Structural Model Validation in Measurement Error Models

Naomi Giertych

Joint work with: Jonathan Williams and Sujit Ghosh

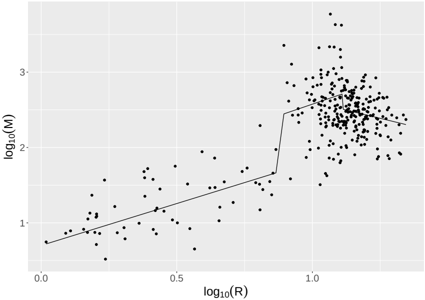
08/07/23

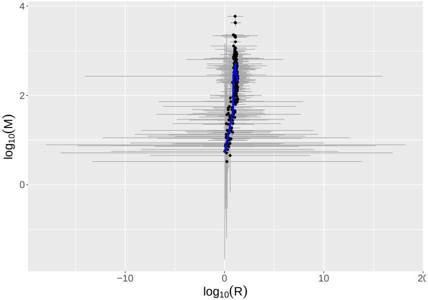
Motivation

- Astronomers interested in answering questions about exoplanet formation and structure
- Build models where mass and radius are used as inputs
- For instance, Dorn et al. (2018) characterize the interiors of the planets in the TRAPPIST-1 system using mass, radius and stellar irradiation
- Challenge: mass measurements are significantly more difficult to obtain (Ma & Ghosh, 2019; Weiss & Marcy, 2014; Wolfgang et al., 2016)
- Solution: predict mass from observed radius
- But....

Challenges in Modeling the M-R Relationship

- Piecewise-linear on log-base-10-scale
- Measurement errors in mass and radius with
  - unknown, heteroskedastic variance
  - unknown distribution





Model and Statistical Problem

Measurement error model with intrinsic scatter:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + v_i \\ d_i &= y_i + w_i \\ z_i &= x_i + q_i \end{aligned} \tag{1}$$

where

$$\begin{array}{lll} E(v_i) = 0 & E(w_i) = 0 & E(q_i) = 0 \\ V(v_i) = \sigma^2 & V(w_i) = \sigma^2_{w_i} & V(q_i) = \sigma^2_{q_i} \\ & l_{w_i} < \sigma^2_{w_i} < u_{w_i} & l_{q_i} < \sigma^2_{q_i} < u_{q_i} \end{array}$$

$(v_1, w_1, q_1), \ldots, (v_n, w_n, q_n)$ mutually independent from some elliptical distribution and $\{l_{w_i}, u_{w_i}, l_{q_i}, u_{q_i}\}_{i=1}^n$ are known.

Goal: Determine fit of model (1)

Proposed Method

- Use new tools from conformal prediction to construct a measure of discrepancy between data and assumed model which
  - accounts for heteroscedastic measurement errors
  - is robust to misspecification of error distribution
- Calculate the measure of discrepancy on two subsets of the data
- Perform a formal goodness-of-fit test using the Anderson-Darling two-sample test

Conformal Prediction

- Transformation that measures how different one observation is from the others (Shafer & Vovk, 2008)

- Can be used to construct prediction intervals with exact Type I error coverage if non-conformity scores i.i.d (Shafer & Vovk, 2008)

- Re-write equation (2) as:

$$d_i = \beta_0 + \beta_1 z_i + e_i$$

where $e_i = v_i + w_i - \beta_1 q_i$, $E(e_i) = 0$, $V(e_i) = \eta_i = (\sigma^2 + \sigma_{w_i}^2 + \beta_1^2 \sigma_{q_i}^2)$

- Our non-conformity scores: $(d_i - \beta_0 - \beta_1 z_i)^2 / \eta_i$

Proposed Method

1. **Parameter estimation**
2. Goodness-of-fit test

Parameter Estimation: FGLS for ME Models

- Initialize variance estimates at mid-point of bounds
- Use any parameter estimation method that accounts for known measurement error variances
- Using parameter estimate, estimate variances
- Iterate until convergence

Moment-Corrected Estimator

Buonaccorsi (2010) defines moment-corrected estimators:

$$\hat{\beta}_0 = \bar{d} - \hat{\beta}_1 \bar{z}$$

$$\hat{\beta}_1 = \frac{S_{ZD}}{S_{ZZ} - \sigma_q^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (d_i - \hat{\beta}_0 - \hat{\beta}_1 z_i)^2 - \sigma_w^2 - \hat{\beta}_1^2 \sigma_q^2$$

where

$$\sigma_w^2 = \sum_{i=1}^{n} \sigma_{w_i}^2 / n \qquad \sigma_q^2 = \sum_{i=1}^{n} \sigma_{q_i}^2 / n \qquad \bar{d} = \sum_{i=1}^{n} d_i / n \qquad \bar{z} = \sum_{i=1}^{n} z_i / n$$

$$S_{ZD} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(d_i - \bar{d})}{n-1} \qquad S_{ZZ} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})^2}{n-1}$$

Least Squares Variance Component Estimation

Suppose

$$E(y) = X\beta \qquad V(y) = \sum_{k=1}^{n} s_k Q_k$$

Let

- B be a basis matrix for $\text{null}(X')$
- $y_{vh} = \text{vech}(B' y y' B)$
- $X_{vh} = \begin{bmatrix} \text{vech}(B' Q_1 B) & \cdots & \text{vech}(B' Q_n B) \end{bmatrix}$

Then $\hat{s} = (X_{vh}' X_{vh})^{-1} X_{vh}' y_{vh}$ and $E(\hat{s}) = s$ (Teunissen & Amiri-Simkooei, 2008)

LS-VCE with QP

Re-write equation (1) as

$$d_i = \beta_0 + \beta_1 z_i + e_i \tag{2}$$

where $e_i = v_i + w_i - \beta_1 q_i$, $E(e_i) = 0$, $V(e_i) = \eta_i = (\sigma^2 + \sigma^2_{w_i} + \beta_1^2 \sigma^2_{q_i})$

Obtain upper and lower bounds for η_i using $\{l_{w_i}, u_{w_i}, l_{q_i}, u_{q_i}\}$ and

$$l_{\sigma^2} = \sum_{i=1}^{n} [(d_i - \beta_0 - \beta_1 z_i)^2 - u_{w_i} - \beta_1^2 u_{q_i}]/n$$
$$u_{\sigma^2} = \sum_{i=1}^{n} [(d_i - \beta_0 - \beta_1 z_i)^2 - l_{w_i} - \beta_1^2 l_{q_i}]/n$$

Proposed Method

1. Parameter estimation

2. **Goodness-of-fit test**

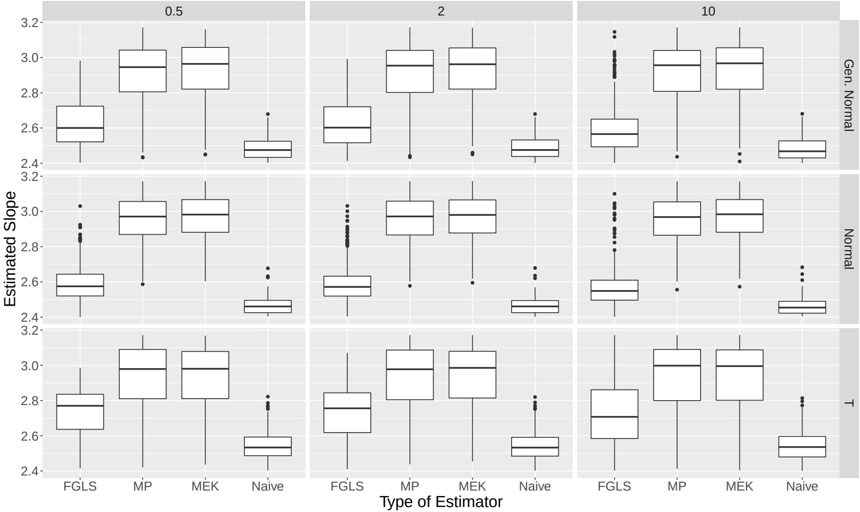
Anderson-Darling Two-Sample Goodness-of-Fit test

- Non-parametric test for equality of distributions, i.e $H_0 : F_X = F_Y$
- Less sensitive to small sample sizes unlike Kolmogorov-Smirnov Test
- Goodness-of-fit test:
  - Split data into train, calibration, and test sets
  - Estimate parameters using the training data
  - Calculate non-conformity scores for calibration and test sets
  - Use non-conformity scores in Anderson-Darling Two-Sample test
  - If the two sets of non-conformity scores have the same distribution, then we fail to reject the null hypothesis and conclude that the model is correctly specified.

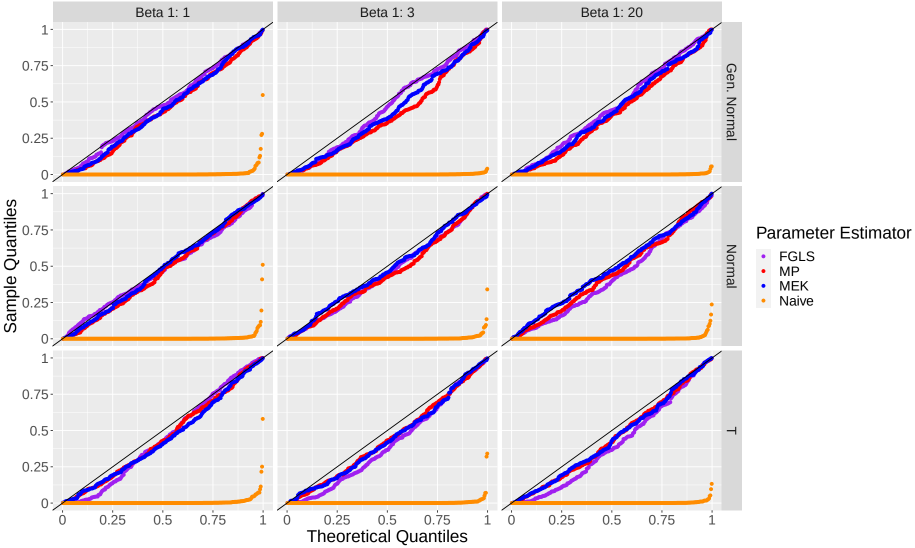
Simulations Setup

- Training: $x_i \overset{i.i.d}{\sim} Unif(0, 30)$, 200 observations
- Calibration: $x_i \overset{i.i.d}{\sim} Unif(0, 30)$, 50 observations
- Test: $x_i \overset{i.i.d}{\sim} Unif(30, 40)$, 50 observations
- $\beta_0 = 1$, $\beta_1 \in \{1, 3, 20\}$
- $\sigma^2 \in \{0.5, 2, 10\}$
- $l_{w_i} = l_{q_i} = 0.06x_i^2$
- $u_{w_i} = u_{q_i} = 0.07x_i^2$
- $\sigma^2_{w_i}, \sigma^2_{q_i} = 0.3l_{w_i} + 0.7u_{w_i}$

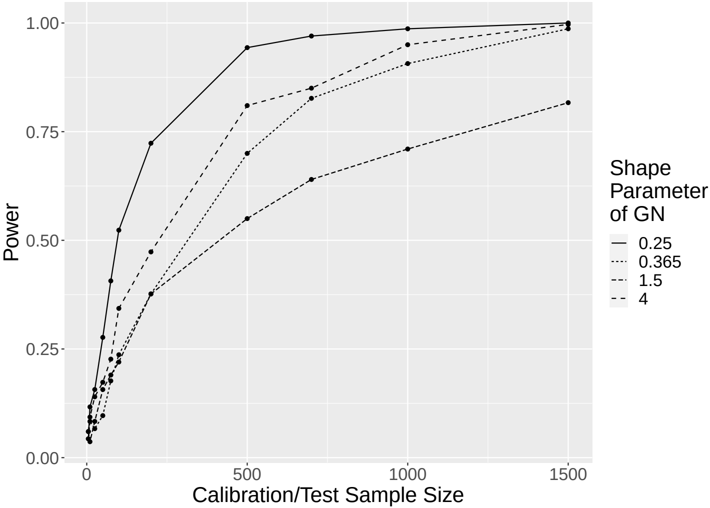
Simulation: Slope Estimates



QQ Plots of Observed AD P-Values vs Theoretical U(0,1)



Power


1.00
0.75
0.50
0.25
0.00
Power
0
500
1000
1500
Calibration/Test Sample Size
Shape Parameter of GN
0.25
0.365
1.5
4

Application to Mass-Radius Relationship

- Followed Weiss & Marcy (2014) and Wolfgang et al. (2016) and considered planets with $1.5R_{\oplus} < R^{obs} < 4R_{\oplus}$ and high SNR ($R^{obs} > 3\sigma_R^{obs}$, $M^{obs} > 3\sigma_M^{obs}$)
- 320 observations total
- Fit relationship on log-base-10 scale

$$\begin{aligned}
\widetilde{M}_i &= \widetilde{C} + \frac{\gamma}{\ln(10)}\widetilde{R}_i + \frac{v_i}{\ln(10)} \\
\widetilde{M}_i^{\text{obs}} &= \widetilde{M}_i + \frac{w_i}{\ln(10)} \\
\widetilde{R}_i^{\text{obs}} &= \widetilde{R}_i + \frac{q_i}{\ln(10)}
\end{aligned} \tag{3}$$

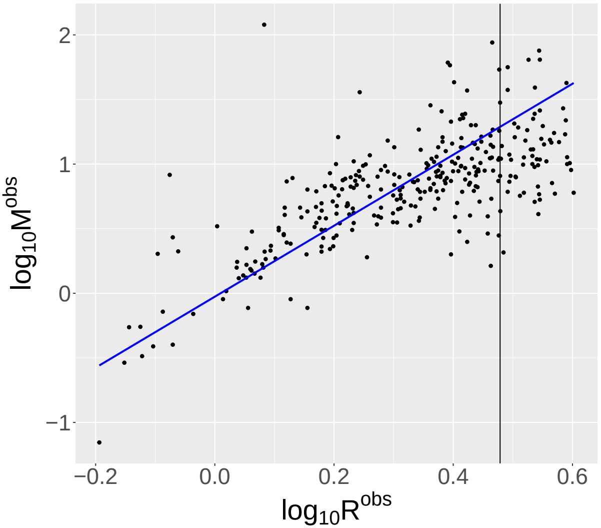
- Training and calibration sets contained observations with $R^{obs} \leq 3R_{\oplus}$
- Test set contained remaining observations

Application to Mass-Radius Relationship

	$\widetilde{C}$	$\gamma / ln(10)$	AD p-value
Our Analysis (Midpoint)	-0.026	2.747	< 0.001
Weiss & Marcy (2014)*	0.430	0.404	-
Wolfgang et al. (2016)*	0.431	0.565	-

Table 1: $\cdot^*$ Did not consider multiplicative measurement errors

Application to Mass-Radius Relationship



Thank you!

https://ngierty.github.io/

References

Buonaccorsi, J. P. 2010, Measurement error: models, methods, and applications (Chapman and Hall/CRC)

Dorn, C., Mosegaard, K., Grimm, S. L., & Alibert, Y. 2018, The Astrophysical Journal, 865, 20

Ma, Q., & Ghosh, S. K. 2019, The Astronomical Journal, 158, 86

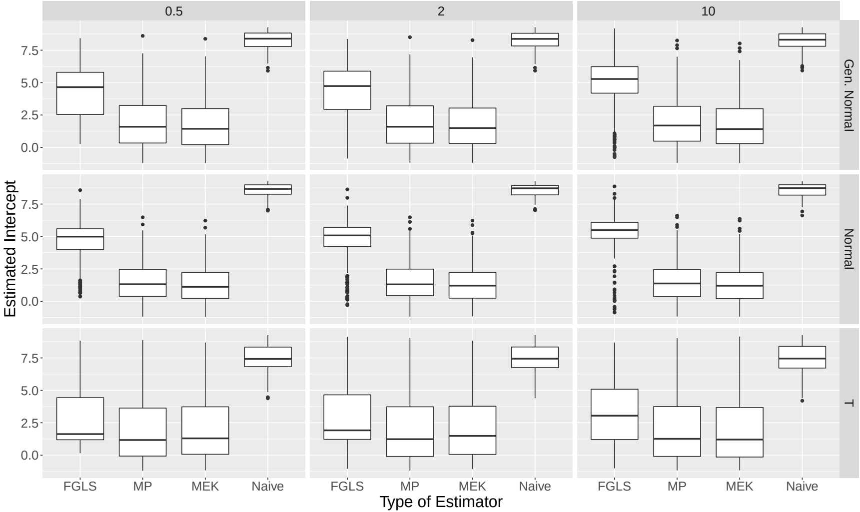
Shafer, G., & Vovk, V. 2008, Journal of Machine Learning Research, 9

Teunissen, P. J., & Amiri-Simkooei, A. 2008, Journal of geodesy, 82, 65

Weiss, L. M., & Marcy, G. W. 2014, The Astrophysical Journal Letters, 783, L6

Wolfgang, A., Rogers, L. A., & Ford, E. B. 2016, The Astrophysical Journal, 825, 19

Simulation: Intercept Estimates



Simulation: Prediction Interval Coverage