Parameter Estimation for Approximate Regression Models with Heteroskedastic Errors

James Long

MD Anderson Department of Biostatistics

September 12, 2018

Outline

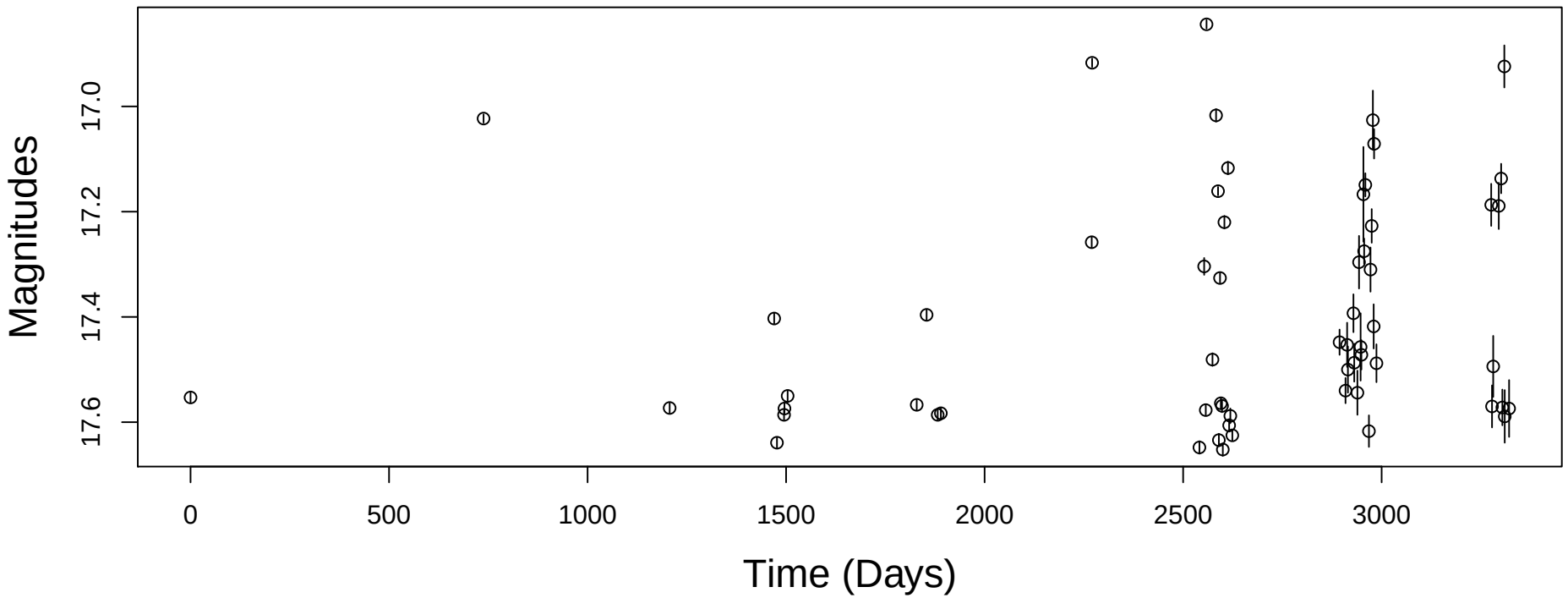
Outline

Models are Approximations

- "All models are wrong, but some are useful."
  – George Box
- But theoretical guarantees underpinning inference tools usually assume model is correct.
  - Coverage probabilities for confidence intervals
  - Asymptotic Efficiency of MLE
  - Bayesian credible sets

Result: Theory suggests using suboptimal procedures because the model is an approximation.

Example: Estimate Period of Function



- Function observed at irregular intervals over ≈ 3000 days.
- Function is periodic with period < 1 day.
- Vertical bars are measurement uncertainties, heteroskedastic.
- **Goal:** Estimate period of function to within 1% of truth.

Notation and Approximate Model

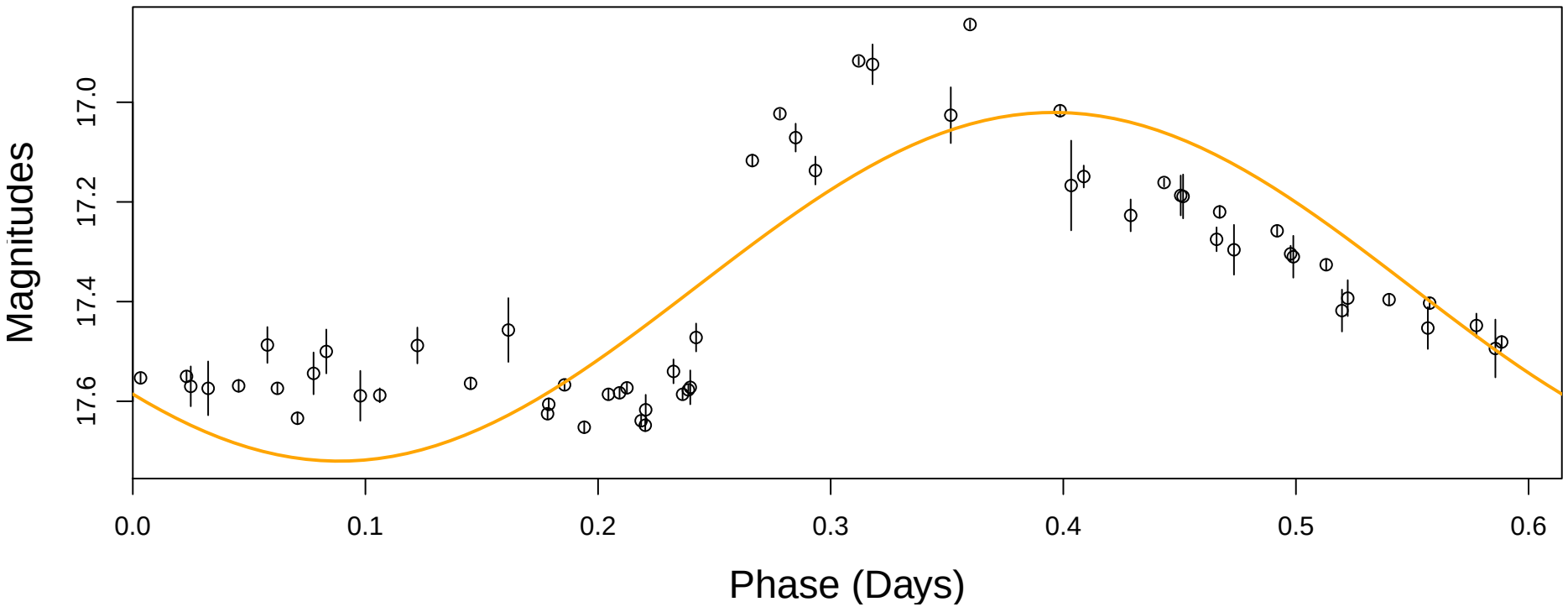
- Observe y_i at time t_i with uncertainty σ_i: $(t_i, y_i, \sigma_i)_{i=1}^n$
- Model with sinusoid with frequency ω:

$$y_i = \beta + a\sin(\omega t_i + \phi) + \sigma_i \epsilon_i$$

  with $\epsilon_i \sim N(0,1)$ independent. Four parameters (β, a, ω, ϕ).
- The MLE (and weighted least squares) frequency estimate is:

$$\widehat{\omega} = \operatorname*{argmin}_{\omega} \min_{a,\beta,\phi} \sum_{i=1}^{n} \left(\frac{y_i - \beta - a\sin(\omega t_i + \phi)}{\sigma_i} \right)^2$$

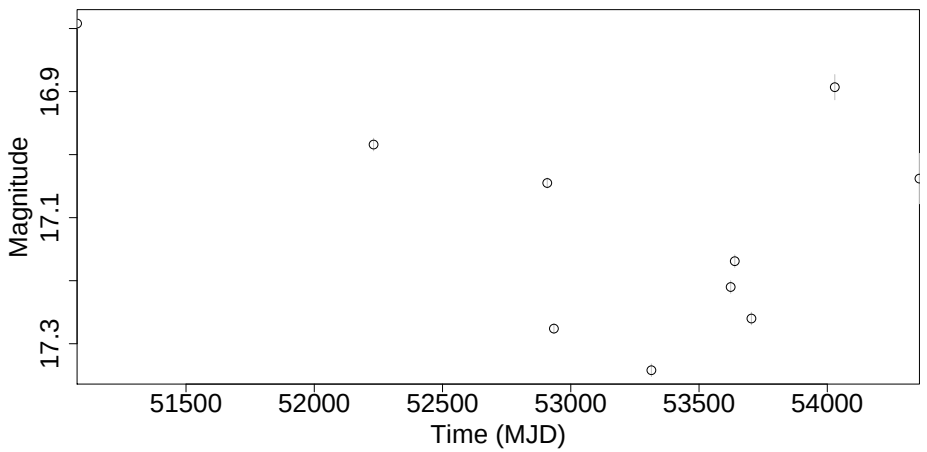
Function Folded on the Estimated Period



- Period estimate ≈ 0.61 days.
- Above is the folded function (time replaced by time modulo estimated period) with best fitting model.
- The sinusoid model is wrong (function not actually a sinusoid), but useful (best fit period is close to truth).

More Challenging Data Regime

- Astronomers are collecting 100M+ of these function.
- Many functions are very sparsely observed, difficult to estimate period correctly:



- **Question:** Is *weighting by the inverse of the observation variances* optimal when the model is an approximation?

Possible Weightings

- Inverse observations variance (MLE of approximate model):

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \min_{a,\beta,\phi} \sum_{i=1}^{n} \frac{(y_i - \beta - a\sin(\omega t_i + \phi))^2}{\sigma_i^2}$$

- Unweighted:

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \min_{a,\beta,\phi} \sum_{i=1}^{n} (y_i - \beta - a\sin(\omega t_i + \phi))^2$$

- Some other weighting which we estimate from data:

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \min_{a,\beta,\phi} \sum_{i=1}^{n} w(\sigma_i)(y_i - \beta - a\sin(\omega t_i + \phi))^2$$

Outline

Section Overview

1. Linear model fitting with heteroskedastic errors.
2. Define what approximate models are estimating.
3. Sandwich estimators, non-standard MLEs, results from Huber and White.

Linear Model with Heteroskedasticity

- Observe $(y_i, x_i, \sigma_i)_{i=1}^n$ i.i.d. where

$$y_i = x_i^T \beta + \sigma_i \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$, $\text{Var}(\epsilon_i) = 1$, $x_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.

- Let $w(\sigma)$ be some weight function (we choose)
- Weighted least squares (WLS) estimator is

$$\begin{aligned}\widehat{\beta}(w) &= \underset{\beta}{\text{argmin}} \sum_{i=1}^n w(\sigma_i)(y_i - x_i^T \beta)^2 \\ &= (X^T W X)^{-1} X^T W Y\end{aligned}$$

where

- W is diagonal weight matrix, $W_{ii} = w(\sigma_i)$.
- $Y = (y_1, \ldots, y_n)^T$ is response
- $X = (x_1^T, \ldots, x_n^T)^T$ is design

Two Common Cases

- Ordinary Least Squares (OLS): $w(\sigma) = 1$, $W = I$

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

- "Standard" Weighted Least Squares (WLS):

$$w(\sigma) = \sigma^{-2}$$
$$W_{ii} = \sigma_i^{-2}$$

In upcoming slides we write $w(\sigma)$ as w.

Asymptotics with Correct Model

Under some regularity conditions on weights and error variances

$$\sqrt{n}(\widehat{\beta}(w) - \beta) \to_d N(0, \underbrace{\mathbb{E}[wxx^T]^{-1}\mathbb{E}[w^2\sigma^2 xx^T]\mathbb{E}[wxx^T]^{-1}}_{\equiv \nu(w)})$$

Implications:

- Any weight function produces a consistent, asymptotically normal estimator.
- The optimal weight function is:

$$w^*(\sigma) = \frac{1}{\sigma^2},$$

  i.e. $0 \preceq \nu(w) - \nu(w^*)$ for any weight function w.

Note:

- Standard WLS is the MLE when $\epsilon_i \sim N(0, 1)$ and is asymptotically efficient.

Least Squares for Approximate Models

- Observe $(y_i, x_i, \sigma_i)_{i=1}^n$ i.i.d. where

$$y_i = f(x_i) + \sigma_i \epsilon_i$$

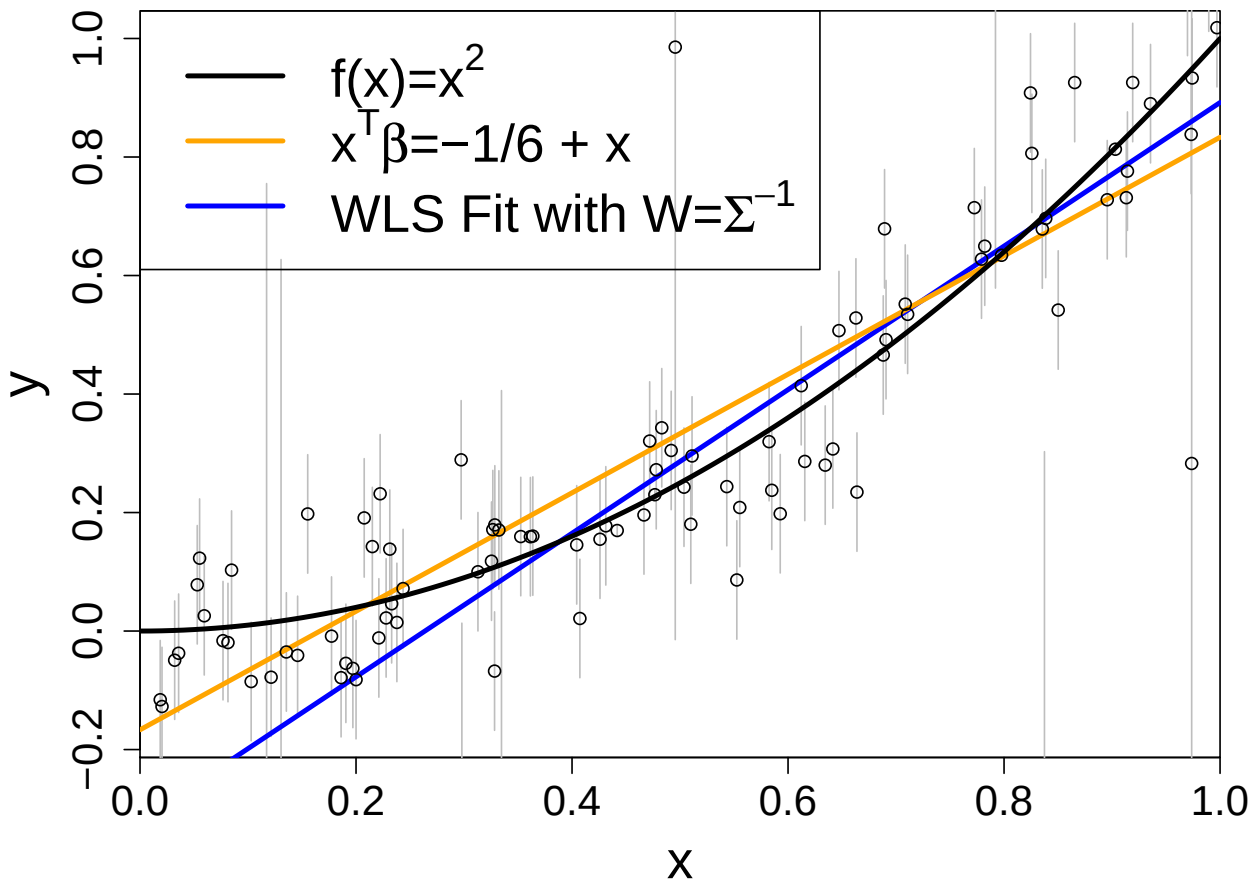
  with $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = 1$, independent across i.
- We use weighted least squares (WLS) estimator:

$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y$$

What are we estimating?

Define the "true" β as the best linear approximation to f i.e.

$$\beta \equiv \underset{\beta}{\operatorname{argmin}}\, \mathbb{E}[(f(x) - x^T\beta)^2] = \mathbb{E}[xx^T]^{-1}\mathbb{E}[xf(x)]$$



Best linear approximation to $f(x) = x^2$ is $-1/6 + x$.

Problem Background

- Huber [2]: Derived asymptotics of MLE when model is an approximation.
  - MLE consistently estimates distribution in model which minimizes KL divergence with true distribution.
  - Derived asymptotic normality of this non-standard MLE.
- White [4, 5]: Derived asymptotic variance and standard errors for OLS when model is an approximation.
  - Huber-white sandwich estimators.
- M-estimator Approach: Define estimators as maximizers of random functions.
  - Many of the above results can be derived through this framework.

Sandwich Estimators

- Often error variances are unknown, so the OLS estimator is used

$$\widehat{\beta}(I) = (X^T X)^{-1} X^T Y$$

- Under some regularity conditions:

$$\sqrt{n}(\widehat{\beta}(I) - \beta) \to_d N(0, \underbrace{\mathbb{E}[xx^T]^{-1}\mathbb{E}[(g^2(x) + \sigma^2)xx^T]\mathbb{E}[xx^T]^{-1}}_{\equiv \nu}).$$

  where $g(x) = f(x) - x^T\beta$.
- Standard asymptotic variance $\sigma^2\mathbb{E}[xx^T]$ is incorrect.
- Estimators of ν are known as "sandwich estimators." They are robust to heteroskedasticity and non–linearity.
  - Various sample based estimators for ν developed, [4, 5, 3]
  - See Buja [1] for a review.

WLS Estimators for Approximate Models

- Existing theory only for OLS, likely due to observation variances being unknown in many fields.
- In astronomy, observations variances generally are known, or at least well approximated.
- Potentially useful to incorporate these variances as weights to construct estimators with lower asymptotic variance.

Outline

How Should We Weight Observations?

Assumptions (Weight Matrix)

Consider diagonal, positive definite weight matrices W where

$$W_{ii} = w(\sigma_i)$$

for $w(\cdot) > 0$ and $\mathbb{E}[w(\sigma)^4] < \infty$.

Important Cases:

- $W_{ii} = w(\sigma_i) = 1$ (OLS)
- $W_{ii} = w(\sigma_i) = \sigma_i^{-2}$ ("Standard" WLS)

Asymptotic Variance

Theorem

Under regularity conditions, $x_i \perp\!\!\!\perp \sigma_i$*, and Weight Matrix Assumptions,*

$$\sqrt{n}(\widehat{\beta}(W) - \beta) \xrightarrow{d} N(0, \nu(w))$$

where $g(x) = f(x) - x^T\beta$ *and*

$$\nu(w) = \frac{\mathbb{E}[w^2] \overbrace{\mathbb{E}[xx^T]^{-1}\mathbb{E}[g^2(x)xx^T]\mathbb{E}[xx^T]^{-1}}^{\equiv A} + \mathbb{E}[\sigma^2 w^2] \overbrace{\mathbb{E}[xx^T]^{-1}}^{\equiv B}}{\mathbb{E}[w]^2}.$$

Special Case: If model is linear $g(x) = 0$ and

$$\nu(w) = \frac{\mathbb{E}[\sigma^2 w^2]\mathbb{E}[xx^T]^{-1}}{\mathbb{E}[w]^2},$$

which is minimized by $w^*(\sigma) = \sigma^{-2}$ (ie $\nu(w^*) \preceq \nu(w)$).

Optimal Weight Function

General Case: No weight function w^* such that $\nu(w^*) \preceq \nu(w) \, \forall \, w$.

Minimize function of asymptotic variance: Consider linear function

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

such that $\Gamma(C) > 0$ for all C positive definite.

Examples:

- $\Gamma(C) = C_{11}$ $\quad (\text{Var}(\widehat{\beta}_1))$
- $\Gamma(C) = Tr(C)$ $\quad (\sum \text{Var}(\widehat{\beta}_j))$

Theorem (Optimal Weighting)

$$w^*(\sigma) = \underset{w}{\text{argmin}} \; \Gamma(\nu(w)) = \frac{1}{\sigma^2 + \Gamma(A)\Gamma(B)^{-1}}$$

Note: argmin is set, but w^* is unique up to rescaling by constant

Adaptive Estimation of Optimal Weighting

Theorem (Adaptivity)

Under regularity conditions, there exist estimators $\widehat{A}$ *and* $\widehat{B}$ *and*

$$\widehat{W}_{ii} = \frac{1}{\sigma_i^2 + \Gamma(\widehat{A})\Gamma(\widehat{B})^{-1}}$$

such that

$$\sqrt{n}(\widehat{\beta}(\widehat{W}) - \beta) \xrightarrow{d} N(0, \nu(w^*)).$$

Theorem

Under regularity conditions

$$\Gamma(\nu(\widehat{\beta}(\widehat{W}))) \leq \min(\Gamma(\nu(\widehat{\beta}(I))), \Gamma(\nu(\widehat{\beta}(\Sigma^{-1}))))$$

with strict inequality if $\mathbb{E}[g^2(x)xx^T] \succ 0$ *and* $P(\sigma = c) \neq 1$ *for any* c.

Simulation: Model $f(x) = x^2$ with Linear Function

Simulation Parameters:

$$n = 100$$

$$x_i \sim Unif[0, 1]$$

$$\sigma_i \sim f_\sigma(\sigma) = \begin{cases} 0.05 & : \sigma = 0.01, 1.0 \\ 0.9 & : \sigma = 0.1 \end{cases}$$

$$\epsilon_i \sim N(0, 1)$$

$$y_i = x_i^2 + \epsilon_i \sigma_i$$

Best Linear Approximation:

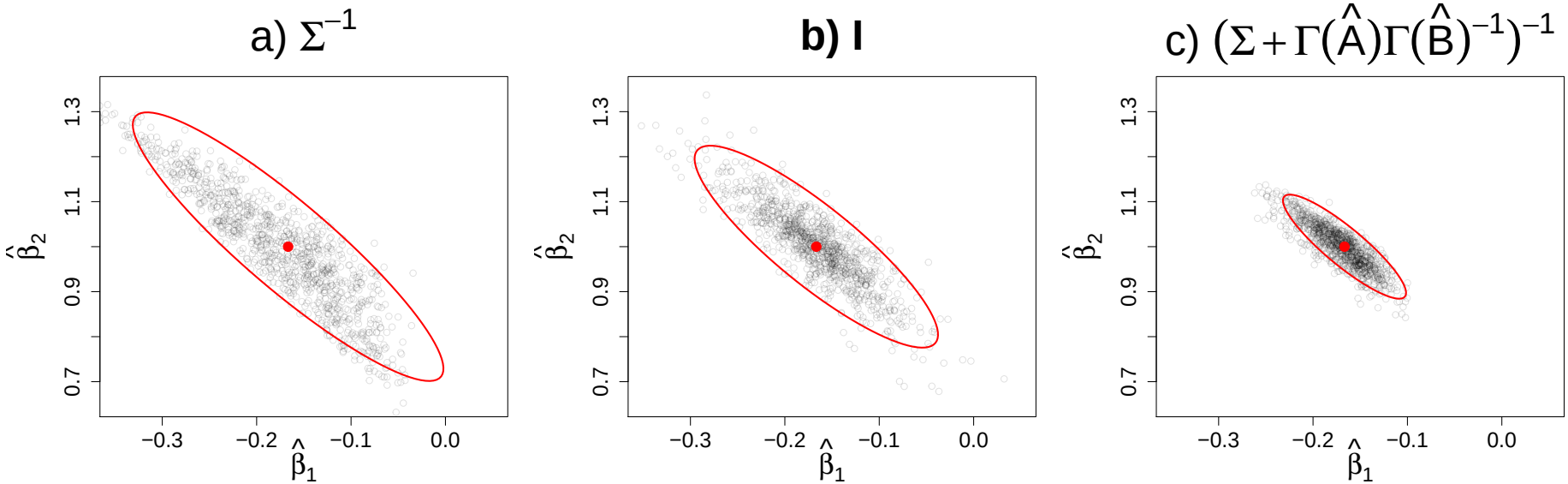
$$\beta = (\beta_1, \beta_2) = (\text{y-int.}, \text{slope}) = (-1/6, 1)$$

Compare Sampling Distribution of:

$$\widehat{\beta}(\widehat{W}) = (X^T \widehat{W} X)^{-1} \widehat{W} X Y$$

for $\widehat{W} =$ a) Σ^{-1} b) I c) $(\Sigma + \Gamma(\widehat{A})\Gamma(\widehat{B})^{-1})^{-1}$

Simulation: Model $f(x) = x^2$ with Linear Function



- $\Gamma(C) = Trace(C)$
- Asymptotic variance for $\widehat{W} = (\Sigma + \Gamma(\widehat{A})\Gamma(\widehat{B})^{-1})^{-1}$ smallest.
- OLS is better than standard WLS.

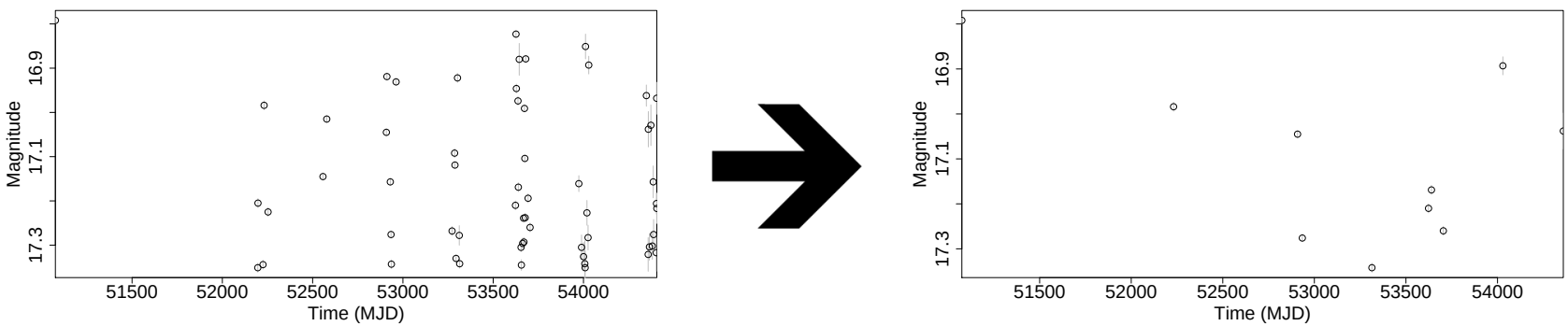
Notes and Extensions

- x, σ Dependence: When the errors and the design are dependent, WLS estimators are typically not consistent for best approximation to f.
- OLS is consistent (for best linear approximation of f) regardless of error–design dependence.
- Unknown σ case: Standard WLS ($w(\sigma) = 1/\sigma^2$) may have larger asymptotic variance than OLS. Thus even good estimates of σ are of questionable value when used in standard WLS procedure with approximate models.

Outline

Application to Variable Star Period Estimation

- Collected 238 periodic function from SDSS Stripe 82 RR Lyrae (g–filter data only)
- Functions are initially well sampled, so true period is known.
- Downsample functions to 10, 20, 30, and 40 observations.



- Downsampling simulates quality of other astronomy data sets.

Methods

Model Class:

$$\widehat{\omega}(w) = \underset{\omega}{\operatorname{argmin}} \min_{a,\beta,\phi} \sum_{i=1}^{n} w(\sigma_i)(m_i - \beta - \sum_{k=1}^{K} a_k \sin(k\omega t_i + \phi_k))^2$$

Weights:

Method	Weight
WLS (Σ^{-1})	$w(\sigma) = \sigma^{-2}$
OLS (I)	$w(\sigma) = 1$
Adaptive (Δ)	$w(\sigma) = (\sigma^2 + \Gamma(\widehat{A})\Gamma(\widehat{B})^{-1})^{-1}$

Harmonics: We fit with fourier models with $K = 1, 2, 3$ ($p = 4, 6, 8$ parameters) harmonics to represent increasing accurate models. However the more complex models (e.g. $K = 3$) may suffer from overfitting with limited data.

Fraction of Period Estimates with 1% of Truth

	$K=1$			$K=2$			$K=3$		
	Σ^{-1}	I	Δ	Σ^{-1}	I	Δ	Σ^{-1}	I	Δ
10	0.09	0.16	0.15	0.13	0.11	0.11	0.03	0.03	0.03
20	0.46	0.58	0.59	0.63	0.68	0.69	0.69	0.77	0.77
30	0.64	0.78	0.79	0.71	0.82	0.83	0.82	0.86	0.85
40	0.75	0.79	0.79	0.80	0.85	0.85	0.87	0.92	0.92

Table: Fraction of periods estimated correctly using different weightings for models with $K = 1, 2, 3$ harmonics. Ignoring the observation uncertainties (I) in the fitting is superior to using them (Σ^{-1}). The strategy for determining an optimal weight function (Δ) does not provide much improvement over ignoring the weights. The standard deviation on these accuracies is no larger than $\sqrt{0.5(1-0.5)/238} \approx 0.032$.

For All the Details

- "A Note on Parameter Estimation for Misspecified Regression Models with Heteroskedastic Errors." J.P. Long. *Electronic Journal of Statistics, 2017.* `projecteuclid.org/euclid.ejs/1492567402`
- `github.com/longjp/hetero_approx`: 100% of code, data, and paper text.

Thank you. Questions?

Bibliography I

[1] Andreas Buja, Richard A Berk, Lawrence D Brown, Edward I George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang.
Models as approximations-a conspiracy of random regressors and model deviations against classical inference in regression.
Statistical Science, page 1, 2015.

[2] Peter J Huber.
The behavior of maximum likelihood estimates under nonstandard conditions.
In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

[3] James G MacKinnon and Halbert White.
Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties.
Journal of econometrics, 29(3):305–325, 1985.

[4] Halbert White.
Using least squares to approximate unknown regression functions.
International Economic Review, pages 149–170, 1980.

[5] Halbert White.
Consequences and detection of misspecified nonlinear regression models.
Journal of the American Statistical Association, 76(374):419–433, 1981.