Not your grandparents' confidence intervals

Gregor Reich (NHH) and Ken Judd (Hoover)

March 23, 2021

How can we determine the statistical properties of our estimators?

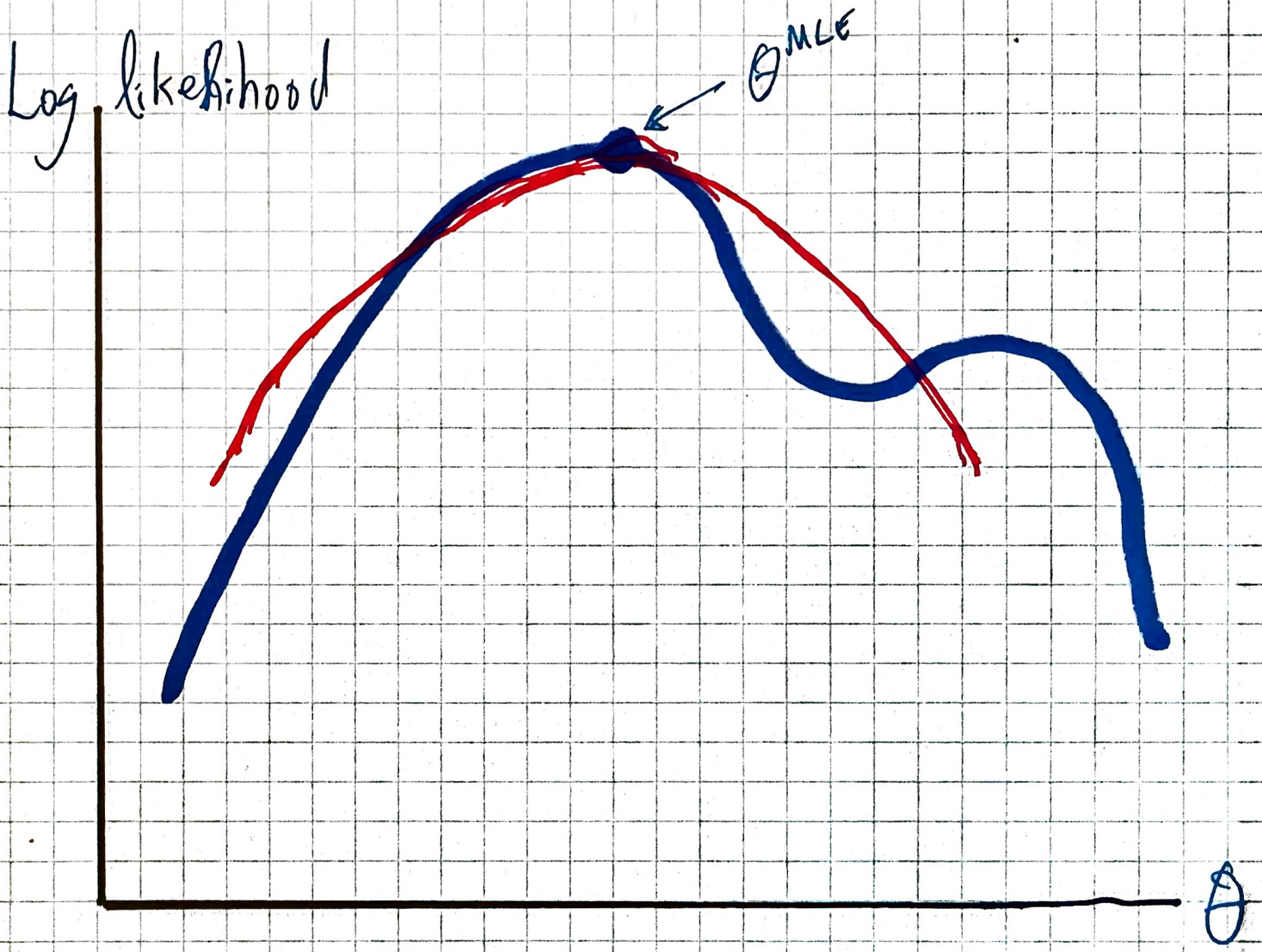
- Repeat the experiment many times
  - Pick true parameters
  - Generate synthetic data sets of various sizes
  - Apply procedures
  - Record results and fit to some class of distributions
  - Who needs theoretical econometricians?
- Problem: we aren't allowed to get the required computer power
  - We didn't build the bomb
  - Current users do not want new users
- Econometricians to the rescue
  - They develop theories
  - They use asymptotic properties to derive useful statistical tests

Problem 1

- Asymptotically, we are all dead

Problem 2

- We have finite sample problems during our finite life


Log likelihood
θ^{MLE}
θ

Today

- Review basic statistics to remind ourselves of the subtle differences in concepts
- Describe Reich-Judd approach which avoids some of the approximations typically used
- Describe application to ... what else .... Zurcher bus model

Statistics and estimates

- Let $f(x;\theta)$ denote a family of probability masses or density functions over S – potentially multivariate – parameterized by θ. Suppose the random variables $X^1, \ldots, X^n$ are independently and identically distributed according to some $f(x;\theta_0)$ for some θ_0. Let $X^{1:n} = (X^1, \ldots, X^n)$ denote a collection of random variables for some n. The data matrix $x^{1:n} = (x^1, \ldots, x^n)$ is called a *realization* of the *random sample of size n*
- Consider a real function $h(\cdot)$; the random variable $T_n = h(X^{1:n})$ with realization $t_n = h(x^{1:n})$ is called a *statistic*. If a sequence of statistics, T_n, is used to infer an unknown parameter θ, it is called an *estimator*; when appropriate, it can be denoted by $\hat{\theta} = \hat{\theta}(X^{1:n})$. A concrete value for such an estimator based on $x^{1:n}$ is called an *estimate*, either denoted by t or θ.

Standard error and confidence interval

- Let T_n be an estimator of θ, and V be a consistent estimator of its variance $\text{var}(T_n)$. Then, the *standard error*, $(T) \equiv \sqrt{V}$, is a consistent estimator of its standard deviation $\sqrt{\text{var}(T)}$.
- Given a fixed $\gamma \in (0, 1)$, the two statistics $T_{n,l}$ and $T_{n,b}$ form the boundaries of a $\gamma \cdot 100\%$ *confidence interval* if $P(T_{n,l} \leq \theta \leq T_{n,l}) = \gamma \; \forall \theta \in \Theta$; γ is called the *confidence level*, or alternatively the *coverage probability*.
- Comments
  - The main difficulty with standard errors is obtaining a consistent estimator V of the variance of the estimator T_n
  - Finding a statistic that fulfills the coverage condition is generally nontrivial. Most of the time, general statistics that rely on asymptotics will be used.
  - The correct interpretation of a confidence interval is that if the random sampling in the population were to be repeated, $\gamma \cdot 100\%$ of the confidence intervals obtained would cover the true parameter θ.
  - It is *not* correct to say that given a sample, the confidence interval contains the true parameter with $\gamma \cdot 100\%$ probability, as there is no randomness involved anymore once the sample is taken.

z-Statistic, Asymptotic Normality, Wald Confidence Interval

- If T_n is a consistent estimator for θ, the z-statistic will—under appropriate regularity conditions—by asymptotically (standard) normal distributed:

$$Z(\theta) \equiv \frac{T_n - \theta}{(T_n)}(0, 1).$$

- The two statistics $T_n \pm z_{\frac{1+\gamma}{2}}$ form the boundary of an approximate $\gamma \cdot 100\%$ confidence interval, also referred to as the Wald confidence interval, where z is the corresponding quantile of the standard normal distribution.

Likelihood function

- The *likelihood (function)* is the joint probability or the joint density of the data, given a particular value of the parameter, written as a function of the parameter (fixing the data): $L(\theta; x) \equiv f(x; \theta)$
- The *maximum likelihood estimate* is defined as $\theta^{ML} = \arg\max_{\theta \in \Theta} L(\theta; x)$, and the *maximum likelihood estimator* as $\hat{\theta}^{ML} = \arg\max_{\theta \in \Theta} L(\theta; X)$; both objects might be abbreviated by MLE.
- Due to the independence of the draws, the likelihood function for the sample is the product of the individual likelihoods: $L(\theta; x^{1:n}) = \prod_1^n f(x^i; \theta)$.
- Every monotone transformation of L has the same extremal values
  - We often use the natural logarithm of the likelihood, called the *log-likelihood*: $l(\theta; x) \equiv \log(L(\theta; x))$.
  - Since the log of a product is a sum, maximizing the log-likelihood avoids problems of underflow.

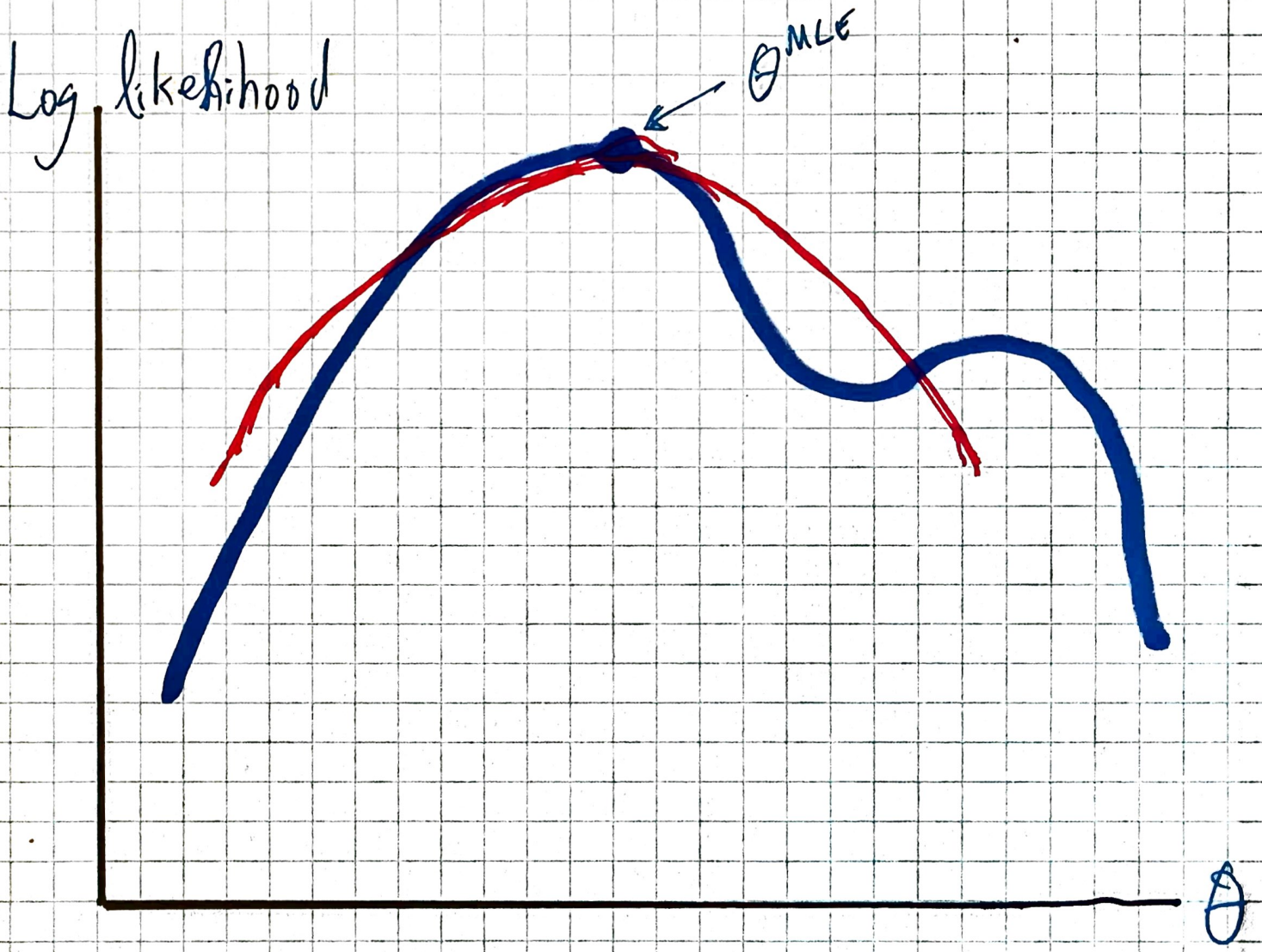
Relative likelihood

- The relative likelihood is defined by $\tilde{L}(\theta; x) = \frac{L(\theta;x)}{L(\hat{\theta}_{ML};x)}$. In particular, $0 \leq \tilde{L}(\theta; x) \leq 1$.
- The following definitions give names to the first and second derivatives of the log-likelihood function:
  - *Score function*: $S(\theta; X) \equiv \frac{dl(\theta;x)}{d\theta}$
  - *(Ordinary) Fisher information*: $I(\theta; X) \equiv -\frac{d^2 l(\theta;x)}{d\theta^2} = -\frac{dS(\theta;X)}{d\theta}$
  - *Expected Fisher information*: $\mathbb{E}(I(\theta_0; X))$, where the expectation is taken with respect to X (this implies that the expectation is integrated with against $f(X, \theta_0)$ at the true parameter value).
  - *Observed Fisher information*: $I(\hat{\theta}^{ML}; X^{1:n})$ (at the ML estimator)
- Asymptotic Normality of ML Estimator Suppose $\hat{\theta}^{ML}$ is a consistent estimator for the true parameter θ_0, and the Fisher regularity conditions hold. Then,

$$\sqrt{n \cdot J(\theta_0)}(\hat{\theta}^{ML} - \theta_0)(0, 1)$$

Wald statistics and confidence interval

- To test $H_0 : \theta_0 = \tilde{\theta}_0$, the *Wald statistic* is defined by $\sqrt{I(\hat{\theta}^{ML}; X^{1:n})}(\hat{\theta}^{ML} - \tilde{\theta}_0)$, which is asymptotically (standard) normal distributed.
- The bounds of the $\gamma \cdot 100\%$ *Wald confidence interval* are obtained as $\hat{\theta}^{ML} \pm z_{\frac{1+\gamma}{2}}(\hat{\theta}^{ML})$
- The Wald confidence interval is generally considered to be "too large" for a given γ.
- It is not invariant to non-linear transformations because the Wald statistic is based on a second order approximation of likelihood, and does not involve the likelihood function itself

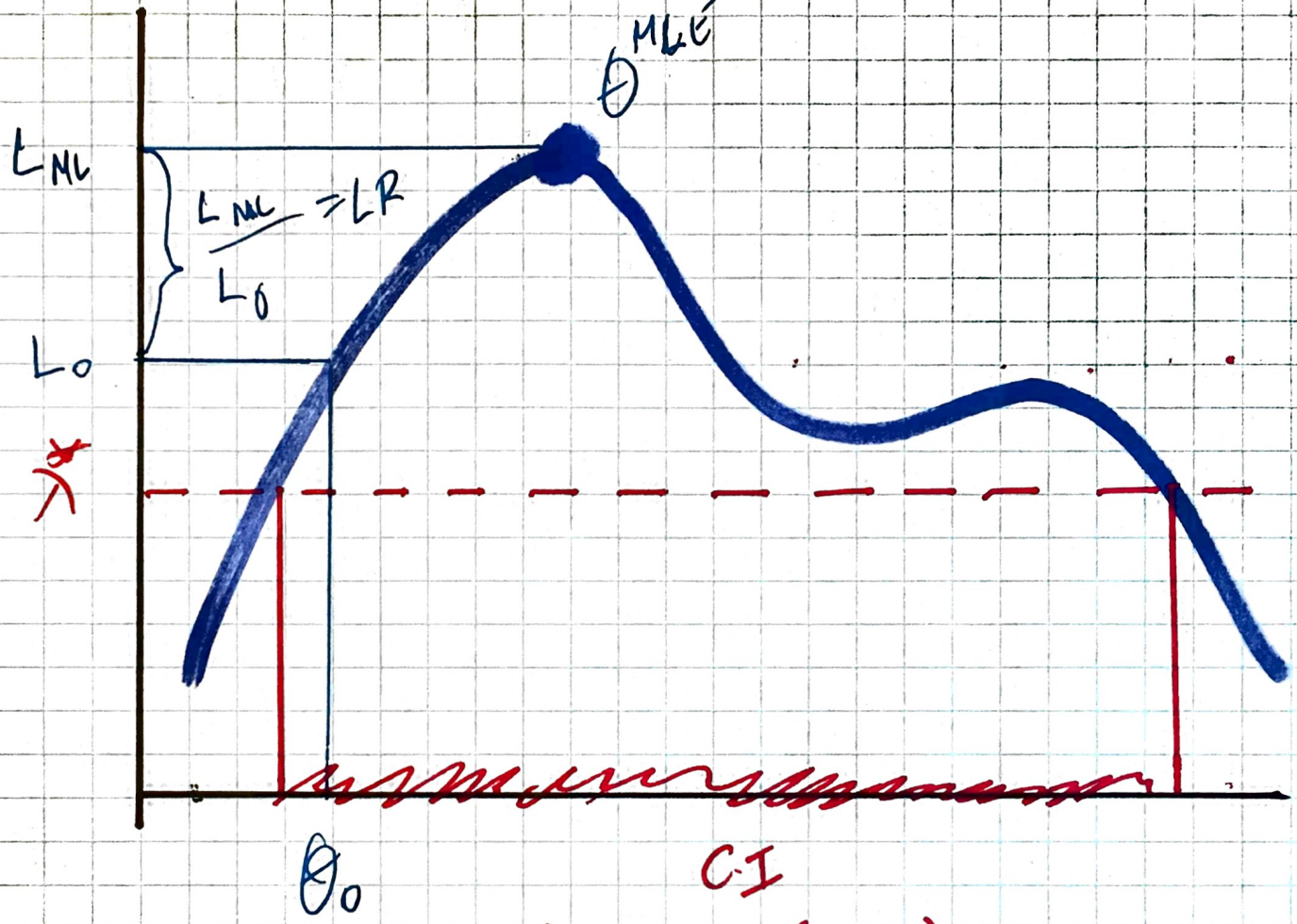

Log likelihood
θ^{MLE}
θ

Likelihood ratio statistics

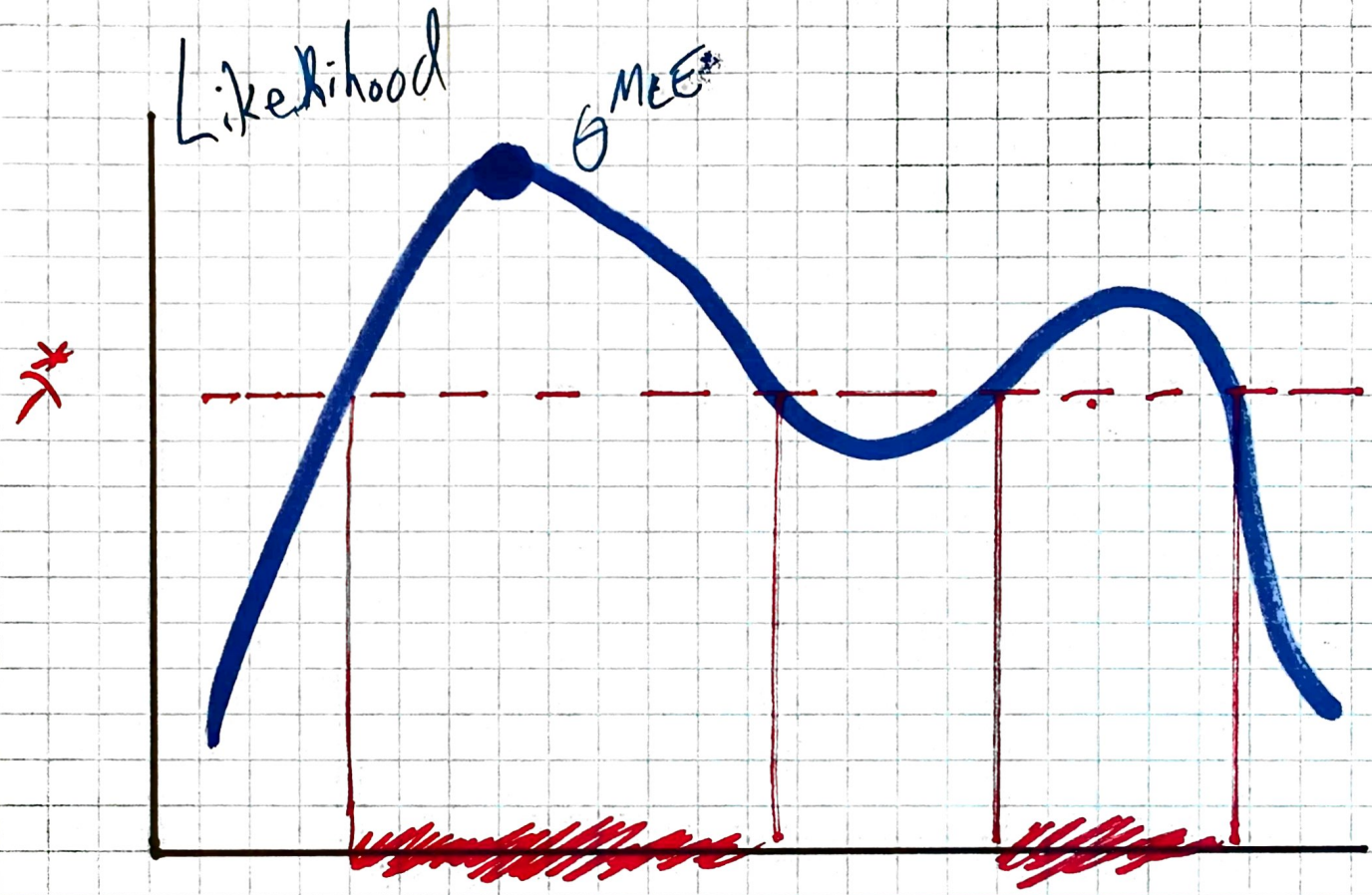
- The likelihood ratio statistic asymptotically follows a Chi-squared distribution with one degree of freedom:

$$-2(l(\hat{\theta}^{ML}; X) - l(\theta_0; X)) \equiv -2\tilde{l}(\theta_0; X)\chi^2(1).$$

- The set $\left\{\theta : \tilde{l}(\theta; X) \geq -0.5\chi^2_\gamma(1)\right\}$ forms the $\gamma \cdot 100\%$ *likelihood ratio (LR) confidence interval* for θ, where $\chi^2_\gamma(1)$ is the corresponding quantile of the Chi-squared distribution with one degree of freedom.
- The likelihood ratio confidence interval defines a *manifold*.
  - Numerical methods are required to approximate its *boundary*
  - In one dimension, finding the boundary boils down to finding an even number (usually 2) of solutions to a one dimensional equation.
  - Computing these confidence intervals as the solution to the likelihood ratio statistic equaling a quantile of the Chi-squared distribution is also referred to as *test inversion*, because one seeks the one value of the likelihood ratio such that the inequality holds strictly.
- *Wilk's theorem* generalizes this to multiple dimensions



$$\lambda^* = \Lambda(LR)$$


Likelihood
θ^{MLE}
y^*

Reich and Judd idea

- Let's compute the likelihood level set of the Chi-squared quantile
- Yes, let's compute manifolds
- Today we stay with one dimensional manifolds
- Perhaps you will see the multidimensional version next year

The Likelihood Ratio Test

- Setup
  - Model $\mathcal{M}$: Structural parameters $\theta \in \Theta$, states $x \in \mathcal{S}$, "outcomes" $y \in \mathcal{Y}$, policy/endogenous variables $\sigma \in \Sigma$
  - Model solution conditions $h(x; \sigma, \theta) = 0, \ \forall x \in \mathcal{S}$
  - Data set $\{\hat{x}_t, \hat{y}_t\}_{t=1}^T$
  - Log-likelihood function $L(\theta; \sigma) \equiv \log(P_{\mathcal{M}}(\{\hat{x}_t, \hat{y}_t\}_{t=1}^T; \sigma, \theta))$
- Estimation of θ (here: MPEC, but "nesting" NFXP):

$$\hat{\theta}, \hat{\sigma} = \arg \max_{\theta \in \Theta, \sigma \in \Sigma} L(\theta; \sigma)$$
$$\text{s.t. } h(x; \sigma, \theta) = 0, \ \forall x \in \mathcal{S}$$

- Likelihood ratio test
  - Hypothesis function: $\tau : \Theta \to \mathbb{R}$, $\tau \in \mathcal{C}^1$
  - Hypotheses: $H_0 : \tau(\theta) = 0$ against $H_1 : \tau(\theta) \neq 0$ (two-sided)
  - Test statistic: If H_0 is true, $2(L(\hat{\theta}; \hat{\sigma}) - L(\theta_0; \sigma_0)) \overset{a}{\sim} \chi_1^2$, where

$$\theta_0, \sigma_0 = \arg \max_{\theta \in \Theta, \sigma \in \Sigma} L(\theta; \sigma)$$
$$\text{s.t. } h(x; \sigma, \theta) = 0, \ \forall x \in \mathcal{S}$$
$$\tau(\theta) = 0$$

Test Inversion and Confidence Intervals

- Set of hypothesis values a which would *not* be rejected, given $L(\hat{\theta};\hat{\sigma})$

$$\mathcal{A}^{\alpha} \equiv \{a \in \mathbb{R} : \exists \theta, \sigma : h(x;\sigma,\theta) = 0 \text{ and } H_0 : \tau(\theta) = a \text{ not rejected at level } \alpha\}$$

  - Convex hull: $\mathcal{A}^{\alpha} \subseteq [\min(\mathcal{A}^{\alpha}), \max(\mathcal{A}^{\alpha})] \equiv [\underline{a}, \bar{a}]$
  - $\mathcal{A} \neq \emptyset$ because $\tau(\hat{\theta}) \in \mathcal{A}^{\alpha}$; not a singleton if $L \in \mathcal{C}^0$ and $\alpha > 0$
- Computation of $\underline{a}$ ($\bar{a}$ analogously as max problem, or min $-\tau(\theta)$):

$$\begin{aligned} \hat{\underline{a}} = \min_{\theta \in \Theta, \sigma \in \Sigma} & \tau(\theta) \\ \text{s.t. } & h(x;\sigma,\theta) = 0, \ \forall x \in \mathcal{S} \\ & L(\theta;\sigma) \geq L(\hat{\theta};\hat{\sigma}) - 0.5\chi^2_1(1-\alpha) \end{aligned}$$

- $\mathcal{A}^{\alpha}$ forms a $(1-\alpha)\cdot 100\%$ confidence interval for $\tau(\theta)$
  - In repeated sampling experiments and estimations of θ, $\mathcal{A}^{\alpha}$ would contain the "true" value of θ in $(1-\alpha)\cdot 100\%$ of the times
  - "Duality of hypothesis testing and confidence intervals"
  - Dimension-wise confidence intervals of θ using $\tau : \theta \mapsto \theta_k$

The Bus Engine Replacement Model (Rust, 1987)

- Dynamic machine renewal problem
  - Payoff function

$$u(x, i; \theta) + \varepsilon(i) = \begin{cases} \theta_{RC} + \varepsilon(1) & i = 1 \\ \theta_1 \cdot x + \varepsilon(0) & i = 0 \end{cases}$$

  - Law of motion of the states:
    - $Pr(x'|x, i; \theta)$, with $Pr(x' < x|i = 0; \theta) = 0$ and $Pr(x' = 0|x, i = 1; \theta) > 0$
    - $\varepsilon \sim EV1$ i.i.d.
  - (Integrated) Bellman equation

$$\begin{aligned} EV(x, i) &\equiv \mathbb{E}[V(x', \varepsilon')|x, i] \\ &= \iint \max\{u(x', i'; \theta) + \varepsilon'(i') + \beta EV(x', i')\} Pr(x'|x, i; \theta) q(\varepsilon') d\varepsilon' dx' \\ &\equiv T[EV; \theta](x, i) \end{aligned}$$

- Estimate θ from data $\{x_t, i_t\}_{t,i}$ (here: MPEC, but "nesting" NFXP)

$$\begin{aligned} \hat{\theta}, \widehat{EV} = \arg \max_{\theta \in \Theta, EV} & L(\theta; EV) \\ \text{s.t. } & EV(x, i) = T[EV; \theta](x, i), \ \forall x \in \mathcal{S}, i \in \{0, 1\} \end{aligned}$$

Confidence Intervals

- $(1-\alpha)\cdot 100\%$ Confidence intervals for $\tau = (\theta_{RC}, \theta_1, \theta_{RC}/\theta_1)$ (and $-\tau$)

$$\min_{\theta \in \Theta, EV} \tau_k$$
$$\text{s.t. } EV(x,i) = T[EV_\theta; \theta](x,i),\ \forall x \in \mathcal{S}, i \in \{0,1\}$$
$$L(\theta; EV) \geq L(\hat{\theta}; \widehat{EV}) - 0.5\chi^2_1(1-\alpha)$$

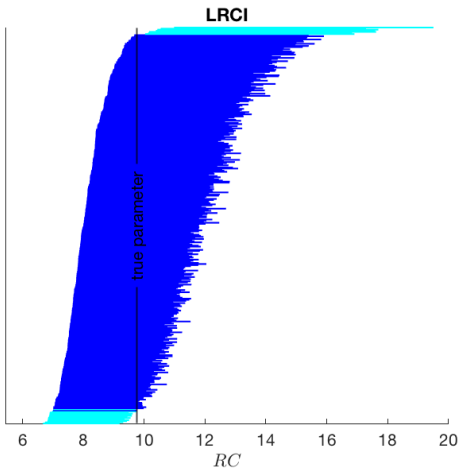
- Coverage analysis:
  - Simulate data sets under $\tilde{\theta}$
  - Estimate $\hat{\tilde{\theta}}$ and its confidence intervals
  - Check for inclusion of $\tilde{\theta}$
- Comparison:
  - Two different data set sizes (8,112 and 780)
  - Various types of confidence intervals
    - Likelihood ratio confidence intervals (LRCI)
    - Wald/SE (with delta method for mapped parameters)
    - Bootstrapping (sample quantiles)

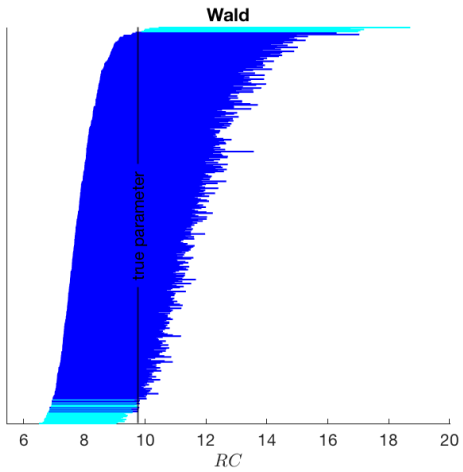
Confidence Intervals: Coverage Analysis (1)

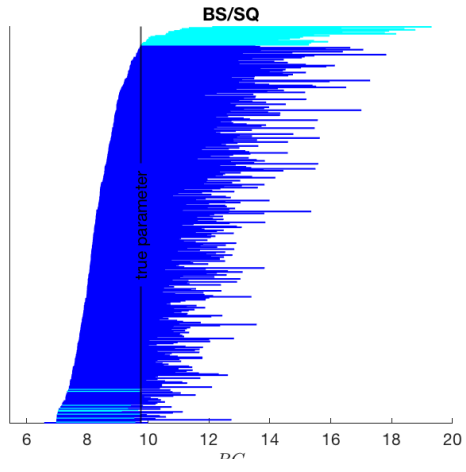
	LRCI					
	Sample size: 8,112			Sample size: 780		
	coverage	min	max	coverage	min	max
θ_{RC}	**0.961**	6.465	21.77	**0.958**	4.333	153.7
θ_1	**0.953**	0.558	7.888	**0.938**	7e-16	73.33
θ_{RC}/θ_1	**0.942**	2.348	12.07	**0.911**	1.305	4e07
	Wald/SE (with delta method)					
θ_{RC}	**0.952**	6.367	20.85	**0.955**	**-42.53**	132.8
θ_1	**0.928**	0.450	7.404	**0.935**	**-22.60**	61.00
θ_{RC}/θ_1	**0.962**	2.212	10.30	**0.791**	**-8e04**	8e04
	Bootstrap (sample quantiles)					
θ_{RC}	**0.928**	5.736	20.56	**0.675**	4.709	350.0
θ_1	**0.939**	0.273	7.723	0.813	1e-12	167.4
θ_{RC}/θ_1	**0.939**	2.231	11.11	0.880	1.181	5e12

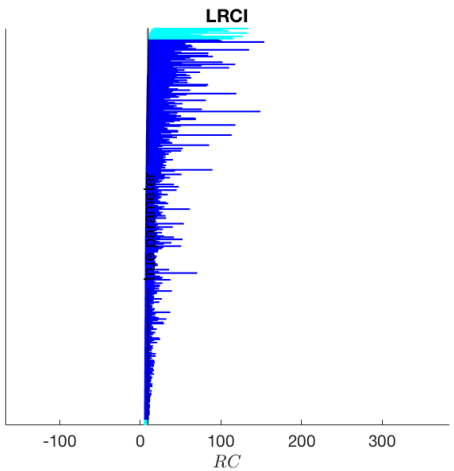
	LRCI	Wald	Bootstrap
time (sec)	288	12	6,305

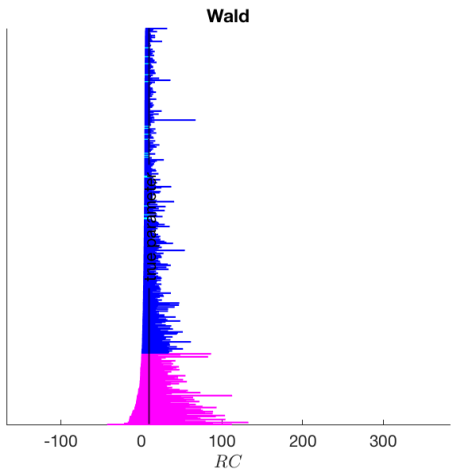
Confidence Intervals: Coverage Analysis (2)

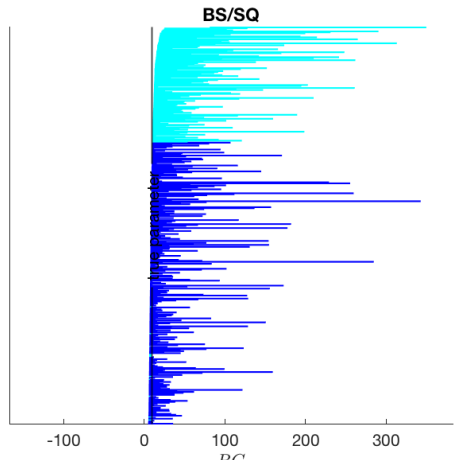












Counter-Factuals: Demand Estimation in Rust (1987)

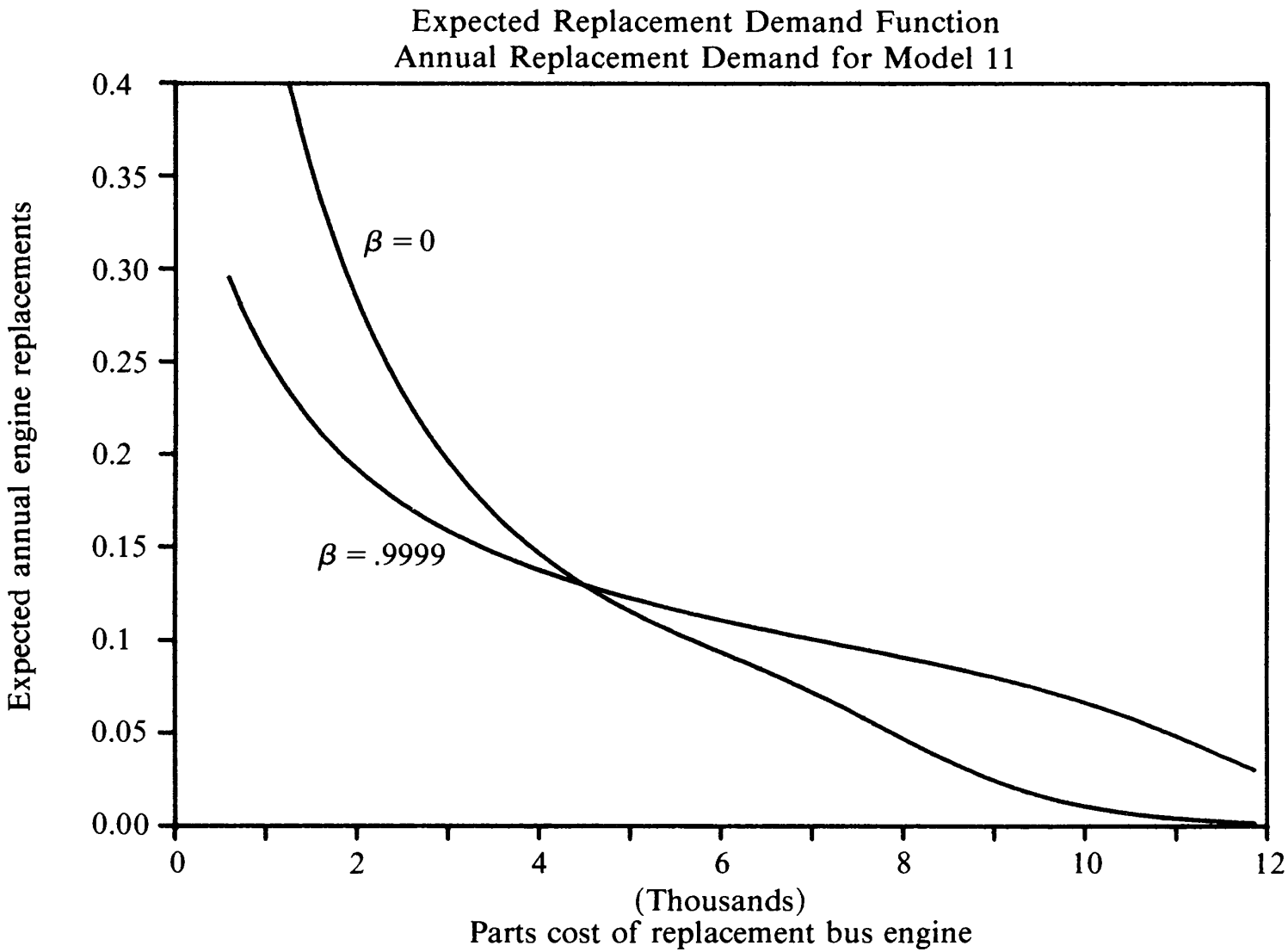
- Counter-factual: Use *estimated* model to carry out "policy experiments", e.g. by simulating/integrating the model variants to obtain and compare some derived quantity.
  - Assumption: Structural parameters are *policy-invariant*.
  - Goal: Analyze how estimation error propagates to derived quantities.
- **Counter-factual is a map of the parameters**, but its derivative is not always straightforward to compute (needed for delta method)
- Demand function estimation in Rust (1987)

$$d(\theta_{RC}) \equiv \int \pi_\theta(x, i = 1) dx$$

where the stationary distribution is defined as

$$\pi(x, i) = \iint Pr(i|x; EV_\theta) Pr(x|x', i'; \theta) \pi(x', i') dx' di',$$

Demand Curve in Rust (1987)



Confidence Intervals for Demand Curve (1)

- Confidence interval for $d(\theta_{RC})$ (θ_{RC} fix)

$$\underline{\hat{d}}(\theta_{RC}) = \arg \min_{\theta_1, \tilde{\theta}_{RC}, \pi, EV, \widetilde{EV}} \int \pi(x, i=1) dx$$

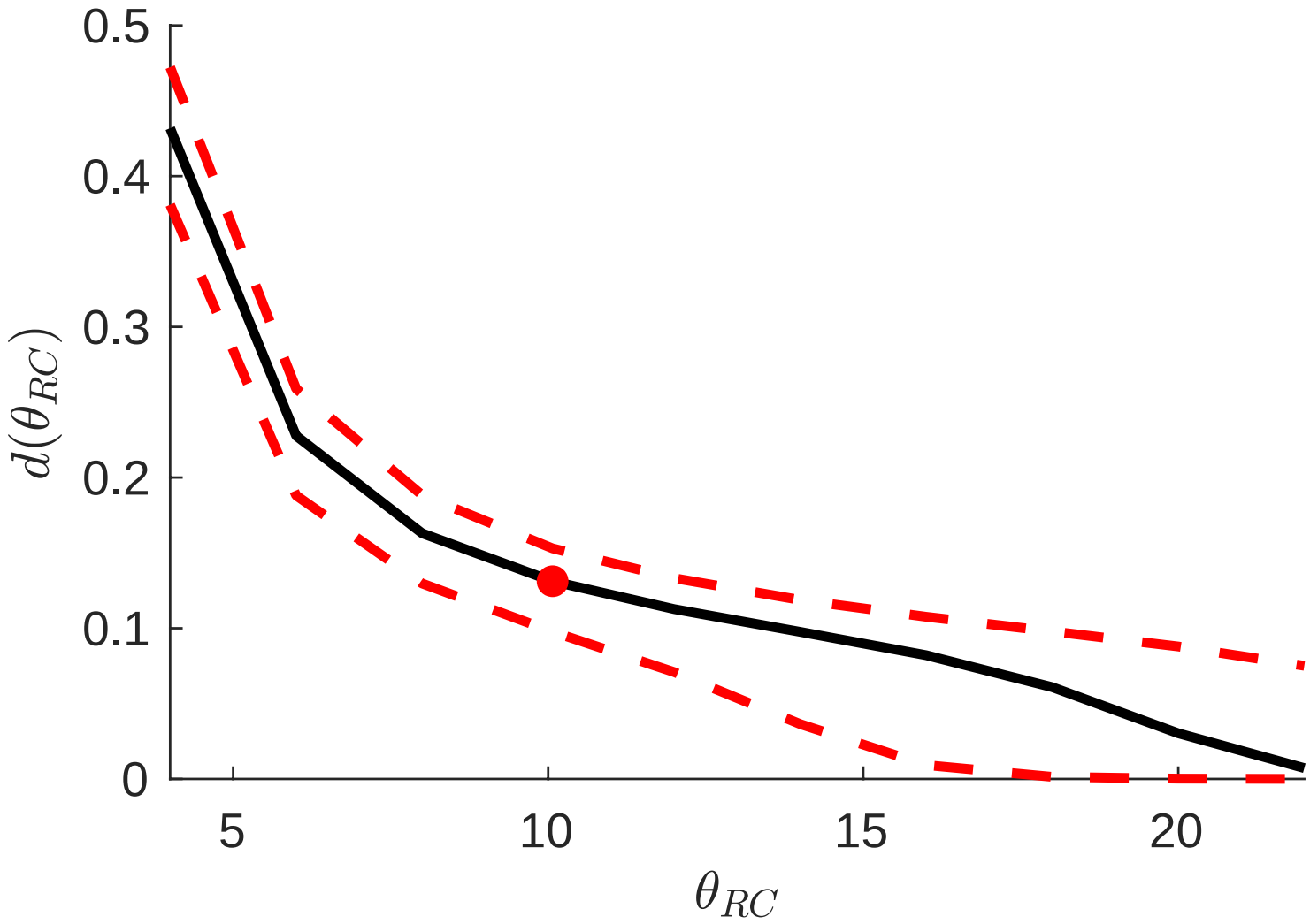
$$\text{s.t. } \pi(x,i) = \iint Pr(i|x; EV) Pr(x|x', i'; \theta_{RC}, \theta_1) \pi(dx', di'), \forall x, i$$

$$EV(x,i) = T[EV; \theta_{RC}, \theta_1](x,i), \ \forall x, i$$

$$\widetilde{EV}(x,i) = T[\widetilde{EV}; \tilde{\theta}_{RC}, \theta_1](x,i), \ \forall x, i$$

$$L(\tilde{\theta}_{RC}, \theta_1; \widetilde{EV}) \geq L(\hat{\theta}; \widehat{EV}) - 0.5\chi^2_1(1-\alpha)$$

Confidence Intervals for Demand Curve (2)



Conclusions

- We propose an efficient and easy-to-implement way to compute likelihood ratio confidence intervals (LRCI) for structural parameters—and mappings thereof—using constrained optimization
- We demonstrate that LRCI have very competitive coverage properties, in particular for mappings and smaller data sets; runtime performance is somewhere in between standard error based CIs and bootstrapping approaches
- We demonstrate the applicability to counter-factuals—a specific kind of mapping—which would otherwise be hard to assess for estimation error