Wald (and score) tests

- MLEs have an approximate multivariate normal sampling distribution for large samples (Thanks Mr. Wald.)
- Approximate mean vector = vector of true parameter values for large samples
- Asymptotic variance-covariance matrix is easy to estimate
- H_0: $\mathbf{C}\boldsymbol{\theta} = \boldsymbol{h}$ (Linear hypothesis)

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$$

$$\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h} \text{ is multivariate normal as } n \rightarrow \infty$$

Leads to a straightforward chisquare test

- Called a Wald test
- Based on the full model (unrestricted by any null hypothesis)
- Asymptotically equivalent to the LR test
- Not as good as LR for smaller samples
- Very convenient sometimes

Example of H_0: $\boldsymbol{C\theta=h}$

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots \theta_7)$, with

$$H_0 : \theta_1 = \theta_2,\ \theta_6 = \theta_7,\ \frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6)$$

$$\left[\begin{array}{rrrrrrr} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{array} \right] \left[\begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right]$$

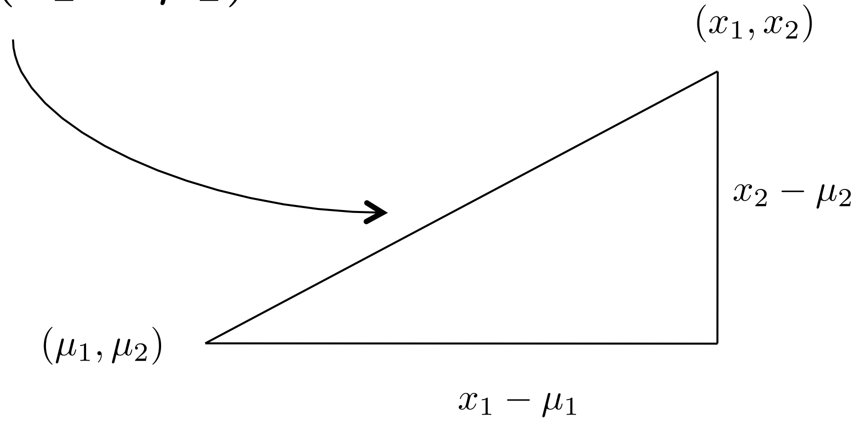
Multivariate Normal

- Univariate Normal
  - $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$
  - $\frac{(x-\mu)^2}{\sigma^2}$ is the squared Euclidian distance between x and μ, in a space that is stretched by σ^2.
  - $\frac{(X-\mu)^2}{\sigma^2} \sim \chi^2(1)$
- Multivariate Normal
  - $f(\mathbf{x}) = \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{k}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$
  - $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is the squared Euclidian distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, in a space that is warped and stretched by $\mathbf{\Sigma}$.
  - $(\mathbf{X}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2(k)$

Distance: Suppose $\boldsymbol{\Sigma}$ = $\mathbf{I}_2$

$$\begin{aligned}
d^2 &= (\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \\
&= \left[\begin{array}{cc} x_1-\mu_1, & x_2-\mu_2 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right] \\
&= \left[\begin{array}{cc} x_1-\mu_1, & x_2-\mu_2 \end{array}\right] \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right] \\
&= (x_1-\mu_1)^2+(x_2-\mu_2)^2
\end{aligned}$$

$$d = \sqrt{(x_1-\mu_1)^2+(x_2-\mu_2)^2}$$



The multivariate normal reduces to the univariate normal when $p = 1$. Other properties of the multivariate normal include the following.

1. $E(\mathbf{X}) = \boldsymbol{\mu}$
2. $V(\mathbf{X}) = \boldsymbol{\Sigma}$
3. If $\mathbf{c}$ is a vector of constants, $\mathbf{X} + \mathbf{c} \sim N(\mathbf{c} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$
4. If $\mathbf{A}$ is a matrix of constants, $\mathbf{AX} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
5. All the marginals (dimension less than p) of $\mathbf{X}$ are (multivariate) normal, but it is possible in theory to have a collection of univariate normals whose joint distribution is not multivariate normal.
6. For the multivariate normal, zero covariance implies independence. The multivariate normal is the only continuous distribution with this property.
7. The random variable $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom.
8. After a bit of work, the multivariate normal likelihood may be written as

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})\right\}, \quad \text{(A.15)}$$

   where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ is the sample variance-covariance matrix (it would be unbiased if divided by $n - 1$).

Approximately, for large *n*

$$\widehat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, \mathbf{V}(\boldsymbol{\theta})) \qquad \mathbf{C}\widehat{\boldsymbol{\theta}} \sim N_k(\mathbf{C}\boldsymbol{\theta}, \mathbf{CVC}')$$

If $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$ is true,

$$(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})'(\mathbf{CVC}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h}) \sim \chi^2(r)$$

$\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ is unknown, but

$$\begin{aligned} W &= (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})'(\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h}) \\ &\sim \chi^2(r) \end{aligned}$$

Wald Test Statistic

$$W = (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})'(\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})$$

- Approximately chi-square with $df = r$ for large N if H_0: $\boldsymbol{C\theta}=\boldsymbol{h}$ is true
- Matrix $\boldsymbol{C}$ is $r \times k$, $r \leq k$, rank r
- Matrix $\boldsymbol{V(\theta)}$ is called the “Asymptotic Covariance Matrix” of $\widehat{\boldsymbol{\theta}}$
- $\widehat{\mathbf{V}}$ is the *estimated* Asymptotic Covariance Matrix
- How to calculate $\widehat{\mathbf{V}}$?

Fisher Information Matrix $\mathcal{J}$

- Element (*i,j*) is $-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta},\mathbf{Y}),\ \text{where}$

- The log likelihood is

$$\ell(\boldsymbol{\theta},\mathbf{Y}) = \sum_{i=1}^{N} \log f(Y_i;\boldsymbol{\theta}).$$

- This is sometimes called the *observed* Fisher information – based on the observed data $Y_1, \ldots, Y_N$

For a random sample $Y_1, \ldots, Y_n$ (No x values)

- Independent and identically distributed
- Fisher information in a single observation is

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[E[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\boldsymbol{\theta})]\right]$$

- Approximate the expected value with a sample mean

$$\widehat{\boldsymbol{\mathcal{I}}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y_i;\boldsymbol{\theta})$$

Fisher Information in the whole sample

- $N \cdot \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$
- Estimate it with the *observed information* $N \cdot \widehat{\boldsymbol{\mathcal{I}}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\theta})$
- Evaluate this at the MLE and we have a statistic: $$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})$$
- Call it the **Fisher Information.** Technically it's the observed Fisher information evaluated at the MLE.
- Applies when there are x values.

For a simple logistic regression

- $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_0, \beta_1)$

- $\ell(\boldsymbol{\beta}, \mathbf{y}) = \beta_0 \sum_{i=1}^{N} y_i + \beta_1 \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} \log(1 + e^{\beta_0 + \beta_1 x_i})$

$$\begin{aligned}
\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\beta}}) &= -\left[\begin{array}{cc} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{array}\right]\Bigg|_{\beta_0 = \widehat{\beta}_0, \beta_1 = \widehat{\beta}_1} \\
&= \left[\begin{array}{cc} \sum_{i=1}^{N} \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{(1+e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i})^2} & \sum_{i=1}^{N} \frac{x_i e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{(1+e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i})^2} \\ \sum_{i=1}^{N} \frac{x_i e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{(1+e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i})^2} & \sum_{i=1}^{N} \frac{x_i^2 e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{(1+e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i})^2} \end{array}\right]
\end{aligned}$$

The asymptotic covariance matrix is the inverse of the Fisher Information

Meaning that the estimated asymptotic covariance matrix of the MLE is the inverse of the observed Fisher information matrix, evaluated at the MLE.

$$\widehat{\mathbf{V}} = \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1}, \text{ where } \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}) = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}, \mathbf{Y}) \right] \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

Connection to Numerical Optimization

- Suppose we are minimizing the minus log likelihood by a direct search.
- We have reached a point where the gradient is close to zero. Is this point a minimum?
- Hessian is a matrix of mixed partial derivatives. If its determinant is positive at a point, the function is concave up there.
- It's *the* multivariable second derivative test.
- The Hessian at the MLE is <u>exactly</u> the Fisher Information:

$$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} - \ell(\boldsymbol{\theta}, \mathbf{Y})\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

Asymptotic Covariance Matrix $\widehat{\mathbf{V}}$ is Useful

- Square roots of diagonal elements are standard errors – Denominators of Z-test statistics. Also used for confidence intervals.
- Diagonal elements converge to the respective Cramér-Rao lower bounds for the variance of an estimator: "Asymptotic efficiency"
- And of course there are Wald tests

$$W = (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})^\prime (\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}^\prime)^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{h})$$

Score Tests

- $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, size $k \times 1$
- $\widehat{\boldsymbol{\theta}}_0$ is the MLE under H_0, size $k \times 1$
- $\mathbf{u}(\boldsymbol{\theta}) = (\frac{\partial \ell}{\partial \theta_1}, \ldots \frac{\partial \ell}{\partial \theta_k})'$ is the gradient.
- $\mathbf{u}(\widehat{\boldsymbol{\theta}}) = 0$
- If H_0 is true, $\mathbf{u}(\widehat{\boldsymbol{\theta}}_0)$ should also be close to zero.
- Under H_0 for large N, $\mathbf{u}(\widehat{\boldsymbol{\theta}}_0) \sim N_k(\mathbf{0}, \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}))$, approximately.
- And,

$$S = \mathbf{u}(\widehat{\boldsymbol{\theta}}_0)' \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_0)^{-1} \mathbf{u}(\widehat{\boldsymbol{\theta}}_0) \sim \chi^2(r)$$

Where *r* is the number of restrictions imposed by H_0