2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

CHAPTER 2: FREQUENTIST INFERENCE

Approaches to Inference

We distinguish three approaches to inference:

- Estimating functions.
- Likelihood.
- Bayes.

Likelihood inference can be posed as estimating function inference, but the rationale is subtly different.

Let:

- $\boldsymbol{\theta}$ be a $p \times 1$ vector of unknown parameters, and
- $\boldsymbol{Y}$ an $n \times 1$ data vector.

Approaches to Inference

Estimating Functions: A frequentist approach, with consistency being a primary objective.

A *p*-dimensional function of the parameter, i.e., of the same dimensionality as the parameter, and data, $\boldsymbol{G}(\theta, \boldsymbol{Y})$, is considered for which

$$\mathrm{E}[\boldsymbol{G}(\theta, \boldsymbol{Y})] = \mathbf{0}.$$

The expectation is over the data, which are considered random.

The estimator is then found as the solution to the **estimating equation**

$$\boldsymbol{G}(\hat{\theta}, \boldsymbol{Y}) = \mathbf{0}.$$

For inference, the frequency properties of the estimating function are derived (and this is often relatively straightforward) and these are transferred to the resultant estimator.

Approaches to Inference

Simplest example of an estimating function: $\mathsf{E}[\overline{Y} - \theta] = 0$.

Estimating equation:

$$\overline{Y} - \widehat{\theta} = 0.$$

How do we find an estimating function?

- Method of moments is one route.
- Often the estimating function is derived from a likelihood, specifically the score (derivative of log likelihood) is an estimating function.

Approaches to Inference

Likelihood:

Specify a probability model for the data $p(\boldsymbol{y}|\boldsymbol{\theta})$.

Viewed as a function of $\boldsymbol{\theta}$, this is known as the likelihood function.

The MLE is defined as that value of $\boldsymbol{\theta}$ that gives the highest probability to the observed data.

Frequency properties then derived.

Approaches to Inference

Bayesian:

Specify in addition to $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, a prior distribution $\pi(\boldsymbol{\theta})$.

Then via Bayes theorem derive the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\boldsymbol{y})}.$$

All inference follows from the posterior distribution.

Notice the differences in the three procedures, in terms of what we are required to specify.

Example: Linear Regression

Assume we have k covariates and:

(A) $\mathrm{E}[\boldsymbol{Y}|\boldsymbol{x}] = \boldsymbol{x}\beta$ and

(B) $\mathrm{var}(\boldsymbol{Y}|\boldsymbol{x}) = \sigma^2 \boldsymbol{I}_n$.

where

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal$ is $n \times 1$,
- $\boldsymbol{x}$ is $n \times (k+1)$ and
- β is $(k+1) \times 1$.

For simplicity assume σ^2 known.

Example: Linear Regression

Likelihood:

With the additional assumption of

(c) independent normal errors

we have the likelihood function

$$L(\boldsymbol{\beta}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right].$$

The log-likelihood is

$$l(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}).$$

Example: Linear Regression

Likelihood:

To maximize we calculate the *score* equation

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y} - 2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}\beta \\ &= 2\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{x}\beta).\end{aligned}$$

Hence, the MLE is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y},$$

which corresponds to an estimator with a long history, the

Least Squares (LS) Estimator.

Estimating Function:

$$\boldsymbol{G}(\beta, \boldsymbol{Y}) = \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{x}\beta).$$

Example: Linear Regression

Bayesian:

For computational convenience we assume that σ^2 is known, and the **prior** distribution

(D) $\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{m}, \sigma^2 \boldsymbol{v})$ with $\boldsymbol{m}, \boldsymbol{v}$ known.

The **posterior** distribution is then

$$\boldsymbol{\beta}|\boldsymbol{y} \sim \mathrm{N}_p\{(\boldsymbol{I} - \boldsymbol{w})\boldsymbol{m} + \boldsymbol{w}\hat{\boldsymbol{\beta}}, \boldsymbol{w}(\boldsymbol{x}^\top \boldsymbol{x})^{-1}\sigma^2\},$$

where $\boldsymbol{w} = (\boldsymbol{x}^\top \boldsymbol{x} + \boldsymbol{v}^{-1})^{-1}\boldsymbol{x}^\top \boldsymbol{x}$ so that the posterior mean is a weighted combination of the LS estimate, and the prior mean.

Limiting cases:

- $\boldsymbol{v}^{-1} \rightarrow \mathbf{0}$, $\boldsymbol{w} \rightarrow \boldsymbol{I}$, $\mathrm{E}[\boldsymbol{\beta}|\boldsymbol{y}] \rightarrow \hat{\boldsymbol{\beta}}$.
- As $n \rightarrow \infty$, $\boldsymbol{w} \rightarrow \boldsymbol{I}_{k+1}$ (under conditions on $\boldsymbol{x}^\top \boldsymbol{x}$).

Frequentist Estimation Criteria

In the frequentist view of statistics it is the behavior of procedures under **repeated sampling** that is considered.

So far as the estimation of a $p \times 1$ vector of parameters is concerned, the **sampling distribution** of a proposed estimator $\hat{\theta}(\boldsymbol{Y})$ is the relevant quantity.

For notational ease we consider a univariate parameter, the extension to the multivariate case is straightforward.

Frequentist Estimation Criteria

A fundamental criteria is:

1. **Consistency**: Consider an estimator $\hat{\theta}_n$ based on a sample of size n. Weak consistency states that as $n \to \infty$, $\widehat{\theta}_n \to_p \theta$, that is

   $$\Pr(\mid \widehat{\theta}_n - \theta \mid > \epsilon) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for any} \quad \epsilon > 0.$$

   From a frequentist viewpoint this undoubtedly seems a meaningful property.

A second meaningful frequentist quantity to consider as a criteria is the **mean squared error** which is defined (for a univariate parameter θ) as

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{E}_{Y|\theta}\left[(\hat{\theta} - \theta)^2\right] = \mathsf{var}(\hat{\theta}) + \mathsf{bias}(\hat{\theta})^2. \tag{1}$$

The following specific frequentist criteria are considered because they are mathematically tractable:

2. **Unbiased Estimation**: The search for an unbiased estimator (UE), that is for an estimator for which

$$\mathsf{E}[\hat{\theta}] = \theta.$$

Often unbiasedness is only available asymptotically as $n \to \infty$.

3. **Minimum variance unbiased estimation (MVUE)**:
   Within the class of unbiased estimators the search for that with the smallest variance.

   This leads to the concept of *efficiency* by which the variance of an estimator is judged relative to that achieved by the 'best possible' (i.e., that with the smallest variance).

   For example, the relative efficiency of two unbiased estimators $\widetilde{\theta}$ and $\widehat{\theta}$ is given by
   $$\frac{\mathrm{var}(\widetilde{\theta})}{\mathrm{var}(\widehat{\theta})}.$$

These latter two criteria are often examined in an asymptotic scenario, so it is the asymptotic bias and variance (to give the **asymptotic relative efficiency** (ARE)) that are considered.

An estimator with minimum MSE is not necessarily that which is the MVUE since an estimator with a small bias may have a small variance which compensates in (1).

Estimating Functions: General Case

We give the general theory first, before giving specific recipes for the construction of estimating functions.

We assume that Y_i, $i = 1, \ldots, n$ are i.i.d.

An **estimating function** is a function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) \tag{2}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \mathbf{0}. \tag{3}$$

The estimating function $\boldsymbol{G}(\boldsymbol{\theta})$ is a random variable because it is a function of $\boldsymbol{Y}$.

The corresponding **estimating equation** that defines the estimator $\widehat{\boldsymbol{\theta}}_n$ has the form

$$\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}_n, Y_i) = \mathbf{0}. \tag{4}$$

An Important Result

Outline: Derive asymptotic properties of the estimating function, and then transfer these to the estimator.

Result 2.1: Suppose that $\widehat{\theta}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e., $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}$ (consistency) and

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \mathrm{N}_p(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\intercal -1}) \tag{5}$$

(asymptotic normality) where

$$\boldsymbol{A} \;=\; \boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} \;=\; \boldsymbol{B}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\intercal}] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}, Y)\}.$$

An Important Result

Proof: In Chapter 2 of book, but key conceptual steps are:

Expand $G_n(\theta)$ in a Taylor series around the true value θ:

$$0 = G_n(\widehat{\theta}_n) = G_n(\theta) + (\widehat{\theta}_n - \theta)\ \left.\frac{dG_n}{d\theta}\right|_{\theta} + \frac{1}{2}(\widehat{\theta}_n - \theta)^2\ \left.\frac{d^2 G_n}{d\theta^2}\right|_{\tilde{\theta}}, \tag{6}$$

where $\widetilde{\theta}$ is a point between $\widehat{\theta}_n$ and θ.

Rewrite (6) as

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) = \frac{-\sqrt{n}\ G_n(\theta)}{\left.\frac{dG_n}{d\theta}\right|_{\theta} + \frac{1}{2}(\widehat{\theta}_n - \theta)\ \left.\frac{d^2 G_n}{d\theta^2}\right|_{\tilde{\theta}}} \tag{7}$$

By a central limit theorem:

$$\sqrt{n}\ G_n(\theta)\ \to_d\ \mathrm{N}\left[0, B(\theta)\right]. \tag{8}$$

Now transfer the properties of the estimating function to the estimator $\widehat{\theta}_n$ via (7).

Notes:

- In the independent but not identically distributed case (e.g., regression) we have

$$(\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{\top -1})^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\rightarrow_d\; \mathrm{N}_p(\mathbf{0}, \boldsymbol{I}_p) \tag{9}$$

  where

$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{G}_n(\boldsymbol{\theta})\right]$$

  and

$$\boldsymbol{B}_n = \mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})\boldsymbol{G}_n(\boldsymbol{\theta})^{\top}] = \mathrm{cov}\{\boldsymbol{G}_n(\boldsymbol{\theta})\}.$$

- In the notation of the independent only case we have $\boldsymbol{A}_n = n\boldsymbol{A}$ and $\boldsymbol{B}_n = n\boldsymbol{B}$.
- Form of the variance – a "sandwich" – will recur frequently.
- We discuss various recipes for deriving estimating functions.

Likelihood

We assume the data are conditionally independent given $\boldsymbol{\theta}$ so that we have

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}).$$

> **Definition:** *Viewing $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ gives the* ***likelihood function****, which we denote by $L(\boldsymbol{\theta})$.*

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, denoted $\widehat{\boldsymbol{\theta}}$, is known as the **Maximum Likelihood Estimator** (MLE).

It is convenient to consider the **log-likelihood** function that is,

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(Y_i \mid \boldsymbol{\theta}).$$

The Score Function

The **score** function is

$$\boldsymbol{S}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{10}$$

$$= \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p}\right]^{\mathsf{T}}$$

$$= [\boldsymbol{S}_1(\boldsymbol{\theta}), \dots, \boldsymbol{S}_p(\boldsymbol{\theta})]^{\mathsf{T}},$$

which is a $p \times 1$ vector. As we now show the score satisfies the requirements of an estimating function upon which inference may be based.

Definition: **Fisher's expected information** *in a sample of size n is,*

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{T}} l(\boldsymbol{\theta})\right] = -E\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right],$$

a $p \times p$ matrix.

An Estimating Function Constructed from the Likelihood

Result: Under suitable regularity conditions:

$$\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathsf{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \tag{11}$$

and, under the model,

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left[\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^{\mathsf{T}}\right] = -E\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right]. \tag{12}$$

Proof: In Chapter 2 of book.

We may take the score as the basis for the estimating function,

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}),$$

and the MLE satisfies $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$.

Inference for the MLE

We have already seen that

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0},$$

and to apply Result 2.1 we require

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{G}(\theta, Y)\right] = \mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\log p(Y \mid \boldsymbol{\theta})\right],$$

and

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{\theta}) &= \mathrm{E}\left[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^{\mathsf{T}}\right] \\ &= \mathrm{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)^{\mathsf{T}}\right],\end{aligned}$$

and we have just shown that

$$\boldsymbol{I}_1(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta}) = -\boldsymbol{B}(\boldsymbol{\theta})$$

Inference for the MLE

Hence, from Result 2.1

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\rightarrow_d\; \mathrm{N}_p\left(\, \mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\theta})^{-1} \,\right). \tag{13}$$

For independent (but not necessarily identically distributed) random variables $Y_1, \ldots, Y_n$ we have, under the model,

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -\boldsymbol{A}_n(\boldsymbol{\theta}) = \boldsymbol{B}_n(\boldsymbol{\theta}),$$

and so

$$\boldsymbol{I}_n(\boldsymbol{\theta})^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\rightarrow_d\; N_p(\mathbf{0}, \boldsymbol{I}_p), \tag{14}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Notes on MLE

- Many statisticians love likelihood because

The MLE is consistent and asymptotically efficient

- Inference for functions of inference from the delta method.

Suppose

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \; \mathrm{N}_p\left[\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta})\right].$$

Then, by the delta method,

$$\sqrt{n}\,\left[g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right] \rightarrow_d \mathrm{N}\left[0, g'(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})g'(\boldsymbol{\theta})^{\intercal}\right],$$

where $g'(\boldsymbol{\theta})$ is the $1 \times p$ vector of derivatives of $g(\cdot)$ with respect to elements of $\boldsymbol{\theta}$.

Example: Lung Cancer and Radon

We examine the association between

- counts of lung cancer incidence, Y_i, and
- the average residential radon, x_i, in county i .

with $i = 1, \ldots, 85$, indexing the counties.

We examine the association using the Poisson loglinear model,

$$\begin{aligned} Y_i | \boldsymbol{\beta} &\sim \text{Poisson}(E_i \theta_i) \\ \log \theta_i &= \beta_0 + \beta_1 x_i. \end{aligned} \tag{15}$$

where E_i is the expected count.

- Recall that, $\text{SMR}_i = Y_i / E_i$ is the standardized mortality ratio in county i, and is a summary measure that controls for the differing age and sex population sizes across counties.
- We can write the model as

$$\log \text{E}[\text{SMR}_i \mid x_i] = \beta_0 + \beta_1 x_i.$$

Example: Lung Cancer and Radon

We take as our parameter of interest $\exp(\beta_1)$ which is the multiplicative change in risk associated with a one pCi/liter increase in radon.

In the epidemiological literature this parameter is referred to as the **relative risk**.

Here it corresponds to the risk ratio for two areas whose radon exposures, *x*, differ by one unit.

Example: Lung Cancer and Radon

Consider the log-linear Poisson model

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}(\mu_i),$$

with $\mu_i = E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})$, $\boldsymbol{x}_i = [1, x_i]$, $i = 1, \dots, n$, and $\boldsymbol{\beta} = [\beta_0, \beta_1]^\intercal$.

The sampling distribution of $\boldsymbol{y}$ is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log y_i! \right)$$

to give log-likelihood

$$l(\boldsymbol{\beta}) = \boldsymbol{\beta}^\intercal \sum_{i=1}^{n} \boldsymbol{x}_i^\intercal Y_i - \sum_{i=1}^{n} E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}).$$

Example: Lung Cancer and Radon

This leads to the 2 × 1 score vector (estimating function)

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \frac{\partial l}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} \left[Y_i - E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right] \\ &= \boldsymbol{x}^{\mathsf{T}} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right], \end{aligned} \tag{16}$$

where $\boldsymbol{x} = [\boldsymbol{x}_1^{\mathsf{T}}, \ldots, \boldsymbol{x}_n^{\mathsf{T}}]^{\mathsf{T}}$, $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^{\mathsf{T}}$, and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^{\mathsf{T}}$.

(16) is obviously unbiased!

The equation

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

does not, in general, have a closed-form solution but, pathological datasets aside, numerical solution is straightforward.

Example: Lung Cancer and Radon

Asymptotic inference is based on

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \to_d \ \mathrm{N}_2(\mathbf{0}, \mathrm{I}_2),$$

where the information matrix is,

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}_n) = \mathrm{var}(\boldsymbol{S}) = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} \mathrm{var}(Y_i) \boldsymbol{x}_i = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{V} \boldsymbol{x},$$

with $\boldsymbol{V}$ the diagonal matrix with diagonal elements

$$\mathrm{var}(Y_i) = E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}),$$

$i = 1, \ldots, n$.

In this case the expected and observed information coincide.

In practice, the information is estimated by replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}_n$.

Example: Lung Cancer and Radon

An important observation is that the score, (16), is a consistent estimator of zero, and $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}$.

This doesn't say anything about the mean specified being an appropriate summary, it just says we are consistently estimating the parameters in this mean function.

If the data do not conform to $\text{var}(Y_i) = \mu_i$, we still have a consistent estimator, but the standard errors will be incorrect.

For the lung cancer data we have $n = 85$, and the MLE is $\widehat{\boldsymbol{\beta}} = [0.17, -0.036]^{\mathsf{T}}$ with

$$\boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1} = \begin{bmatrix} 0.027^2 & -0.95 \times 0.027 \times 0.0054 \\ -0.95 \times 0.027 \times 0.0054 & 0.0054^2 \end{bmatrix}.$$

Example: Lung Cancer and Radon

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.027 and 0.0054, respectively, and an asymptotic 95% confidence interval for β_1 is $[-0.047, -0.026]$.

Leaning on asymptotic normality is appropriate with the large sample size here.

A useful inferential summary is an asymptotic 95% confidence interval for the area-level relative risk associated with a one-unit increase in residential radon, which is

$$\exp(-0.036 \pm 1.96 \times 0.0054) = [0.954, 0.975].$$

This interval suggests the decrease in lung cancer incidence associated with a one-unit increase in residential radon is between 2.5% and 4.6% – this is an area-level summary, it does not refer to individuals in an area (the sampling units and the model are specified at the area-level).

Quasi-Likelihood

An alternative to MLE, when we do not wish to commit to specifying the full distribution of the data is Quasi-MLE.

Suppose we are willing to assume:

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{\beta}] &= \boldsymbol{\mu}(\boldsymbol{\beta}) \\ \mathrm{cov}(\boldsymbol{Y} \mid \boldsymbol{\beta}) &= \alpha \boldsymbol{V}\{\boldsymbol{\mu}(\boldsymbol{\beta})\} \end{aligned}$$

where

- $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta})]^\mathsf{T}$ represents the regression function and
- $\boldsymbol{V}$ is a diagonal matrix (so the observations are uncorrelated), with

  $$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V\{\mu_i(\boldsymbol{\beta})\},$$

  and $\alpha > 0$ a scalar which is independent of $\boldsymbol{\beta}$.

How should one proceed with **estimation?**

Quasi-Likelihood

Consider the sum of squares

$$(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha, \tag{17}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta})$.

To minimize this sum of squares there are two ways to proceed.

One approach: Differentiate (17) and obtain

$$-2\boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha + (\boldsymbol{Y} - \boldsymbol{\mu})^{\mathsf{T}} \frac{\partial \boldsymbol{V}^{-1}}{\partial \boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha,$$

where $\boldsymbol{D}$ is the $n \times p$ matrix of derivatives with elements $\partial \mu_i / \partial \beta_j$, $i = 1, \ldots, n; j = 1, \ldots, p$.

Unfortunately the expectation of this expression is not zero in general, because of the second term, and so an inconsistent estimator of $\boldsymbol{\beta}$ will result.

Quasi-Likelihood

Second approach: Pretend $\boldsymbol{V}$ is not a function of $\boldsymbol{\beta}$, differentiate and set equal to zero to obtain,

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^{\intercal}\boldsymbol{V}(\widehat{\boldsymbol{\beta}})^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\}/\alpha=\mathbf{0}.$$

As shorthand we write this estimating function as

$$\boldsymbol{U}(\boldsymbol{\beta})=\boldsymbol{D}^{\intercal}\boldsymbol{V}^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}\}/\alpha. \tag{18}$$

Quasi-Likelihood

The estimating function (quasi-score),

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{D}^{\top}\boldsymbol{V}^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}\}/\alpha$$

is linear in the data and so its properties are straightforward to evaluate.

In particular:

1. $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta})] = \mathbf{0}$.
2. $\mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^{\top}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$.
3. $-\mathrm{E}\left[\frac{\partial U}{\partial \beta}\right] = \boldsymbol{D}^{\top}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha = \mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\}$.

Applying Result 2.1:

$$(\boldsymbol{D}^{\top}\boldsymbol{V}^{-1}\boldsymbol{D})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\to_d\; \mathrm{N}_p(\mathbf{0}, \alpha \boldsymbol{I}_p),$$

where, do far, we have assumed that α is known.

Quasi-Likelihood: Unknown α

Suppose $\boldsymbol{Z}$ is an $n \times 1$ random variable with $\mathrm{E}[\boldsymbol{Z}] = \boldsymbol{m}$, $\mathrm{var}(\boldsymbol{Z}) = \boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a symmetric $n \times n$ matrix. Then,

$$\mathrm{E}[\boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z}] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{m}^\top \boldsymbol{A}\boldsymbol{m}. \tag{19}$$

In our context, $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{\mu}, \boldsymbol{m} = \boldsymbol{0}, \boldsymbol{\Sigma} = \alpha \boldsymbol{V}$ and take $\boldsymbol{A} = \boldsymbol{V}^{-1}$, then applying (19),

$$\mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})] = \mathrm{tr}(\boldsymbol{V}^{-1}\alpha \boldsymbol{V}) = n\alpha.$$

Therefore, an unbiased estimator of α is

$$\widehat{\alpha} = (\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})/n.$$

For diagonal $\boldsymbol{V}$:

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where the use of $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$ means we no longer have an unbiased estimator.

Quasi-Likelihood: Unknown α

A degrees of freedom corrected (but not in general, unbiased) estimate is given by the Pearson statistic divided by its degrees of freedom:

$$\widehat{\alpha} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

The asymptotic distribution that is used in practice is therefore

$$(\widehat{\boldsymbol{D}}^{\top} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}}/\widehat{\alpha})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I}_p).$$

You’ve been using something that’s equivalent to quasi-likelihood all your life:

normal/least squares regression with α the measurement error variance!

Why "Quasi"?

Integration of the quasi-score (18) gives

$$l(\mu, \alpha) = \int_y^\mu \frac{y-t}{\alpha V(t)} \mathrm{d}t$$

which, if it exists, behaves like a log-likelihood[1].

As an example, for the model $\mathrm{E}[Y] = \mu$ and $\mathrm{var}(Y) = \alpha\mu$ we have

$$l(\mu, \alpha) = \int_y^\mu \frac{y-t}{\alpha t} \mathrm{d}t = \frac{1}{\alpha}[y \log \mu - \mu + c],$$

where $c = -y \log y - y$ and $y \log \mu - \mu$ is the log likelihood of a Poisson random variable.

[1] Peter McCullagh personal communication: "No good reason to choose y to be the lower limit. I just wanted $l(y, \alpha) = 0$ to make it like a sum of squares (deviance)"

Why "Quasi"?

The word "quasi" refers to the fact that the score does not correspond to a unique probability model and may or not even correspond to a probability model.

For example, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution.

Example: Lung Cancer and Radon

Assume the quasi-likelihood model

$$\text{E}[Y_i \mid \boldsymbol{\beta}] = E_i \exp(\boldsymbol{x}_i\boldsymbol{\beta}), \quad \text{var}(Y_i \mid \boldsymbol{\beta}) = \alpha \text{E}[Y_i \mid \boldsymbol{\beta}].$$

Fitting this model yields identical point estimates to the MLEs and $\widehat{\alpha} = 2.81$ so that the quasi-likelihood standard errors are $\sqrt{\widehat{\alpha}} = 1.68$ times larger than the Poisson model-based standard errors.

The variance-covariance matrix is

$$(\widehat{\boldsymbol{D}}^{\mathsf{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}})^{-1} \widehat{\alpha} = \begin{bmatrix} 0.045^2 & -0.95 \times 0.045 \times 0.0090 \\ -0.95 \times 0.045 \times 0.0090 & 0.0090^2 \end{bmatrix}.$$

An asymptotic 95% confidence interval for the relative risk associated with a one-unit increase in radon is $[0.947, 0.982]$.

Beyond Quasi-Likelihood (though not necessarily a good idea)

In the quasi-likelihood method, we had “separable” mean and variance models, that is, $\text{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V_i(\mu_i)$.

The estimating equation has the same solution for $\boldsymbol{\beta}$, regardless of the value of α.

Suppose we have,

$$\begin{aligned} \text{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \text{var}(Y_i \mid \boldsymbol{\beta}) &= V_i(\boldsymbol{\alpha}, \boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{\alpha}$ is a $k \times 1$ vector of parameters that appear only in the variance model.

The estimating equation is

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^{\mathsf{T}} \boldsymbol{V}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})^{-1} \{ \boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}) \} = \mathbf{0}.$$

Beyond Quasi-Likelihood

Let $\widehat{\boldsymbol{\alpha}}_n$ be a consistent estimator of $\boldsymbol{\alpha}$.

We state without proof the following result.

Result: The estimator $\widehat{\boldsymbol{\beta}}_n$ that satisfies

$$\boldsymbol{G}(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n) = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)\left\{\boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_n)\right\} \tag{20}$$

has asymptotic distribution

$$(\widehat{\boldsymbol{D}}^{\top}\widehat{\boldsymbol{V}}^{1/2}\widehat{\boldsymbol{D}})^{-1}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_p(\mathbf{0}, \boldsymbol{I}_p) \tag{21}$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$.

Computation

Previously we assumed $\text{var}(Y_i) = \alpha V_i(\mu_i)$, and the estimating function did not depend on α and so, correspondingly, $\widehat{\boldsymbol{\beta}}$ did not depend on α (though the standard errors did).

In general, iteration is convenient to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate. Then set $j = 0$ and iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(j)}) = \mathbf{0}$ to give $\widehat{\boldsymbol{\beta}}^{(j+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(j+1)}$ with $\widehat{\mu}_i = \mu_i\left(\widehat{\boldsymbol{\beta}}^{(j+1)}\right)$. Set $j \rightarrow j + 1$ and return to Step 1.

Example: Lung Cancer and Radon

Consider the model

$$\begin{aligned} \mathsf{E}[\,Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \mathsf{var}(Y_i \mid \alpha, \boldsymbol{\beta}) &= \mu_i(\boldsymbol{\beta})\{1 + \mu_i(\boldsymbol{\beta})/\alpha\}. \end{aligned} \tag{22}$$

which suggests the estimating function for $\boldsymbol{\beta}$ (with α assumed known):

$$\sum_{i=1}^{n} \boldsymbol{D}(\boldsymbol{\beta})_i^{\top} \boldsymbol{V}_i^{-1}(\alpha, \boldsymbol{\beta})\{y_i - \mu_i(\boldsymbol{\beta})\}$$

Hence, for a fixed α we can solve this estimating equation to obtain an estimator $\widehat{\boldsymbol{\beta}}$.

If the variance function is incorrect we still obtain a consistent estimator of $\boldsymbol{\beta}$ but correct standard errors (and efficiency) are out the window.

Note that this is the score of a negative binomial model, with α assumed known.

Example: Lung Cancer and Radon

We describe a method-of-moments estimator for α for the quadratic variance model . We have

$$\text{var}(Y_i \mid \boldsymbol{\beta}, \alpha) = \mathsf{E}[(Y_i - \mu_i)^2] = \mu_i(1 + \mu_i/\alpha),$$

and so

$$\alpha^{-1} = \mathsf{E}\left[\frac{(Y_i - \mu_i)^2 - \mu_i}{\mu_i^2}\right],$$

for $i = 1, \dots, n$, which suggests the method-of-moments estimator:

$$\widehat{\alpha} = \left[\frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2 - \widehat{\mu}_i}{\widehat{\mu}_i^2}\right]^{-1}. \qquad (23)$$

If we have a consistent estimator, $\widehat{\alpha}$, valid inference follows from

$$(\widehat{\boldsymbol{D}}^{\mathsf{T}}\widehat{\boldsymbol{V}}(\widehat{\alpha})^{-1}\widehat{\boldsymbol{D}})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p).$$

If the variance function is incorrect then standard errors will be incorrect.

Example: Lung Cancer and Radon

We fit the quadratic variance model to the lung cancer and radon data.

The estimates (standard errors) are $\widehat{\beta}_0 = 0.090\ (0.047)$ and $\widehat{\beta}_1 = -0.030\ (0.0085)$.

The latter point estimate differs a little from the MLE (and MQLE) of -0.036, reflecting the different variance weighting in the estimating function.

The moment-based estimator is $\widehat{\alpha} = 57.8$.

The MLE under the negative binomial model is 61.3 and so close to this moment estimate.

Example: Lung Cancer and Radon

An asymptotic 95% confidence interval for the relative risk $\exp(\beta_1)$ is [0.955,0.987], so that the upper limit is closer to unity than the intervals we have seen previously.

In terms of the first two moments, the difference between quasi-likelihood and the negative binomial model is that the variances are, respectively, linear and quadratic functions of the mean.

In Figure 1 we plot the estimated linear and quadratic variance functions over the range of the mean for these data.

To produce a clearer plot the log of the variance is plotted against the log of the mean, and the log of the observed counts, y_i, $i = 1, \ldots, 85$, are added to the plot (with a small amount of jitter).

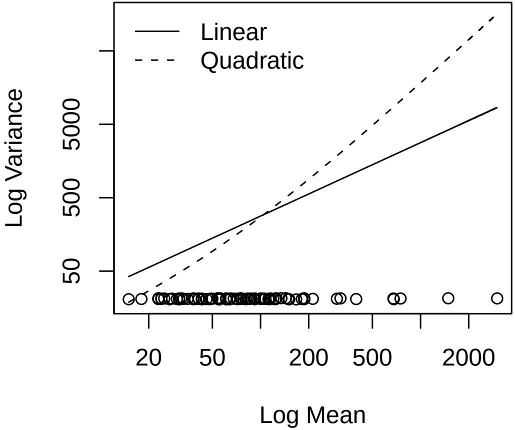

Figure 1: Linear and quadratic variance functions for the lung cancer data.

Example: Lung Cancer and Radon

- Over the majority of the data the two variance functions are similar, but for large values of the mean in particular the variance functions are considerably different which leads to the differences in inference, since these observations are being weighted very differently by the two variance functions.
- Based on this plot, we might expect even greater differences.
- However, closer examination of the data reveals that the x's associated with the large y values are all in the mid range, and hence these points are not influential.
- Examination of the residuals give some indication that the quadratic mean-variance model is more appropriate for these data.
- It is typically very difficult to distinguish between the two models, unless there are sufficient points across a large spread of mean values – the linear variance model is preferred on purely statistical grounds.

Sandwich Estimation

A general method of avoiding stringent modeling conditions when the variance of an estimator is calculated is provided by sandwich estimation.

Given an estimating function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i),$$

based on independent and identically distributed observations we have

$$n \times \text{var}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\mathsf{T}-1} \tag{24}$$

where

$$\boldsymbol{A} = \text{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} = \text{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathsf{T}}]$$

where for (24) to be asymptotically appropriate the expectations need to be evaluated under the true model.

Sandwich Estimation

So far we have used an assumed model to calculate the expectations.

An alternative is to evaluate $\boldsymbol{A}$ and $\boldsymbol{B}$ empirically via

$$\widehat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{G}(\widehat{\theta}, Y_i),$$

and

$$\widehat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}(\widehat{\theta}, Y_i)\boldsymbol{G}(\widehat{\theta}, Y_i)^{\mathsf{T}}.$$

By a law of large numbers $\widehat{\boldsymbol{A}} \rightarrow_d \boldsymbol{A}$ and $\widehat{\boldsymbol{B}} \rightarrow_d \boldsymbol{B}$ and so

$$n \times \text{var}(\widehat{\theta}_n) = \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^{\mathsf{T}-1}, \tag{25}$$

is a consistent estimator of the variance.

Sandwich Estimation

The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations.

For small sample sizes, the sandwich estimator may be highly unstable, and in terms of mean squared error model-based estimators may be preferable for small to medium sized *n*.

Hence empirical is a better description of the estimator than robust.

The second consideration is that if the model is correct, then model-based estimators are more efficient.

Sandwich Estimation

We now consider the situation in which the estimating function arises from the score and suppose we have independent and identically distributed data.

Then

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta}),$$

with $l_i(\boldsymbol{\theta}) = \log p(Y_i \mid \boldsymbol{\theta})$.

In this case,

$$\boldsymbol{A} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} l(\boldsymbol{\theta}) \right]$$

and

$$\boldsymbol{B} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) \right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) \right)^{\mathsf{T}} \right]$$

where $l(\boldsymbol{\theta}) = \log p(Y \mid \boldsymbol{\theta})$.

Sandwich Estimation

Then under the model

$$\boldsymbol{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta}), \tag{26}$$

so that

$$n \times \text{var}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\mathsf{T}-1} = \boldsymbol{I}_1(\boldsymbol{\theta})^{-1}.$$

The sandwich estimator (25) is based on

$$\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}} l_i(\boldsymbol{\theta})\bigg|_{\widehat{\boldsymbol{\theta}}}$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial}{\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)^{\mathsf{T}}\Bigg|_{\widehat{\boldsymbol{\theta}}}.$$

These are empirical averages of functions of Y!

The uncertainty in $\widehat{\boldsymbol{\theta}}$, arises from uncertainty in the estimating equation $\boldsymbol{G}$ and from uncertainty in the transformation from $\boldsymbol{G} \to \boldsymbol{\theta}$.

Example: Poisson Mean

Data simulated from the model

$$\begin{aligned} Y_i \mid \lambda_i &\sim_{ind} \text{Poisson}(\lambda_i) \\ \lambda_i &\sim_{iid} \text{Gamma}(\theta b, b) \end{aligned}$$

$i = 1, \ldots, n$.

This gives marginal moments

$$\begin{aligned} \mathsf{E}[Y_i] &= \mathsf{E}\left[\mathsf{E}[Y_i|\lambda_i]\right] = \mathsf{E}[\lambda_i] = \theta \times \frac{b}{b} = \theta \\ \text{var}(Y_i) &= \mathsf{E}[\text{var}(Y_i|\lambda_i)] + \text{var}(\mathsf{E}[Y_i|\lambda_i]) \\ &= \mathsf{E}[\lambda_i] + \text{var}(\lambda_i) \\ &= \theta + \frac{\theta b}{b^2} = \theta\left(1 + \frac{1}{b}\right) = \theta \times \alpha. \end{aligned}$$

So the variance is linear in the mean.

Example: Poisson Mean

Note: An alternative Poisson-Gamma model has:

$$\begin{aligned} Y_i \mid \delta_i &\sim_{ind} \text{Poisson}(\theta\delta_i), \\ \delta_i &\sim_{iid} \text{Gamma}(c, c), \end{aligned}$$

$i = 1, \ldots, n$.

This gives marginal moments

$$\begin{aligned} \mathsf{E}[Y_i] &= \mathsf{E}\left[\mathsf{E}[Y_i|\delta_i]\right] = \mathsf{E}[\theta\delta_i] = \theta \times \frac{c}{c} = \theta \\ \text{var}(Y_i) &= \mathsf{E}[\text{var}(Y_i|\delta_i)] + \text{var}(\mathsf{E}[Y_i|\delta_i]) \\ &= \mathsf{E}[\theta\delta_i] + \text{var}(\theta\delta_i) \\ &= \theta + \frac{\theta^2}{c} = \theta\left(1 + \frac{\theta}{c}\right). \end{aligned}$$

So the variance is quadratic in the mean.

Aside: Which one is the most natural?? Not obvious...

Example: Poisson Mean

We examine three different variance estimators arising from likelihood, quasi-likelihood and sandwich estimation.

We first assume a Poisson model which gives log-likelihood,

$$l(\theta) = -\theta + Y \log \theta$$

and

$$\frac{dl}{d\theta} = \frac{Y - \theta}{\theta}, \quad \frac{d^2 l}{d\theta^2} = \frac{-Y}{\theta^2},$$

and so $\widehat{\theta} = \overline{Y}$.

In the estimating function notation,

$$\begin{aligned} G(\theta) &= \frac{1}{n} \boldsymbol{D}(\theta)^{\mathsf{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \mathbf{1}\theta) \\ &= \frac{1}{n} [1 1 \cdots 1] \begin{bmatrix} 1/\theta & \cdots & \cdots & 0 \\ 0 & 1/\theta & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 1/\theta \end{bmatrix} \begin{bmatrix} Y_1 - \theta \\ Y_2 - \theta \\ \cdots \\ Y_n - \theta \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i - \theta}{\theta} \end{aligned}$$

Example: Poisson Mean

Under a likelihood approach,

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \text{N}\left(0, \frac{B}{A^2}\right).$$

where

$$A = \text{E}\left[\frac{d^2}{d\theta^2} l(\theta)\right], \quad B = \text{E}\left[\left(\frac{d}{d\theta} l(\theta)\right)^2\right].$$

Hence,

$$A = -\text{E}\left[\frac{d^2 l}{d\theta^2}\right] = I_1(\theta) = \frac{1}{\theta},$$

and the Poisson model-based variance estimator is

$$\widehat{\text{var}}(\widehat{\theta}) = \frac{1}{nI_1(\widehat{\theta})} = \frac{\overline{Y}}{n}.$$

Note that

$$B = \text{cov}\left(\frac{(Y - \theta)^2}{\theta^2}\right) = \frac{\text{var}(Y)}{\theta^2} = \frac{1}{\theta},$$

under the assumption that $\text{var}(Y) = \theta$.

Example: Poisson Mean

The quasi-likelihood estimator is based on

$$U = \sum_{i=1}^{n} \frac{Y_i - \theta}{\alpha\theta} = \frac{n\overline{Y} - n\theta}{\alpha\theta},$$

and

$$\text{var}(\widehat{\theta}) = (\widehat{\boldsymbol{D}}^{\mathsf{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}})^{-1} \widehat{\alpha} = \frac{\widehat{\theta}}{n} \times \widehat{\alpha}$$

where

$$\widehat{\alpha} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\theta})^2}{\widehat{\theta}}.$$

This gives

$$\widehat{\text{var}}(\hat{\theta}) = \frac{s^2}{n},$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \widehat{\theta})^2.$$

Example: Poisson Mean

For the sandwich estimator,

$$\widehat{A} = -\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\widehat{\theta}^2} = -\frac{1}{\overline{Y}},$$

and

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\theta})^2}{\widehat{\theta}^2} = \frac{(n-1)s^2}{n\widehat{\theta}^2}.$$

Hence,

$$\widehat{\text{var}}(\widehat{\theta}) = \frac{s^2(n-1)/n}{n}. \tag{27}$$

Table 1 gives the 95% confidence interval coverage for the model-based, quasi-likelihood and sandwich estimator variance estimates for various values of n and α.

Example: Poisson Mean

	Overdispersion α								
	1			2			3		
n	M	Q	S	M	Q	S	M	Q	S
5	95	87	84	83	87	84	74	86	83
10	94	92	90	83	91	90	73	91	89
15	95	93	92	84	92	92	75	92	91
20	95	93	93	83	93	93	73	93	92
25	95	94	93	83	94	93	74	93	93
50	95	94	94	83	94	94	74	94	94
100	95	95	94	83	95	94	74	95	94

Table 1: Confidence interval coverage for Poisson mean example, based on 50,000 simulations; the nominal coverage is 0.95. The overdispersion is given by $\alpha = \text{var}(Y)/\text{E}[Y]$. M=Model-based, Q=Quasi-likelihood, S=Sandwich. The mean $\theta = 10$.

Example: Poisson Mean

This example shows the efficiency-robustness trade-off:

- if the model is correct (corresponding here to $\alpha = 1$), then the model-based approach performs well.
- The sandwich and quasi-likelihood approaches are more robust to variance mis-specification.
- Hence, the choice as to which variance to use depends crucially on how much we believe the model.
- Using the Poisson model is a risky enterprise since it does not contain an additional variance parameter.

Example: Lung Cancer and Radon

Returning to the lung cancer and radon example, we calculate sandwich standard errors, assuming that counts in different areas are uncorrelated.

We take as “working model” a Poisson likelihood, with maximum likelihood estimation of $\boldsymbol{\beta}$.

The estimating function is

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{x}^{\mathsf{T}} (\boldsymbol{Y} - \boldsymbol{\mu}) \, ,$$

as defined previously, (16).

Under this model,

$$(\boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{\mathsf{T}})^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \to_d \ \mathrm{N}_2(\mathbf{0}, \mathrm{I}_2),$$

with sandwich ingredients

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \boldsymbol{D} \\ \boldsymbol{B} &= \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \mathrm{var}(\boldsymbol{Y}) \boldsymbol{V}^{-1} \boldsymbol{D}. \end{aligned}$$

Example: Lung Cancer and Radon

The estimators are,

$$\begin{aligned}
\widehat{\boldsymbol{A}} &= \widehat{\boldsymbol{D}}^{\mathsf{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}} \\
\widehat{\boldsymbol{B}} &= \widehat{\boldsymbol{D}}^{\mathsf{T}} \widehat{\boldsymbol{V}}^{-1} \begin{bmatrix} \widehat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \widehat{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \widehat{\sigma}_n^2 \end{bmatrix} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}}
\end{aligned}$$

and with squared residuals,

$$\widehat{\sigma}_i^2 = (Y_i - \widehat{\mu}_i)^2,$$

for $i = 1, \ldots, n$.

Example: Lung Cancer and Radon

Substitution of the required data quantities yields the sandwich variance-covariance matrix,

$$\left[\begin{array}{cc} 0.043^2 & -0.87 \times 0.043 \times 0.0080 \\ -0.87 \times 0.043 \times 0.0080 & 0.0080^2 \end{array} \right].$$

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.043 and 0.0080, respectively, and are 60% and 49% larger than their likelihood counterparts, though slightly smaller than the quasi-likelihood versions.

An asymptotic 95% confidence interval for the relative risk associated with a one-unit increase in radon is $[0.949, 0.980]$.

Example: Lung Cancer and Radon

We have a linear exponential family likelihood (so score is linear) and so a consistent estimator of the loglinear association between lung cancer incidence and radon, as is clear from (16).

If the outcomes are independent then the large sample size will yield a consistent sandwich variance estimator.

However, in the context of these data, we may have residual spatial dependence, particularly since we have not controlled for confounders such as smoking which may have spatial structure (and hence will induce spatial dependence), and sandwich standard errors do not account for such dependence (unless we can lean on replication across time).

Although the loglinear association is consistently estimated, this of course says nothing about causality, or about the appropriateness of the mean model.

Behavior of MLEs Under Misspecification

We first sketch a consistency proof for the MLE.

Define,

$$\mathsf{KL}(f,g) = \int \log \frac{f(y)}{g(y)} f(y)\,\mathrm{d}y,$$

as the Kullback-Leibler measure of the "distance" between the densities f and g.

By Jensen's inequality:

$$\begin{aligned} -\mathsf{KL}(f,g) &= \int \log \frac{g(y)}{f(y)} f(y)\,\mathrm{d}y \\ &= \mathsf{E}\left[\log \frac{g(y)}{f(y)}\right] \\ &\leqslant \log \mathsf{E}\left[\frac{g(y)}{f(y)}\right] = 0, \end{aligned}$$

with equality if and only if $f = g$.

Behavior of MLEs Under Misspecification

Maximizing $\log p(\boldsymbol{Y}|\boldsymbol{\theta})$ is equivalent to maximizing

$$\log p(\boldsymbol{Y}|\boldsymbol{\theta}) - \log p(\boldsymbol{Y}|\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$.

Then,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(Y_i|\boldsymbol{\theta})}{p(Y_i|\boldsymbol{\theta}_0)} \quad &\rightarrow_{a.s.} \quad \mathrm{E}_0\left[\log\frac{p(Y|\boldsymbol{\theta})}{p(Y|\boldsymbol{\theta}_0)}\right] \\ &= \quad \mathrm{KL}\{p(Y|\boldsymbol{\theta}), p(Y|\boldsymbol{\theta}_0)\} \leqslant 0.\end{aligned}$$

Also, for the MLE $\widehat{\boldsymbol{\theta}}_n$,

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(Y_i|\widehat{\boldsymbol{\theta}}_n)}{p(Y_i|\boldsymbol{\theta}_0)} = \sup\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(Y_i|\boldsymbol{\theta})}{p(Y_i|\boldsymbol{\theta}_0)} \geqslant 0,$$

for all n.

Hence, $\widehat{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}_0$.

Behavior of MLEs Under Misspecification

We let $p^A(y \mid \boldsymbol{\theta})$ and $p^T(y)$ denote the assumed and true densities, respectively.

The log-likelihood is such that

$$\frac{1}{n}\sum_{i=1}^{n} \log p^A(Y_i \mid \boldsymbol{\theta}) \rightarrow_{a.s.} \mathsf{E}_T[\log p^A(Y \mid \boldsymbol{\theta})], \qquad (28)$$

and so the MLE asymptotically maximizes the expectation of the assumed log-likelihood under the true model.

Hence, $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_T$ where the latter defines the $\boldsymbol{\theta}$ we are estimating.

We write,

$$\begin{aligned} \mathsf{E}_T[\log p^A(Y \mid \boldsymbol{\theta})] &= \mathsf{E}_T\left[\log p^T(Y) - \log p^T(Y) + \log p^A(Y \mid \boldsymbol{\theta})\right] \\ &= \mathsf{E}_T[\log p^T(Y)] - \mathsf{KL}(p^T, p^A). \end{aligned}$$

The MLE minimizes $\mathsf{KL}(p^T, p^A)$, and is therefore that value of $\boldsymbol{\theta}$ that makes the assumed model closest to the true model.

Behavior of MLEs Under Misspecification

Result: Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{S}_n^A(\boldsymbol{\theta}) = \mathbf{0}$, i.e., $\boldsymbol{S}_n^A(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_T) \rightarrow_d \mathrm{N}_p(\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{\mathsf{T}-1}) \tag{29}$$

where

$$\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}_T) = \mathrm{E}_T\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\log p^A(Y \mid \boldsymbol{\theta}_T)\right],$$

and

$$\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\theta}_T) = \mathrm{E}_T\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p^A(Y \mid \boldsymbol{\theta}_T)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p^A(Y \mid \boldsymbol{\theta}_T)\right)^{\mathsf{T}}\right].$$

Proof: Section 2.4.3 of Wakefield (2013).

Example: Exponential: Assumed Model, Gamma: True Model

Suppose that the assumed model is exponential with mean θ, but that the true model is $\text{Ga}(\alpha, \beta)$.

Minimizing the Kullback-Leibler distance with respect to θ corresponds to maximizing (28), that is

$$\mathsf{E}_{\mathsf{T}}\left[-\log\theta - \frac{Y}{\theta}\right] = \log\theta - \frac{\alpha/\beta}{\theta},$$

so that $\theta_{\mathsf{T}} = \alpha/\beta$ is the quantity that is being estimated by the MLE.

Hence, the closest exponential distribution to the gamma distribution, in a Kullback-Leibler sense, is the one that possesses the same mean, and it is this mean we are estimating when we assume an exponential model.

Bootstrap Methods

The bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997; Chernick, 2011) is a popular resampling technique that is useful in situations in which:

- we wish to relax the assumptions of a parametric modeling approach,
- the asymptotic sampling distribution of the estimator is difficult to derive.

Though there is a Bayesian bootstrap (Rubin, 1981), we focus on the frequentist version, which is far and away the most common form used.

Bootstrap Methods

We may be interested in bootstrap methods for:

- Estimating bias of parameter estimators.
- Constructing CIs for parameters.
- Testing.

For simplicity, suppose we are in a population setting where we have iid sampling, and denote the cumulative distribution function of Y by F.

Bootstrap Methods: Known F

If F is known, it is straightforward to mimic frequentist inference.

For $b = 1, \ldots, B$ bootstrap samples:

- Generate a random sample $y_1^{(b)}, \ldots, y_n^{(b)} \sim_{iid} F$.
- Compute $\widehat{\theta}^{(b)}$ using $y_1^{(b)}, \ldots, y_n^{(b)}$.

Use $\widehat{\theta}^{(b)}, b = 1, \ldots, B$, to estimate the sampling distribution of $\widehat{\theta}$ (which we can examine using a histogram or a density estimate).

If $B \to \infty$, we approach the theoretical sampling distribution of $\widehat{\theta}$.

Bootstrap Methods: Unknown F

In practice, of course, we never know F.

Two obvious choices for $\widehat{F}$:

- If one has some faith in the assumed model then we may use this model, call this F_θ, where the notation emphasizes that the distribution function is fully specified by the parameter θ.
- Alternatively one may use the empirical cumulative distribution function of the data, which we denote F_n – this is by far the most common approach used.

We may:

- Sample data from $F_{\hat{\theta}}$ to give the **parametric** bootstrap.
- Sample data with replacement from $\widehat{F}_n$ to give the **non-parametric** bootstrap.

An Example Quantity of Interest

Suppose we are interested in,

$$\Pr(L < \widehat{\theta}(\boldsymbol{Y}) < U \mid F), \tag{30}$$

where for simplicity we assume that θ is one-dimensional.

Then we can substitute an estimate $\widehat{F}$ for F, to give

$$\Pr(L < \widehat{\theta}(\boldsymbol{Y}) < U \mid \widehat{F}). \tag{31}$$

An empirical estimate of (31) is then provided by

$$\frac{1}{B}\sum_{b=1}^{B} 1(L < \widehat{\theta}^{(b)} < U) \tag{32}$$

where $\widehat{\theta}^{(b)} = \widehat{\theta}(\boldsymbol{Y}^{(b)})$ and the indicator function $I(A)$ equals 1 if the event A occurs and is 0 otherwise.

Bootstrap Methods

Theoretical: The population is to the sample

Bootstrap: The sample is to the bootstrap sample

As an example, the theoretical bias of as estimator is

$$\mathrm{E}[\widehat{\theta}(\boldsymbol{Y})|F] - \theta.$$

The bootstrap estimates this by

$$\underbrace{\mathrm{E}[\widehat{\theta}(\boldsymbol{Y})|\widehat{F}]}_{\substack{\text{Average over}\\ \text{Samples}}} - \underbrace{\theta}_{\text{Population}}.$$

And in practice this is estimated by

$$\underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}}_{\substack{\text{Average over}\\ \text{Bootstrap Samples}}} - \underbrace{\widehat{\theta}(\boldsymbol{y})}_{\text{Sample Estimate}}.$$

Bootstrap Methods

There are two approximations/sources of error in the bootstrap:

- Statistical: $\widehat{F} \neq F$ – careful thought can help here.
- Simulation: $B \neq \infty$, but we can take B large.

For the first, if n is small, the approximation will be poorer when we use F_n.

For the second, for some targets such as the bias and variance, we can get away with smaller B (e.g., $B \geqslant 200$), but for others such as confidence intervals we need larger B (e.g., $B \geqslant 1000$).

Bootstrap Methods

Care is required in dependent data situations, such as with time series, spatial or survey data with a complex design– the bootstrap samples must be consistent with the dependence structure in the data (which is present by design in the survey case).

The bootstrap method does not work for all functions of interest:

- In particular it fails in situations when the tail behavior is not well-behaved, for example a bootstrap for the maximum $Y_{(n)}$ will be disastrous – the limiting bootstrap distribution converges to a different distribution to the true distribution.

Bootstrap Methods: Variance and CI Estimation

Obvious estimator:

$$\widehat{V} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}(\boldsymbol{y}^{(b)}) - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}(\boldsymbol{y}^{(b)})\right)^2.$$

If n is sufficiently large that asymptotic normality of the estimator may be appealed to and a CI estimate may be based upon

$$\widehat{\theta} + \widehat{\text{Bias}} \pm z_{1-\alpha/2} \times \widehat{V}^{1/2}$$

where $\widehat{\text{Bias}}$ is the estimated bias $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$ and $Z \sim \text{N}(0,1)$.

Bootstrap Methods: CI Estimation

For CI construction, many improvements on the above normal-based method have been suggested:

- The percentile method – pick the appropriate sample quantiles.
- Various bias corrected versions have been proposed.

Figure 2 is from Section 3.1 of Chernick (2011) .

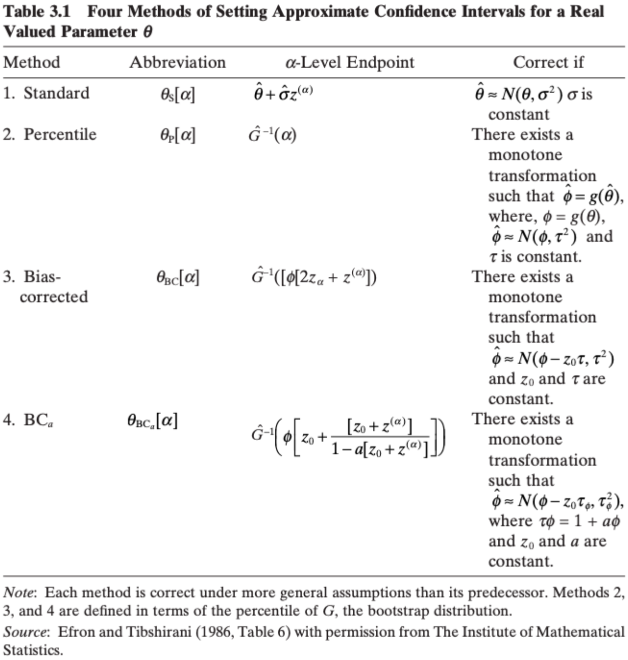

Table 3.1 Four Methods of Setting Approximate Confidence Intervals for a Real Valued Parameter θ

Method	Abbreviation	α-Level Endpoint	Correct if
1. Standard	$\theta_S[\alpha]$	$\hat{\theta} + \hat{\sigma} z^{(\alpha)}$	$\hat{\theta} \approx N(\theta, \sigma^2)$ σ is constant
2. Percentile	$\theta_P[\alpha]$	$\hat{G}^{-1}(\alpha)$	There exists a monotone transformation such that $\hat{\phi} = g(\hat{\theta})$, where, $\phi = g(\theta)$, $\hat{\phi} \approx N(\phi, \tau^2)$ and τ is constant.
3. Bias-corrected	$\theta_{BC}[\alpha]$	$\hat{G}^{-1}([\phi[2z_\alpha + z^{(\alpha)}])$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau, \tau^2)$ and z_0 and τ are constant.
4. BC$_a$	$\theta_{BC_a}[\alpha]$	$\hat{G}^{-1}\left(\phi\left[z_0 + \frac{[z_0 + z^{(\alpha)}]}{1 - a[z_0 + z^{(\alpha)}]}\right]\right)$	There exists a monotone transformation such that $\hat{\phi} \approx N(\phi - z_0\tau_\phi, \tau_\phi^2)$, where $\tau\phi = 1 + a\phi$ and z_0 and a are constant.

Note: Each method is correct under more general assumptions than its predecessor. Methods 2, 3, and 4 are defined in terms of the percentile of G, the bootstrap distribution.

Source: Efron and Tibshirani (1986, Table 6) with permission from The Institute of Mathematical Statistics.

Bootstrap Methods: Regression

In a regression context an important distinction is between:

- resampling residuals (model-based resampling) and
- resampling cases.

We illustrate the difference by considering the model

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i,$$

where the residuals ϵ_i are such that $\mathrm{E}[\epsilon_i] = 0$, $i = 1, \ldots, n$, and are assumed uncorrelated.

Bootstrap Methods: Regression

The two methods are characterized according to whether we take F to be the distribution of Y only, or of $\{Y, \boldsymbol{X}\}$.

In the resampling residuals approach the covariates $\boldsymbol{x}_i$ are considered as fixed and bootstrap datasets are formed as

$$Y_i^{(b)} = f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) + \epsilon_i^{(b)},$$

where a number of options are available for sampling $\epsilon_i^{(b)}$, $b = 1, \dots, B$, $i = 1, \dots, n$.

The simplest non-parametric version is to sample $\epsilon_i^{(b)}$ with replacement from

$$e_i = y_i - f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) - \left(\frac{1}{n} \sum_{i=1}^{n} [y_i - f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})] \right).$$

If we are willing to assume (say) that $\epsilon_i \sim_{ind} \mathrm{N}(0, \sigma^2)$ then a parametric resampling residuals method samples $\epsilon_i^{(b)} \sim \mathrm{N}(0, \widehat{\sigma}^2)$.

Bootstrap Methods

The above methods have the advantage of respecting the “design”, that is the $\boldsymbol{x}_i$’s.

A disadvantage is that an assumption of homoscedasticity of errors is made by the non-parametric version and so this method cannot be used for data in which it is known that the variance changes with the mean (e.g., Poisson regression).

Bootstrap Methods

The resampling cases method forms bootstrap datasets by sampling with replacement from $\{Y_i, \boldsymbol{X}_i,\ i = 1, \ldots, n\}$.

Again parametric and non-parametric versions are available but the latter is preferred since the former requires a model for the joint distribution of the responses and covariates, which is likely to be difficult to specify.

Bootstrap Methods

When cases are resampled the design in each bootstrap sample will not in general correspond with that in the original dataset, which is not ideal, since even if the $\boldsymbol{x}$'s were not fixed by design they are conditioned upon in the analysis stage.

However, having random $\boldsymbol{x}$ will lead to wider interval estimates (and hence is conservative) and in practice inference will rarely be affected by this aspect, but can be awkward in some designed experiments.

Relationship between the Bootstrap and Sandwich Estimation

As detailed in Section 2.7.3 of Wakefield (2013):

- the variance of an estimator arising from the non-parametric bootstrap is approximately equal to that arising from sandwich estimation.

The Jackknife

The jackknife was introduced by Quenouille (1949, 1956) to reduce bias and Tukey (1958) suggested it could be used to estimate variances and calculate CIs.

The jackknife is another technique by which replicate samples are created from the original sample, but these samples are not generated randomly but rather systematically.

Here, the delete-1 jackknife will be described, in which *n* replicate datasets are created from the original by deleting a single observation at a time.

We lean heavily on Shao and Tu (2012) in the following short description.

The Jackknife

Let $\widehat{\theta}$ be an estimator of an unknown parameter.

Denote by $\widehat{\theta}_{(i)}$ an estimator of the same form as $\widehat{\theta}$ but with observation i removed.

Example: If $\widehat{\theta} = \overline{y}$, then

$$\widehat{\theta}_{(i)} = \overline{y}_{(i)} = \sum_{k \neq i} \frac{y_k}{n-1}.$$

The bias of $\widehat{\theta}$ is defined as

$$\mathrm{Bias}(\widehat{\theta}) = \mathrm{E}[\widehat{\theta}] - \theta.$$

Quenouille's jackknife bias estimator is

$$b_{\mathrm{JACK}} = (n-1)(\overline{\theta} - \widehat{\theta}),$$

where

$$\overline{\theta} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{(i)}.$$

The Jackknife

Then, the bias-reduced jackknife estimator of θ is

$$\widehat{\theta}_{\text{JACK}} = \widehat{\theta} - b_{\text{JACK}} = n\widehat{\theta} - (n-1)\overline{\theta}.$$

A heuristic justification for b_{JACK} and $\widehat{\theta}_{\text{JACK}}$ is the following.

Suppose that

$$\text{Bias}(\widehat{\theta}) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right),$$

where a and b are unknown, and do not depend on n.

Since $\widehat{\theta}_{(i)}$, $i = 1, \ldots, n$, are identically distributed,

$$\text{Bias}(\widehat{\theta}_{(i)}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right), \tag{33}$$

and Bias($\overline{\theta}$) has the same expression as in (33).

The Jackknife

Therefore,

$$\begin{aligned}
\mathrm{E}[b_{\mathrm{JACK}}] &= (n-1)\mathrm{E}[(\overline{\theta}-\theta)-(\widehat{\theta}-\theta)] = (n-1)[\mathrm{Bias}(\overline{\theta})-\mathrm{Bias}(\widehat{\theta})] \\
&= (n-1)\left[\left(\frac{1}{n-1}-\frac{1}{n}\right)a+\left(\frac{1}{(n-1)^2}-\frac{1}{n^2}\right)b\right]+O\left(\frac{1}{n^3}\right) \\
&= \frac{a}{n}+\frac{(2n-1)b}{n^2(n-1)}+O\left(\frac{1}{n^3}\right),
\end{aligned}$$

which means that b_{JACK}, the estimator of the bias of $\overline{\theta}$, is correct up to order n^{-2}.

It follows that

$$\begin{aligned}
\mathrm{Bias}(\widehat{\theta}_{\mathrm{JACK}}) &= \mathrm{E}[\widehat{\theta}_{\mathrm{JACK}}]-\theta \\
&= \mathrm{E}[\widehat{\theta}-b_{\mathrm{JACK}}]-\theta \\
&= \mathrm{Bias}(\widehat{\theta})-\mathrm{E}[b_{\mathrm{JACK}}] \\
&= -\frac{b}{n(n-1)}+O\left(\frac{1}{n^3}\right),
\end{aligned}$$

so the bias of $\widehat{\theta}_{\mathrm{JACK}}$ is of the order n^{-2}.

The Jackknife

Define the delete-1 jackknife variance estimator as

$$\widehat{V}_{\mathrm{JACK}}(\widehat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{(i)} - \widehat{\theta})^2.$$

Note that various other possibilities are available here, such as

$$\widehat{V}^{\star}_{\mathrm{JACK}}(\widehat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{(i)} - \overline{\theta})^2.$$

The Jackknife

To see why $(n-1)/n$ look at $\widehat{\theta} = \overline{y}$:

$$\overline{y}_{(i)} = \frac{1}{n-1}\sum_{k\neq i} y_k = \frac{1}{n-1}\left(\sum_{k=1}^{n} y_k - y_i\right) = \overline{y} - \frac{1}{n-1}(y_i - \overline{y}).$$

Then

$$\sum_{i=1}^{n}(\overline{y}_{(i)} - \overline{y})^2 = \sum_{i=1}^{n}\frac{(y_i - \overline{y})^2}{(n-1)^2} = \frac{s^2}{n-1}.$$

So

$$\widehat{V}_{\text{JACK}}(\widehat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\overline{y}_{(i)} - \overline{y})^2 = \frac{s^2}{n}.$$

The jackknife is generally less computer-intensive than the bootstrap.

A disadvantage is that the estimator may be inconsistent for a statistic that is not very smooth (this inconsistency can be rectified by using the delete-d jackknife).

Example: Lung Cancer and Radon

For the lung cancer and radon example we implement the non-parametric bootstrap resampling, with replacement, $B = 1000$ sets of n case triples $[Y^{\star}_{bi}, E^{\star}_{bi}, x^{\star}_{bi}]$, $b = 1, \dots, B$, $i = 1, \dots, n$.

Figure 3 displays the histogram of estimators arising from the bootstrap samples, along with the asymptotic normal approximations to the sampling distribution of the estimator under the Poisson and quasi-Poisson models.

We see that the distribution under the quasi-likelihood model is much wider than that under the Poisson model.

This is not surprising since we have already seen that the lung cancer data are overdispersed relative to a Poisson distribution.

The bootstrap histogram and quasi-Poisson sampling distribution are very similar, however.

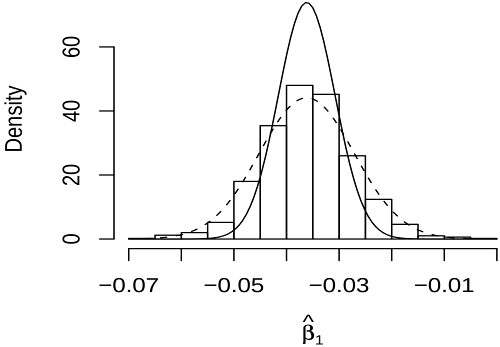

FIGURE 3: Sampling distribution of $\widehat{\beta}_1$ arising from the non-parametric bootstrap samples. The solid curve is the asymptotic distribution of the MLE under the Poisson model, and the dashed line is the asymptotic distribution under the quasi-Poisson model.

Example: Lung Cancer and Radon

Table 2 summarizes inference for β_1 for a number of different methods, and again confirms the similarity of asymptotic inference and the parametric bootstrap under the Poisson model.

The parametric bootstrap cannot be used with a quasi-likelihood model since there is no probability distribution for the data.

Point estimates from the Poisson, quasi-likelihood and sandwich approaches are identical.

Example: Lung Cancer and Radon

Inferential Method	$\widehat{\beta}_1$	s.e.$(\widehat{\beta}_1)$	95% confidence
	$(\times 10^3)$	$(\times 10^4)$	interval for $e^{10\beta_1}$
Poisson	-0.036	0.0054	0.954, 0.975
Quasi-Likelihood	-0.036	0.0090	0.947, 0.982
Quadratic Variance	-0.030	0.0085	0.955, 0.987
Sandwich Estimation	-0.036	0.0080	0.949, 0.980
Bootstrap Normal	-0.036	0.0087	0.948, 0.981
Bootstrap Percentile	-0.036	0.0087	0.949, 0.981

Table 2: Comparison of inferential approaches for the lung cancer example.

Hypothesis Testing

Two distinct aims:

- Determining whether a particular set of data is consistent with a particular hypothesis.
- Making a decision as to which of two hypotheses is best supported by the data.

A hypothesis is simple if the values of all parameters are completely specified, i.e. $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for a fixed value $\boldsymbol{\theta}_0$, otherwise it is composite.

We assume there exists a test statistic $T = T(\boldsymbol{Y})$ the distribution of which we wish to evaluate under H_0, and that large values of T suggest departures from H_0.

Hypothesis Testing

We evaluate the p-value, or significance level, as

$$p = p(\boldsymbol{Y}) = \Pr\{T(\boldsymbol{Y}) > T(\boldsymbol{y}) \mid \boldsymbol{\theta}_0\},$$

so that if this probability is "small" the data are inconsistent with H_0.

If $T(\boldsymbol{Y})$ is continuous then under H_0, $p(\boldsymbol{Y}) \sim U(0,1)$ so that the significance level is the observed $p(\boldsymbol{y})$.

The distribution of $T(\boldsymbol{Y})$ under H_0 may be known, or may be simulated to give a Monte Carlo test.

Hypothesis Testing

We distinguish between three procedures:

1. A pure significance test evaluates p but does not reject H_0, and is often viewed as more of an exploratory tool.

2. A *test of significance* sets a cut-off value α (e.g. $\alpha = 0.05$) and rejects H_0 if $p < \alpha$ (and has a corresponding critical region $T > T_\alpha$).

3. A hypothesis test goes one step further and specifies an alternative hypothesis, H_1, and take a decision as to which of H_0 and H_1 are chosen. In this context α is known as the Type I error. A Type II error occurs when H_0 is not rejected when it is false; to evaluate the Type II error specific alternative values of the parameters need to be considered. The power is defined as the probability of rejecting H_0 when it is false, we hope that under H_1 p tends to concentrate close to 0.

Hypothesis Testing

An important point to emphasize is that the consistency of the data with H_0 is being evaluated and we are not saying anything about the probability that the null hypothesis is true.

As usual in frequentist inference, H_0 is a fixed unknown and probability statements cannot be placed upon it.

Similarly p-values cannot be converted to probabilities of H_0 since they do not involve alternatives.

Score Tests

We begin with the simple hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. We have

$$\boldsymbol{I}(\boldsymbol{\theta})^{-1/2}\boldsymbol{S}_n(\boldsymbol{\theta}) \ \to_d \ \mathrm{N}_p(\mathbf{0}, \boldsymbol{I}_p),$$

where $\boldsymbol{S}_n$ is the score statistic.

Under the null hypothesis we have the test statistic

$$\boldsymbol{S}_n(\boldsymbol{\theta}_0)^{\intercal}\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0) \ \to_d \ \chi^2_p. \tag{34}$$

Wald Tests

The Wald statistic is based upon the asymptotic distribution

$$\boldsymbol{I}^{1/2}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \ \to_d \ \mathrm{N}_p(\mathbf{0}, \boldsymbol{I}_p), \tag{35}$$

Under the simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, the Wald statistic is given by

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ \to_d \ \chi^2_p \tag{36}$$

the quadratic form based on (35).

An alternative form of the statistic

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{I}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ \to_d \ \chi^2_p,$$

which follows because $\boldsymbol{I}(\widehat{\boldsymbol{\theta}}_n)$ is a consistent estimator of $\boldsymbol{I}(\boldsymbol{\theta}_0)$ by a weak law of large numbers.

Wald Tests

Under a composite null hypothesis

$$\begin{aligned}
(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^\top \boldsymbol{I}_{11.2}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) &\quad \to_d \quad \chi^2_r \\
(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^\top \boldsymbol{I}_{11.2}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) &\quad \to_d \quad \chi^2_r.
\end{aligned}$$

where

$$\boldsymbol{I}_{11.2} = \boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21}.$$

Likelihood Ratio Statistic

The likelihood ratio statistic is given by

$$R_n(\boldsymbol{\theta}) = \frac{L_n(\boldsymbol{\theta})}{L_n(\widehat{\boldsymbol{\theta}}_n)},$$

where $L_n(\boldsymbol{\theta})$ is the likelihood and $\widehat{\boldsymbol{\theta}}_n$ is the MLE so that $R(\boldsymbol{\theta}) \leqslant 1$.

A second order Taylor expansion of $l_n(\boldsymbol{\theta}_0)$ about $\widehat{\boldsymbol{\theta}}$ gives

$$l_n(\boldsymbol{\theta}_0) = l_n(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n)^{\mathsf{T}} \left.\frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}_n} + \frac{1}{2}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n)^{\mathsf{T}} \left.\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right|_{\boldsymbol{\theta}^*} (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n),$$

where $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$.

The middle term of the right hand side is zero, and

$$\left.\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right|_{\boldsymbol{\theta}^*} \to_p -\boldsymbol{I}(\boldsymbol{\theta}_0).$$

Likelihood Ratio Statistic

Hence,

$$-2\{l_n(\boldsymbol{\theta}_0) - l_n(\widehat{\boldsymbol{\theta}}_n)\} \approx (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\intercal} \boldsymbol{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

and so

$$-2\{l_n(\boldsymbol{\theta}_0) - l_n(\widehat{\boldsymbol{\theta}}_n)\} \ \rightarrow_d \ \chi^2_p. \tag{37}$$

Under a composite null hypothesis we obtain the likelihood ratio statistic,

$$-2 \log R(\widehat{\boldsymbol{\theta}}_0) = 2\{\log L(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) - \log L(\boldsymbol{\theta}_{10}, \widehat{\boldsymbol{\theta}}_{20})\} \rightarrow_d \chi^2_r.$$

Comparison of Test Statistics

The score, Wald, and likelihood ratio test statistics are asymptotically equivalent but are not equally well-behaved in finite samples.

An advantage of the Wald statistic is that confidence intervals can be derived directly from the statistic and so there is a direct link between estimation and testing. Interpretation is also more straightforward.

A major drawback of the Wald statistic is that it is not invariant under reparameterization.

The score test statistic derived from the estimating function is invariant under reparameterization, providing the expected, rather than the observed, information is used.

Comparison of Test Statistics

The score statistic may be evaluated without second derivatives if $\boldsymbol{S}_n(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0)^\intercal$ is used; this may be useful if the second derivatives are complex or unavailable.

The score statistic requires the value of the score at the null but the MLE is not required.

If $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ are close then the three statistics will tend to agree.

The likelihood ratio statistic is invariant under reparameterization.

Confidence intervals derived from likelihood ratio tests always preserve the support of the parameter, unlike score- and Wald-based intervals.

Example: Poisson Mean

Suppose we have data

$$Y_i \mid \lambda \sim_{iid} \text{Poisson}(\lambda),$$

$i = 1, \ldots, n$, and we are interested in $H_0 : \lambda = \lambda_0$.

We have

$$\begin{aligned} l_n(\lambda) &= -n\lambda + n\overline{Y} \log \lambda, \\ S_n(\lambda) &= -n + \frac{n\overline{Y}}{\lambda} = \frac{n(\overline{Y} - \lambda)}{\lambda}, \\ I_n(\lambda) &= \frac{n}{\lambda}. \end{aligned}$$

The score and Wald statistics are given by

$$\frac{n(\lambda_0 - \overline{Y})^2}{\lambda_0} \rightarrow_d \chi_1^2.$$

The likelihood ratio statistic is given by

$$2n\{(\lambda_0 - \overline{Y}) + \overline{Y}(\log \overline{Y} - \log \lambda_0)\} \rightarrow_d \chi_1^2.$$

Suppose we observe $\sum_{i=1}^{20} y_i = 12$ events in $n = 20$ trials so that $\overline{y} = 0.6$ and we are interested in testing the null hypothesis $H_0 : \lambda_0 = 1.0$.

The score and Wald statistics are 3.20 and the likelihood ratio statistic is 3.74 with associated observed significance levels 7.3% and 5.4%.

The Three Statistics

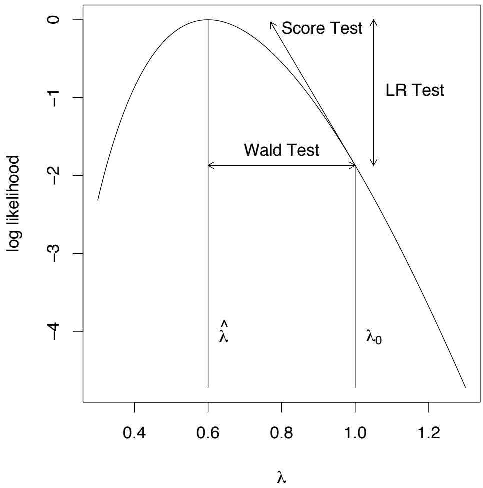

Figure 4: Geometric interpretation of score, Wald and likelihood ratio (LR) statistics, for a test of $H_0 : \lambda_0 = 1$ with data resulting in $\widehat{\lambda} = \overline{y} = 0.6$.

Critique of Hypothesis Testing

There are a number of difficulties with hypothesis testing:

- The first is that statistical rather than practical significance is being assessed. A particular ramification of this is that for large *n* the estimate may be very close to the null value, but the null will be rejected because there is high power and small differences will be highly significant. In this respect confidence intervals are appealing inferential tools since the size of the estimate is apparent.
- A second and more fundamental problem is the difficulty of deciding upon a significance level α, in particular the interpretation of an observed significance level is impossible without knowledge of the sample size.
- As *n* increases α should decrease but specific prescriptions for this change are not routinely used.
- So, the use of the Neyman-Pearson lemma with a fixed significance level is also not to be recommended since one should always try to balance Type I and Type II errors.

Discussion

- Frequentist inference focusses on the sampling distribution of an estimator over hypothetical repeated draws from the population of interest, under the design used to collect the data.
- Asymptotic normality is heavily leaned under so that point estimates and standard errors are sought.
- Likelihood: Very popular, and provides efficient inference under correct model specification.
- Many flavors of likelihood: quasi-likelihood, profile likelihood, modified profile likelihood, partial likelihood, marginal likelihood, conditional likelihood,...
- So many versions may be seen as an advantage (tune to context) or a disadvantage (what the hell am I supposed to do?).

Discussion

- Estimating Equations: A strong emphasis on obtaining estimates with desirable properties (consistency, accurate coverage of CIs) under conditions other than those used to derive the estimates.

- Sandwich estimation is favored by many in biostatistics and economics (White, 1982), in particular.

- Resampling methods such as the bootstrap and jackknife are also ubiquitous, since they avoid tedious theory, and can reduce the reliance on assumptions.

- But if you have small *n*, inference is tricky, and model checking difficult.

- Bayesian methods allow “exact” inference, even for small *n*, but at the cost of assumptions (and model checking will be difficult).

References

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Rubin, D. B. (1981). The Bayesian bootstrap. *The annals of statistics*, pages 130–134.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. In *Annals of Mathematical statistics*, volume 29, pages 614–614.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.