How should we do linear regression?

Sara El Khamlichi

October 13, 2025

Supervisors: Elliot Young and Richard Samworth

Linear Models

Model

$$Y = X^\top \beta_0 + \varepsilon, \qquad X, \beta_0 \in \mathbb{R}^d, \quad \varepsilon \perp\!\!\!\perp X.$$

Ordinary Least Squares (OLS)

$$\hat{\beta}^{\mathrm{OLS}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2.$$

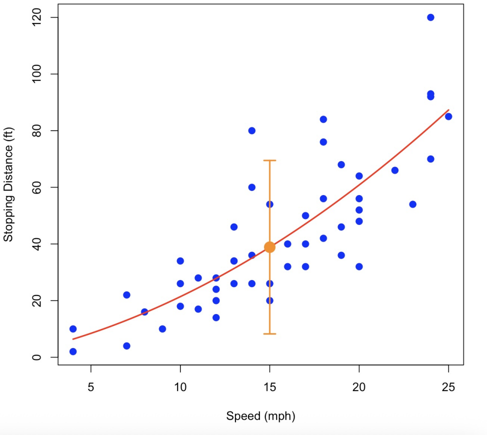
- **Gauss–Markov theorem:** OLS is the *Best Linear Unbiased Estimator (BLUE)*.
- The `lm()` function builds on OLS: hypothesis tests, confidence and prediction intervals.

$\Rightarrow$ But:

- Optimality is only among linear, unbiased estimators.
- Inference tools assume Gaussian errors.

Example: What is a prediction interval?

- We observe the speed of cars X_i and their stopping distance Y_i, for $i = 1, \dots, n$.
- **Model**: $Y = \beta_1 X + \beta_2 X^2 + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X$ where $\varepsilon \sim N(0, \sigma^2)$.
- **Goal:** Predict Y^* for a new speed $x^* = 15$ mph.



Prediction Intervals — Definitions

Setup

We fit a linear model to data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ of the form $Y_i = X_i^\top \beta_0 + \varepsilon_i$. We have a new independent pair (X^*, Y^*).

Prediction Interval

A level $(1-\alpha)$ prediction interval is a random interval $C_n(x^*;\alpha) \subset \mathbb{R}$ such that

$$\mathbb{P}\big(Y^* \in C_n(x^*;\alpha) \,\big|\, \mathcal{D}_n, X^* = x^*\big) = 1 - \alpha.$$

Asymptotic Validity

A sequence of intervals $\{C_n(x^*;\alpha)\}_{n\geq 1}$ is *asymptotically valid* if

$$\mathbb{P}\big(Y^* \in C_n(x^*;\alpha) \,\big|\, \mathcal{D}_n, X^* = x^*\big) \xrightarrow{p} 1 - \alpha, \quad \text{as } n \to \infty.$$

How Wrong Can OLS Be?

Lemma

In the linear model

$$Y = X^\top \beta_0 + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X, \quad \mathbb{E}(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2,$$

as $n \to \infty$,

$$\mathbb{P}\left(Y^* \in C_n^{\mathrm{OLS}}(x^*; \alpha) \,\middle|\, \mathcal{D}_n\right) \xrightarrow{p} 1 - \mathbb{P}\left(|\varepsilon| \geq \sigma z_{1-\alpha/2}\right) \geq \max\left\{1 - \frac{1}{z_{1-\alpha/2}^2}, 0\right\}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of $N(0,1)$, and this lower bound is tight.

Target Coverage	**Worst-Case OLS Coverage**
90%	63%
95%	74%

M-estimation

Model

$$Y = X^\top \beta + \varepsilon, \quad X, \beta \in \mathbb{R}^d, \quad \varepsilon \perp\!\!\!\perp X, \quad \varepsilon \sim p_0$$

M-estimation

- For a loss function $l : \mathbb{R} \to \mathbb{R}$ with $\psi := -l'$

$$\hat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(Y_i - X_i^\top \beta).$$

- Under some regularity conditions, (van der Vaart, 2000, Theorem 5.21),

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N_d\left(0, V_{p_0}(\psi) \cdot \left[\mathbb{E}(X_1 X_1^\top)\right]^{-1}\right).$$

Antitonic Score Matching (ASM)

Model

$$Y = X^\top \beta + \varepsilon \quad X, \beta \in \mathbb{R}^d, \quad \varepsilon \perp\!\!\!\perp X$$

where $\varepsilon \sim p_0$ and p_0 is unknown.

ASM Estimator

- Constructs a data-driven convex loss function l.
- Minimises the asymptotic variance over all convex l :

$$\sqrt{n}(\hat{\beta}_n^{\mathrm{ASM}} - \beta_0) \xrightarrow{d} N_d\left(0, i^*(p_0) \cdot \left[\mathbb{E}(X_1 X_1^\top)\right]^{-1}\right)$$

where $i^*(p_0) = \min\limits_{l \text{ convex}} V(\psi)$.

[0] Feng, O., Kao, Y., Xu, M. and Samworth, R. (2025). *Optimal Convex M-Estimation via Score Matching. Annals of Statistics* (to appear).

Where does the uncertainty come from?

Prediction decomposition

For a new pair (x^*, Y^*) independent of the data $\mathcal{D}_n$:

$$Y^* - x^{*\top}\hat{\beta}_n^{\mathrm{ASM}} = \underbrace{\varepsilon^*}_{\text{irreducible noise}} + \underbrace{x^{*\top}(\beta_0 - \hat{\beta}_n^{\mathrm{ASM}})}_{\text{estimation error}}.$$

- The first term, ε^*, captures the randomness in Y given X — the irreducible noise.
- The second term arises because $\hat{\beta}_n^{\mathrm{ASM}}$ is estimated from finite data — the estimation error.

How we model the errors

Two components

1. Estimate the **noise distribution** p_0 from the residuals:

$$\hat{\varepsilon}_i = Y_i - X_i^\top \hat{\beta}_n^{\mathrm{ASM}}, \quad i \in [n].$$

2. Approximate the **estimation error** using asymptotic normality (Feng et al., 2025):

$$x^{*\top}(\hat{\beta}_n^{\mathrm{ASM}} - \beta_0) \overset{d}{\approx} \mathcal{N}\left(0, \tfrac{1}{n} x^{*\top} \widehat{\Sigma}_n x^*\right).$$

Then, combine samples from these two distributions to form a prediction interval.

Constructing the prediction intervals

(A) Noise

$\varepsilon_b \sim \tilde{p}_n$

(B) Estimation Error

$Z_b \sim \mathcal{N}\left(0, \frac{1}{n} x^{*\top} \widehat{\Sigma}_n x^*\right)$

$+$

Combine: $S_b = \varepsilon_b + Z_b, \;\; b \in [B_n].$

Take empirical quantiles of $\{S_b\}_{b \in [B_n]}$.

ASM Prediction Interval

$$C_n^{\mathrm{ASM}}(x^*; \alpha) = \left[x^{*\top} \hat{\beta}_n + \tilde{Q}_{n,B_n}(\alpha/2; x^*),\, x^{*\top} \hat{\beta}_n + \tilde{Q}_{n,B_n}(1 - \alpha/2; x^*) \right],$$

where $\tilde{Q}_{n,B_n}(\tau; x^*)$ is the τ–quantile of $\{S_b\}_{b \in [B_n]}$.

Asymptotic validity of ASM prediction intervals

Theorem (Asymptotic consistency of ASM PIs)

Under regularity conditions, for P_X-almost every $x^* \in \mathbb{R}^d$,

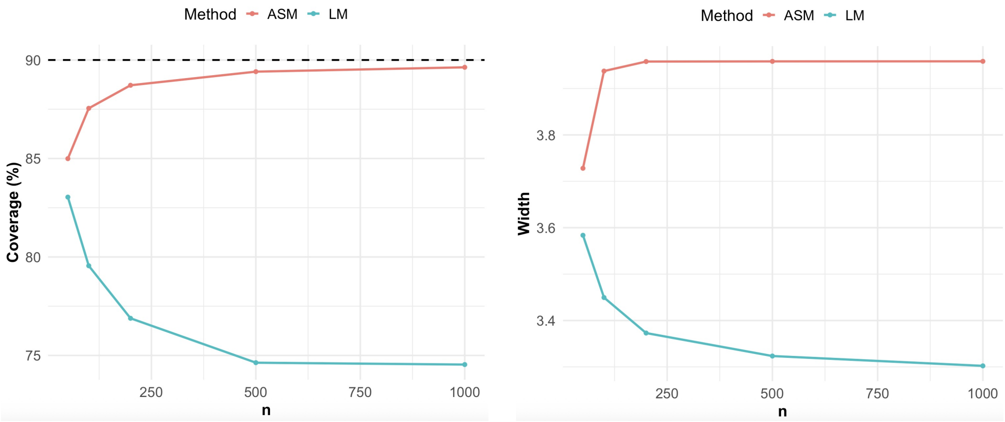
$$\mathbb{P}\big(Y^* \in C_n^{\mathrm{ASM}}(x^*; \alpha) \mid \mathcal{D}_n, X^* = x^*\big) \xrightarrow{p} 1 - \alpha.$$

In contrast:

$$\texttt{predict.lm()} \Rightarrow \mathbb{P}\big(Y^* \in C_n^{\mathrm{OLS}}(x^*; \alpha) \mid \mathcal{D}_n, X^* = x^*\big) \xrightarrow{p} 1 - \mathbb{P}\big(|\varepsilon| \geq \sigma z_{1-\alpha/2}\big),$$

which can drop to as low as **74%** for a 95% PI.

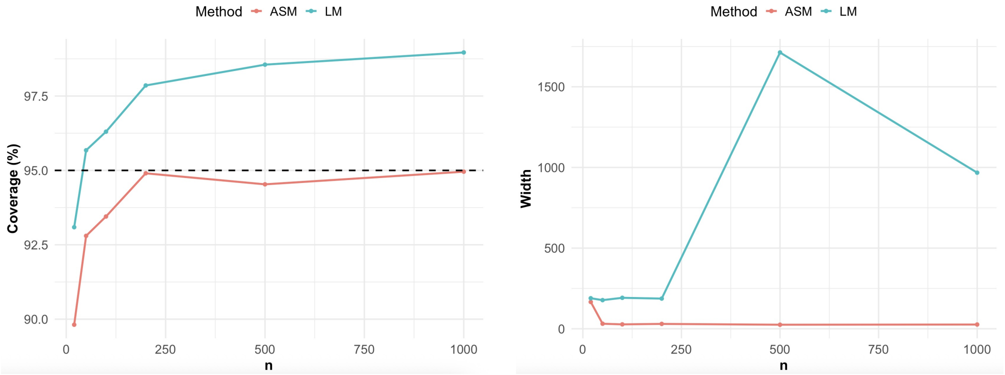
Gaussian Location Mixture (Multimodal Distribution)



Left: Coverage Right: PI width.

- `predict.lm()` doesn't capture the multimodal structure $\Longrightarrow$ PIs undercover.

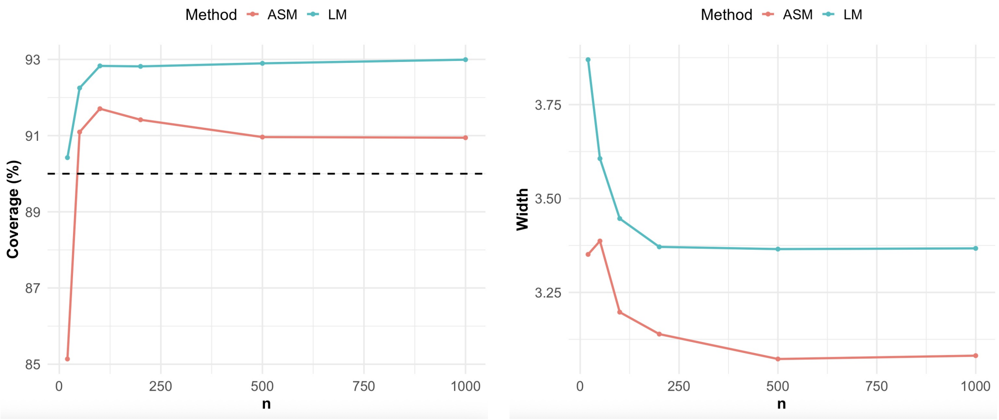
Cauchy Distribution (Heavy Tails)



Left: Coverage. *Right: PI width.*

- `predict.lm()` tries to estimate the variance of the errors $\Longrightarrow$ PIs overcover.

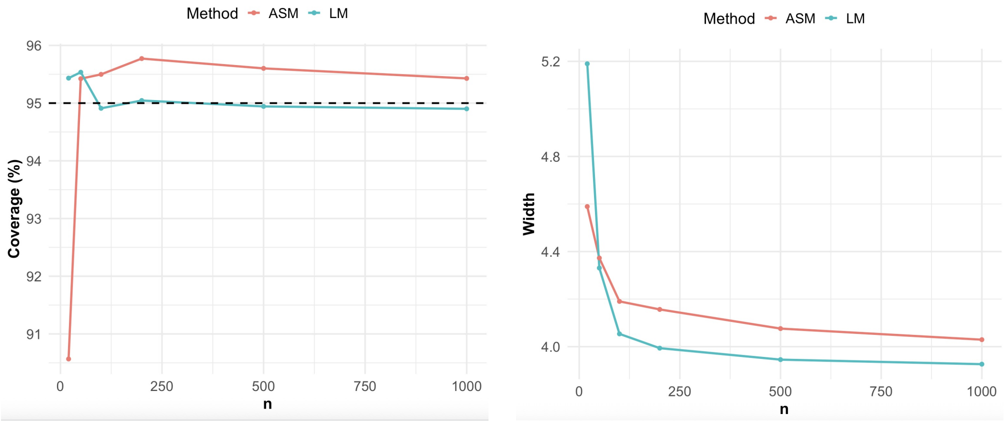
Smoothed Exponential (Skewed Distribution)



Left: Coverage. Right: PI width.

- `predict.lm()` gives symmetric PIs $\Longrightarrow$ PIs are wider than necessary.

Gaussian



Left: Coverage Right: PI width.

- `predict.asm()` performs comparably to `predict.lm()`.

Summary

Our contribution

- A framework for constructing asymptotically valid prediction intervals under general, non-Gaussian noise.
- `predict.asm()` — a more reliable alternative to `predict.lm()`.

Open questions

- What conditions are required for convergence in expected length of the PIs?
- Over what class of distributions is this convergence uniform?

References

O. Feng, Y. Kao, M. Xu, and R. Samworth. Optimal convex M-estimation via score matching. *Annals of Statistics*, 2025. to appear.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 2000.