Including systematic errors in Poisson regression

M. Bonamente, University of Alabama in Huntsville

Outline

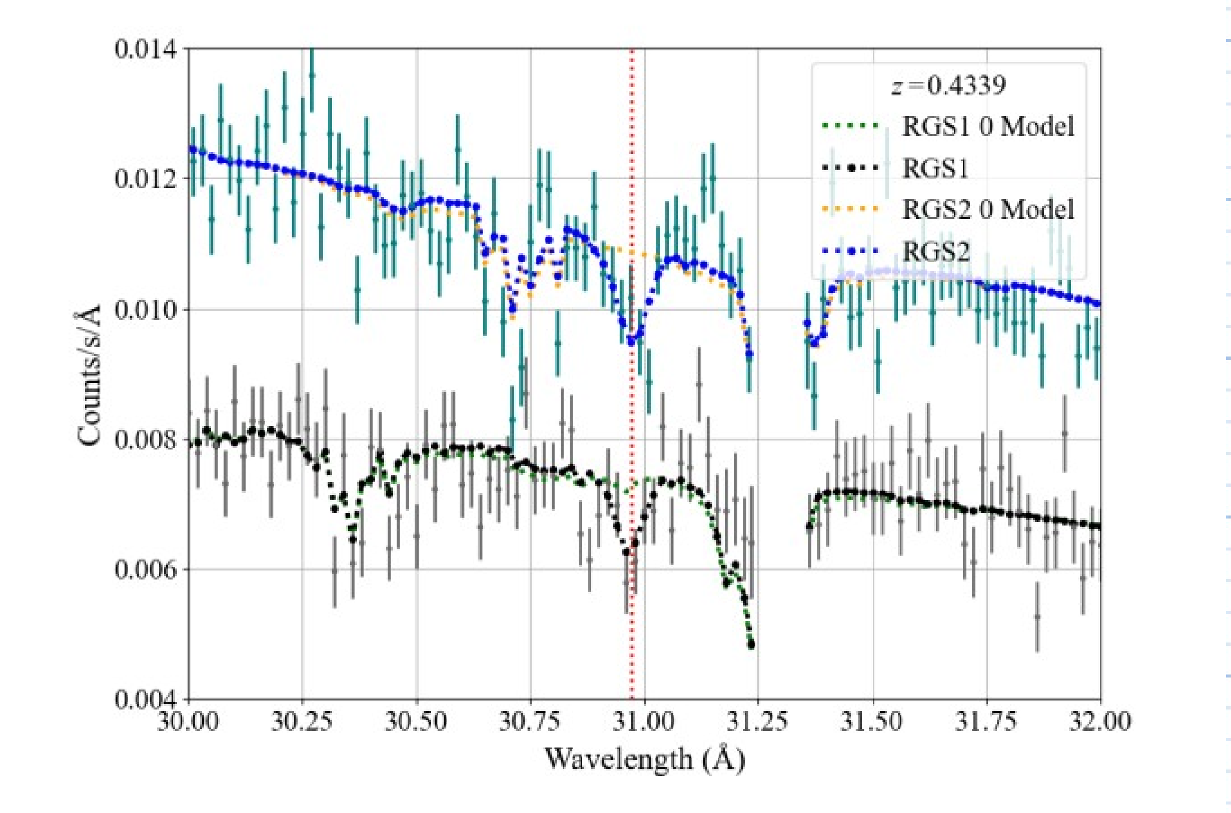
1. Introduction to systematic errors and some methods to account for them

2. A new model of systematic errors based on the intrinsic model variance

3. Use of this method for (a) estimating systematic errors; or (b) hypothesis testing.

1. Systematic errors: An introduction

It is a common situation that the goodness of fit, in this case C_{min}, is not acceptable, yet the model generally follows the data without *systematic* trends. In such cases, it is possible to consider whether there are other sources of variance in the data that have not been considered.



These spectra are from Spence+2023 for the quasar 1ES 1553+113.

Spence, D., Bonamente, M. et al. (2023). *A search for the missing baryons with X-ray absorption lines towards the blazar 1ES 1553+113*, MNRAS 532(2), 2329.

Some quotes on “systematic errors”:

Eisenhart (1962): ‘… *Systematic error, precision, and accuracy are inherent characteristics of a measurement process and not of a particular measurement yielded by the process*’,

Jeffries (1966): ‘*The usual physical practice is to distinguish between ”accidental” errors, which are reduced according to the usual rule when many observations are combined, and ”systematic” errors, which appear in every observation and persist in the mean.’*

A review by van Dyk & Lyons (2023) summarizes some of the avaliable methods: one-parameter-at-a-time error propagation,several parameters simultaneously, nuisance parameters etc., sometimes requiring ancillary data for their determination.

C. Eisenhart (1962), Realistic evaluation of the precision and accuracy of instrument calibration systems, Journal of Research of the National Bureau of Standards-C, Engineering and Instrumentation 67C(2)
H. Jeffreys (1966), Theory of Probability, Third Ed.
D. van Dyk and L. Lyons (2023), How to incorporate systematic effects into parameter determination.

For Gaussian data regression, $y_i \sim \mathrm{Gauss}(\mu_i,\sigma^2_i)$, the traditional route is to identify additional sources of variance, σ^2_{sys} , and typically add these variances prior to the ML regression,

$$\sigma^2_{new,i} = \sigma^2_i + \sigma^2_{sys} \qquad (12)$$

and then carry on with the chi-squared distributed goodness-of-fit statistic

$$S = \Sigma \ (y_i - \mu_i)^2/ \ \sigma^2_{new,i}$$

For Poisson data, $y_i \sim \mathrm{Poiss}(\mu_i)$, there is no *direct* way to provide additional variance, unlike in the case of Gaussian data. This is an intrinsic limitation of the Poisson regression.

One could alternatively chose other integer-values distributions that result from the compounding of the Poisson with other distributions, such as the negative binomial (e.g., Hilbe 2011) or the Poisson inverse Gaussian. However, retaining the Poisson distribution is generally preferred, especially in astronomy, primarily for its simplicity.

Hilbe, J.M. (2011). *Negative Binomial Regression*, Cambridge Univ. Press,

There are a number of considerations to take into account when considering a method to account for systematic errors

(a) The need for ancillary observations. Often nuisance parameters require additional data for their likelihood, and sometimes this is not possible or undesirable.

(b) Overall complexity of the computations. For example, Bayesian methods may require an integration (usually numerical)over the prior, which may be expensive.

(c) Does the method yield a goodness-of-fit measure? Bayesian methods would tend to use *relative* information criteria (Bayes factors), e.g. AIC (Akaike, 1974), or BIC (Schwartz, 1978). Sometimes an absolute measure is preferred.

I argue that, at present, there is no simple method that yields a goodness-of-fit statistic, and without the need for ancillary data, for the regression of Poisson or count data, similar to the one for normal data.

H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974), pp. 716–723.
G. Schwarz, Estimating the Dimension of a Model, The Annals of Statistics 6 (1978), pp. 461 – 464. Available at https://doi.org/10.1214/aos/1176344136

2. A new model of systematic errors for regression applications

Motivated by this limitation, I have developed a method to account for systematic errors in the Poisson ML regression (preliminary results were in Bonamente 2023).

Data model: $y_i \sim \text{Poiss}(\mu_i)$, $i=1,..,N$ and $\mu_i=f(x_i;\theta)$with m free parameters.

The method is based on *treating the ML estimate of each data point as a random variable M_i*, according to

$$\hat{f}(x_i) \overset{d}{:=} M_i, \text{ with } \mathrm{E}(M_i) = \hat{\mu}_i, \ \mathrm{Var}(M_i) := \sigma^2_{int,i}.$$

This introduces an *intrinsic model variance* $\sigma^2_{i,int}$ associated with the model itself, while retaining the Poisson distribution for the data.

This means that the ML estimate, for example

$$M_i \sim \text{Gauss}(\widetilde{\mu}_i, \sigma^2_{int,i}) \qquad (13)$$

is no longer a fixed number, but a random variable whose mean is the measured $\widetilde{\mu}_i$. Other distributions (e.g., gamma) can/should be used instead.

Bonamente, M. (2023). *Hypothesis testing with the Cash statistic for overdispersed Poisson count data*. MNRAS 522 (2), 1987.

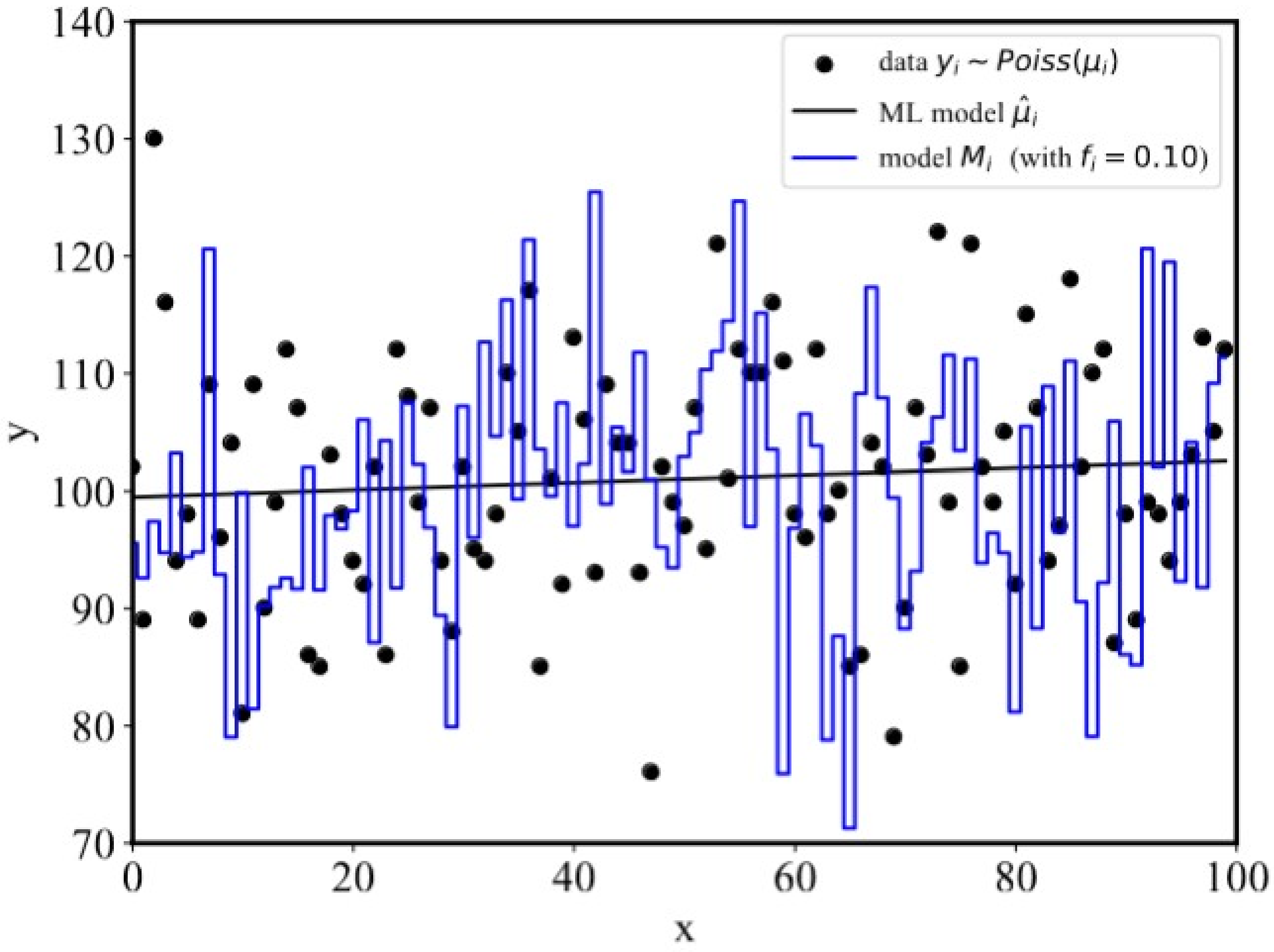


Illustration of the model of systematic errors: The best-fit model (solid curve) is obtained via usual ML methods, and a post-fit randomization according to M_i (blue curve) is used to determine the C_{min} statistic, which is explained in the following.

As a result of this assumption, the goodness-of-fit statistic becomes:

$$C_{\text{min,sys}} = C_{\text{min}}(M_i) := 2\sum_{i=1}^{N}\left(y_i \ln\left(\frac{y_i}{M_i}\right) - (y_i - M_i)\right)$$

(the usual C_{min} statistic has $\tilde{\mu}_i$ in place of M_i).
This statistic can be written as the sum of two separate components:

$C_{\text{min,sys}} = X + Y$ with $X = C_{\text{min}}$ and $Y \sim N(\tilde{\mu}_C,\tilde{\sigma}^2_C)$

where $\tilde{\mu},\tilde{\sigma}^2_C$ can be estimated from the data. Y is defined by:

$$Y := Z - X = C_{\text{min}}(M_i) - C_{\text{min}}(\hat{\mu}_i)$$

It is argued that X and Y are in fact *independent*, under the null hypothesis H_0. In fact, the distribution of $X=C_{\text{min}}$ is independent of model parameterization (per Wilks' theorem); and Y, by construction, is independent of y_i.

Bonamente, M. (2023). *Hypothesis testing with the Cash statistic for overdispersed Poisson count data*. MNRAS 522 (2), 1987.

For small values of the systematic errors, $f_i=\sigma_{int,i}/\mu_i$, the Y variable is

$$Y = 2\sum_{i=1}^{N}(M_i - \hat{\mu}_i) - y_i \ln\left(\frac{M_i}{\hat{\mu}_i}\right)$$

$$\simeq 2\sum_{i=1}^{N}(M_i - \hat{\mu}_i)\left(1 - \frac{y_i}{\hat{\mu}_i}\right) + \sum_{i=1}^{N} y_i \left(\frac{M_i - \hat{\mu}_i}{\hat{\mu}_i}\right)^2$$

The mean and variance of Y can also be easily calculated:

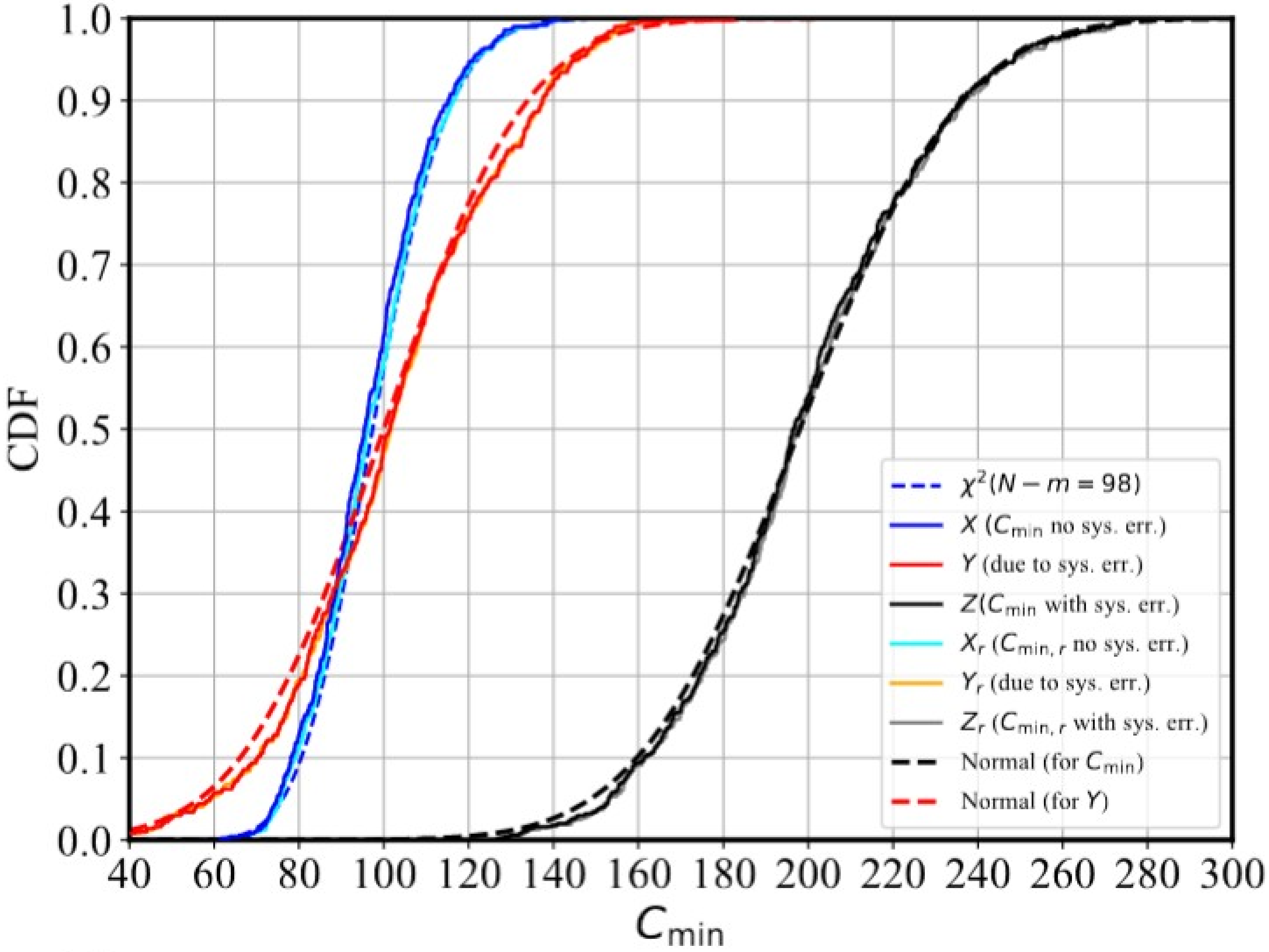
$$\mathrm{E}(Y) = \sum_{i=1}^{N} \mu_i \left(\frac{\sigma_{i,int}}{\hat{\mu}_i}\right)^2 \simeq \sum_{i=1}^{N} y_i \left(\frac{\sigma_{i,int}}{\hat{\mu}_i}\right)^2$$

$$\mathrm{Var}(Y) = 4\sum_{i=1}^{N} \mu_i f_i^2 + \sum_{i=1}^{N} (\mu_i^2 + \mu_i) f_i^4 \cdot \mathrm{kurt}(M_i) - \sum_{i=1}^{N} \mu_i^2 f_i^4$$

These results have been tested with several numerical simulations to be presented in Bonamente+24.

Bonamente, M. et al. (2024) *Maximum-likelihood Poisson regression with systematic errors: Methodology and applications to the goodness-of-fit statistic*, to be submitted.

Simulation with f_i=0.1, N=100 data points; the CDF of the Y component has E[Y]=100, which means that the goodness-of-fit statistic has E[$C_{min,sys}$]= E[C_{min}]+E[Y]≈ 200.



The distribution of Z is therefore the convolution of a normal and a chi-squared distribution, in the asymptotic large-mean limit, which is referred to as an *overdispersed chi-squared* distribution,

$$C_{\mathrm{min,sys}} \sim B(\nu, \hat{\mu}_C, \hat{\sigma}_C^2)$$

In the extensive data limit of large N, it is the convolution of two normal distribution.

As an aside: In Bonamente and Zimmerman (2024) we report an analytical form for the convolution of a gamma and normal distribution, which generalizes the convolution problem. This leads the the *gamma-normal* distribution

$$f_{\mathrm{GN}}(z; \alpha, r, \mu, \sigma^2) = \frac{(\alpha\sigma)^r}{\sqrt{2\pi\sigma^2}} D_{-r}(\zeta) \cdot E(z),$$

where $D_{-p}(x)$ is a parabolic cylinder function, and E(x) is an exponential function of the parameters. This generalizes earlier studies on the exponential-normal (e.g., Grushka 1972) that have been used expensively in biology.

Bonamente, M. and Zimmerman, D. (2024). , *The univariate normal-gamma and related probability distributions*, *Submitted to METRON*.
Gruskha, E. (1972). *Characterization of exponentially modified Gaussian peaks in chromatography*. Analytical chemistry 44, 1733–1738

3. Practical uses of this model for systematic errors

3.1 To estimate systematic errors from the data

In this case, one assumes that the model *is* correct, and E[Z]=(N-m)+ μ_C , thus leading to an estimate

$$\hat{\mu}_C = C_{\min} - (N - m) > 0$$

From this, it is immediate to estimate the f parameter:

$$\widehat{f} = \sqrt{\frac{C_{\min} - (N - m)}{\sum_{i=1}^{N} y_i}}.$$

(confidence intervals on $\widetilde{f}$ can also be easily obtained. For the Spence+23 spectra, this method gives reasonable results. Given the parameters

N=1526, m=48 (it was a spline model) and a measured C_{min}=1862.7,

the method estimated f=0.018±0.02 which is the expected level of systematic errors for the XMM data.

3.2 To do hypothesis testing

The natural use of the method is to do hypothesis testing, assuming an a priori estimate for $f_i=\sigma_{int,i}/\mu_i$. In this case, f_i leads to an estimate of $E[Y]=\tilde{\mu}_C$ (technically it is not a 'hat' quantity).

The goodness-of-fit statistic, under H_0 and in the extensive data limit, N → ∞, is a normally distributed

$$C_{\mathrm{min,sys}} \sim N(N-m+\hat{\mu}_C, 2(N-m)+\hat{\sigma}^2_C)$$

and usual hypothesis testing follows immediately, as usual.

Note: In the low-mean regime, it may be possible to use the latest results by Li+24, which guarantees asymptotic normality of the Poisson goodness-of fit. The E[X] and Var(X)would have to be modified accordingly.

Li, X., Chen, Y. Meng, X., Kashyap, V. and Bonamente, M. (2024), *Comparison of Goodness-of-fit Assessment Methods with C statistics in Astronomy*, to be submitted