# Inferring Model Parameters in a High-Dimensional Logistic Regression

Qian Zhao

Febuaray 25th, 2022

# Table of Contents


1. Logistic Regression
2. Inferring Model Parameters
3. Current Theoretical Results
4. Non-Gaussian Covariates
5. The Resized Bootstrap Procedure
6. Future Work
7. Appendix

# Logistic Model

- Covariates $X \in \mathbb{R}^p$, binary response $Y \in \{0, 1\}$
- $\mu(x) = \mathrm{P}(Y = 1 \mid X = x) = 1/(1 + \exp(-x^\top \beta))$ Equivalently, the log-odds is a linear function of $X$,

$$\log\left(\frac{\mu(X)}{1 - \mu(x)}\right) = x^\top \beta.$$

# Logistic Model

- Covariates $X \in \mathbb{R}^p$, binary response $Y \in \{0, 1\}$
- $\mu(x) = \mathrm{P}(Y = 1 \,|\, X = x) = 1/(1 + \exp(-x^\top \beta))$ Equivalently, the log-odds is a linear function of $X$,

$$\log\left(\frac{\mu(X)}{1 - \mu(x)}\right) = x^\top \beta.$$

**Model Parameter**

- Logistic regression estimates $\beta$ by minimizing the negative log-likelihood of observing $(x_i, y_i)$, $i = 1, \ldots, n$,

$$\hat{\beta} = \mathrm{argmin}_{b \in \mathbb{R}^p} \log(1 + e^{-y_i x_i^\top b})$$

- E.g. $X$ is measurement at each SNP and $Y$ is a binary trait.

# An example with synthetic gene expression data

- Model gene expression through a Hidden Markov Model (HMM).
- Generate one sample:
  - $X_i \in \mathbb{R}^p$ ($p = 1454$) from a HMM. $X_i$ are standardized to have 0 mean and variance equal to $1/n$.
  - Sample true coefficients $\beta$ by randomly pick 100 to be non-nulls and sample non-null $\beta_j \sim \mathcal{N}(0, 10)$.
  - $Y_i \in \{0, 1\}$ from a logistic model.
- Each data consists of $n = 5000$ samples generated as above.
- Fit a logistic regression to compute the MLE $\hat{\beta}$ for each data.
- We study the distribution of $\hat{\beta}_j$ by repeat this process 1000 times.

# Table of Contents

1 Logistic Regression

2 Inferring Model Parameters

3 Current Theoretical Results

4 Non-Gaussian Covariates

5 The Resized Bootstrap Procedure

6 Future Work

7 Appendix

# Classical maximum likelihood theory

**Theorem 5.21 (van der Vaart)**

If $p$ is fixed and $n$ goes to infinity, then under mild regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_\beta^{-1})$$

where $\mathcal{I}_\beta$ is the Fisher information matrix evaluated at $\beta$.

# Classical maximum likelihood theory

**Theorem 5.21 (van der Vaart)**

If $p$ is fixed and $n$ goes to infinity, then under mild regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\beta}^{-1})$$

where $\mathcal{I}_{\beta}$ is the Fisher information matrix evaluated at $\beta$.

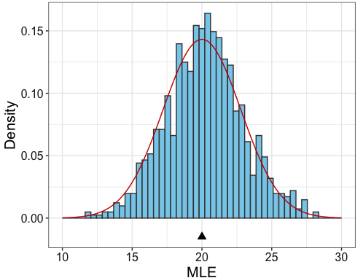
For a logistic regression,

$$\mathcal{I}_{\beta} = \mathbb{E}\left[(X^{\top} W X)^{-1}\right],$$

where $W = \mathrm{diag}(w_1, w_2, \ldots, w_n)$, $w_i = 1/\{(1 + e^{-x_i^{\top}\beta})(1 + e^{x_i^{\top}\beta})\}$.

# An Example When $p/n$ is Small

- Number of observations $n = 5000$
- Number of variables $p = 50$
- $X$ is the first 50 variables in the HMM model, standardized to have zero mean and variance equal to $1/n$.
- $\beta_j \in \{-20, 20\}$.
- $Y \mid X$ is from a logistic model




In 1000 simulations, the MLE is centered at 20.2 (the true coefficient is $\beta_j = 20$). The empirical Std. Dev is 2.75 and the estimate by glm function is 2.79 in one data.

# Classical maximum likelihood theory

**Theorem 5.21 (van der Vaart)**

If $p$ is fixed and $n$ goes to infinity, then under mild regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_\beta^{-1})$$

where $\mathcal{I}_\beta$ is the Fisher information matrix evaluated at $\beta$.

- The classical theory also holds when $p^2/n \to 0$
- But, the classical theory does not hold if $p/n \to \kappa > 0$! (Huber ,1973)
- We will study the high-dimensional setting when $p, n \to \infty$ while $p/n \to \kappa \in (0, 1)$

# An example with synthetic gene expression data (Cont'd)

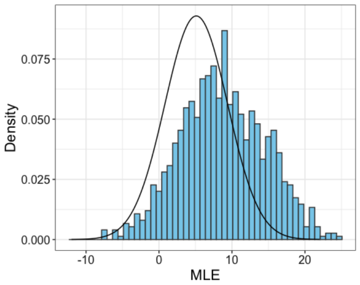



Figure: Histogram of a non-null MLE; the black line shows the estimated density by classical theory in one data.pval

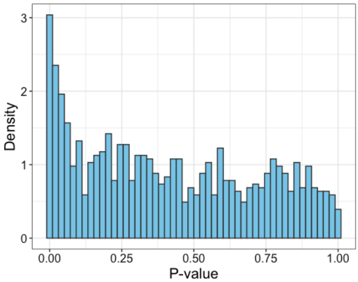



Figure: Histogram of a null P-value for testing $\mathcal{H}_0 : \beta_j = 0$. The P-values are far from $\mathrm{Unif}(0, 1)$! We falsely reject a true null hypothesis more often than we should.

# Table of Contents


1. Logistic Regression
2. Inferring Model Parameters
3. Current Theoretical Results
4. Non-Gaussian Covariates
5. The Resized Bootstrap Procedure
6. Future Work
7. Appendix

# Distribution of the MLE

Suppose $X \sim \mathcal{N}(0, \Sigma)$ and $Y \mid X$ is from a logistic model with coefficients $\beta$. Let

$$p/n \to \kappa, \quad \text{and} \quad \mathrm{Var}(X^\top \beta) = \gamma^2$$

and assume $(\kappa, \gamma)$ are in the region where the MLE exists asymptotically.

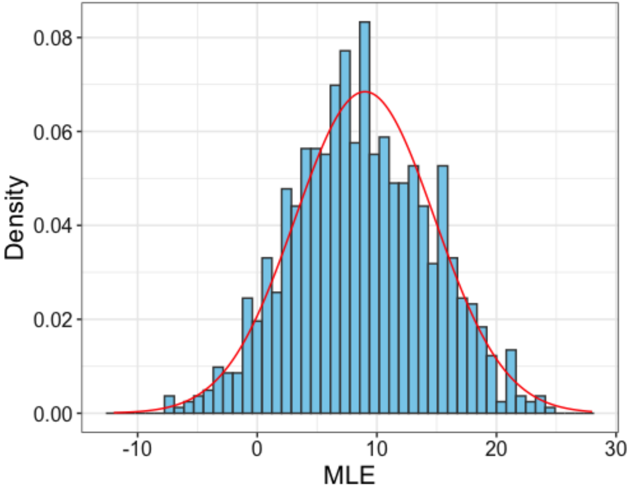
**Theorem 1 MLE distribution**

Let $\tau_j^2 = \mathrm{Var}(X_j \mid X_{-j})$. If $\sqrt{n}\tau_j\beta_j = O(1)$, then

$$\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2/\tau_j^2)$$

The parameters $\alpha_\star$ and $\sigma_\star$ depends on the ratio $\kappa$ and the signal strength $\gamma$.

# Empirical Acuracy



# Testing $\mathcal{H}_0 : \beta_j = 0$

**Corollary 1 Null distribution of the MLE**

If $\beta_j = 0$, then

$$\sqrt{n}\hat{\beta}_j \xrightarrow{d} \mathcal{N}(0, \sigma_*^2/\tau_j^2),$$

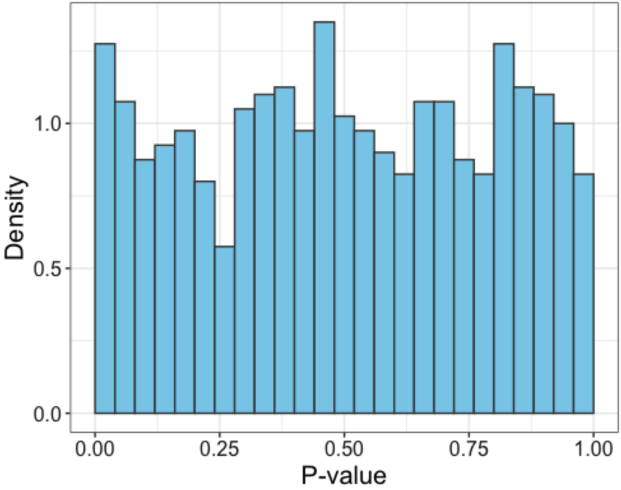
where $\tau_j^2 = \mathrm{Var}(X_j \mid X_{-j})$.

A two-sided p-value for testing $\mathcal{H}_0 : \beta_j = 0$ is given by

$$p_j = 2 \times \Phi(-\sqrt{n}\tau_j|\hat{\beta}_j|/\sigma_\star),$$

where $\Phi$ is the normal cdf.

# Histogram of a Null P-Value



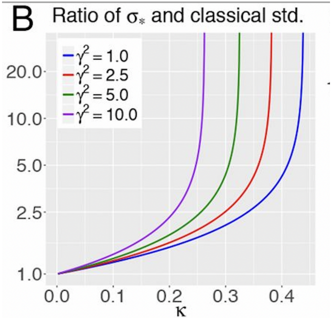
# Previous research

Previous studies have proved the null distribution of $\hat{\beta}_j$ when $\Sigma = I$.

**Theorem 3 (Sur and Candès, 2019)**

If $\beta_j = 0$, then as $n, p \to \infty$ while $p/n \to \kappa$,

$$\sqrt{n}\hat{\beta}_j \xrightarrow{d} \mathcal{N}(0, \sigma_*^2)$$




$\sigma_\star$ is larger than classical theory unless $\kappa \to 0$.

# Deriving the Null Distribution of the MLE

The MLE minimizes the negative log-likelihood

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top b}),$$

where $X_i \sim \mathcal{N}(0, \Sigma)$. WLOG, we let $j = p$. Let $L^\top L = \Sigma$ be the Cholesky decomposition of $\Sigma$, i.e. $L$ is a lower triangular matrix. Then,

$$\sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top b}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \underbrace{\boxed{x_i^\top L^{-\top}}}_{\boxed{L^{-1}X_i \sim \mathcal{N}(0, I)}} L^\top b}).$$

In addition, $Y_i \mid L^{-1}X_i$ is from a logistic model with coefficients $L^\top \beta$. In particular, the last coefficient is $L_{j,j}\beta_j$! This means

$$\sqrt{n} L_{j,j} \hat{\beta}_j \xrightarrow{d} \mathcal{N}(0, \sigma_\star^2)$$

# Previous Research

- Logistic regression (Sur and Candès (2019))
  - Precisely characterized of the condition when the MLE exists
  - Derived the exact MLE distribution of a null variable
  - Established the "bulk" distribution of the MLE
- Robust regression (El Karoui (2013), Donoho and Montanari (2016))
  - Derived the exact MLE distribution
- The LASSO regression
  - Studies the distribution of LASSO coefficients and how to construct CI for model coefficients when the model is sparse $s_0 = o(n/\log p)$ (Zhang and Zhang, van der Geer, Javanmard and Montanari) and when the model is not sparse (Bellec and Zhang, Celentano et.al)

# Summary and Extensions

- We derived the asymptotic distribution of the MLE of a logistic regression model when $X \sim \mathcal{N}(0, \Sigma)$,

$$\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2/\tau_j^2),$$

  where $\tau_j^2 = \mathrm{Var}(X_j \mid X_{-j})$ and this holds when $\sqrt{n}\tau_j\beta_j = O(1)$
- We are able to construct valid confidence intervals.
- Extensions:
  - We developed procedures to esimate the signal strength in practice.
  - We extended the theory to include the case when there is a non-zero intercept.

# Table of Contents


1. Logistic Regression
2. Inferring Model Parameters
3. Current Theoretical Results
4. Non-Gaussian Covariates
5. The Resized Bootstrap Procedure
6. Future Work
7. Appendix

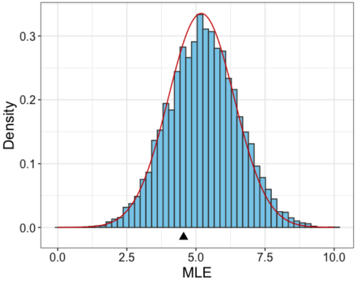
# An Example With Non-Gaussian Covariates

The asymptotic distribution of the M-estimators of robust regression depends on the covariate distribution (El Karoui, 2018).

- $n = 2000$, $p = 200$
- $X \sim \mathrm{MVT}(0, \Sigma)$ with $\nu = 8$ degrees of freedom. $\Sigma$ is a circular matrix $\Sigma_{ij} = 0.5^{\min(|i-j|, p+1-|i-j|)}$. Standardize $X$ to have variance equal to $1/p$.
- Sample 20 variables to be non-nulls. The non-null $\beta_j$ are i.i.d. from $\mathcal{N}(\pm 5, 1)$.
- $Y \mid X$ is sampled from a logistic model.

# An Example With Non-Gaussian Covariates (Result)

- The high-dimensional theory slightly under-estimates the Std. Dev.
  - The empirical bias is 1.16 and the theoretical prediction is 1.14.
  - The empirical Std. Dev is 1.27 while the theoretical prediction is 1.19.
- The CI slightly undercovers $\beta_j$
  - Theoretical 95% CI covers approximately 93.3% times.



# Can We Use the Bootstrap Method?

- The bootstrap is a resampling method to estimate the sampling distribution of a statistics.
- Two standard sampling methods:
  - The parametric bootstrap
  - The nonparametric (pairs) bootstrap
- Bootstrap confidence intervals: percentile, bootstrap-$t$, ...

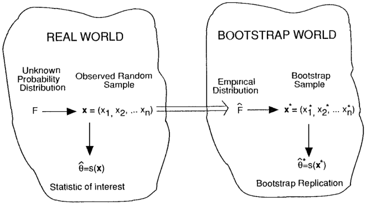
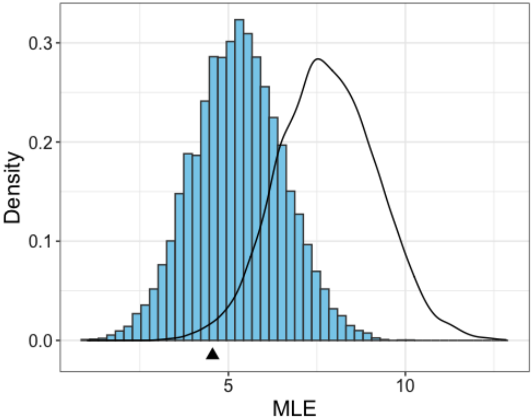



Figure: Figure 8.1 in *An Introduction to the Bootstrap*, by Efron and Tibshirani

# Parametric Bootstrap

1. Given observed data $(X_i, Y_i)$, $i = 1, \ldots, n$.
2. Fit a logistic regression to obtain the MLE $\hat{\beta}$.
3. Construct $B$ bootstrap samples. The $b$th bootstrap sample:
    1. Fix the covariates at observed $X_i$.
    2. Sample $Y_i^b$ using $X_i$ as the covariates and $\hat{\beta}$ as the model coefficient.
    3. Compute the MLE for the bootstrap sample $\hat{\beta}^b$.
4. The percentile bootstrap $(1-c)$-CI for $\beta_j$ is $[\hat{\beta}_j^b[c/2], \hat{\beta}_j^b[1-c/2]]$ where $\hat{\beta}_j^b[c/2]$ is the $c/2$ quantile of the bootstrap samples $(\hat{\beta}_j^1, \ldots, \hat{\beta}_j^B)$.

# Can We Use the Parametric Bootstrap?




The standard parametric bootstrap do not work in high-dimensions!

# Why Does Parametric Bootstrap Fail in High-Dimensions?

When $X \sim \mathcal{N}(0, \Sigma)$, the MLE is approximately

$$\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j) \xrightarrow{d} \mathcal{N}(0, \sigma_\star / \tau_j)$$

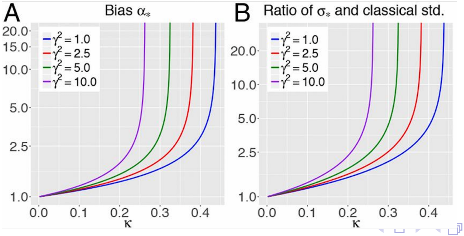
variance depends on the ratio $\kappa = p/n$ and the signal strength $\gamma = \mathrm{Var}(X^\top \beta)^{1/2}$.

We can compute that

$$\mathrm{Var}(X^\top \hat{\beta}) \approx \alpha_\star^2 \gamma^2 + \kappa \sigma_\star^2 > \gamma$$

Both $\alpha_\star$ and $\sigma_\star$ increases as $\gamma$ increases.

This suggests that we can apply the parametric bootstrap at a coefficient different from the MLE.



# Table of Contents


1. Logistic Regression
2. Inferring Model Parameters
3. Current Theoretical Results
4. Non-Gaussian Covariates
5. The Resized Bootstrap Procedure
6. Future Work
7. Appendix

# The Resized Bootstrap Method

- We generate new responses using the parametric bootstrap, but choosing $\beta^s$ that satisfies

$$\mathrm{Var}(X^\top \beta^s) = \gamma^2.$$

- We set $\beta^s = s \times \hat{\beta}$ by "resizing" the MLE.
- After $B$ repetitions, we obtain $B$ bootstrap MLE

$$(\hat{\beta}^1, \ldots, \hat{\beta})^B$$

# Estimating Bias and Std.Dev using the resized bootstrap

- Estimate $\hat{\sigma}_j$ by the standard deviation of the bootstrap MLE.
- Estimate $\hat{\alpha}_j$ by the regression coefficient of average bootstrap MLE onto $\beta^s$, and weight $j$th variable proportional to $1/\hat{\sigma}_j^2$.

# Constructing CI for $\beta_j$

Method 1 (Gaussian approximation):
Suppose

$$\frac{\hat{\beta}_j - \alpha_j \beta_j}{\sigma_j} \approx \mathcal{N}(0, 1),$$

and use the bootstrap MLE $\hat{\beta}^1, \ldots, \hat{\beta}^B$ to estimate $\alpha_j$ and $\sigma_j$.
Method 2 (Bootstrap-$t$):
Suppose the MLE are not gaussian, we use the bootstrap MLE to estimate the distribution of the MLE, i.e. assuming

$$\frac{\hat{\beta}_j - \hat{\alpha}_j \beta_j}{\hat{\sigma}_j} \overset{d}{\approx} \frac{\hat{\beta}_j^b - \hat{\alpha} \beta_j^s}{\hat{\sigma}_j}$$

and estimate the RHS by the quantiles of the bootstrap MLE.

# Coverage Probability of a Null Variable

| | Theoretical CI | Resized bootstrap | |
|---|---|---|---|
| Nominal | High-Dim | I. Boot-$g$ | II. Boot-$t$ |
| 95% | 93.4% | 95.2 % (0.87%) | 94.8% (0.92%) |
| 90% | 87.5% | 89.3% (1.26%) | 89.8% (1.24%) |
| 80% | 77.5% | 79.3% (1.66%) | 78.7% (1.68%) |

Table: Coverage probability of a single **null** variable in $N = 600$ samples. The standard deviations are shown in the parentheses.

# Coverage Probability of a Non-Null Variable

| | Theoretical CI | Resized bootstrap | |
|---|---|---|---|
| Nominal | High-Dim | I. Boot-$g$ | II. Boot-$t$ |
| 95% | 92.7 % | 95.4 % (0.86%) | 94.6% (0.93%) |
| 90% | 87.0% | 90.2% (1.22%) | 90.5% (1.20%) |
| 80% | 76.1% | 82.9% (1.55%) | 82.6% (1.56%) |

Table: Coverage probability of a single **non-null** variable in $N = 600$ samples. The standard deviations are shown in the parentheses.

# Conclusion

- We develop a resized bootstrap method which combines the high-dimensional theory with parametric bootstrap to infer parameters in a high-dimensional GLM.
- The resized bootstrap can be used to construct CI and the CI achieve reasonable coverage for moderate $n$.
- The resized bootstrap applies to other GLM and different covariate distributions.

# Table of Contents

1 Logistic Regression

2 Inferring Model Parameters

3 Current Theoretical Results

4 Non-Gaussian Covariates

5 The Resized Bootstrap Procedure

6 Future Work

7 Appendix

# Future Work

We derived the theoretical distribution of the logistic MLE when the covariates are multivariate Gaussian.

| $\tau_j \beta_j / \gamma$ | Std. Dev |
|---|---|
| 0.15 | 2.91 |
| 0.3 | 3.28 |
| 0.5 | 4.11 |

The variance of the MLE increase with the magnitude of $\beta$.
Can we characterize the variance as a function of $\beta$?

# Future Work (Cont'd)

- We developed a resized bootstrap method to estimate the MEL distribution. Can we analyze the procedure to justify or improve it?
- Can the idea of resized bootstrap be applied to other high-dimensional problems?

# References I


Huber, P.
Robust Regression: Asymptotics, Conjectures and Monte Carlo, *Ann. Statist.*, 1973

Vaart, A. W. van der
Asymptotic Statistics, *Cambridge Series in Statistical and Probabilistic Mathematics*, 1998

Donoho, D., Montanari, A.
High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing, *Probab. Theory Relat. Fields*, 2016

El Karoui, N.
On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators *Probab. Theory Relat. Fields*, 2018

El Karoui, N. and Purdom, E.
Can We Trust the Bootstrap in High-Dimensions? The Case of Linear Models, *J. Mach. Learn. Res.*, 2018

# References II


Sur, P. and Candès

A modern maximum-likelihood theory for high-dimensional logistic regression, *PNAS*, 2019

Candès, E. and Sur, P.

The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression *Ann. Statist.*, 2020

# Table of Contents


1. Logistic Regression
2. Inferring Model Parameters
3. Current Theoretical Results
4. Non-Gaussian Covariates
5. The Resized Bootstrap Procedure
6. Future Work
7. Appendix

# Estimating $\tau_j$

We estimate $\tau_j^2 = \mathrm{Var}(X_j \mid X_{-j})$ by the residual sum of squares of regressing $X_j$ onto $X_{-j}$:

$$\hat{\tau}_j^2 = \frac{\mathrm{RSS}_j}{n-p},$$

is an unbiased estimator of $\tau_j^2$.

# Main results: distribution of a single MLE coordinate

**Theorem**

*Let* $\tau_j^2 = \mathrm{Var}(x_{i,j} \mid \boldsymbol{x}_{i,-j})$. *If* $\sqrt{n}\tau_j\beta_j = O(1)$, *then*

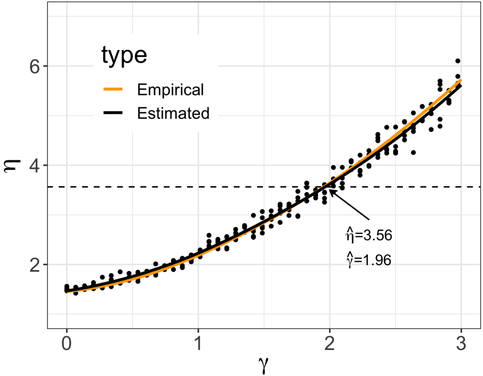
$$\frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star\beta_j)}{\sigma_\star/\tau_j} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof: wlog, assume $j = p$,

$$\begin{aligned}
\frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star\beta_j)}{\sigma_\star/\tau_j} &= \frac{\sqrt{n}(\hat{\theta}_j - \alpha_\star\theta_j)}{\sigma_\star} \quad \text{from previous Lemma} \\
&= \left(\sqrt{n}\frac{\hat{\theta}_j - \alpha(n)\theta_j}{\sigma(n)} + \sqrt{n}\theta_j\frac{(\alpha(n) - \alpha_\star)}{\sigma(n)}\right)\frac{\sigma(n)}{\sigma_\star} \\
&= \mathcal{N}(0,1) + o_P(1)
\end{aligned}$$

**Conjecture:** $\sqrt{n}(\alpha(n) - \alpha_\star) = O_P(1)$, so only need $\tau_j\beta_j = O(1)$.

# Estimating $\gamma$




- Use the one-to-one correspondence between $\mathrm{Var}(X_{\mathrm{new}}^{\top}\hat{\beta})$ with $\gamma$.
- Use the SLOE estimator to estimate $\mathrm{Var}(X_{\mathrm{new}}^{\top}\hat{\beta})$ from the MLE.
- Apply the parametric bootstrap to compute $\mathrm{Var}(X_{\mathrm{new}}^{\top}\hat{\beta})$ when $\beta = s \times \hat{\beta}$.