model (H_0), obtaining $\tilde{\underline{\beta}} = (\tilde{\underline{\beta}}^{(0)}, \underline{0})$
$$\tilde{\ell}(\text{restricted}) = \ell(\tilde{\underline{\beta}}^{(0)}, \underline{0}) = \sum_{i=1}^n y_i\log\tilde{\pi}_i + \sum_{i=1}^n(1-y_i)\log(1-\tilde{\pi}_i) \quad \text{with} \quad \tilde{\pi}_i = \frac{e^{\tilde{x}_i^T\tilde{\beta}}}{1+e^{\tilde{x}_i^T\tilde{\beta}}}$$

The observed value of the test is
$$w^{obs} = 2(\hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted})) =$$
$$= 2\left(\sum_{i=1}^n y_i\log\hat{\pi}_i + \sum_{i=1}^n(1-y_i)\log(1-\hat{\pi}_i) - \sum_{i=1}^n y_i\log\tilde{\pi}_i - \sum_{i=1}^n(1-y_i)\log(1-\tilde{\pi}_i)\right) =$$
$$= 2\left(\sum_{i=1}^n y_i\log\frac{\hat{\pi}_i}{\tilde{\pi}_i} + \sum_{i=1}^n(1-y_i)\log\frac{(1-\hat{\pi}_i)}{(1-\tilde{\pi}_i)}\right)$$

If the null hypothesis is true, $\hat{\ell}(\text{model}) \approx \tilde{\ell}(\text{restricted}) \Rightarrow w^{obs} \approx 0$
If the null hypothesis is not true, $\hat{\ell}(\text{model}) > \tilde{\ell}(\text{restricted}) \Rightarrow w^{obs} \gg 0 \rightarrow$ reject for large values

The reject region will comprise large values of the test

- fixed significance level α
  $\alpha = \mathbb{P}_{H_0}(W > \chi^2_{p-p_0;1-\alpha})$
  $R = (\chi^2_{p-p_0,1-\alpha}; +\infty)$
  $\chi^2_{p-p_0,1-\alpha}$: $(1-\alpha)$-quantile of a χ^2 distribution with $p-p_0$ d.o.f.

- pvalue
  $\alpha^{obs} = \mathbb{P}_{H_0}(W \geq w^{obs})$
  with $W \overset{H_0}{\sim} \chi^2_{p-p_0}$

• TEST about the OVERALL SIGNIFICANCE

We compare the proposed model with the null model
$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$

We use the previous test with $p_0 = 1$

$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$ $\quad \beta_1 = \beta^{(0)} \in \mathbb{R}$, $\underline{\beta}^{(1)} \in \mathbb{R}^{p-1}$

We need to compute the maximum of the log-likelihood under the null model
Under H_0 $\quad Y_i \sim Bern(\pi_i)$

$$\pi_i = \frac{e^{\beta_1}}{1+e^{\beta_1}} = \pi \qquad \text{reparameterization:} \quad \pi = \frac{e^{\beta_1}}{1+e^{\beta_1}} \iff \beta_1 = \log\frac{\pi}{1-\pi}$$

$\Rightarrow$ we can compute the estimate $\tilde{\pi}$ and automatically obtain the MLE of β as $\tilde{\beta}_1 = \log\frac{\tilde{\pi}}{1-\tilde{\pi}}$

$$L(\pi) = \prod_{i=1}^n \pi^{y_i}(1-\pi)^{1-y_i}$$
$$\ell(\pi) = \sum_{i=1}^n\{y_i\log\pi + (1-y_i)\log(1-\pi)\}$$
$$\ell_*(\pi) = \sum_{i=1}^n\left\{\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}\right\} = \sum_{i=1}^n\left\{\frac{y_i - \cancel{\pi y_i} - \pi + \cancel{\pi y_i}}{\pi(1-\pi)}\right\}$$

$\ell_*(\pi) = 0 \Rightarrow \sum_{i=1}^n y_i - n\pi = 0 \Rightarrow \tilde{\pi} = \bar{y}$ $\quad$ MLE of $\pi = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$ under the null model is the sample mean: PROPORTION OF SUCCESSES

$$\ell_{**}(\pi) = \sum_{i=1}^n\left\{-\frac{y_i}{\pi^2} - \frac{(1-y_i)}{(1-\pi)^2}\right\} = -\frac{n\bar{y}}{\pi^2} - \frac{n-n\bar{y}}{(1-\pi)^2}$$
$$\ell_{**}(\tilde{\pi}) = -\frac{n\cancel{\bar{y}}}{\bar{y}^{\cancel{2}}} - \frac{n\cancel{(1-\bar{y})}}{(1-\bar{y})^{\cancel{2}}} = -\frac{n}{\bar{y}} - \frac{n}{(1-\bar{y})} < 0$$

Hence $\tilde{\ell}(\text{null}) = \ell(\tilde{\pi}) = \ell(\tilde{\beta}_1)$
$$= \sum_{i=1}^n\{y_i\log\tilde{\pi} + (1-y_i)\log(1-\tilde{\pi})\} = \sum_{i=1}^n\{y_i\log\bar{y} + (1-y_i)\log(1-\bar{y})\}$$
$$= n\bar{y}\log\bar{y} + n(1-\bar{y})\log(1-\bar{y})$$

As usual, we then estimate the full model (H_1), obtaining $\hat{\underline{\beta}} = (\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)})$ and
$$\hat{\ell}(\text{model}) = \ell(\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)}) = \sum_{i=1}^n y_i\log\hat{\pi}_i + \sum_{i=1}^n(1-y_i)\log(1-\hat{\pi}_i) \quad \text{with} \quad \hat{\pi}_i = \frac{e^{\tilde{x}_i^T\hat{\beta}}}{1+e^{\tilde{x}_i^T\hat{\beta}}}$$

The LR test in this case is $W = 2(\hat{\ell}(\text{model}) - \tilde{\ell}(\text{null})) \mathrel{\dot\sim} \chi^2_{p-1}$ under H_0

$$w^{obs} = 2\left\{\sum_{i=1}^n\left[y_i\log\hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i) - y_i\log\bar{y} - (1-y_i)\log(1-\bar{y})\right]\right\}$$
$$= 2\left\{\sum_{i=1}^n\left[y_i\log\frac{\hat{\pi}_i}{\bar{y}} + (1-y_i)\log\frac{(1-\hat{\pi}_i)}{(1-\bar{y})}\right]\right\}$$

And the test is then completed as usual.

• DEVIANCE for Bernoulli regression

We defined the deviance as the LR test to compare the SATURATED MODEL and the proposed "full" model with $p < n$ parameters.
(Saturated model: a model with n parameters, one for each observation.
We obtain a model with a perfect fit, since we are interpolating the n points).

What happens to the Bernoulli log-likelihood when we compute it for the saturated model?
We have $Y_i \sim Bernoulli(\pi_i)$ with a separate π_i $\forall i$

$$L(\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$
$$\ell(\pi_i) = y_i\log\pi_i + (1-y_i)\log(1-\pi_i) = \begin{cases} \log\pi_i & \text{if } y_i = 1 \\ \log(1-\pi_i) & \text{if } y_i = 0 \end{cases}$$
$$\ell_*(\pi_i) = \frac{y_i}{\pi_i} - \frac{(1-y_i)}{1-\pi_i} = \frac{(1-\pi_i)y_i - \pi_i(1-y_i)}{\pi_i(1-\pi_i)}$$

$\ell_*(\pi_i) = 0 \Rightarrow y_i - \cancel{\pi_i y_i} - \pi_i + \cancel{\pi_i y_i} = 0$
$\tilde{\pi}_i^S = y_i \in \{0, 1\}$ $\quad$ under the saturated model, we estimate a probability that is equal to 1 if $y_i = 1$, and equal to 0 if $y_i = 0$.

The log-likelihood evaluated at $\tilde{\pi}_i^S$ is
if $y_i = 1 \Rightarrow \tilde{\pi}_i^S = 1 \Rightarrow \ell(\tilde{\pi}_i^S) = \log 1 = 0$
if $y_i = 0 \Rightarrow \tilde{\pi}_i^S = 0 \Rightarrow \ell(\tilde{\pi}_i^S) = \log(1-0) = 0$

the loglikelihood for the saturated model is always equal to 0.

Hence $D = \text{deviance}(\text{model}) = 2\{\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})\} = -2\hat{\ell}(\text{model})$

$$D = -2\left(\sum_{i=1}^n y_i\log\hat{\pi}_i + \sum_{i=1}^n(1-y_i)\log(1-\hat{\pi}_i)\right)$$
$$= \sum_{i=1}^n \underbrace{-2(y_i\log\hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i))}_{\text{individual contribution } D_i} = \sum_{i=1}^n D_i$$

if $y_i = 1$ $\quad D_i = -2\log\hat{\pi}_i$
if $y_i = 0$ $\quad D_i = -2\log(1-\hat{\pi}_i)$

When we assume a Bernoulli distribution for Y_i, the deviance is not useful to evaluate the goodness of fit of the model.

However, we can still derive the test about the overall significance as:
$$D(\text{null}) - D(\text{model}) = -2\tilde{\ell}(\text{null}) - (-2\hat{\ell}(\text{model}))$$
$$= 2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{null})\} = \text{LR test between null and proposed model}$$

"null deviance" - "residual deviance"

Also the analysis of the residuals in this setting is not useful.