

# Score, Wald, and Likelihood Ratio

Patrick Breheny

October 29, 2025

# Introduction

- We've now covered the most important theoretical properties of the MLE: it is consistent, asymptotically normal, and efficient
- Today, we turn our attention to a different problem: likelihood-based inference
- Specifically, we will go beyond the likelihood as a mechanism for simply producing point estimates and look at how we can use the likelihood function to construct (frequentist) confidence intervals and carry out hypothesis tests

# The holy trinity

- There are three widely used approaches for carrying out likelihood-based inference:
  - Wald (Abraham Wald)
  - Score (C.R. Rao)
  - Likelihood ratio (Jerzy Neyman / Egon Pearson / Samuel Wilks)
- We'll be discussing all three approaches, and considering two different scenarios:
  - Simple null hypotheses:  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$
  - Composite null hypotheses:  $H_0 : \boldsymbol{\theta} \in \Theta_0$

# Nuisance parameters

- The second case is particularly important in the multivariate setting, as we are usually interested in testing something like  $H_0 : \theta_j = 0$ , which means  $H_0 : \boldsymbol{\theta} \in \{\boldsymbol{\theta} : \theta_j = 0\}$
- So, to be more specific, we won't necessarily consider composite null hypotheses in their full generality, but rather focus on the setting where  $\boldsymbol{\theta}$  can be divided into parameters of interest,  $\boldsymbol{\theta}_1$ , and nuisance parameters,  $\boldsymbol{\theta}_2$ , with  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top \boldsymbol{\theta}_2^\top)^\top$ , with  $r$  denoting the length of  $\boldsymbol{\theta}_1$  and  $d - r$  the length of  $\boldsymbol{\theta}_2$
- Our composite tests, then, will be of the form  $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ , with  $\boldsymbol{\theta}_2$  left unspecified by the null hypothesis
- (I'm describing these ideas in terms of tests, but everything applies to confidence intervals as well)

# Wald approach

- The Wald approach is perhaps the simplest to understand
- It is based on the result that  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}^{-1}(\boldsymbol{\theta}^*))$  and simply uses the standard tools for the normal distribution to carry out inference
- **Proposition:** If consistency assumptions (A)-(D)<sup>1</sup> hold,

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathcal{J}_n(\boldsymbol{\theta}^*) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \chi_d^2$$

- This can be inverted to find confidence regions for  $\boldsymbol{\theta}$

---

<sup>1</sup>If one assumes (A)-(C) only, the result still holds, but for the consistent sequence of roots (which may or may not be the MLE); this applies to all of the theorems in this lecture

## Which information?

- As alluded to previously, we could use either the Fisher or expected information here and the result would still hold
- In fact, we have even more choices; all of the following hold:

$$(\hat{\theta} - \theta^*)^\top \mathcal{J}_n(\theta^*) (\hat{\theta} - \theta^*) \xrightarrow{d} \chi_d^2$$

$$(\hat{\theta} - \theta^*)^\top \mathcal{J}_n(\hat{\theta}) (\hat{\theta} - \theta^*) \xrightarrow{d} \chi_d^2$$

$$(\hat{\theta} - \theta^*)^\top \mathcal{I}_n(\theta^*) (\hat{\theta} - \theta^*) \xrightarrow{d} \chi_d^2$$

$$(\hat{\theta} - \theta^*)^\top \mathcal{I}_n(\hat{\theta}) (\hat{\theta} - \theta^*) \xrightarrow{d} \chi_d^2$$

$$(\hat{\theta} - \theta^*)^\top \mathbf{V}_n(\hat{\theta}) (\hat{\theta} - \theta^*) \xrightarrow{d} \chi_d^2,$$

where  $\mathbf{V}_n(\theta) = \sum_i \mathbf{u}_i(\theta) \mathbf{u}_i(\theta)^\top$

- In practice, Wald approaches typically use  $\mathcal{I}_n(\hat{\theta})$  or  $\mathcal{J}_n(\hat{\theta})$

## Nuisance parameters

- Testing  $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$  is also rather straightforward with the Wald approach
- **Proposition:** If (A)-(D) hold and  $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1^*$  (i.e., if  $H_0$  is true), then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{V}_{11}),$$

where  $\mathcal{V}_{11} = \mathcal{I}^{11}$  is the (1,1) block of the inverse of  $\mathcal{I}(\boldsymbol{\theta}^*)$

- Again, recall that  $\mathcal{V}_{11}^{-1} = \mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}$ , so that  $\mathcal{V}_{11}^{-1} \preceq \mathcal{I}_{11}$  and  $\mathcal{V}_{11} \succeq \mathcal{I}_{11}^{-1}$ ; the presence of unknown nuisance parameters increases the variance of our estimator

## Wald confidence intervals

- If our parameter of interest is a scalar, then we have simple closed-form expressions for confidence intervals:

$$\hat{\theta}_j \pm z_{1-\alpha/2} \sqrt{\mathcal{V}_{jj}^n(\hat{\boldsymbol{\theta}})}$$

is an approximate  $1 - \alpha$  confidence interval for  $\theta_j$

- Again, this is *not* the same thing as

$$\hat{\theta}_j \pm \frac{z_{1-\alpha/2}}{\sqrt{\mathcal{I}_{jj}^n(\hat{\boldsymbol{\theta}})}};$$

this second approach is incorrect, as it fails to account for the impact of nuisance parameters and produces confidence intervals that are too narrow

## Remarks on the Wald approach

- The ease with which confidence intervals can be constructed is the primary advantage of the Wald approach
- As we will see, confidence intervals are considerably more cumbersome in the score and likelihood ratio approaches
- The primary disadvantage of the Wald approach is that it tends to provide the least accurate approximation of the three approaches

## Score approach: Simple null

- Next, let's consider the score approach: as the name implies, this method revolves around the score vector
- **Proposition:** If (A)-(C)<sup>2</sup> hold,

$$\mathbf{u}(\boldsymbol{\theta}^*)^\top \mathcal{I}_n^{-1}(\boldsymbol{\theta}^*) \mathbf{u}(\boldsymbol{\theta}^*) \xrightarrow{d} \chi_d^2$$

- Again, we can use any consistent estimator of  $\mathcal{I}(\boldsymbol{\theta}^*)$  in place of the Fisher information; score approaches typically use  $\mathcal{I}_n(\boldsymbol{\theta}_0)$  or  $\mathcal{J}_n(\boldsymbol{\theta}_0)$
- In principle, this can be inverted to find a confidence region, but in practice, doing so is usually not straightforward

---

<sup>2</sup>Don't need (D) here since the MLE doesn't appear

# Nuisance parameters

- What about testing  $H_0 : \theta_1 = \theta_0$ ?
- This is less straightforward than the Wald case
- We need to evaluate the score and information, but for what value of  $\theta$ ?
- Setting  $\theta_1 = \theta_0$  seems obvious, but for  $\theta_2$ , we are going to have to maximize the likelihood under the restriction imposed by  $H_0$

## Restricted MLEs

- Specifically, let us define the restricted, or constrained, MLE  $\hat{\theta}_2(\theta_0)$  as the value of  $\theta_2$  that maximizes  $L(\theta)$  under the restriction that  $\theta_1 = \theta_0$ , with  $\hat{\theta}_0 = (\theta_0^\top \hat{\theta}_2(\theta_0)^\top)^\top$
- The following lemma will prove useful to us (its proof is essentially identical to the case for the unrestricted MLE  $\hat{\theta}$ )
- **Lemma:** If (A)-(D) hold and  $\theta_0 = \theta_1^*$ , then

$$\sqrt{n}(\hat{\theta}_2(\theta_0) - \theta_2^*) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}_{22}^{-1});$$

note that here we *do* have convergence to  $\mathcal{J}_{22}^{-1}$ , not  $\mathcal{V}_{22}$ , as under  $H_0$ , we are not affected by uncertainty regarding  $\theta_1$

- Note that this only works if  $H_0$  is true: if it isn't,  $\hat{\theta}_2(\theta_0)$  may converge to something very different from  $\theta_2^*$

## Score test with nuisance parameters

- **Theorem:** If (A)-(D) hold and  $\theta_0 = \theta_1^*$ , then

$$\mathbf{u}_1(\hat{\theta}_0)^\top \mathcal{V}_{11}^n(\hat{\theta}_0) \mathbf{u}_1(\hat{\theta}_0) \xrightarrow{d} \chi_r^2,$$

where  $\mathcal{V}^n = \mathcal{J}_n^{-1}$

- In the special case where the parameter of interest is  $\theta_j$ , we have  $u_j(\hat{\theta}_0) \sqrt{\mathcal{V}_{jj}^n(\hat{\theta}_0)} \sim N(0, 1)$
- Unfortunately, inverting this test to obtain a confidence interval is not trivial, as every time we change  $\theta_0$ , we would need to re-solve for  $\hat{\theta}_2(\theta_0)$

## Remarks on the score approach

- The difficulty of obtaining confidence intervals is the biggest drawback of the score approach
- Conversely, it is often the easiest *test* to carry out, which is its biggest advantage
- In particular, we don't even need to solve for the MLE in order to carry out the test

## Example: Linear regression

- For example, consider the linear regression model  $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ ; for the purposes of this exercise, we'll treat  $\sigma^2$  as known
- Suppose we have fit a baseline model involving a number of covariates that we know we want to adjust for, and are considering including an additional predictor  $\mathbf{x}_j$  in the model
- The score test  $H_0 : \beta_j = 0$  is

$$z_j = \frac{\mathbf{x}_j^\top \mathbf{r}}{\sigma \sqrt{\mathbf{x}_j^\top \mathbf{x}_j - \mathbf{x}_j^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{x}_j}},$$

where  $\mathbf{r}$  is the vector of residuals from the baseline fit and  $z_j \sim N(0, 1)$  under the null hypothesis

## Example: Linear regression (cont'd)

- In particular, note that the “expensive” part of this calculation,  $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}$ , only needs to be computed once, and the rest of the calculations are simple
- This makes score tests very attractive if you are, say, carrying out a genetic association study in which you want to adjust for some baseline characteristics such as age, sex, etc., then test for associations between a clinical outcome and hundreds of thousands of genetic markers
- To apply the Wald or likelihood ratio tests, we would need to fit hundreds of thousands of models; the score tests involve dramatically less computational burden

# Likelihood ratio approach

- Finally, let's consider the likelihood ratio approach
- **Theorem:** If (A)-(D) hold, then

$$2 \log \frac{L(\hat{\boldsymbol{\theta}})}{L(\boldsymbol{\theta}^*)} \xrightarrow{d} \chi_d^2$$

- Note that the likelihood ratio test does not involve calculating any derivatives (score or information), only the likelihood function itself

# LRT with nuisance parameters

- Like the score test, when nuisance parameters are involved we must solve for restricted MLEs
- **Theorem (Wilks):** If (A)-(D) hold and  $\theta_0 = \theta_1^*$ , then

$$2 \log \frac{L(\hat{\theta})}{L(\hat{\theta}_0)} \xrightarrow{d} \chi_r^2$$

- Again, this can be inverted to find confidence intervals for  $\theta_j$  (a root-finding problem), but this involves repeatedly re-solving for  $\hat{\theta}_0$

## Example: Gamma distribution

- As an example of how all these tests work, let's apply them to the gamma distribution
- As you have already derived on assignment 8,

$$\mathbf{u} = \begin{bmatrix} n \log \beta - n\psi_0(\alpha) + \sum \log x_i \\ n\alpha/\beta - \sum x_i \end{bmatrix}$$
$$\mathcal{I}_n = \begin{bmatrix} n\psi_1(\alpha) & -n/\beta \\ -n/\beta & n\alpha/\beta^2 \end{bmatrix}$$

- Let's derive confidence intervals for the rate parameter  $\beta$  (you may recall that  $\beta^* = 1$  and  $\hat{\beta} = 1.66$ )

## Wald approach

- First, the Wald approach
- The diagonal element of  $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$  corresponding to  $\beta$  is 0.118, so an approximate 95% confidence interval is given by

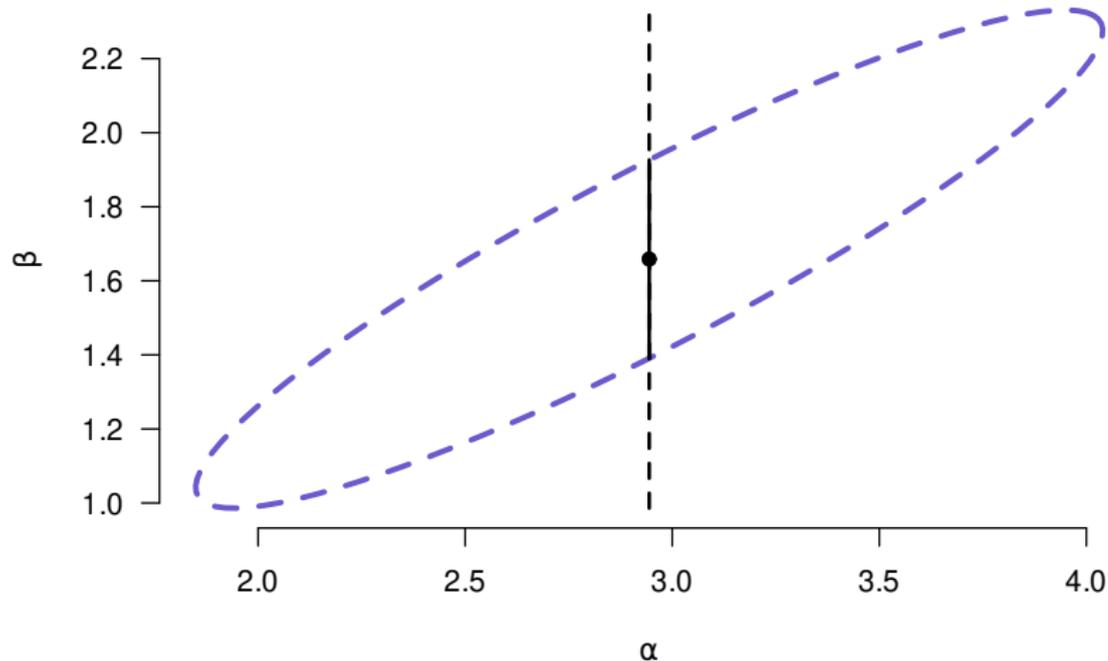
$$\hat{\theta}_2 \pm z_{1-\alpha/2} \sqrt{\mathcal{V}_{22}(\hat{\boldsymbol{\theta}})} = (0.99, 2.33)$$

- Note that this is much wider than the incorrect interval we get from just inverting  $\mathcal{I}_{22}(\hat{\boldsymbol{\theta}})$ :

$$\hat{\theta}_2 \pm \frac{z_{1-\alpha/2}}{\sqrt{\mathcal{I}_{22}(\hat{\boldsymbol{\theta}})}} = (1.39, 1.93);$$

as we have said several times, this second interval does not account for uncertainty in  $\alpha$

## Wald: Correct and incorrect intervals



# Score

- Obtaining score intervals for  $\beta$  is considerably more computer-intensive, as we must repeatedly solve for  $\hat{\alpha}(\beta)$ , the MLE of  $\alpha$  under the constraint that the rate is equal to  $\beta$
- The endpoints of the confidence interval, then, can be found by finding the two solutions of

$$u_2(\hat{\alpha}(\beta), \beta)^2 \mathcal{V}_{22}^n(\hat{\alpha}(\beta), \beta) = \chi_{1,1-\alpha}^2$$

- This yields the confidence interval (0.99, 2.33); not identical to the Wald interval, but equal up to 2 decimal places
- Again, failing to account for uncertainty by using the MLE  $\hat{\alpha}$  instead of the restricted MLE  $\hat{\alpha}(\beta)$  produces an interval that is much too narrow: (1.39, 1.93)

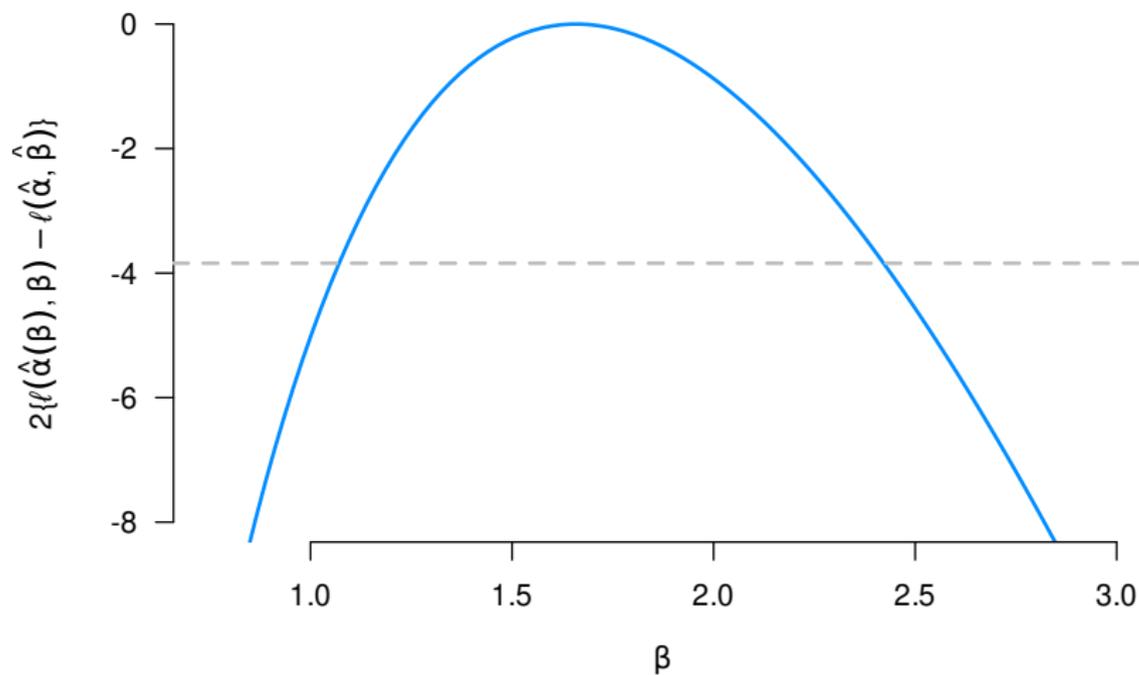
# Likelihood ratio

- Similarly, finding the endpoints of the likelihood ratio confidence interval involves finding the roots of

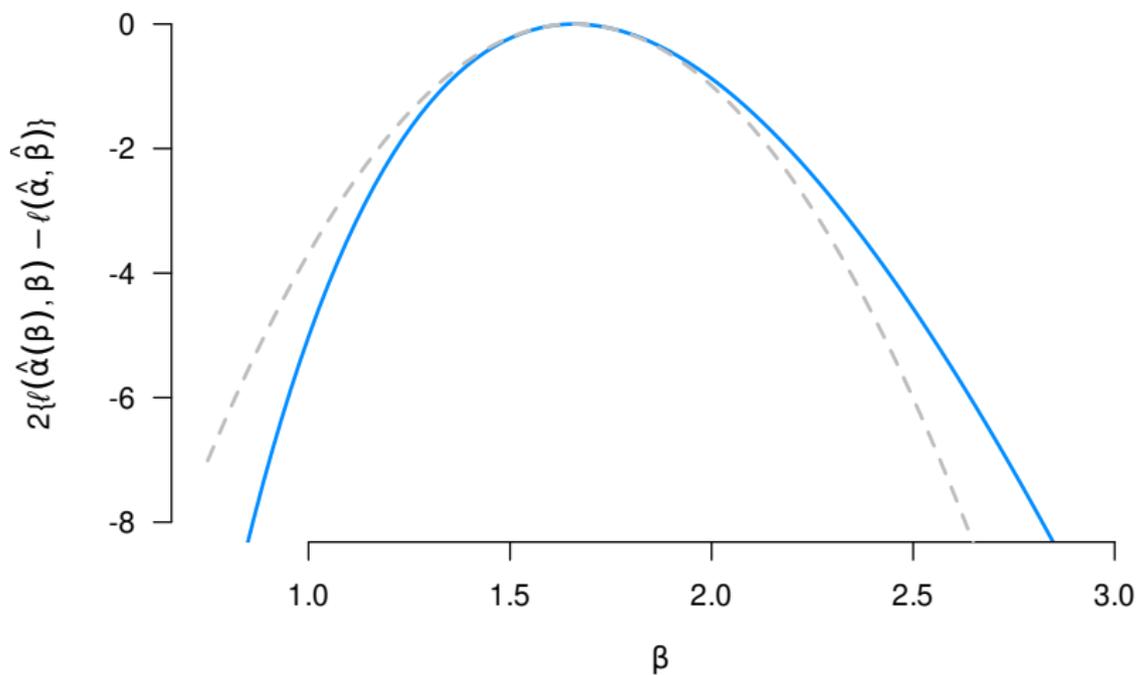
$$2\{\ell(\hat{\alpha}, \hat{\beta}) - \ell(\hat{\alpha}(\beta), \beta)\} = \chi_{1,1-\alpha}^2$$

- This yields the interval (1.07, 2.42)

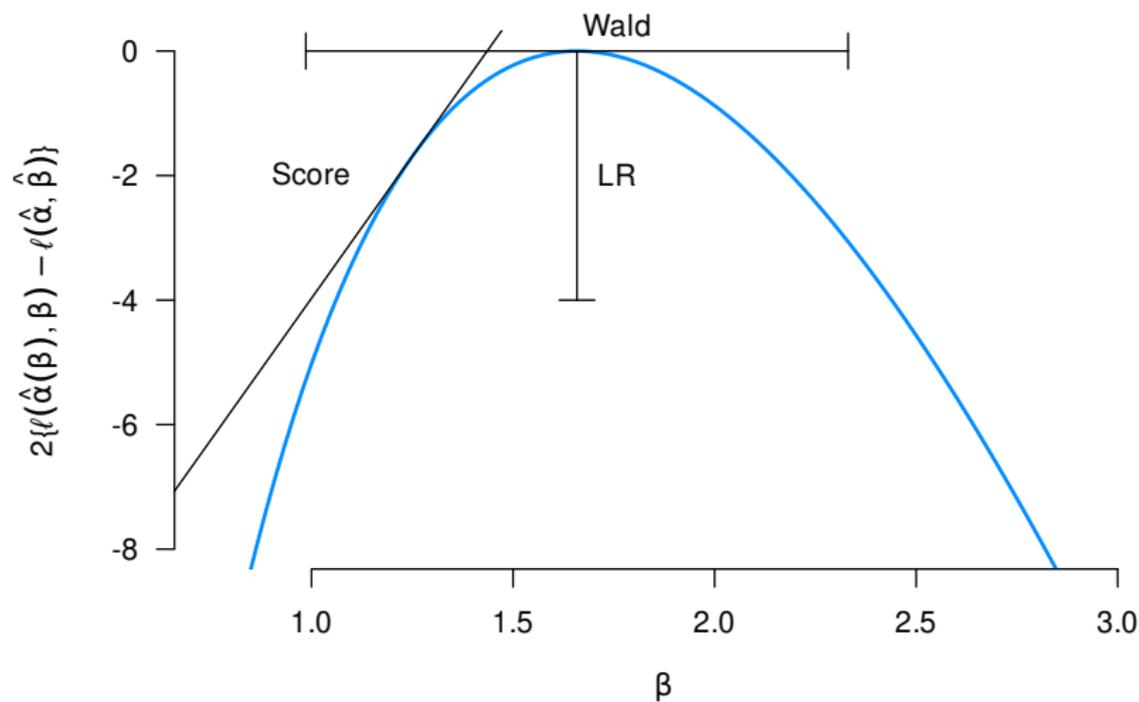
## Likelihood ratio plot



# Quadratic approximation



## Visualization of all three methods



## Final remarks

- All three approaches are asymptotically equivalent; letting  $W_n$  denote the Wald test statistic,  $S_n$  the score test statistic, and  $LR_n$  the likelihood ratio test statistic,

$$LR_n = W_n + o_p(1)$$

$$LR_n = S_n + o_p(1),$$

and indeed, all three approaches are potentially useful and widely used, depending on the context

- However, this potentially gives the wrong impression that all three approaches are equally accurate in terms of approximation inference

## Superiority of the likelihood ratio approach

- This is not true – the likelihood ratio approach is the most accurate of the three approaches
- This has been shown repeatedly in many theoretical and simulation studies, but it is also intuitive
- The Score and Wald approaches depend on derivatives, and thus, can change substantially if we reparameterize the model (e.g., if we consider  $\theta = \log \lambda$ )
- In other words, the best-case scenario for Score and Wald is that we find a normalizing transformation, in which case the results are simply equivalent to the LR
- Conversely, Score and Wald can be much worse approximations than LR if we choose a bad transformation