

---

# Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data

**Oliver Kuss**

Institute of Medical Epidemiology, Biostatistics and Informatics, Medical Faculty  
University of Halle-Wittenberg,  
[Oliver.Kuss@medizin.uni-halle.de](mailto:Oliver.Kuss@medizin.uni-halle.de)

## **Programme:**

1. The Logistic Regression Model
2. Checking Goodness-of-Fit
3. The Problem of Sparse Data
4. Solutions
5. Which Solution is the Best???
6. Simulation Results
7. Conclusions
8. Literature

---

# 1. The Logistic Regression Model

Standard method for the regression of binary responses

## Reasons:

- Easy interpretation of parameters as odds-ratios
- One can predict response probabilities
- Software is available
- Valid in prospective and retrospective designs
- Methodologically sound (loglinear model, GLM, nonlinear regression model)

## Notation:

$N$  independent observations **grouped by covariate patterns**  $(y_i, x_i)$ ,  $i=1, \dots, N$

$x_i$  : vector of  $p+1$  covariates,

$y_i$  : number of successes,  $Y_i \sim B(m_i, \pi_i)$ ,

$m_i$  : number of trials,

$M = \sum_{i=1}^N m_i$  : Number of individual observations

---

**Data:**

		Response		
		1	0	
Covariate Pattern	1	$Y_1$	$m_1 - Y_1$	$m_1$
	2	$Y_2$	$m_2 - Y_2$	$m_2$
	:	:	:	:
	N	$Y_N$	$m_N - Y_N$	$m_N$

**Example:**Continuous covariate(s):  $N=M$  ( $m_i \equiv 1$ )

		Response		
		1	0	
Covariate Pattern	1	1	0	1
	2	0	1	1
	:	:	:	:
	N	1	0	1

**Model equation:**

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^p x_{ij} \beta_j$$

with  $\beta_j = (\beta_0, \dots, \beta_p)$  vector of regression parameters.Estimate parameters  $\beta_j$  via ML.

---

## 2. Checking Goodness-of-Fit

Statistical modelling consists of two steps (Hosmer et al., 1991):

### **Model building** and **Model checking**

**Model building:** How can I explain the variation in response values in terms of the covariates (*systematical component*)

**Model checking:** Assess all variation that is not explained by the systematical component by comparing observations and prognoses from the model (*error component*)

Systematical component describes the "average" value of the response, error component describes the deviation from this "average" value

### **Two different kinds of model checking:**

1) Consider deviations from prognoses for each single observations (numerically and graphically):

⇒ **Residual analysis**

2) Calculate goodness-of-fit measures and assess model fitting by a single number and a statistical test:

⇒ **Goodness-of-fit tests**

---

## Goodness-of-Fit Tests

- **Specific:** Embed the logistic model in a wider class of models and test the parameter that describes the standard model:

e.g.: Pregibon, 1980

$$g(\pi_i, \lambda) = \log \left( \frac{(1/(1 - \pi_i))^\lambda - 1}{\lambda} \right)$$

Testing  $\lambda=1$  yields a specific goodness-of-fit test.

- **Global:** Testing of a unspecific null hypothesis „The model fits“

From global goodness-of-fit tests we get no advice how to improve the model in case of a bad model fit

However, it is dangerous to expect this from specific tests

---

## Two classical Goodness-of-Fit tests:

### Pearson statistic:

$$X^2 = \sum_{i=1}^N \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}$$

### Residual deviance:

$$D = 2 \sum_{i=1}^N y_i \ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (m_i - y_i) \ln\left(\frac{m_i - y_i}{m_i - \hat{\pi}_i}\right)$$

Large values of  $X^2$ ,  $D$  indicate lack-of-fit

**Statistical test:** Compare  $X^2$ ,  $D$  to a  $\chi^2$ -distribution with  $N-p-1$  df

---

### 3. The Problem of Sparse Data

**The  $\chi^2$ -limiting distribution is only valid for large  $m_i$   
( $N$  fixed,  $m_i \rightarrow \infty$  for all  $i$ )**

Unrealistic with a large number of covariates or with continuous covariates

#### **A disaster:**

In the extreme case of  $m_i \equiv 1$   $D$  degenerates to

$$D = 2 \sum_{i=1}^N \hat{\pi}_i \ln \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \ln(1 - \hat{\pi}_i),$$

is independent of the  $y_i$ , and contains absolutely no information about the model fit.

Not much better, for  $X^2$  there is:  $X^2 \approx N$

---

## The problem is well-known, ...

- The  $X^2$  and D goodness-of-fit statistics do not have approximate chi-squared distributions when applied to logistic regression models with a continuous covariate, unless there are many observations at each level of the covariate. (Agresti, 1990)
- Neither  $X^2$  nor D is appropriate in the many strata standard asymptotic model, because under this model there is no  $\chi^2$ -limiting distribution. (Santner/Duffy, 1989)
- Thus, p-values calculated for  $X^2$  and D when  $M \approx N$ , using the  $\chi^2$ -distribution, are incorrect. (Hosmer/Lemeshow, 1989)
- The effect of sparseness is noticed mainly on D and  $X^2$ , which fail to have the properties required for goodness-of-fit statistics. (McCullagh/Nelder, 1989)

---

## ... but what is the solution???

- In principle it would seem preferable to accept the failure of the chi-square limit and to use a more accurate approximation to the null distribution without accumulating cells. (Lloyd, 1999)
- Thus, to analyze lack of fit when explanatory variables are continuous, we apply goodness-of-fit statistics and related residual measures by grouping observed and fitted values for a partition of the space of explanatory variable values. (Agresti, 1989)
- It is good statistical practice, however, not to rely on either  $D$  or  $X^2$  as an absolute measure of goodness of fit in these circumstances. It is much better to look for specific deviations from the model of a type that is easily understood scientifically. (McCullagh/Nelder, 1989)

---

## 4. Solutions

### 4.1 Modify limiting distribution

- $X^2$ ,  $D$  are asymptotically normal under  $n, m_i \rightarrow \infty$   
(Osious/Rojek, 1992; McCullagh, 1986)

### 4.2 Grouping observations

- Hosmer-Lemeshow test (Hosmer/Lemeshow, 1980)  
Maybe the standard test with sparse data nowadays,  
but it has some deficiencies (Hosmer et al, 1997,  
Bertolini et al., 2000)

### 4.3 Use other tests statistics

- $X_F^2$  (Farrington, 1996)

$$X_F^2 = X^2 + \sum_{i=1}^N \frac{-(1 - 2\hat{\pi}_i)}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)} (y_i - m_i \hat{\pi}_i)$$

Approximate moments:

$$E(X_F^2 | \hat{\beta}) = N - p - 1 + \sum_{i=1}^N \hat{\pi}_i (1 - \hat{\pi}_i) \hat{Q}_{ii}$$

$$\text{Var}(X_F^2 | \hat{\beta}) = 2 \left( 1 - \frac{p+1}{N} \right) \sum_{i=1}^N \frac{m_i - 1}{m_i}$$

with  $\hat{Q} = X(X^t \hat{W} X)^{-1} X^t$ ,  $\hat{W} = \text{diag}(m_i \hat{\pi}_i (1 - \hat{\pi}_i))$ .

**Test:**

$$Z_F = \frac{X_F^2 - E(X_F^2 | \hat{\beta})}{\text{Var}(X_F^2 | \hat{\beta})^{1/2}}$$

is standard normal under the hypothesis

**Problem:** For  $m_i \equiv 1$  we have  $X_F^2 = N$

---

**IM-Test** (White, 1982; Orme, 1988)

Information matrix equation:

$$-E\left(\frac{\partial^2 L}{\partial \beta \partial \beta'}\right) = E\left(\frac{\partial L}{\partial \beta} \frac{\partial L}{\partial \beta'}\right)$$

Estimate both matrices, summation of the main diagonal elements yields the  $((p+1) \times 1)$  vector

$$\hat{d} = \sum_{i=1}^M (y_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i)z_i$$

with  $z_i = (1, x_{i1}^2, \dots, x_{ip}^2)^t$

**Test:**

$$IM = \frac{1}{M} \hat{d}' \hat{V}^{-1} \hat{d}$$

is  $\chi^2$ -distributed with  $(p+1)$  df

and

$$\hat{V} = \frac{1}{M} \left[ Z^{*t} \left( I - X^* (X^{*t} X^*)^{-1} X^{*t} \right) Z^* \right],$$

$$X^* = \sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)} X,$$

$$Z^* = \sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)} (1 - 2\hat{\pi}_i) Z,$$

$Z$  as the matrix with the  $z_i$  as rows.

---

$\mathbf{R}_C$  (Copas, 1986, Hosmer et al., 1997)

$$R_C = \sum_{i=1}^M (y_i - m_i \hat{\pi}_i)^2$$

Summation of the raw Pearson residuals

Asymptotical moments:

$$E\left(R_C - \sum_{i=1}^M \hat{\pi}_i (1 - \hat{\pi}_i)\right) = 0$$

$$\text{Var}\left(R_C - \sum_{i=1}^M \hat{\pi}_i (1 - \hat{\pi}_i)\right) = (1 - 2\hat{\pi})' (\hat{W} - \hat{W}\hat{Q}\hat{W}) (1 - 2\hat{\pi})$$

with  $\hat{Q} = X(X'\hat{W}X)^{-1}X'$ ,  $\hat{W} = \text{diag}(m_i \hat{\pi}_i (1 - \hat{\pi}_i))$ .

**Test:**

$$Z_C = \frac{R_C - \sum_{i=1}^M \hat{\pi}_i (1 - \hat{\pi}_i)}{\text{Var}\left(R_C - \sum_{i=1}^M \hat{\pi}_i (1 - \hat{\pi}_i)\right)^{1/2}}$$

is standard normal under the hypothesis

---

**Example:**

Occupational hand eczema in hairdresser apprentices,  
M=574 (340 „successes“),

Several covariates (p=6): genetical disposition, work  
hazards, confounders,

N=334,

Distribution of the  $m_i$ :

$m_i$	Frequency
1	205 (61%)
2	68 (20%)
3	35 (11%)
>3	26 (8%)

Assessing goodness-of-fit:

	p-value
$X^2$	0,053
D	0,012
$X_O^2$	0,044
$X_M^2$	0,031
HL-Test	0,451
$X_F^2$	0,408
IM-Test	0,365
$R_C$	0,062

**Who is right???**

---

## 5. Which Solution is the Best???

Up to now there is only one large, systematic investigation of global goodness-of-fit tests in logistic regression (Hosmer et al., 1997)

### **Results:**

$R_C$  and  $X_M^2$  were "winners"

### **But some need of supplement:**

- Add new tests
- Varying  $m_i$

---

## 6. Simulation Results

### 6.1. Null hypothesis

Single continuous covariate  $x_1$  with  $x_1 \sim N(0,1)$ ,  
 $\beta_0=0$ ,  $\beta_1=0,693$ ,  $M=500$ , 1000 runs,  $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1-10$	$m_i \equiv 10$
$X^2$	0,000	0,010	0,002	0,046
D	1,000	0,977	0,585	0,114
$X_O^2$	0,061	0,043	0,040	0,041
$X_M^2$	0,063	0,052	0,045	0,052
HL test	0,055	0,051	0,054	0,052
$X_F^2$	0,000	0,051	0,055	0,062
IM test	0,057	0,049	0,045	0,049
$R_C$	0,053	0,052	0,046	0,051

Three continuous covariates  $x_i$  with  $x_i$  iid  $N(0,1)$ ,  
 $\beta_0=0$ ,  $\beta_1=0.693$ ,  $\beta_2=0.405$ ,  $\beta_3=0.223$ ,  $M=500$ , 1000 runs,  
 $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1-10$	$m_i \equiv 10$
$X^2$	0,000	0,001	0,009	0,043
D	1,000	0,959	0,866	0,118
$X_O^2$	0,074	0,039	0,042	0,026
$X_M^2$	0,078	0,052	0,059	0,057
HL-Test	0,049	0,049	0,052	0,042
$X_F^2$	0,000	0,052	0,058	0,058
IM-Test	0,051	0,049	0,058	0,051
$R_C$	0,058	0,046	0,048	0,049

---

## 6.2. Alternative hypothesis

### Overdispersion

Continuous covariate  $x_1$  with  $x_1 \sim U(-6,6)$ ,  $\beta_0=0$ ,  
 $\beta_1=0.405$ ,

**Misspecification:**  $\beta_0$  random,  $E(\beta_0)=0$ ,  $\text{Var}(\beta_0)=0.323$ ,  
 $M=500$ , 1000 runs,  $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1-10$	$M_i \equiv 10$
$X_O^2$	0,045	0,211	0,201	0,645
$X_M^2$	0,047	0,230	0,230	0,694
HL-Test	0,046	0,052	0,121	0,231
$X_F^2$	0,000	0,232	0,464	0,699
IM-Test	0,043	0,040	0,086	0,123
$R_C$	0,045	0,053	0,061	0,107

### Misspecified Link Function

Continuous covariate  $x_1$  with  $x_1 \sim U(-6,6)$ ,

**Misspecification:**  $\log(-\log(1-\pi_i))=0.405x_1$

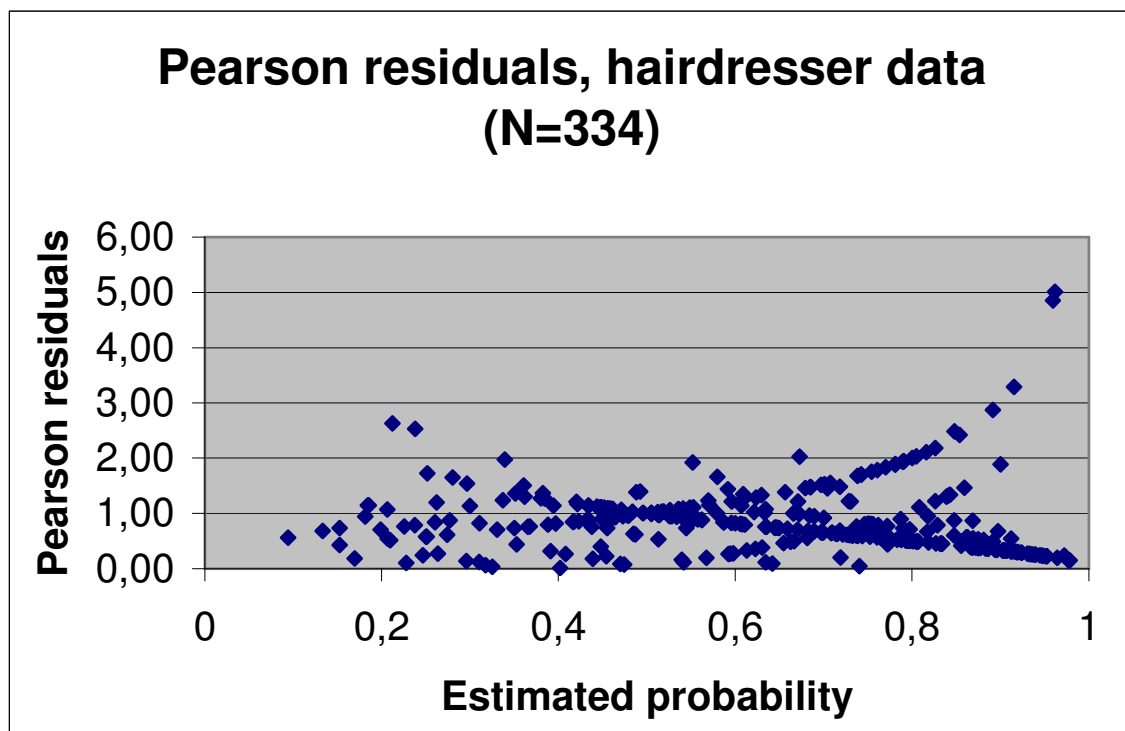
$M=500$ , 1000 runs,  $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1-10$	$M_i \equiv 10$
$X_O^2$	0,000	0,001	0,000	0,025
$X_M^2$	0,000	0,001	0,000	0,037
HL-Test	0,200	0,197	0,204	0,195
$X_F^2$	0,000	0,067	0,059	0,126
IM-Test	0,541	0,545	0,527	0,517
$R_C$	0,275	0,277	0,289	0,267

---

## Back to the example:

	p-Wert	p*-Wert
$X^2$	0,053	0,391
D	0,012	0,033
$X_O^2$	0,044	0,511
$X_M^2$	0,031	0,458
HL-Test	0,451	0,299
$X_F^2$	0,408	0,427
IM-Test	0,365	0,873
$R_C$	0,062	0,734



---

## 7. Conclusion

- $X^2$  and D are no valid goodness-of-fit tests in logistic regression with sparse data.
- There are alternatives to this test, even the Hosmer-Lemeshow test can be outperformed, calculation of these is straightforward.
- However, for extreme sparseness ( $m_i \equiv 1$ ) and small sample size the alternative tests have low power
  - Global goodness-of-fit are a valuable tool, but a sound analysis of lack-of-fit should not be considered as adequate after calculating a single goodness-of-fit statistic.

### **The fundamental dilemma remains:**

A non-significant result of a goodness-of-fit test doesn't tell you that your model is correct.

### **Software:**

SAS/IML macro %GOFLOGIT

Write to [Oliver.Kuss@medizin.uni-halle.de](mailto:Oliver.Kuss@medizin.uni-halle.de)

---

## 8. Literature

- Agresti A. *Categorical data analysis*. John Wiley & Sons, 1990.
- Bertolini G et al. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidem Biostat*, 5:251-253, 2000.
- Copas JB. Unweighted Sum of Squares Test for Proportions. *Appl Statist*, 38:71-80, 1989.
- Farrington CP. On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *J R Statist Soc B*, 58:349-360, 1996.
- Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Statist - Theor Meth*, 9:1043-1069, 1980.
- Hosmer DW, Lemeshow S. *Applied logistic regression*. John Wiley & Sons, 1989.
- Hosmer DW, Taber S, Lemeshow S. The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *Am J Public Health*, 81:1630-1635, 1991.
- Hosmer DW et al. A comparison of goodness-of-fit tests for the logistic regression model. *SiM*, 16:965-980, 1997.
- Lloyd CJ. *Statistical Analysis of Categorical Data*. John Wiley & Sons, 1999.
- McCullagh P. On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models. *International Statistical Review*, 53:61-67, 1985.
- McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall, 1989.
- Osius G, Rojek D. Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom. *JASA*, 87:1145-1152, 1992.
- Orme C. The calculation of the information matrix test for binary data models. *The Manchester School*, 54:370-376, 1988.
- Pregibon D. Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15-24, 1980.
- Santner TJ, Duffy DE. *The statistical analysis of discrete data*. Springer, 1989.
- White H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1-25, 1982.