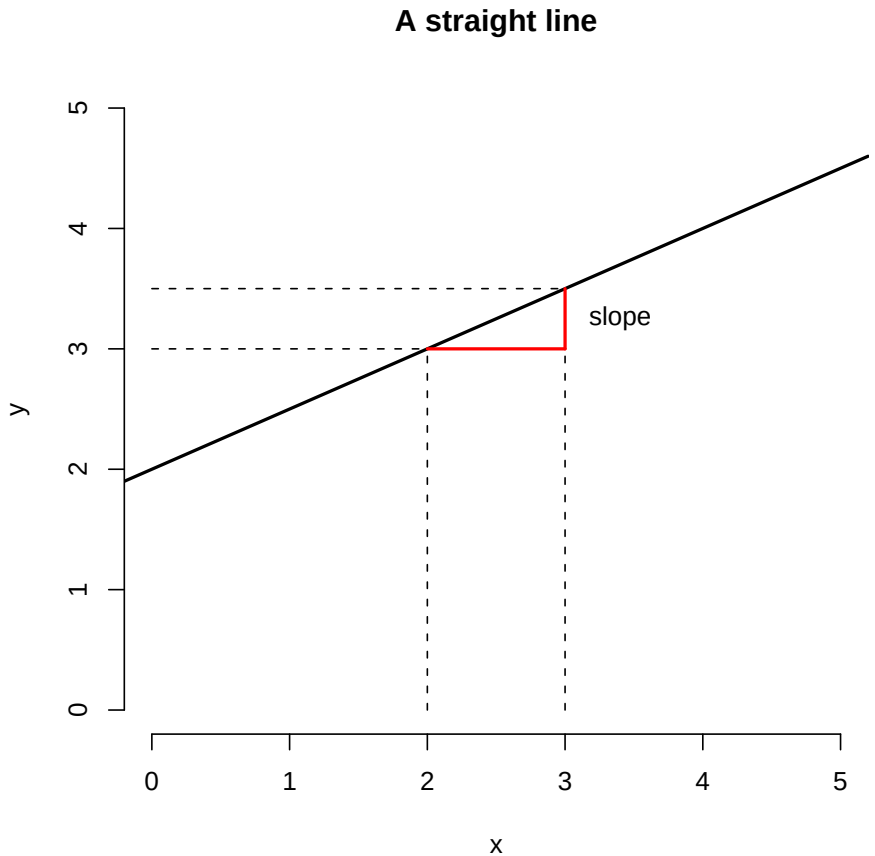
Derivatives of straight lines

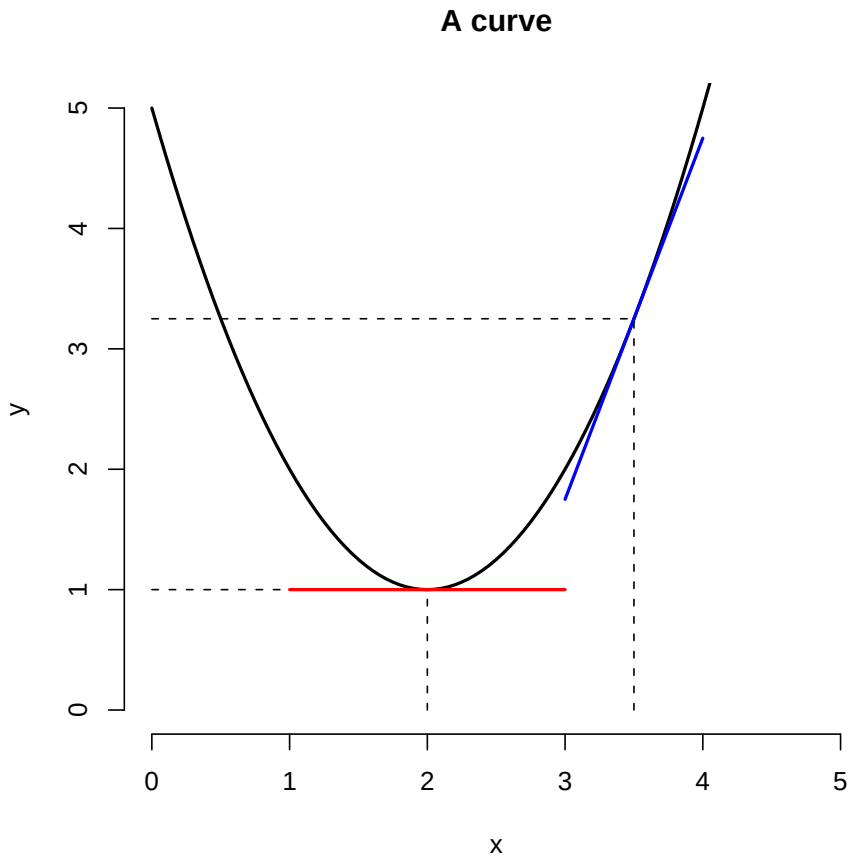


$$y = \frac{1}{2}x + 2$$

$$\frac{dy}{dx} = \frac{1}{2}$$

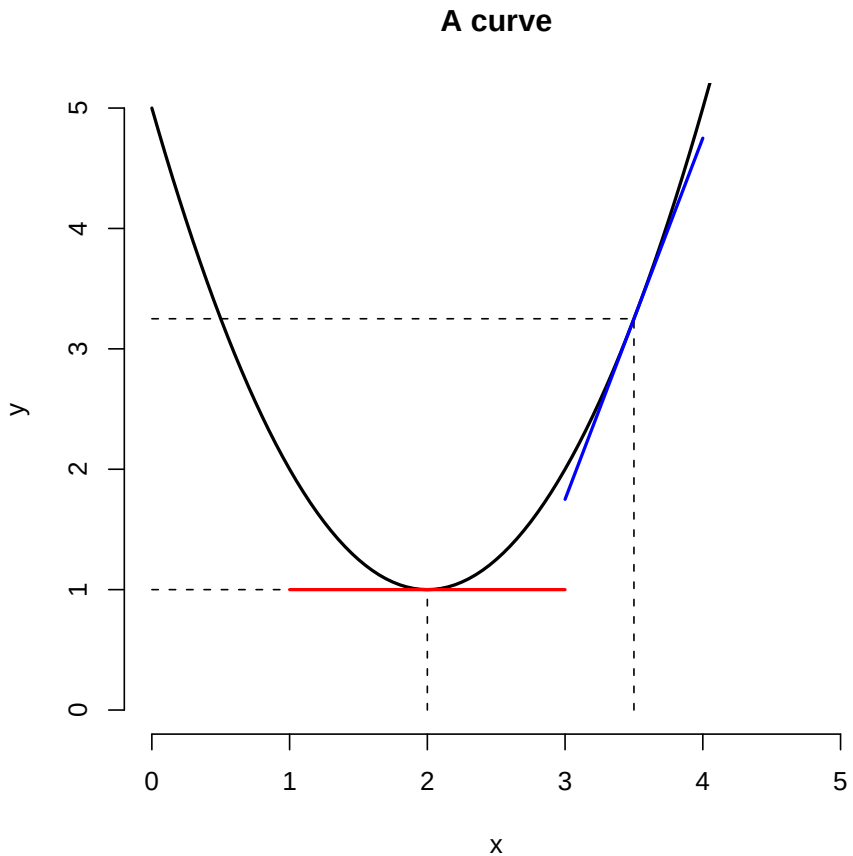
Derivatives of curves

$$y = x^2 - 4x + 5$$



Derivatives of curves

$$y = x^2 - 4x + 5$$

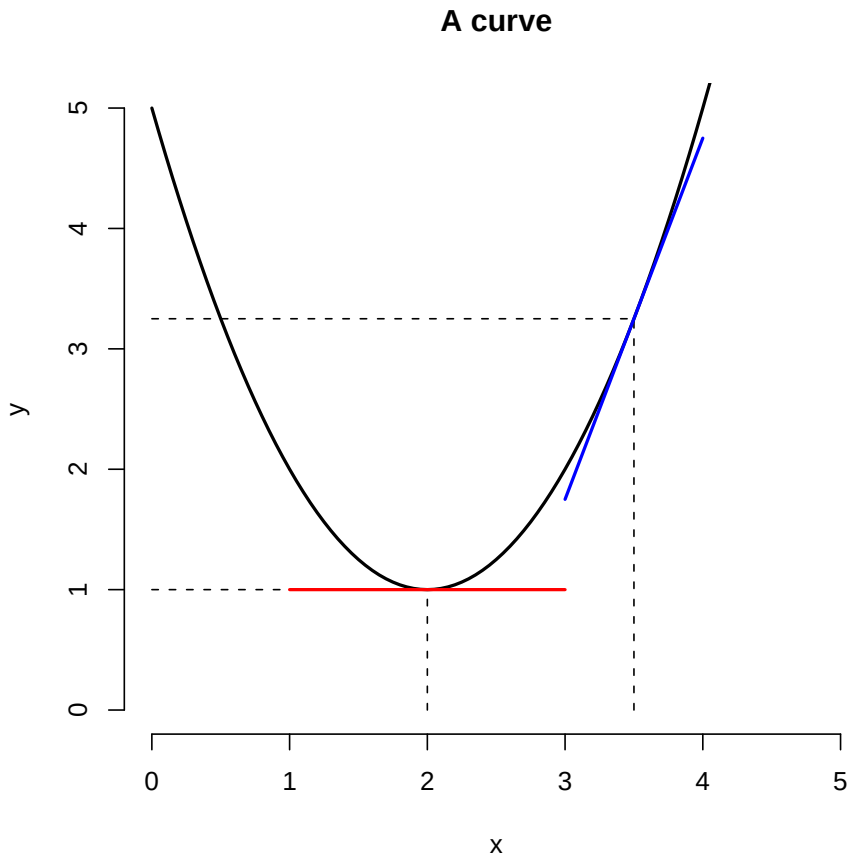


$$\frac{d(ax^b)}{dx} = bax^{b-1}$$

$$\frac{d(a+b)}{dx} = \frac{da}{dx} + \frac{db}{dx}$$

Derivatives of curves

$$y = x^2 - 4x + 5$$



$$\frac{d(ax^b)}{dx} = bax^{b-1}$$

$$\frac{d(a+b)}{dx} = \frac{da}{dx} + \frac{db}{dx}$$

$$\frac{dy}{dx} = 2x - 4$$

Finding location of peaks and valleys

A maximum or minimum value of a curve can be found by setting the derivative equal to zero.

Finding location of peaks and valleys

A maximum or minimum value of a curve can be found by setting the derivative equal to zero.

$$y = x^2 - 4x + 5$$
$$\frac{dy}{dx} = 2x - 4 = 0$$
$$2x = 4$$
$$x = \frac{4}{2} = 2$$

Finding location of peaks and valleys

A maximum or minimum value of a curve can be found by setting the derivative equal to zero.

$$y = x^2 - 4x + 5$$
$$\frac{dy}{dx} = 2x - 4 = 0$$
$$2x = 4$$
$$x = \frac{4}{2} = 2$$

So at $x = 2$ we will find a (local) maximum or minimum value - the plot shows that in this case it is a minimum.

Derivatives in regression

This concept of finding the minimum or maximum is crucial for regression analysis.

- With **ordinary least squares** (OLS), we estimate the β-coefficients by finding the minimum of the sum of squared errors.
- With **maximum likelihood** (ML), we estimate the β-coefficients by finding the maximum point of the (log)likelihood function.

Derivatives of matrices

The derivative of a matrix consists of the matrix containing the derivatives of each element of the original matrix.

Derivatives of matrices

The derivative of a matrix consists of the matrix containing the derivatives of each element of the original matrix.

$$\mathbf{M} = \begin{bmatrix} x^2 & 2x & 3 \\ 2x+4 & 3x^3 & 2x \\ 3 & 2 & 4x \end{bmatrix}$$

$$\frac{d\mathbf{M}}{dx} = \begin{bmatrix} 2x & 2 & 0 \\ 2 & 9x^2 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

Derivatives of matrices

$$\frac{d(\mathbf{v}'\mathbf{x})}{d\mathbf{x}} = \mathbf{v}$$
$$\frac{d(\mathbf{x}'\mathbf{A}\mathbf{x})}{d\mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$$
$$\frac{d(\mathbf{B}\mathbf{x})}{d\mathbf{x}} = \mathbf{B}'$$
$$\frac{d^2(\mathbf{x}'\mathbf{A}\mathbf{x})}{d\mathbf{x}d\mathbf{x}'} = \mathbf{A} + \mathbf{A}'$$

Preliminaries

Ordinary least squares and its variations (GLS, WLS, 2SLS, etc.) is very flexible and applicable in many circumstances, but for some models (e.g. limited dependent variable models) we need more flexible estimation procedures.

The most prominent alternative estimators are **maximum likelihood** (ML) and **Bayesian** estimators. This lecture is about the former.

Logarithms

Some rules about logarithms, which are important for understanding ML:

$$\log ab = \log a + \log b$$
$$\log e^a = a$$
$$\log a^b = b \log a$$
$$\log a > \log b \quad \text{iff} \quad a > b$$
$$\frac{d(\log a)}{da} = \frac{1}{a}$$
$$\frac{d(\log f(a))}{da} = \frac{d(f(a))/da}{f(a)}$$

Likelihood: intuition

Dice example.

Likelihood: intuition

Dice example.

Imagine a dice is thrown with unknown number of sides k. We know after throwing the dice that the outcome is 5.

How likely is this outcome if $k = 0$? And $k = 1$? Continue until $k = 10$. Plot of this is the likelihood function.

Likelihood

"Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes."

(Wolfram Mathworld)

Preliminaries

Looking first at just univariate models, we can express the model as the **probability density function** (PDF) $f(\mathbf{y}, \boldsymbol{\theta})$, where $\mathbf{y}$ is the dependent variable, $\boldsymbol{\theta}$ the set of parameters, and $f(\cdot)$ expresses the model specification.

Preliminaries

Looking first at just univariate models, we can express the model as the **probability density function** (PDF) $f(\mathbf{y}, \boldsymbol{\theta})$, where $\mathbf{y}$ is the dependent variable, $\boldsymbol{\theta}$ the set of parameters, and $f(\cdot)$ expresses the model specification.

ML is purely **parametric** - we need to make strong assumptions about the shape of $f(\cdot)$ before we can estimate $\boldsymbol{\theta}$.

Likelihood

Given $\boldsymbol{\theta}$, $f(\cdot, \boldsymbol{\theta})$ is the PDF of $\mathbf{y}$.

Likelihood

Given $\boldsymbol{\theta}$, $f(\cdot, \boldsymbol{\theta})$ is the PDF of $\mathbf{y}$.

Given $\mathbf{y}$, $f(\mathbf{y}, \cdot)$ cannot be interpreted as a PDF and is therefore called the **likelihood function**.

Likelihood

Given $\boldsymbol{\theta}$, $f(\cdot, \boldsymbol{\theta})$ is the PDF of $\mathbf{y}$.

Given $\mathbf{y}$, $f(\mathbf{y}, \cdot)$ cannot be interpreted as a PDF and is therefore called the **likelihood function**.

$\hat{\boldsymbol{\theta}}^{ML}$ is the $\boldsymbol{\theta}$ that maximizes this likelihood function.

(Davidson & MacKinnon 1999, 393-396)

(Log)likelihood function

Generally, observations are assumed to be **independent**, in which case the joint density of the entire sample is the product of the densities of the individual observations.

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})$$

(Log)likelihood function

Generally, observations are assumed to be **independent**, in which case the joint density of the entire sample is the product of the densities of the individual observations.

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})$$

It is easier to work with the log of this function, because sums are easier to deal with than products and the logarithmic transformation is **monotonic**:

$$\ell(\mathbf{y}, \boldsymbol{\theta}) \equiv \log f(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(y_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(y_i, \boldsymbol{\theta})$$

(Davidson & MacKinnon 1999, 393-396)

Example: exponential distribution

For example, take the PDF of $\mathbf{y}$ to be

$$f(\mathbf{y}, \theta) = \theta e^{-\theta \mathbf{y}} \qquad y_i > 0, \quad \theta > 0$$

Example: exponential distribution

For example, take the PDF of $\mathbf{y}$ to be

$$f(\mathbf{y}, \theta) = \theta e^{-\theta \mathbf{y}} \qquad y_i > 0, \quad \theta > 0$$

This distribution is useful when modeling something which is by definition positive.

(Davidson & MacKinnon 1999, 393-396)

Example: exponential distribution

Deriving the loglikelihood function:

$$f(\mathbf{y}, \theta) = \prod_{i=1}^{n} f(y_i, \theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i}$$

$$\ell(\mathbf{y}, \theta) = \sum_{i=1}^{n} \log f(y_i, \theta) = \sum_{i=1}^{n} \log(\theta e^{-\theta y_i})$$

$$= \sum_{i=1}^{n} (\log \theta - \theta y_i) = n \log \theta - \theta \sum_{i=1}^{n} y_i$$

(Davidson & MacKinnon 1999, 393-396)

Maximum Likelihood Estimates

The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which maximizes the (log)likelihood function $\ell(\mathbf{y}, \boldsymbol{\theta})$.

Maximum Likelihood Estimates

The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which maximizes the (log)likelihood function $\ell(\mathbf{y}, \boldsymbol{\theta})$.

In some cases, such as the above exponential distribution or the linear model, we can find this value analytically, but taking the derivative, and setting equal to zero.

Maximum Likelihood Estimates

The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which maximizes the (log)likelihood function $\ell(\mathbf{y}, \boldsymbol{\theta})$.

In some cases, such as the above exponential distribution or the linear model, we can find this value analytically, but taking the derivative, and setting equal to zero.

In many other cases, there is no such analytical solution, and we use numerical search algorithms to find the value.

Example: exponential distribution

To find the maximum, we take the derivative and set this equal to zero:

$$\ell(\mathbf{y}, \theta) = n \log \theta - \theta \sum_{i=1}^{n} y_i$$

$$\frac{d(n \log \theta - \theta \sum_{i=1}^{n} y_i)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} y_i$$

$$\frac{n}{\hat{\theta}^{ML}} - \sum_{i=1}^{n} y_i = 0$$

$$\hat{\theta}^{ML} = \frac{n}{\sum_{i=1}^{n} y_i}$$

(Davidson & MacKinnon 1999, 393-396)

Numerical methods

In many cases, closed form solutions like the examples above do not exist and numerical methods are necessary. A computer algorithm searches for the value that maximizes the loglikelihood.

Numerical methods

In many cases, closed form solutions like the examples above do not exist and numerical methods are necessary. A computer algorithm searches for the value that maximizes the loglikelihood.

In R: `optim()`

Newton's Method

$$\boldsymbol{\theta}_{(m+1)} = \boldsymbol{\theta}_{(m)} - \mathbf{H}^{-1}_{(m)}\mathbf{g}_{(m)}$$

(Davidson & MacKinnon 1999, 401)

Linear model

Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Linear model

Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

We assume ε to be independently distributed and therefore, conditional on $\mathbf{X}$, $\mathbf{y}$ is assumed to be.

$$f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}$$

(Davidson & MacKinnon 1999, 396-398)

Linear model: loglikelihood function

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}$$

$$\ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}\right)$$

Linear model: loglikelihoodfunction

To "zoom in" on one part of this function:

$$\begin{aligned}
\log \frac{1}{\sqrt{2\pi\sigma^2}} &= \log \frac{1}{\sigma\sqrt{2\pi}} \\
&= \log 1 - \log(\sigma\sqrt{2\pi}) \\
&= 0 - (\log \sigma + \log \sqrt{2\pi}) \\
&= -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \log 2\pi, \text{using} \\
\log \sigma &= \frac{1}{2} \log \sigma^2
\end{aligned}$$

Linear model: loglikelihood function

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}$$

$$\ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}\right)$$

Linear model: loglikelihood function

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}$$

$$\ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log f_i(y_i, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2\right)$$

$$= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2$$

$$= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Linear model: estimating σ^2

$$\frac{\partial \ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{n}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$n = \frac{1}{\hat{\sigma}^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Thus $\hat{\sigma}^2$ is a function of $\boldsymbol{\beta}$.

Linear model: estimating σ^2

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\mathbf{e}'\mathbf{e}}{n}$$

Note that this is almost the equivalent of the variance estimator of OLS ($\frac{\mathbf{e}'\mathbf{e}}{n-k}$). This estimator is a **biased**, but **consistent**, estimator of $\hat{\sigma}^2$.

Linear model: estimating β

$$\ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\begin{aligned}\ell(\mathbf{y}, \boldsymbol{\beta}) &= -\frac{n}{2}\log(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) - \frac{n}{2}\log 2\pi - \frac{n}{2}\end{aligned}$$

Linear model: estimating β

$$\ell(\mathbf{y}, \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$$\begin{aligned}\ell(\mathbf{y}, \beta) &= -\frac{n}{2}\log(\frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)) - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2(\frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta))}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ &= -\frac{n}{2}\log(\frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)) - \frac{n}{2}\log 2\pi - \frac{n}{2}\end{aligned}$$

Because the middle term is equivalent to minus $n/2$ times the log of SSR, maximizing $\ell(\mathbf{y}, \beta)$ with respect to β is the equivalent to minimizing SSR: $\hat{\beta}^{ML} = \hat{\beta}^{OLS}$.

Linear model

$$\ell(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

```
ll <- function(par, X, y) {
  s2 <- exp(par[1])
  beta <- par[-1]
  n <- length(y)
  r <- y - X %*% beta
  - n/2 * log(s2) - n/2 * log(2*pi) - 1/(2*s2) * sum(r^2)
}
mlest <- optim(c(1,0,0,0), ll, NULL, X, y,
   control = list(fnscale = -1), hessian=TRUE)
```

Linear model

Some practical tricks:

- By default, `optim()` minimizes rather than maximizes, hence the need for `fnscale = -1`.
- To get $\hat{\sigma}^2$, we need to take the exponent of the first parameter - this is done to make sure σ^2 is always assumed positive.

Gradient vector

The **gradient vector** or **score vector** is a vector with typical element:

$$g_j(\mathbf{y}, \boldsymbol{\theta}) \equiv \frac{\partial \ell(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_j},$$

i.e. the vector of partial derivatives of the loglikelihood function towards each parameter in $\boldsymbol{\theta}$.

(Davidson & MacKinnon 1999, 400)

Hessian matrix

$\mathbf{H}(\boldsymbol{\theta})$ is a $k \times k$ matrix with typical element

$$\mathbf{h}_{ij} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \theta_j},$$

i.e. the matrix of second derivatives of the loglikelihood function.

(Davidson & MacKinnon 1999, 401, 407)

Hessian matrix

$\mathbf{H}(\boldsymbol{\theta})$ is a $k \times k$ matrix with typical element

$$\mathbf{h}_{ij} = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \theta_j},$$

i.e. the matrix of second derivatives of the loglikelihood function.

Asymptotic equivalent: $\mathcal{H}(\boldsymbol{\theta}) \equiv \text{plim}_{n\to\infty} \frac{1}{n}\mathbf{H}(\mathbf{y}, \boldsymbol{\theta})$.

(Davidson & MacKinnon 1999, 401, 407)

Information matrix

$$\mathbf{I}(\theta) = \sum_{i=1}^{n} E_\theta((\mathbf{G}_i(\mathbf{y}, \theta))' \mathbf{G}_i(\mathbf{y}, \theta))$$

$\mathbf{I}(\theta)$ is the **information matrix**, or the covariance matrix of the score vector.

(Davidson & MacKinnon 1999, 406-407)

Information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} E_{\boldsymbol{\theta}}((\mathbf{G}_i(\mathbf{y}, \boldsymbol{\theta}))' \mathbf{G}_i(\mathbf{y}, \boldsymbol{\theta}))$$

$\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix**, or the covariance matrix of the score vector.

There is also an asymptotic equivalent, $\mathcal{I}(\boldsymbol{\theta}) \equiv \text{plim}_{n\to\infty} \frac{1}{n}\mathbf{I}(\boldsymbol{\theta})$.

(Davidson & MacKinnon 1999, 406-407)

Information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} E_{\boldsymbol{\theta}}((\mathbf{G}_i(\mathbf{y}, \boldsymbol{\theta}))' \mathbf{G}_i(\mathbf{y}, \boldsymbol{\theta}))$$

$\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix**, or the covariance matrix of the score vector.

There is also an asymptotic equivalent, $\mathcal{I}(\boldsymbol{\theta}) \equiv \text{plim}_{n\to\infty} \frac{1}{n}\mathbf{I}(\boldsymbol{\theta})$.

$\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta})$ - they both measure the **amount of curvature** in the loglikelihood function.

(Davidson & MacKinnon 1999, 406-407)

Asymptotics

Under some weak conditions, ML estimators are **consistent**, **asymptotically efficient** and **asymptotically normally distributed**.

(Davidson & MacKinnon 1999, 402-403)

Asymptotics

Under some weak conditions, ML estimators are **consistent**, **asymptotically efficient** and **asymptotically normally distributed**.

New ML estimators thus do not require extensive proof of this - once it is shown it is an ML estimator, it "inherits" those properties.

(Davidson & MacKinnon 1999, 402-403)

Variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{ML}$

$$V(plim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\theta})) = H^{-1}(\boldsymbol{\theta})I(\boldsymbol{\theta})H^{-1}(\boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta}),$$

the latter being true iff $\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta})$ is true.

(Davidson & MacKinnon 1999, 409-411)

Variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{ML}$

$$V(plim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\theta})) = H^{-1}(\boldsymbol{\theta})I(\boldsymbol{\theta})H^{-1}(\boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta}),$$

the latter being true iff $\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta})$ is true.

One estimator for the variance is: $V(\hat{\boldsymbol{\beta}}^{ML}) = -\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}}^{ML})$. For alternative estimators see reference below.

(Davidson & MacKinnon 1999, 409-411)

Estimating $V(\hat{\beta}^{ML})$

```
mlest <- optim(..., hessian=TRUE)
V <- solve(-mlest$hessian)
sqrt(diag(V))
```

Exercise

Imagine, we take a random sample of N colored balls from a vase, with replacement. In the vase, a fraction p of the balls are yellow and the other ones blue. In our sample, there are q yellow balls. Assuming that $y_i = 1$ if the ball is yellow and $y_i = 0$ for blue, the probability of obtaining our sample is:

$$P(q) = p^q(1-p)^{N-q},$$

which is the likelihood function with unknown parameter p. Write down the loglikelihood function.

Exercise

Imagine, we take a random sample of N colored balls from a vase, with replacement. In the vase, a fraction p of the balls are yellow and the other ones blue. In our sample, there are q yellow balls. Assuming that $y_i = 1$ if the ball is yellow and $y_i = 0$ for blue, the probability of obtaining our sample is:

$$P(q) = p^q(1-p)^{N-q},$$

which is the likelihood function with unknown parameter p. Write down the loglikelihood function.

$$\ell(p) = \log(p^q(1-p)^{N-q}) = q\log p + (N-q)\log(1-p)$$

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Take derivative towards p.

Set equal to zero and find $\hat{p}^{ML}$.

(Verbeek 2008, 172-173)

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Take derivative towards p.

$$\frac{d\ell(p)}{dp} = \frac{q}{p} + \frac{N - q}{1 - p}$$

Set equal to zero and find $\hat{p}^{ML}$.

(Verbeek 2008, 172-173)

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Take derivative towards p.

$$\frac{d\ell(p)}{dp} = \frac{q}{p} + \frac{N - q}{1 - p}$$

Set equal to zero and find $\hat{p}^{ML}$.

$$\frac{q}{\hat{p}} + \frac{N - q}{1 - \hat{p}} = 0 \qquad \Longrightarrow \qquad \hat{p} = \frac{q}{N}$$

(Verbeek 2008, 172-173)

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Using the loglikelihood function and `optim()`, find p for $N = 100$ and $q = 30$.

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Using the loglikelihood function and `optim()`, find p for $N = 100$ and $q = 30$. ML estimates are "invariant under reparametrization", so we will use the logit transform of p instead of p, to avoid probabilities outside $[0, 1]$:

$$p^* = \frac{1}{1 + e^{-p}}.$$

Exercise

$$\ell(p) = q \log p + (N - q) \log(1 - p)$$

Using the loglikelihood function and `optim()`, find p for $N = 100$ and $q = 30$. ML estimates are "invariant under reparametrization", so we will use the logit transform of p instead of p, to avoid probabilities outside $[0, 1]$:

$$p^* = \frac{1}{1 + e^{-p}}.$$

```
ll <- function(p, N, q) {
  pstar <- 1 / (1 + exp(-p))
  q * log(pstar) + (N - q) * log(1 - pstar)
}
mlest <- optim(0, ll, NULL, N=100, q=30, hessian=TRUE,
   control = list(fnscale = -1))
phat <- 1 / (1 + exp(-mlest$par))
```

Exercise

Using the previous estimate, including the Hessian, calculate a 95% confidence interval around this $\hat{p}$.

Exercise

Using the previous estimate, including the Hessian, calculate a 95% confidence interval around this $\hat{p}$.

```
se <- sqrt(-1/mlest$hessian)
ci <- c(mlest$par - 1.96 * se, mlest$par + 1.96 * se)
ci <- 1 / (1 + exp(-ci))
```

Repeat for $N = 1000$, $q = 300$.

Exercise

Using the `uswages.dta` data, estimate model

$$log(wage)_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \varepsilon_i$$

with the `lm()` function and with the `optim()` function.

- Compare estimates for β, σ^2 and $V(\beta)$.
- Try different optimization algorithms in `optim()`.

Standard tests

Within the ML framework, there are three common tests:

- Likelihood ratio test (LR)
- Wald test (W)
- Lagrange multiplier test (LM)

These are asymptotically equivalent. Given r restrictions, all have asymptotically a $\chi^2(r)$ distribution.

(Davidson & MacKinnon 1999, 414)

Standard tests

Within the ML framework, there are three common tests:

- Likelihood ratio test (LR)
- Wald test (W)
- Lagrange multiplier test (LM)

These are asymptotically equivalent. Given r restrictions, all have asymptotically a $\chi^2(r)$ distribution.

In the following, $\tilde{\boldsymbol{\theta}}^{ML}$ will denote the restricted estimate and $\hat{\boldsymbol{\theta}}^{ML}$ the unrestricted one.

(Davidson & MacKinnon 1999, 414)

Likelihood ratio test

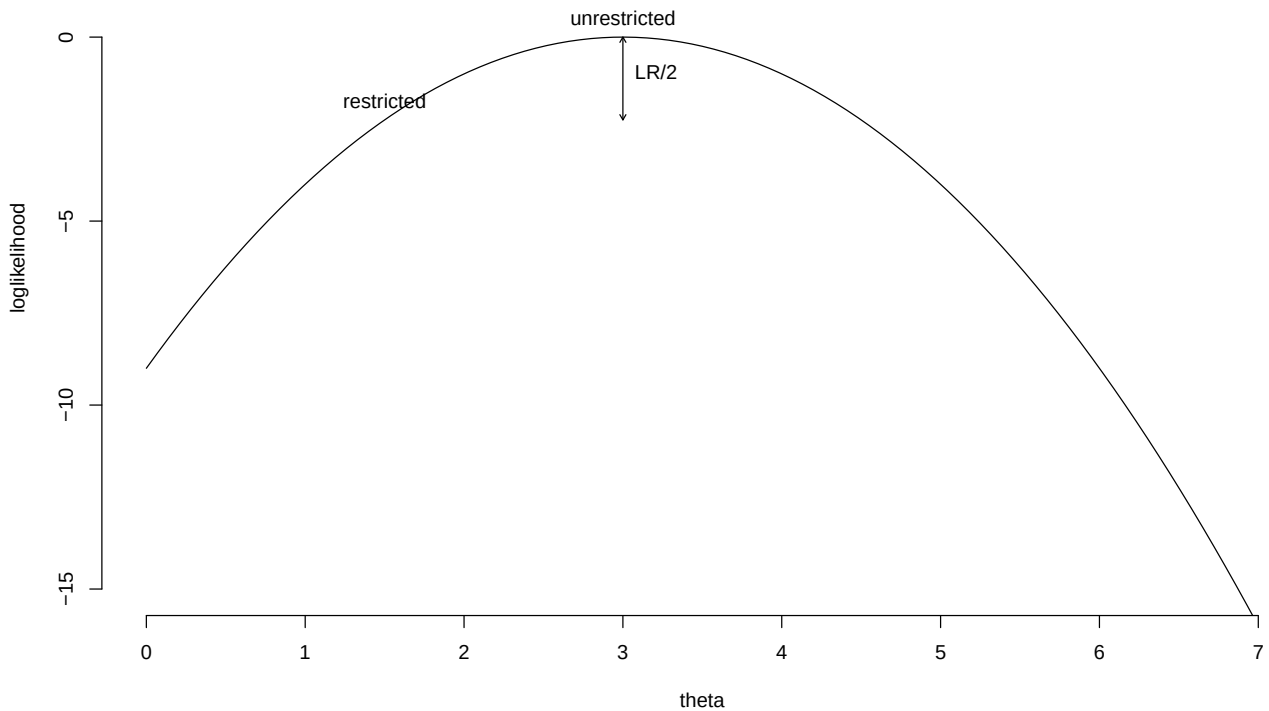
$$LR = 2(\ell(\hat{\boldsymbol{\theta}}^{ML}) - \ell(\tilde{\boldsymbol{\theta}}^{ML})) = 2 \log \frac{L(\hat{\boldsymbol{\theta}}^{ML})}{L(\tilde{\boldsymbol{\theta}}^{ML})},$$

thus twice the log of the ratio of likelihood functions.

If we have two separate estimates, this is thus very easy to compute or even "eye-ball".

(Davidson & MacKinnon 1999, 414-416)

Likelihood ratio test



Wald test

The basic intuition is that the Wald test is quite comparable to an F-test on a set of restrictions. For a single regression parameter the formule would be:

$$W = (\hat{\beta} - \beta_0)^2(-\frac{d^2L(\beta)}{d\beta^2})$$

(Davidson & MacKinnon 1999, 416-418)

Wald test

The basic intuition is that the Wald test is quite comparable to an F-test on a set of restrictions. For a single regression parameter the formule would be:

$$W = (\hat{\beta} - \beta_0)^2(-\frac{d^2L(\beta)}{d\beta^2})$$

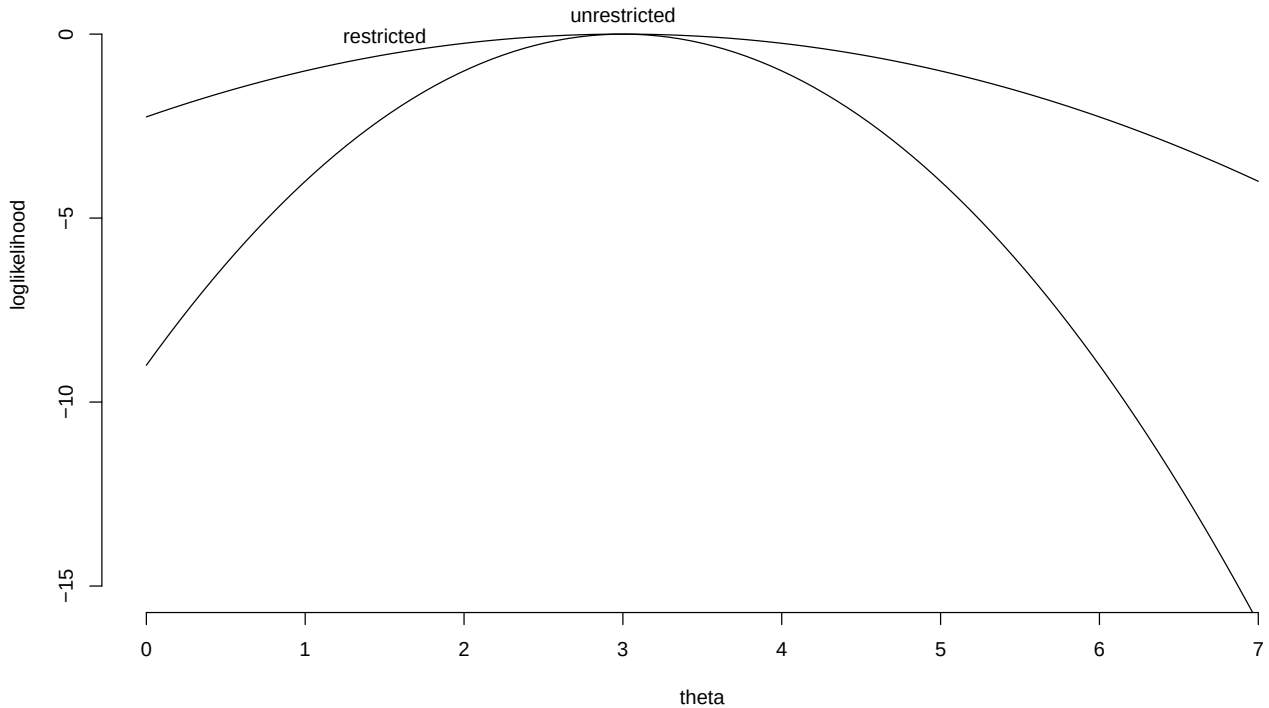
This test does not require an estimate of $\tilde{\boldsymbol{\theta}}^{ML}$. The test is somewhat sensitive to the formulation of $\mathbf{r}$, so that as long as estimating $\tilde{\boldsymbol{\theta}}^{ML}$ is not too costly, it is better to use an LR or LM test.

(Davidson & MacKinnon 1999, 416-418)

Wald test

If the second derivative is higher, the slope of $\ell(\cdot)$ changes faster, thus the difference between the likelihoods at the restricted and unrestricted positions will be larger.

Wald test


unrestricted
restricted
0
−5
−10
−15
loglikelihood
0
1
2
3
4
5
6
7
theta

Lagrange multiplier test

For a single parameter:

$$LM = \frac{(\frac{dL(\tilde{\beta})}{d\tilde{\beta}})^2}{\frac{d^2L(\tilde{\beta})}{d\tilde{\beta}^2}}$$

Thus the steeper the slope of the loglikelihood function at the point of the restricted estimate, the more likely it is significantly different from the unrestricted estimate, unless this slope changes very quickly (the second derivative is high).

Lagrange multiplier test