Pearson residuals in logistic regression: comparison between *logit* and *glm*

Oslo, 24 September 2026

Marta Ponzano[1], Rino Bellocco[2,3]

[1] Department of Life Sciences, Health and Health Professions, Link Campus University, Rome, Italy;

[2] Department of Statistics and Quantitative Methods, University of Milano–Bicocca, Milan, Italy;

[3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden

Logistic regression

- Logistic regression models the relationship between a set of covariates and a binary outcome Y :

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

$$P(Y_i = 1 \mid \mathbf{x}_i) = \pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}$$

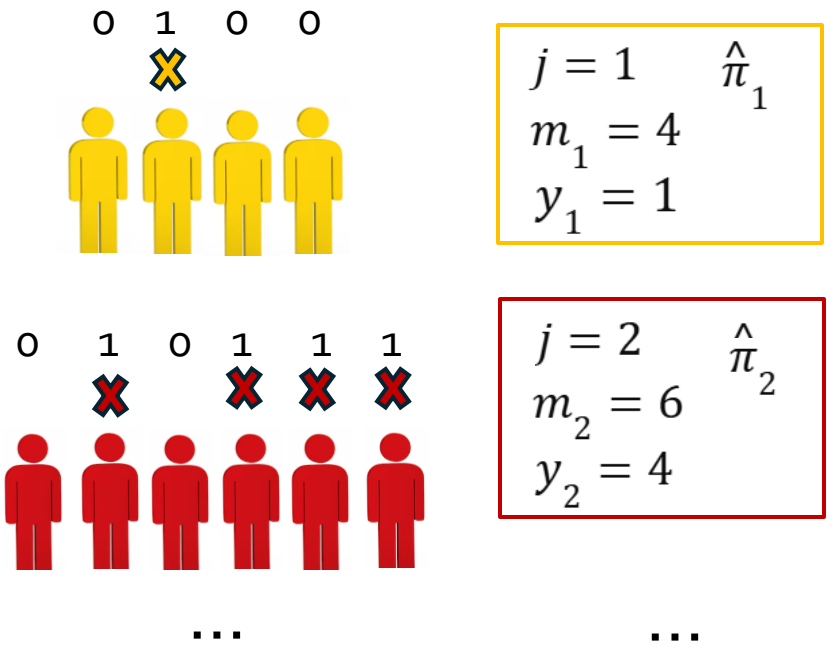
$$\hat{\pi}(\mathbf{x}_i)$$



Covariate pattern

- It consists of observations with the same set of covariates.

- Suppose that there are *J* distinct covariate patterns, for the *j*-th covariate pattern:
  - m_j : number of observations;
  - y_j : number of observed events;
  - $\hat{\pi}_j$: predicted probability.



Pearson Residuals Def. 1 – *covariate pattern level*:

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Computed using all observations with the *j*-th covariate pattern

It does not vary across observations with the *j*-th covariate pattern

$$X^2 = \sum_{j=1}^{J} r_j^2$$

Pearson chi-square statistic for goodness-of-fit

$$\chi^2_{J-(p+1)}$$

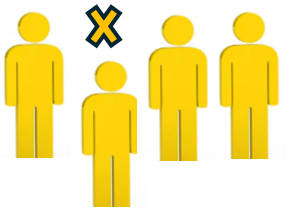
Pearson Residuals Def. 2 – *individual level*:

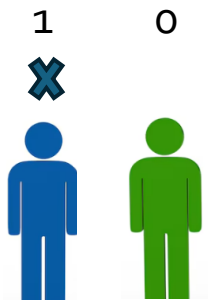
Observed outcome

$$r_i = \frac{y_i - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Computed at the observation level

It can vary across observations with the *j*-th covariate pattern due to the observed outcome





$$m_j = 1$$

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Pearson Residuals– *covariate pattern level*:

- The use of residuals calculated at the covariate pattern level is recommended for assessing the fit of the model

Covariate pattern A ($m_j = 1$, $y_j = 0$, $\hat{\pi}_j = 0.5$)

Observation identifier	y_i	r_i	r_j
A1	0	-1	-1

Covariate pattern B ($m_j = 16$, $y_j = 0$, $\hat{\pi}_j = 0.5$)

Observation identifier	y_i	r_i	r_j
B1	0	-1	-4
B2	0	-1	-4
B3	0	-1	-4
B4	0	-1	-4
B5	0	-1	-4
B6	0	-1	-4
B7	0	-1	-4
B8	0	-1	-4
B9	0	-1	-4
B10	0	-1	-4
B11	0	-1	-4
B12	0	-1	-4
B13	0	-1	-4
B14	0	-1	-4
B15	0	-1	-4
B16	0	-1	-4

Pearson Residuals– *covariate pattern level*:

Covariate pattern C ($m_j = 16$, $y_j = 8$, $\hat{\pi}_j = 0.5$)

Observation identifier	y_i	r_i	r_j
C1	0	-1	0
C2	0	-1	0
C3	0	-1	0
C4	0	-1	0
C5	0	-1	0
C6	0	-1	0
C7	0	-1	0
C8	0	-1	0
C9	1	1	0
C10	1	1	0
C11	1	1	0
C12	1	1	0
C13	1	1	0
C14	1	1	0
C15	1	1	0
C16	1	1	0

Pearson Residuals – *Stata:*

- In *Stata*, logistic regression can be performed using the *logit* command as well as the *glm* command with *family(binomial)* and *link(logit)*.
- Both commands allow for the calculation of Pearson residuals, but the computation follows substantially different approaches

logit postestimation — Def. 1 r_j

```
predict
residuals        Pearson residuals; adjusted for number sharing covariate pattern
```

```
estat gof
```

glm postestimation — Def. 2 r_i

```
predict
pearson          Pearson residuals
```

```
e(deviance_p)
```

Pearson Residuals – *Stata – Illustrative example:*

```
. use https://www.stata-press.com/data/r18/lbw.dta
(Hosmer & Lemeshow data)

. de *

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
id                int     %8.0g                 Identification code
low               byte    %8.0g                 Birthweight<2500g
age               byte    %8.0g                 Age of mother
lwt               int     %8.0g                 Weight at last menstrual period
race              byte    %8.0g       race      Race
smoke             byte    %9.0g       smoke     Smoked during pregnancy
ptl               byte    %8.0g                 Premature labor history (count)
ht                byte    %8.0g                 Has history of hypertension
ui                byte    %8.0g                 Presence, uterine irritability
ftv               byte    %8.0g                 Number of visits to physician during 1st trimester
bwt               int     %8.0g                 Birthweight (grams)
```

M3 M3 M1 M3 M2 M3 M3 M3 M3

Pearson Residuals – *Stata* – *Illustrative example*

Model 1: Single Numerical Covariate

```
. logit low lwt, nolog

Logistic regression                                     Number of obs =    189
                                                        LR chi2(1)    =   5.96
                                                        Prob > chi2   = 0.0146
Log likelihood = -114.35403                             Pseudo R2     = 0.0254

------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lwt |  -.0140371   .0061685    -2.28   0.023    -.0261272    -.001947
       _cons |   .9957626   .7852434     1.27   0.205    -.5432862    2.534811
------------------------------------------------------------------------------

. estat gof

Goodness-of-fit test after logistic model
Variable: low

      Number of observations =    189
Number of covariate patterns =     76
           Pearson chi2(74) =  81.36
                 Prob > chi2 = 0.2609
```

```
. qui glm   low lwt, family(binomial) link(logit)

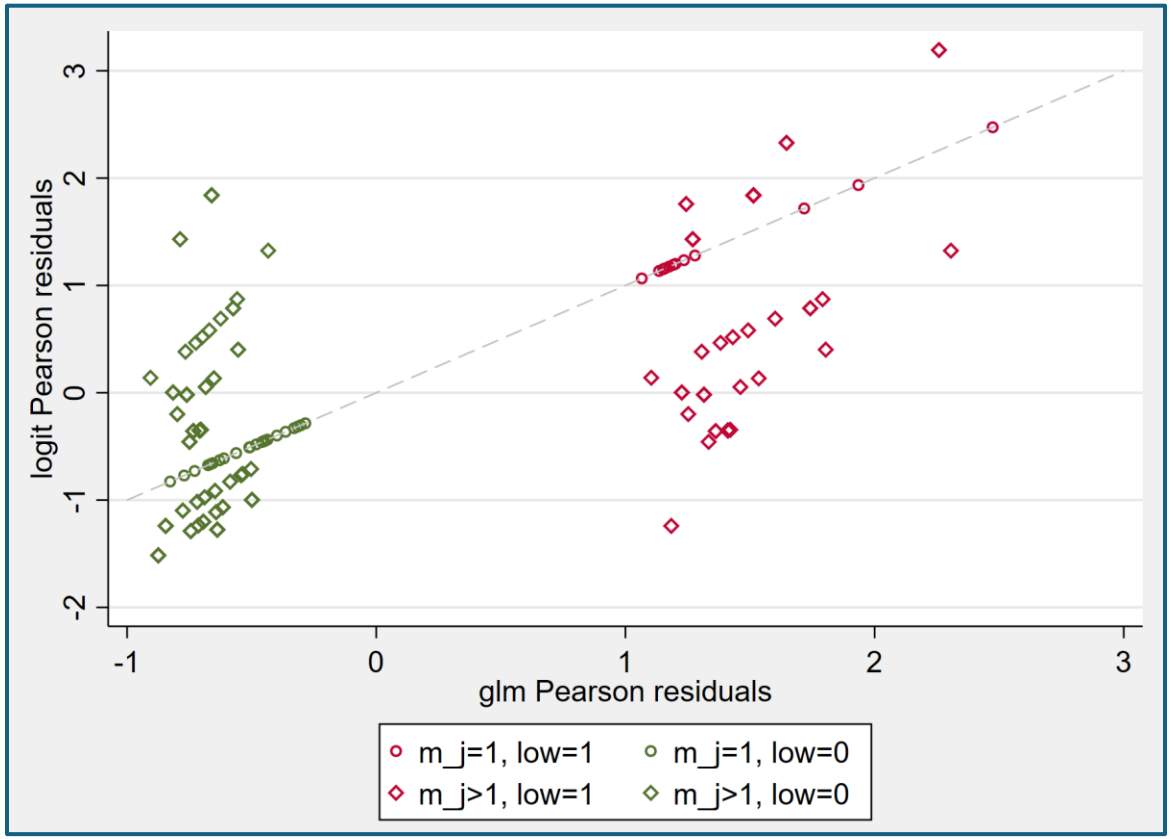
. di e(deviance_p)
189.64023
```

Pearson Residuals – *Stata* – *Illustrative example*

Model 1: Single Numerical Covariate

```
. list  low lwt res_logit_M1 res_glm_M1 if lwt==150 | lwt==250
```

	low	lwt	res_logit_M1	res_glm_M1
149.	0	150	.7876129	-.5741283
150.	0	150	.7876129	-.5741283
151.	0	150	.7876129	-.5741283
152.	1	150	.7876129	1.741771
153.	1	150	.7876129	1.741771
189.	0	250	-.2845752	-.284575



Pearson Residuals – *Stata* – *Illustrative example*

Model 2: Single Categorical Covariate

```
. logit low i.race, nolog

Logistic regression                                     Number of obs =    189
                                                        LR chi2(2)    =   5.01
                                                        Prob > chi2   = 0.0817
Log likelihood = -114.83082                             Pseudo R2     = 0.0214

------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        race |
      Black  |   .8448103   .4634141     1.82   0.068    -.0634647    1.753085
      Other  |   .6361714    .347831     1.83   0.067    -.0455648    1.317908
             |
       _cons |  -1.154965   .2391169    -4.83   0.000    -1.623626   -.6863047
------------------------------------------------------------------------------

. estat gof

Goodness-of-fit test after logistic model
Variable: low

      Number of observations =  189
Number of covariate patterns =    3
            Pearson chi2(0) = 0.00
                Prob > chi2 =    .
```

```
. qui glm   low i.race, family(binomial) link(logit)

. di e(deviance_p)
189
```

Pearson Residuals – *Stata* – *Illustrative example*

Model 2: Single Categorical Covariate

```
. tab race low
```

Race	Birthweight<2500g 0	1	Total
White	73	23	96
Black	15	11	26
Other	42	25	67
Total	130	59	189

```
. tab res_logit_M2 race
```

Pearson residual	Race White	Black	Other	Total
0	96	26	67	189
Total	96	26	67	189

```
. tab res_glm_M2 race
```

Pearson residual	Race White	Black	Other	Total
-.85635	0	15	0	15
-.77152	0	0	42	42
-.56131	73	0	0	73
1.16775	0	11	0	11
1.29615	0	0	25	25
1.78155	23	0	0	23
Total	96	26	67	189

Pearson Residuals – *Stata* – *Illustrative example*

Model 3: Multiple Covariates of Different Types

```
. logit low lwt i.race age i.smoke i.ht i.ui ptl  , nolog

Logistic regression                             Number of obs =    189
                                                LR chi2(8)    =  33.22
                                                Prob > chi2   = 0.0001
Log likelihood = -100.724                       Pseudo R2     = 0.1416
```

low	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
lwt	-.0151508	.0069259	-2.19	0.029	-.0287253	-.0015763
race						
Black	1.262647	.5264101	2.40	0.016	.2309024	2.294392
Other	.8620792	.4391532	1.96	0.050	.0013548	1.722804
age	-.0271003	.0364504	-0.74	0.457	-.0985418	.0443412
smoke						
Smoker	.9233448	.4008266	2.30	0.021	.137739	1.708951
1.ht	1.832518	.6916292	2.65	0.008	.4769494	3.188086
1.ui	.7585135	.4593768	1.65	0.099	-.1418484	1.658875
ptl	.5418366	.346249	1.56	0.118	-.136799	1.220472
_cons	.4612239	1.20459	0.38	0.702	-1.899729	2.822176

```
. estat gof

Goodness-of-fit test after logistic model
Variable: low

      Number of observations =    189
Number of covariate patterns =    182
           Pearson chi2(173) = 179.24
                 Prob > chi2 = 0.3567
```

```
. qui glm   low lwt i.race age i.smoke i.ht i.ui ptl , ///
>       family(binomial) link(logit) nolog

. di e(deviance_p)
182.02334
```

Pearson Residuals – *Stata* – *Illustrative example*

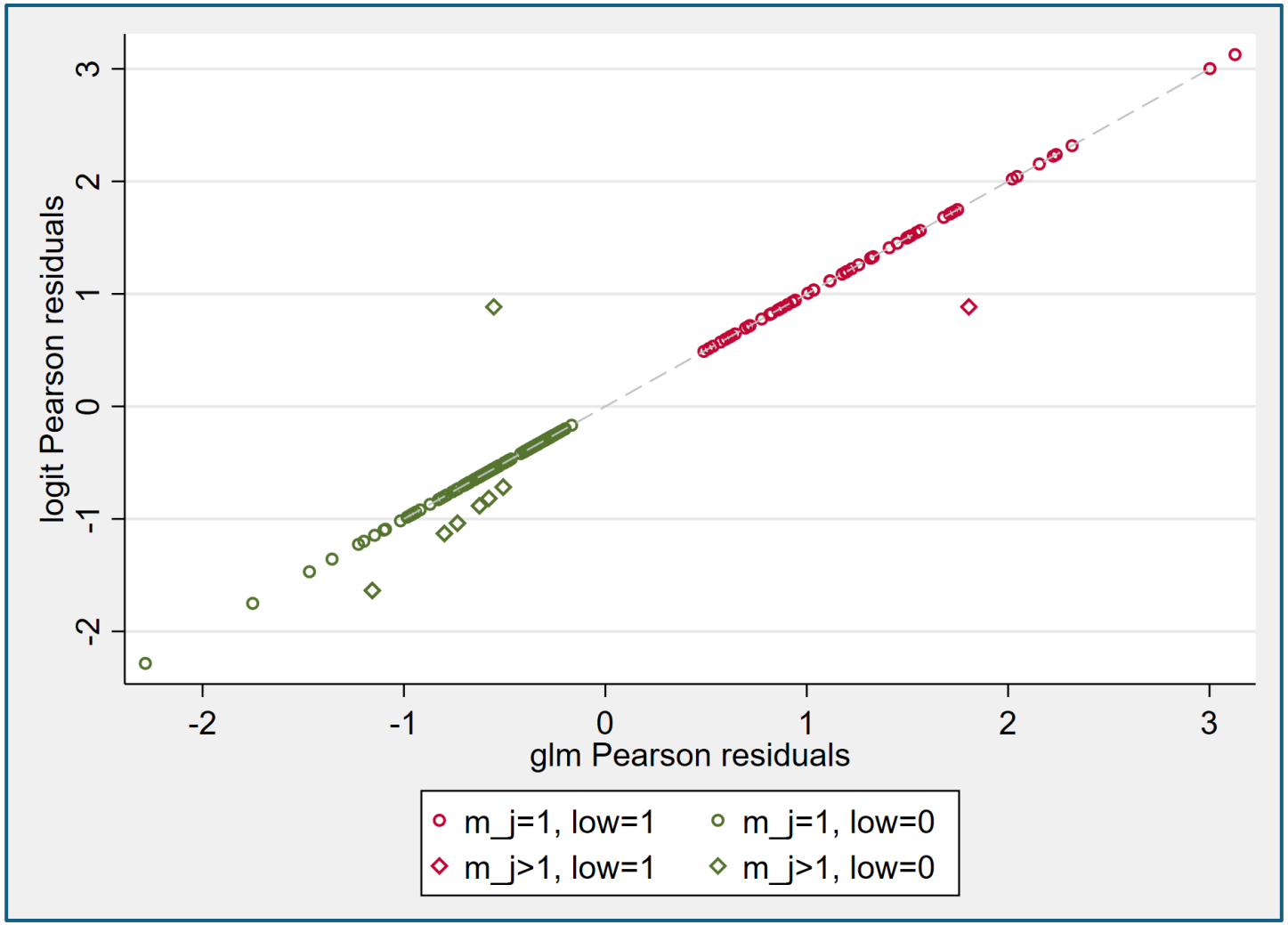
Model 3: Multiple Covariates of Different Types

```
. list  id cov_patt_M3 m_j_M3 low lwt race age smoke ht ui ptl ///
>       res_logit_M3 res_glm_M3 if m_j_M3==2, abbreviate(20)
```

	id	cov_patt_M3	m_j_M3	low	lwt	race	age	smoke	ht	ui	ptl	res_logit_M3	res_glm_M3
5.	132	5	2	0	90	White	18	Smoker	0	1	0	-1.636279	-1.157024
6.	133	5	2	0	90	White	18	Smoker	0	1	0	-1.636279	-1.157024
20.	100	19	2	0	100	White	18	Smoker	0	0	0	-1.038123	-.7340638
21.	101	19	2	0	100	White	18	Smoker	0	0	0	-1.038123	-.7340638
59.	117	57	2	0	113	Black	17	Nonsmoker	0	0	0	-1.129907	-.7989649
60.	116	57	2	0	113	Black	17	Nonsmoker	0	0	0	-1.129907	-.7989649
73.	147	70	2	0	119	Other	17	Nonsmoker	0	0	0	-.8837322	-.624893
74.	148	70	2	0	119	Other	17	Nonsmoker	0	0	0	-.8837322	-.624893
112.	140	108	2	0	130	White	22	Smoker	0	0	0	.8846817	-.5539824
113.	67	108	2	1	130	White	22	Smoker	0	0	0	.8846817	1.805111
134.	189	129	2	0	135	White	16	Smoker	0	0	0	-.8182194	-.5785685
135.	167	129	2	0	135	White	16	Smoker	0	0	0	-.8182194	-.5785685
146.	192	140	2	0	147	White	19	Smoker	0	0	0	-.7173581	-.5072488
147.	193	140	2	0	147	White	19	Smoker	0	0	0	-.7173581	-.5072488

Pearson Residuals – *Stata* – *Illustrative example*

Model 3: Multiple Covariates of Different Types



Pearson Residuals – *Conclusions*

- Distinction between residuals calculated at the covariate pattern level (*logit*) and those computed at the individual level (*glm*).
- When J ≪ n, residuals calculated at the covariate pattern level are recommended for assessing the fit of the model.

- We hope that this contribution helps *Stata* users towards the correct interpretation and practical implementation of Pearson residuals, thereby promoting a more informed assessment of model fit.

Selected references

- Bellocco, R., and Algeri, S. 2013. Goodness-of-fit tests for categorical data. The Stata Journal 13(2): 356–365;
- Hosmer, D. W., Lemeshow, S. and Sturdivant R. X. 2013. Applied Logistic Regression. 3rd ed. John Wiley & Sons;
- Hilbe, J. 1991. sqv1.3: An enhanced Stata logistic regression program. Stata Technical Bulletin 4: 16–18;
- Statalist. 2004. Residuals in Logistic Regression. Discussion, April 9–13, 2004.



ponzano.marta@gmail.com
m.ponzano@unilink.it