Goodness-of-Fit Tests for Binary GLMs in the Case of No/Few Replicates

STAT 851 Final Project

Nikola Surjanovic, Peter Tea

April 6, 2020

Outline

Introduction / Motivation

Why Do We Need GOF Tests?

We should check if GLM assumptions are correct

- Goodness-of-fit (GOF) tests

Why Do We Need GOF Tests?

We should check if GLM assumptions are correct

- Goodness-of-fit (GOF) tests

However... with enough data, we will always reject the null hypothesis!

- What's the point then?

Why Do We Need GOF Tests?

Let's make sure our GOF tests have "stable" performance properties

- Probably don't want to blindly use these tests
- Still, make sure that the theory behind them makes sense!
- Use as an aid

Notation

$Y_i | X_i = x_i \sim Bin(n_i, \pi_i)$, $i = 1, \ldots, J$

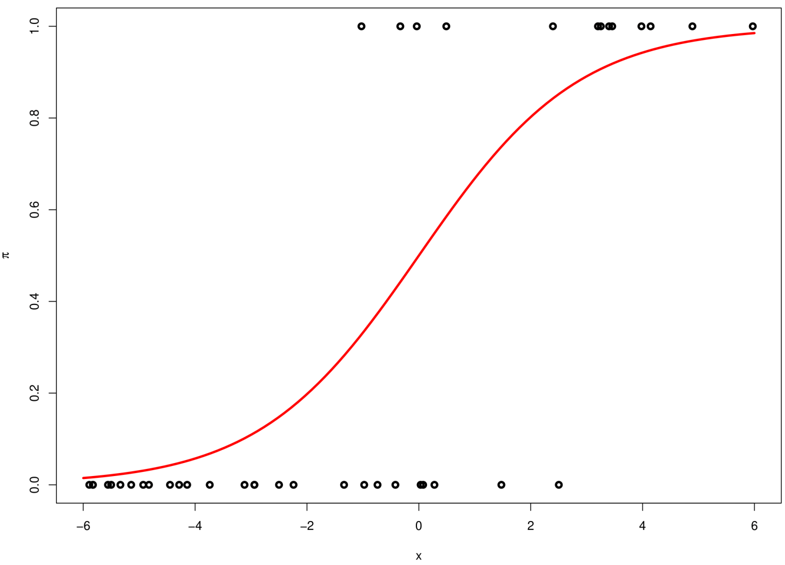
x_i's are p-dimensional

$$q(\pi_i) = \beta^\top x_i,$$

for some p-dimensional vector β

$q(\cdot)$ is specified in advance

- E.g. could be logit$(\cdot)$



1.0
0.8
0.6
0.4
0.2
0.0
π
−6
−4
−2
0
2
4
6
x

Motivation

Consider $Y_i \sim Bin(n_i, \pi_i)$, $i = 1, \ldots, J$

The Pearson chi-squared statistic:

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

Motivation

If J is *fixed*, and each $n_i \to \infty$,

$$X^2 \xrightarrow{d} \chi^2_{J-p},$$

where p is the number of estimated parameters.

Motivation

Consider $Y_i \sim Bin(1, \pi_i)$, $i = 1, \ldots, n$

The Pearson chi-squared statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}$$

Motivation

As $n \to \infty$,

$$X^2 \xrightarrow{d} \chi^2_{n-p} \quad (?)$$

Motivation

As $n \to \infty$,

$$X^2 \xrightarrow{d} \chi^2_{n-p} \quad (?)$$

How can we converge to something that is "increasing"?

Motivation

$$X^2 \xrightarrow{d} \chi^2_{n-p} \quad (?)$$

Idea: Maybe

$$\frac{X^2 - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1) \quad (?),$$

for some μ_n, σ_n^2 by a generalization of CLT

What's Special About Few/No Replicates?

Few or no replicates?

This can happen, for example, when:

- A continuous explanatory variable is present
- A discrete variable that can take on infinitely many values

Test Statistics

Overview of Tests

1. Osius-Rojek test
2. Hosmer-Lemeshow test
3. C_n test statistic
4. Stukel test
5. Other tests

Osius-Rojek Test

Consider $Y_i \sim Bin(1, \pi_i)$, $i = 1, \ldots, n$, again

The Pearson chi-squared statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}$$

Osius-Rojek Test

In general (no replicates, few replicates, many replicates, etc.),

$$\frac{X^2 - \hat{\mu}}{\hat{\sigma}} \xrightarrow{d} N(0, 1),$$

for some $\hat{\mu}, \hat{\sigma}^2$ (see Osius and Rojek, 1992)

Osius-Rojek Test

What are μ, σ^2 for Pearson X^2?

$$\hat{\mu} = \mu = J \quad \text{(number of covariate patterns)}$$

$\hat{\sigma}^2$ can be obtained through the residual sum of squares from a weighted least squares regression on a transformed set of data

Osius-Rojek Test

Question: Is this a one-sided or a two-sided test?

Osius-Rojek Test

Question: Is this a one-sided or a two-sided test?

If $J = n$, should treat as a two-sided test.

- Can be viewed as a type of score test

Beauty of Osius-Rojek Test

Results derived for a generalized distance between Y_i's and $n_i\hat{\pi}_i$'s

Essentially formed infinitely many test statistics indexed by a parameter, λ

- $\lambda = 1$ for Pearson
- $\lambda = 0$ for deviance
- See Read and Cressie (1988)

Drawbacks of Osius-Rojek Test

Not clear how to generalize to other GLMs

- Should be a multinomial response (i.e. binomial is OK)

Hosmer-Lemeshow Test

Idea: Group by fitted values, rather than by covariate pattern

- 1980: David Hosmer and Stanley Lemeshow
- Partition units into g subgroups based on $\hat{\pi}_i$'s

Test Statistic Presentation

	1	2	...	g	
y = 0	1	0		3	
y = 1	0	2		5	

- J = # of distinct covariate patterns
- n_i = # of replicates, with the jth covariate pattern $(i = 1, 2, \ldots, J)$
- y_i = # of successes among the n_i units

Hosmer-Lemeshow Test

$$\widehat{C} = \sum_{k=1}^{g} \left[\frac{\left(o_{1k} - \hat{e}_{1k}\right)^2}{\hat{e}_{1k}} + \frac{\left(o_{0k} - \hat{e}_{0k}\right)^2}{\hat{e}_{0k}} \right] \tag{1}$$

$$o_{1k} = \sum_{j=1}^{c_k} y_j, \qquad o_{0k} = \sum_{I_k} \left(m_j - y_j\right)$$

$$\hat{e}_{1k} = \sum_{I_k} m_j \hat{\pi}_j, \qquad \hat{e}_{0k} = \sum_{I_k} m_j \left(1 - \hat{\pi}_j\right)$$

I_k is the collection of indices of observations that fall in the kth group

Hosmer-Lemeshow Test

Distribution of $\widehat{C}$

- Simulation study by Hosmer and Lemeshow (1980) found that when J = n,

  $\hat{C} \dot{\sim} \chi^2_{g-2}$

  Approximately true when $J \approx n$

Hosmer-Lemeshow Test

Strategies to split units into meaningful subgroups

I. Percentiles of estimated probabilities (a.k.a. “deciles-of-risk” if $g = 10$)

II. Fixed endpoints for the estimated probabilities

Method I is commonly used

Hosmer-Lemeshow Test

Choice of g impacts results

"Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression" - Paul Allison (UPenn)

https://statisticalhorizons.com/hosmer-lemeshow

C_n Test Statistic

Based on sum of squared Pearson residuals

$$\sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i (1 - \hat{\pi}_i)}$$

Can be standardized to be distributed $N(0,1)$ asymptotically

Unstable variance with extreme values of p_i (Chen et al., 2018)

C_n Test Statistic

Stabilized variance (see Chen et al., 2018)

$$C_n = \sum_{i=1}^{n} \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i \left(1 - \hat{\pi}_i\right)}}$$

Under some regularity assumptions,

$$\tilde{C}_n = \frac{n^{-\frac{1}{2}} C_n}{\sigma_n} \xrightarrow{d} N(0,1)$$

Conduct 2-sided GOF test

C_n Test Statistic

No replicates?

One regularity assumption:

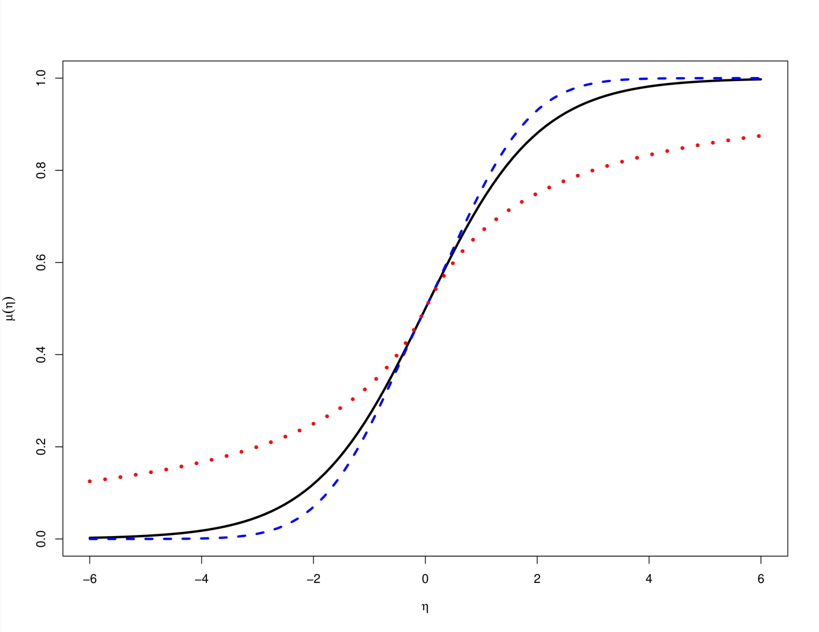
- There exists a finite number M such that $\|x_i\| \leq M, \forall i \in \mathbb{N}$ and $\lim_{n\to\infty} \sum_i x_i x_i^\top$ is a finite non-singular matrix (x_i represents the ith row of the model matrix as a column vector)

Stukel Test

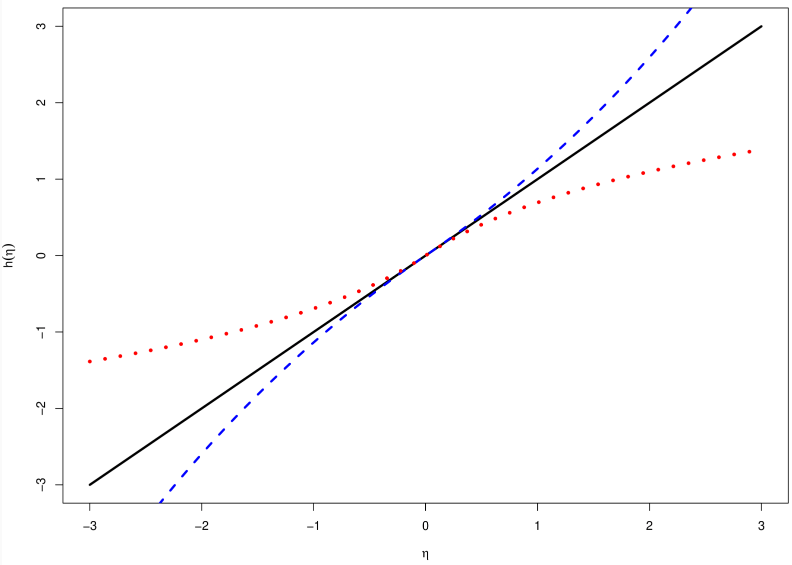
Stukel test: Essentially a test of the link function

- A missing quadratic term might be (approximately) captured by a modified link function

Create a family of link functions



μ(η)
0.0
0.2
0.4
0.6
0.8
1.0
−6
−4
−2
0
2
4
6
η



h(η)
η
−3
−2
−1
0
1
2
3

Stukel Test

A family of link functions, parameterized by $\alpha = (\alpha_1, \alpha_2)^\top$

$$\pi_\alpha(\eta) = \frac{\exp(h_\alpha(\eta))}{1 + \exp(h_\alpha(\eta))}$$

For fixed α, we still have a GLM, with link $h_\alpha^{-1}(\text{logit}(\cdot))$

- We estimate α, though

Stukel Test

When $\alpha = (0, 0)$ we have the regular logit link

$\alpha \approx (.165, .165)$ yields probit

$\alpha \approx (.62, -.037)$ yields complementary log-log

Stukel Test

$$H_0 : \alpha = 0 \text{ vs. } H_1 : \alpha \neq 0$$

Score test essentially tests logit link

- Could probably test $\alpha = (.165, .165)$ (probit), etc.

Other Tests

There are many other tests

1. For example, see Chen et al. (2018) and Hosmer et al. (1997)

Information Matrix Statistic

Information matrix statistic: A good test to keep in mind

- Very general
- Based on comparing two estimates of the Fisher information

Information Matrix Statistic

Under some regularity conditions,

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log(f(X;\theta))\right)\left(\frac{\partial}{\partial\theta}\log(f(X;\theta))\right)^{\top}\right] = -\,\mathrm{E}\left[\left(\frac{\partial^2}{\partial\theta\partial\theta}\log(f(X;\theta))\right)\right]$$

Estimates differ **if model is misspecified**

Information Matrix Statistic

Roughly speaking, check if

$$|\text{Estimate 1} - \text{Estimate 2}| > 0$$

See White (1982)

Application

Tennis Data

Novak Djokovic holding his 8th Australian Open Trophy (17th Grand Slam)

Tennis Data

Context: Observe tennis serve speeds (continuous), and outcome of service point (binary)

- 70% of all points are won in 4 shots or less
- All points begin with a serve
- Fit binary logistic regression, Win Point $\sim$ Serve Speed + Opponent (7 matches total)

Discussion

Pros and Cons

Pros

1. Osius-Rojek: Easy to compute, not restricted to logit link
2. Hosmer-Lemeshow: Simple and intuitive
3. C_n statistic: Also simple
4. Stukel: Tests link function directly

Pros and Cons

Cons

1. Osius-Rojek: Doesn't generalize to non-binary responses
2. Hosmer-Lemeshow: Depends on g
3. C_n statistic: Only works without replicates
4. Stukel: Generalized link function not exact for probit, cloglog, etc.

Summary of Test Statistics

Summary of performances from Chen et al. (2018)

References

Lu Chen, Yihan Sui, Chi Song, and Grzegorz A. Rempala. The sum of standardized residuals: Goodness-of-fit test for binary response models. *Statistics in Medicine*, 37(11):1932–1941, 2018.

David W. Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069, 1980.

David W Hosmer, Trina Hosmer, Saskia Le Cessie, and Stanley Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997.

Gerhard Osius and Dieter Rojek. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(420):1145–1152, 1992.

TR Read and NA C Cressie. Goodness-of-fit statistics for discrete multivariate data. 1988.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Thank You!

Questions?

Bonus Material

Stukel's $h_\alpha(\eta)$

For $\eta \geq 0$,

$$h_\alpha = \begin{cases} 1/\alpha_1(\exp(\alpha_1|\eta|) - 1) & \alpha_1 > 0 \\ \eta & \alpha_1 = 0 \\ -\alpha_1 \log(1 - \alpha_1|\eta|) & \alpha_1 < 0 \end{cases}$$

Stukel's $h_\alpha(\eta)$

For $\eta < 0$,

$$h_\alpha = \begin{cases} -1/\alpha_2(\exp(\alpha_2|\eta|) - 1) & \alpha_2 > 0 \\ \eta & \alpha_2 = 0 \\ \alpha_2 \log(1 - \alpha_2|\eta|) & \alpha_2 < 0 \end{cases}$$

Generalized Distance Between Y_i and $n_i\hat{\pi}_i$

$$
\begin{aligned}
a_1(Y_{i,i}\,\hat{\pi}_i) &= \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} \\
&= \frac{2Y_i}{1(1+1)} \cdot \left[\left(\frac{Y_i}{n_i\hat{\pi}_i} \right)^1 - 1 \right] - \frac{2}{1+1}(Y_i - n_i\hat{\pi}_i)
\end{aligned}
$$

Generalized Distance Between Y_i and $n_i\hat{\pi}_i$

$$a_\lambda(Y_{i,i}\,\hat{\pi}_i) = \frac{2Y_i}{\lambda(\lambda+1)} \cdot \left[\left(\frac{Y_i}{n_i\hat{\pi}_i}\right)^\lambda - 1\right] - \frac{2}{\lambda+1}(Y_i - n_i\hat{\pi}_i)$$

Derivation of Osius-Rojek Test

Test is derived by

1. Taylor series expansion to replace $\hat{\pi}$'s with π_0's
2. Lyapunov CLT (for independent but not identically distributed Y_i's)

Hosmer-Lemeshow Test

Alternate form of Hosmer-Lemeshow test statistic:

$$\widehat{C} = \sum_{k=1}^{g} \frac{\left(o_{1k} - n'_k \bar{\pi}_k\right)^2}{n'_k \bar{\pi}_k \left(1 - \bar{\pi}_k\right)} \tag{2}$$

where

$$\bar{\pi}_k = \frac{1}{n'_k} \sum_{j=1}^{c_k} m_j \hat{\pi}_j$$

and n'_k is the number of units in the kth subgroup