Hypothesis statist test for simulated light curve data

Salar Ghaderi

The Python codes used in this report can be found at the following link: GitHub Repository.

Introduction

In the measurement(simulated data) we have a set of eight time-series light curves. Each light curve consists of 1024 data points and is accompanied by a set of inverse variance measurements (i.e., the inverses of the squares of the uncertainties).

1. A list of 1024 time points, t_i.
2. A matrix of light curve values, $y_{j,i}$ (with $j = 0, 1, \ldots, 7$ for the 8 light curves).
3. A set of 1024 inverse variances, ivars_i.

the main question is to answer *Does each light curve contain a harmonic signal at Earth's sidereal period?*

Hypotheses

We test the following hypotheses for each light curve:

- **Null Hypothesis (H_0):** The light curve is constant in time, i.e.,

  $$y(t) = \mu + \epsilon(t),$$

  where μ is the constant mean brightness and $\epsilon(t)$ is noise that is gaussian.

- **Alternative Hypothesis (H_a):** The light curve contains an additional harmonic component at the sidereal frequency, i.e.,

  $$y(t) = \mu + A\cos(\omega t + \phi) + \epsilon(t),$$

  where A is the amplitude, ω is the angular frequency corresponding to Earth's sidereal period, and ϕ is a phase offset.

1 (Part 1) Test Statistic and Mathematical Formulation

To test H_0, we fit the constant model using the weighted mean:

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i \, \mathrm{ivars}_i}{\sum_{i=1}^{N} \mathrm{ivars}_i}, \quad N = 1024.$$

The chi-squared statistic is then defined as:

$$\chi^2 = \sum_{i=1}^{N} \mathrm{ivars}_i \, (y_i - \bar{y})^2,$$

which, under the null hypothesis, is expected to follow a chi-squared distribution with $\nu = N - 1 = 1023$ degrees of freedom. The p-value is computed as:

$$p = P\left(\chi^2_\nu > \chi^2\right),$$

which quantifies the probability that a chi-squared variable with 1023 degrees of freedom exceeds the observed value. A very low p-value would indicate that the data are unlikely under the null hypothesis, suggesting the presence of a harmonic signal (i.e., supporting H_a).

Data Generation and the Role of the Wrongness Factor(how wrong error estimation in measurements can trick the statistics test

The synthetic light curves are generated by the function:

```
def make_time_series(times, ivars, amps=[0.], wrongness_factor=1.0):
    ys = np.ones_like(times) + rng.normal(size=times.shape) /
         np.sqrt(wrongness_factor * ivars)
    for i, amp in enumerate(amps):
        theta = rng.uniform(0., 2. * np.pi)
        ys += amp * np.cos((i + 1.) * SIDEREAL_OMEGA * times + theta)
    return ys
```

The noise term is scaled as:

$$\sigma = \frac{1}{\sqrt{\text{wrongness\_factor} \times \text{ivars}}},$$

so that when $\text{wrongness\_factor} < 1$, the actual noise level is higher than what would be expected if we assumed $\sigma = 1/\sqrt{\text{ivars}}$. For example, with $\text{wrongness\_factor} = 0.8$, the effective noise becomes

$$\sigma = \frac{1}{\sqrt{0.8}} \cdot \frac{1}{\sqrt{\text{ivars}}} \approx 1.118 \cdot \frac{1}{\sqrt{\text{ivars}}}.$$

This underestimation of the uncertainties leads to an *inflated* chi-squared statistic because the calculation

$$\chi^2 = \sum_{i=1}^{N} \text{ivars}_i \left(y_i - \bar{y}\right)^2$$

assumes the variances are $1/\text{ivars}_i$, not $1/(\text{wrongness\_factor} \times \text{ivars}_i)$. Consequently, even if H_0 is true, the inflated χ^2 may produce a falsely low p-value.

Results

After running the code, we obtained the following results for the eight light curves:

- **Light Curve 0:** $\bar{y} \approx 1.01104$, $\chi^2 \approx 1018.58$ (dof $= 1023$), $p \approx 5.33 \times 10^{-1}$.
  Here, H_0 is not rejected.
- **Light Curve 1:** $\bar{y} \approx 1.02065$, $\chi^2 \approx 1063.36$ (dof $= 1023$), $p \approx 1.85 \times 10^{-1}$.
  H_0 is not rejected.
- **Light Curve 2:** $\bar{y} \approx 0.98941$, $\chi^2 \approx 1597.34$ (dof $= 1023$), $p \approx 5.80 \times 10^{-28}$.
  H_0 is rejected; this light curve shows strong evidence for a harmonic signal.
- **Light Curve 3:** $\bar{y} \approx 0.98902$, $\chi^2 \approx 7313.72$ (dof $= 1023$), $p \approx 0.0$.
  H_0 is strongly rejected.

- **Light Curve 4:** $\bar{y} \approx 0.99500$, $\chi^2 \approx 1232.10$ (dof $= 1023$), $p \approx 6.65 \times 10^{-6}$.
Although H_0 is rejected by the test, this light curve was generated with no harmonic component (i.e., $A = 0$) but with a wrongness_factor $= 0.8$. The underestimated error inflates the chi-squared statistic, leading to a falsely low p-value.
- **Light Curve 5:** $\bar{y} \approx 1.00383$, $\chi^2 \approx 1321.88$ (dof $= 1023$), $p \approx 6.40 \times 10^{-10}$.
H_0 is rejected.
- **Light Curve 6:** $\bar{y} \approx 1.00569$, $\chi^2 \approx 1341.74$ (dof $= 1023$), $p \approx 6.02 \times 10^{-11}$.
H_0 is rejected.
- **Light Curve 7:** $\bar{y} \approx 0.99135$, $\chi^2 \approx 1178.53$ (dof $= 1023$), $p \approx 4.96 \times 10^{-4}$.
H_0 is rejected.

Discussion

The results indicate that:

- **Light Curves 0 and 1:** Their relatively high p-values (0.533 and 0.185, respectively) imply that there is no significant deviation from the constant model. Therefore, these light curves are consistent with the null hypothesis H_0 (i.e., no harmonic signal).
- **Light Curves 2, 3, 5, 6, and 7:** These light curves have extremely low p-values, which means that the constant model is a poor fit. This provides strong evidence in favor of the alternative hypothesis H_a (i.e., a harmonic signal is present).
- **Light Curve 4:** Although the test rejects H_0 (with a p-value of $\sim 6.65 \times 10^{-6}$), this is a false positive. The light curve was generated with no harmonic component ($A = 0$), but the use of a wrongness_factor $= 0.8$ led to an underestimation of the errors. This underestimation inflates the chi-squared statistic and results in a misleadingly low p-value.

This example illustrates the importance of accurate error estimation when performing statistical tests.

2 (part2)Comparison of Constant and Constant-plus-Harmonic Models

As we disscussed above the the second model H_a states the light curve contains an additional harmonic component at the sidereal frequency:

$$y(t) = \mu + A\cos(\omega t + \phi) + \epsilon(t),$$

where A is the amplitude, ω is the angular frequency corresponding to Earth's sidereal period, and ϕ is a phase offset.

Reparameterization of the Harmonic Component

To facilitate linear least-squares fitting, we rewrite the harmonic term using the identity:

$$A\cos(\omega t + \phi) = a\cos(\omega t) + b\sin(\omega t),$$

where the parameters are related via

$$a = A\cos\phi, \quad b = -A\sin\phi, \quad \text{and} \quad A = \sqrt{a^2 + b^2}.$$

Thus, the alternative model becomes:

$$y(t) = \mu + a\cos(\omega t) + b\sin(\omega t),$$

which is linear in the parameters μ, a, and b.

Fitting the Models

Null Model (H_0): The only parameter is μ, which is estimated by the weighted mean:

$$\mu = \frac{\sum_i w_i \, y_i}{\sum_i w_i},$$

where w_i are the inverse variances. The corresponding chi-square statistic is:

$$\chi_0^2 = \sum_i w_i \, (y_i - \mu)^2.$$

Alternative Model (H_a): We construct a design matrix for N observations:

$$X = \begin{pmatrix} 1 & \cos(\omega t_1) & \sin(\omega t_1) \\ 1 & \cos(\omega t_2) & \sin(\omega t_2) \\ \vdots & \vdots & \vdots \\ 1 & \cos(\omega t_N) & \sin(\omega t_N) \end{pmatrix}.$$

Using weighted linear least squares (with weights $\sqrt{w_i}$), we solve for

$$p = \begin{pmatrix} \mu \\ a \\ b \end{pmatrix}$$

by minimizing

$$\chi_h^2 = \sum_{i=1}^{N} w_i \left[y_i - (\mu + a\cos(\omega t_i) + b\sin(\omega t_i)) \right]^2.$$

After obtaining a and b, the best-fit amplitude is computed as:

$$A = \sqrt{a^2 + b^2}.$$

Likelihood Ratio Test

For Gaussian errors, the log-likelihood is (up to an additive constant) given by:

$$\log \mathcal{L} \propto -\frac{1}{2}\chi^2.$$

Thus, the improvement in log-likelihood when moving from H_0 to H_a is:

$$\Delta \log \mathcal{L} = \frac{1}{2}\left(\chi_0^2 - \chi_h^2\right).$$

Defining the likelihood ratio statistic as:

$$\Lambda = \chi_0^2 - \chi_h^2 = 2\,\Delta \log \mathcal{L},$$

and noting that H_a introduces two additional parameters (a and b), under H_0 the statistic Λ is asymptotically chi-square distributed with 2 degrees of freedom. A large Λ (or equivalently, a large $\Delta \log \mathcal{L}$) indicates that adding the harmonic component significantly improves the fit.

Implementation in Code

- **Null Model Fit:** The code computes the weighted mean μ and then evaluates
$$\chi_0^2 = \sum_i w_i \, (y_i - \mu)^2.$$
- **Alternative Model Fit:** The design matrix X is constructed with columns corresponding to the constant term, $\cos(\omega t)$, and $\sin(\omega t)$. The weighted least-squares solution yields the parameters $p = (\mu, a, b)$, and the chi-square is computed as:
$$\chi_h^2 = \sum_i w_i \left[y_i - (\mu + a\cos(\omega t_i) + b\sin(\omega t_i))\right]^2.$$
- **Comparison:** The improvement is quantified via
$$\Delta \log \mathcal{L} = \frac{1}{2}\left(\chi_0^2 - \chi_h^2\right),$$
and the best-fit amplitude is recovered as $A = \sqrt{a^2 + b^2}$.

Results and Interpretation

The following results were obtained for the eight light curves:

- **Light curve 0:** $\chi_0^2 = 1018.58$, $\chi_h^2 = 1015.66$, $\Delta \log \mathcal{L} = 1.46$, $A \approx 0.02145$.
- **Light curve 1:** $\chi_0^2 = 1063.36$, $\chi_h^2 = 1000.16$, $\Delta \log \mathcal{L} = 31.60$, $A \approx 0.09887$.
- **Light curve 2:** $\chi_0^2 = 1597.34$, $\chi_h^2 = 1045.72$, $\Delta \log \mathcal{L} = 275.81$, $A \approx 0.29595$.
- **Light curve 3:** $\chi_0^2 = 7313.72$, $\chi_h^2 = 973.69$, $\Delta \log \mathcal{L} = 3170.01$, $A \approx 0.98991$.
- **Light curve 4:** $\chi_0^2 = 1232.10$, $\chi_h^2 = 1229.98$, $\Delta \log \mathcal{L} = 1.06$, $A \approx 0.01833$.
- **Light curve 5:** $\chi_0^2 = 1321.88$, $\chi_h^2 = 1276.77$, $\Delta \log \mathcal{L} = 22.55$, $A \approx 0.08485$.
- **Light curve 6:** $\chi_0^2 = 1341.74$, $\chi_h^2 = 1276.04$, $\Delta \log \mathcal{L} = 32.85$, $A \approx 0.10255$.
- **Light curve 7:** $\chi_0^2 = 1178.53$, $\chi_h^2 = 1125.77$, $\Delta \log \mathcal{L} = 26.38$, $A \approx 0.09093$.

Interpretation:

- For light curves with significant harmonic signals (e.g., light curves 1, 2, 3, 5, 6, and 7), the addition of the harmonic component leads to a large reduction in χ^2 (i.e., a large $\Delta \log \mathcal{L}$) and a nonzero best-fit amplitude.
- Light curves 0 and 4 exhibit only minor improvements ($\Delta \log \mathcal{L} \approx 1$), indicating that the harmonic term does not significantly improve the fit. In particular, light curve 4 is correctly identified as not containing a significant harmonic signal. This is important because a p-test based solely on the null model's χ^2 was misleading for light curve 4 (due to underestimated uncertainties), whereas the likelihood ratio test—by comparing the two models—reveals the absence of a true harmonic signal.

Conclusion

The likelihood ratio test statistic,
$$\Lambda = \chi_0^2 - \chi_h^2,$$
(or equivalently, $\Delta \log \mathcal{L} = \frac{1}{2}\Lambda$), provides a robust measure for determining whether adding a harmonic term significantly improves the model fit. Unlike the p-test based on the null model alone, this method correctly predicts that light curve 4 does not contain a significant harmonic signal, even when noise is misestimated. Thus, the likelihood ratio test is more reliable for comparing H_0 and H_a in this context.So as in lecture we learned it is more sensitive.

3 (part3) Determining the Number of Harmonic Signals Using the BIC

how many harmonic signals are present in each light curve? We assume that the signals occur at integer multiples of the Earth's sidereal frequency. For each light curve, i have:

$$y(t) = \mu + \sum_{j=1}^{k} \Big[a_j \cos(j\omega t) + b_j \sin(j\omega t) \Big] + \epsilon(t),$$

where:

- μ is the constant mean brightness,
- $\omega = \dfrac{2\pi}{T_{\text{sid}}}$ is the angular frequency corresponding to Earth's sidereal period T_{sid},
- a_j and b_j are the coefficients for the cosine and sine terms of the j^{th} harmonic,
- k is the number of harmonics included, and
- $\epsilon(t)$ is a Gaussian noise term.

Counting the Free Parameters

The total number of free parameters in the model is given by:

$$p = 1 + 2k,$$

since the constant term μ contributes one parameter and each harmonic contributes two parameters (a_j and b_j).

Fitting Procedure: Weighted Least Squares

Given a dataset $\{(t_i, y_i)\}_{i=1}^{N}$ with inverse variances ivars_i, we construct the design matrix X for a model with k harmonics:

$$X = \begin{bmatrix} 1 & \cos(\omega t_1) & \sin(\omega t_1) & \cdots & \cos(k\omega t_1) & \sin(k\omega t_1) \\ 1 & \cos(\omega t_2) & \sin(\omega t_2) & \cdots & \cos(k\omega t_2) & \sin(k\omega t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos(\omega t_N) & \sin(\omega t_N) & \cdots & \cos(k\omega t_N) & \sin(k\omega t_N) \end{bmatrix}.$$

Each row corresponds to one observation at time t_i. The fitting is performed using weighted least squares, where the weight for each observation is $\sqrt{\text{ivars}_i}$. We then minimize the weighted chi-squared:

$$\chi^2 = \sum_{i=1}^{N} \text{ivars}_i \Big[y_i - \hat{y}(t_i) \Big]^2,$$

with $\hat{y}(t_i)$ being the model prediction at time t_i.

The Bayesian Information Criterion (BIC)

The BIC is used for model selection by balancing the quality of the fit against the complexity of the model. It is defined as:

$$\text{BIC} = \chi^2 + p \ln(N),$$

where:

- χ^2 is the chi-squared statistic from the fit,
- $p = 1 + 2k$ is the number of free parameters in the model, and
- N is the number of data points.

A lower BIC value indicates a more favored model.

Procedure and Results

For each light curve, we fit models with $k = 0, 1, 2, 3$ harmonics:

- **For** $k = 0$**:** The model is $y(t) = \mu + \epsilon(t)$ (i.e., a constant model). Here, $p = 1$.
- **For** $k = 1$**:** The model is
  $$y(t) = \mu + a_1 \cos(\omega t) + b_1 \sin(\omega t) + \epsilon(t),$$
  with $p = 3$.
- **For** $k = 2$**:** The model includes two harmonics, so
  $$p = 1 + 2 \cdot 2 = 5.$$
- **For** $k = 3$**:** Three harmonics yield $p = 1 + 2 \cdot 3 = 7$.

For each model, we compute χ^2 via the weighted least squares fit and then calculate:

$$\text{BIC} = \chi^2 + p \ln(N).$$

The model with the lowest BIC is chosen as the best representation for that light curve.

Summary of the Results

The computed BIC values for the light curves are as follows:

Light Curve	Preferred Model Order (k)	BIC Value
0	0	1025.52
1	1	1020.95
2	1	1066.51
3	1	994.49
4	0	1239.03
5	1	1297.56
6	3	1127.84
7	3	1044.06

Discussion

The model selection results are in agreement with the code used to generate the data:

- **Light Curves 0 and 4:** The best model is the constant model ($k = 0$). This indicates that these light curves do not contain a significant harmonic signal, which is consistent with the data generation where no harmonic was injected.
- **Light Curves 1, 2, 3, and 5:** The optimal model has one harmonic ($k = 1$), implying that a single periodic signal at the sidereal frequency is present. This is exactly as expected from the generating process.
- **Light Curves 6 and 7:** The BIC prefers a model with three harmonics ($k = 3$). These light curves were generated with multiple harmonic components, and the BIC correctly identifies the need for additional harmonics to explain the observed variations.

These findings are in accordance with the code that generated the data:

1. Light curves without an injected harmonic signal (0 and 4) are best modeled as constants.
2. Light curves with a single periodic component (1, 2, 3, and 5) require one harmonic term.
3. Light curves with more complex periodic behavior (6 and 7) require three harmonic terms.

4 (part4) Introduction

In *Data Analysis Recipes: Fitting a Model to Data* [1], we focous on **Exercise 8**, in which we compare the standard (matrix-based) weighted least-squares estimates of slope uncertainty with two resampling-based methods: *jackknife* and *bootstrap.*

In particular, we examine two data sets:

- **Exercise 2 Data** (a full or larger set of points), which includes some high-uncertainty or potentially outlier-like measurements.
- **Exercise 1 Data** (a subset presumably more homogeneous).

The key question is whether the standard formula for the slope uncertainty in a weighted linear fit is realistic, or if the resampling methods produce significantly different (often larger) uncertainties, indicating possible mismatch between model assumptions and real data.

Weighted Linear Least-Squares Fit

Model and Objective Function

We assume a linear model for our data $\{x_i,\, y_i\}$:

$$y_i \;=\; b \;+\; m\, x_i \;+\; \epsilon_i, \tag{1}$$

where m is the slope, b is the intercept, and ϵ_i is noise (which we usually assume is Gaussian with known variance $\sigma_{y_i}^2$).

Following the standard *weighted* χ^2 minimization approach, we define

$$\chi^2 \;=\; \sum_{i=1}^{N} \frac{\bigl(y_i \;-\; b \;-\; m\, x_i\bigr)^2}{\sigma_{y_i}^2}. \tag{2}$$

Minimizing χ^2 with respect to m and b leads to a linear system. In matrix form, let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}, \quad C = \mathrm{diag}\bigl(\sigma_{y_1}^2,\; \ldots,\; \sigma_{y_N}^2\bigr). \tag{3}$$

Then Eq. (2) is

$$\chi^2 \;=\; (\,Y - A\,X\,)^{\mathsf{T}}\, C^{-1}\, (\,Y - A\,X\,), \tag{4}$$

where $X = (\,b,\; m\,)^{\mathsf{T}}$ is the parameter vector. The solution is given by

$$X_{\text{best}} \;=\; (\,A^{\mathsf{T}}\, C^{-1}\, A\,)^{-1}\, \bigl(A^{\mathsf{T}}\, C^{-1}\, Y\bigr). \tag{5}$$

The covariance matrix on (b, m) is then

$$\mathrm{Cov}(b, m) \;=\; \bigl(\,A^{\mathsf{T}}\, C^{-1}\, A\,\bigr)^{-1}, \tag{6}$$

and the diagonal elements provide $\mathrm{Var}(b)$ and $\mathrm{Var}(m)$.

Resampling-Based Uncertainties

While Eq. (5) provides a neat formula for the slope uncertainty, it relies on assumptions that the data points truly follow the linear model with correct Gaussian noise variances. In practice, we often suspect outliers, misestimates of σ_{y_i}, or systematic deviations from linearity. Two *resampling* techniques—**jackknife** and **bootstrap**—provide empirical estimates of how sensitive the fitted slope is to the data set.

4.1 Jackknife

The **jackknife** approach leaves out one data point at a time:

- For each $i \in [1, N]$, exclude data point i and fit the slope m_i.
- Collect $\{\, m_i\}$; the sample variance of these m_i values is scaled to give

$$\sigma_m^2 \;=\; \frac{N-1}{N} \sum_{i=1}^{N} \bigl(m_i - \overline{m}\bigr)^2, \tag{7}$$

  where $\overline{m} = \frac{1}{N} \sum_i m_i$ is the average of the leave-one-out slopes.

Large scatter in $\{m_i\}$ implies that certain points heavily influence the slope, leading to bigger σ_m than the simple matrix formula might indicate.

4.2 Bootstrap

The **bootstrap** draws N points with replacement from the original data set, many times (e.g. 2000 trials). For each bootstrap resampling b, we perform the weighted fit and store the slope m_b. Then the variance of the $\{m_b\}$ distribution provides

$$\sigma_m^2 \;=\; \mathrm{Var}\bigl(\{m_b\}\bigr). \tag{8}$$

If outliers or questionable points drastically shift the slope when they appear (or vanish) multiple times in the resampled sets, σ_m can become large, revealing that the data are not as stable as naive assumptions might suggest.

Results and Discussion

Figures 1 and 2 show the final fits and bootstrap lines for the two data sets. Numerically, we find:

- **Exercise 2:** Standard fit might give slope $m \approx 1.14 \pm 0.08$, but jackknife and bootstrap suggest uncertainties in the range 0.64–0.93. This large disparity points to one or more outliers or model mismatch.
- **Exercise 1:** The matrix-based uncertainty (~ 0.12) is smaller than the resampling-based values (~ 0.20–0.27), but the discrepancy is not as drastic. Likely, there are fewer high-leverage points in this subset.

These findings illustrate how relying solely on the standard weighted least-squares formula might understate uncertainties if real data deviate from strict linear + Gaussian assumptions. A robust or outlier-aware approach—or simply a more conservative resampling-based error bar—is often safer for final reporting.

4.2.1 to put in a nutshell i can see:

1. *Resampling methods* (jackknife, bootstrap) can yield larger slope uncertainties than the matrix-based formula when the data contain outliers or mislabeled uncertainties.
2. A big discrepancy between standard and resampling-based errors should prompt suspicion of either unaccounted-for systematic effects or underestimated σ_{y_i}.

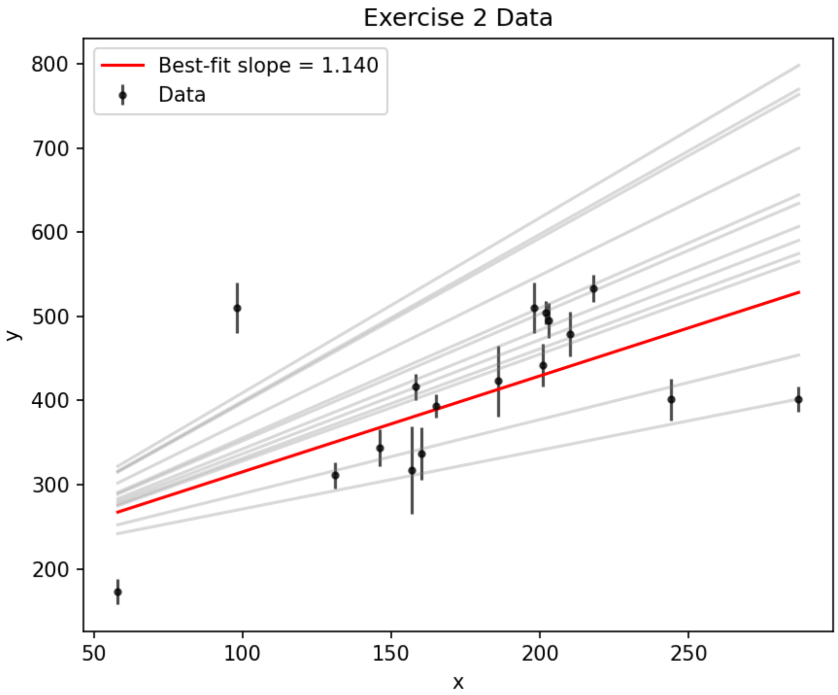

Figure 1: Exercise 2 Data: Weighted fit (red) and 12 randomly chosen bootstrap lines (gray). The difference between standard (matrix-based) and resampling-based uncertainties is substantial, implying possible outliers or misestimated noise.

3. “Exercise 2” data show a stark mismatch, while “Exercise 1” data look more coherent.

References

[1] Hogg, D. W., Bovy, J., and Lang, D. *Data analysis recipes: Fitting a model to data*, `arXiv:1008.4686` (2010).

[2] D. W. Hogg, *Lectures on Data Analysis*, (2011–2023), personal communication and lecture notes.

[3] Wikipedia contributors, *Sidereal year*, `https://en.wikipedia.org/wiki/Sidereal_year`

[4] Wikipedia contributors. *Bootstrapping (statistics)*. Wikipedia, The Free Encyclopedia. Available at: `https://en.wikipedia.org/wiki/Bootstrapping_(statistics)`

[5] Wikipedia contributors. *Bayesian Information Criterion*. Wikipedia, The Free Encyclopedia. Available at: `https://en.wikipedia.org/wiki/Bayesian_information_criterion`

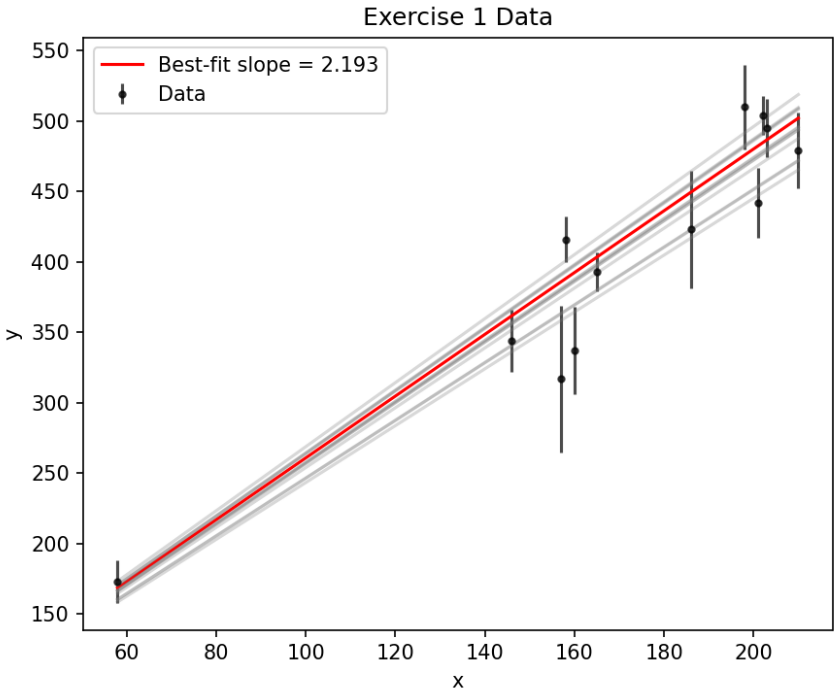

Figure 2: Exercise 1 Data: Weighted fit (red) and 12 randomly chosen bootstrap lines (gray). The gap between matrix-based and resampling-based uncertainties is smaller but still present.