Week 7. Maximum likelihood for multidimensional parameter

Section 6.8 & 6.9

$Y_1, Y_2, ..., Y_n$ is a sample from a pdf $f(y_1, y_2, ..., y_n; \theta_1, \theta_2, ..., \theta_m)$ or $f(\mathbf{y}; \boldsymbol{\theta})$.
$\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m) \in R^m$
Requirements:

- Parameters are identifiable: $f(\mathbf{y}; \boldsymbol{\theta}_1) = f(\mathbf{y}; \boldsymbol{\theta}_2)\ \forall \mathbf{y}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.
- The support of f is independent of $\boldsymbol{\theta}$. For example, uniform distribution with unknown upper limit, $\mathcal{R}(0, \theta)$ does not comply.

Random sample: $Y_i \overset{\text{iid}}{\sim} f(y; \boldsymbol{\theta})$, $i = 1, 2, ..., n$.

Definition 1 *The likelihood and log-likelihood functions*

$$\begin{aligned} L(\boldsymbol{\theta}) &= f(\mathbf{Y}; \boldsymbol{\theta}), \\ l(\boldsymbol{\theta}) &= \ln f(\mathbf{Y}; \boldsymbol{\theta}). \end{aligned}$$

The MLE

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} f(\mathbf{Y}; \boldsymbol{\theta}) \\ \widehat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} \ln f(\mathbf{Y}; \boldsymbol{\theta}) \end{aligned}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)$. *Equivalently, the MLE can be derived as the solution of the* ***score equation***

$$\frac{\partial \ln f(\mathbf{Y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Example 2 *Show that MLE=OLS for the multiple linear regression when the error term* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Solution. We have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

or

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

We have $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

The pdf

$$f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right)$$

and

$$\ln f(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2) = const - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

The score equation

$$\frac{\partial \ln f(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

and

$$\widehat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \widehat{\boldsymbol{\beta}}_{OLS}.$$

Fisher information

Definition 3 *Fisher information matrix in the sample is defined as*

$$I(\boldsymbol{\theta}) = E_{\mathbf{Y}}\left[\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right]. \tag{1}$$

Theorem 4 *The following holds:*

$$E_{\mathbf{Y}}\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathbf{0},$$

and therefore

$$I(\boldsymbol{\theta}) = cov\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right). \tag{2}$$

This fact follows from

$$E(\mathbf{U}\mathbf{U}') = cov(\mathbf{U}) + [E(\mathbf{U})][E(\mathbf{U})]'.$$

Theorem 5 *Fisher information matrix can be derived from second derivative/Hessian,*

$$I(\boldsymbol{\theta}) = -E_{\mathbf{Y}}\left(\frac{\partial^2 \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)$$

where

$$\frac{\partial^2 \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$$

is called Hessian.

Example 6 *Derive the Fisher information matrix for multiple linear regression (σ^2 is known).*

Solution. We have

$$\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(\mathbf{Y};\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

and

$$\begin{aligned} I(\boldsymbol{\theta}) &= cov\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \frac{1}{\sigma^4} cov\left[\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\right] \\ &= \frac{1}{\sigma^4}\mathbf{X}' cov[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})']\mathbf{X} = \frac{1}{\sigma^4}\mathbf{X}' cov[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\mathbf{X} \\ &= \frac{1}{\sigma^4}\mathbf{X}'\sigma^2\mathbf{X} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}. \end{aligned}$$

Using second derivatives

$$\begin{aligned} I(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right) = -E\left(\frac{\partial^2 \ln f(\mathbf{Y};\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}^2}\right) = -E\left(-\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right) \\ &= \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} \end{aligned}$$

Theorem 7 *Optimal asymptotic properties of MLE. Let $Y_1, Y_2, ..., Y_n$ are iid. Then when $n \to \infty$ we have:*

1. MLE is **consistent**, that is, $p\lim \widehat{\boldsymbol{\theta}}_{ML} = \boldsymbol{\theta}$ or $\Pr(\lim \widehat{\boldsymbol{\theta}}_{ML} = \boldsymbol{\theta}) = 1$.

2. The distribution of MLE converges to the multivariate normal distribution, that is, **asymptotically normal**

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{ML} &\simeq \mathcal{N}\left(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})\right), \\
\widehat{\boldsymbol{\theta}}_{ML} &\simeq \mathcal{N}\left(\boldsymbol{\theta}, I^{-1}(\widehat{\boldsymbol{\theta}}_{ML})\right), \\
cov(\widehat{\boldsymbol{\theta}}_{ML}) &\simeq \left(-E\left(\frac{\partial^2 \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)\right)^{-1} = \left(cov\left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\right)^{-1}.
\end{aligned}$$

3. MLE is **asymptotically efficient**: any other estimator of $\boldsymbol{\theta}$ has the limit covariance matrix larger than MLE, $cov(\widetilde{\boldsymbol{\theta}}) \geq cov(\widehat{\boldsymbol{\theta}}_{ML})$.

Application to linear model when errors are not normally distributed

$$\widehat{\boldsymbol{\beta}} \simeq \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Logistic regression

Section 8.8.2

How to model binary events: regression with the binary/dichotomous dependent variable/outcome which takes values either 0 or 1.

The probability of a binary event $Y = y$ is

$$\Pr(Y = y; p) = p^y(1-p)^{1-y}$$

where $y = 0$ or 1 and p is the probability of $Y = 1$.

Now assume that this probability is different for different i:

$$\Pr(Y = y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i}, \quad y_i \in \{0, 1\}.$$

where

$$\pi_i = \mu(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i)$$

where $\mathbf{x}_i$ is a set of covariates/predictors as in linear model. The final model for binary outcome is

$$\Pr(Y_i = 1) = \mu(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_i), \quad i = 1, ..., n$$

where data are independent for different i. This model looks very much as linear model but (a) Y_i is binary and μ is not a linear function.

Logistic regression models a binary RV with the probability as function μ of a linear combination of predictors

Function μ is called the inverse link function (GLM) and has to comply with being a cdf (probability function):

1. The function $\mu = \mu(s)$ is defined for all $s \in (-\infty, \infty)$.
2. $0 < \mu(s) < 1$, $\lim_{s\to-\infty} \mu(s) = 0$, $\lim_{s\to\infty} \mu(s) = 1$.

3. $d\mu/ds = \mu' = \dot{\mu} > 0$.

4. $d^2(\ln \mu)/ds^2 < 0$.

5. Many probability functions are symmetric,

$$\mu(s) = 1 - \mu(-s).$$

The symmetry holds if μ' is an even function.

The popular choices of probability function μ and respective regressions have their own names:

1. **Logistic** regression with the probability function

$$\mu(s) = \frac{e^s}{1 + e^s}$$

is popular in biomedical and epidemiologic studies due to a simple interpretation of the beta-coefficients via the probability of disease using the language of Odds Ratio (OR).

2. **Probit** regression with the probability function

$$\mu(s) = \Phi(s) = (1/\sqrt{2\pi}) \int_{-\infty}^{s} e^{-0.5t^2} dt$$

is frequently used in toxicology and econometrics.

Logistic regression is used to model the dependent variable which takes binary, 0 or 1, values

$$\Pr(Y_i = 1) = \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}.$$

$\boldsymbol{\beta}'\mathbf{x}_i$ is called linear predictor. Vectors $\{\mathbf{x}_i, i = 1, ..., n\}$ are fixed.Logit transformation is called **link**

$$\boldsymbol{\beta}'\mathbf{x}_i = \ln \frac{p}{1 - p}$$

The likelihood function may be derived using an obvious expression

$$\Pr(Y_i = y) = \left(\frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)^y \left(\frac{1}{1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)^{1-y}, \ y = 0, 1.$$

The likelihood function

$$L(Y_1, ..., Y_n; \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)^{Y_i} \left(\frac{1}{1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}\right)^{1-Y_i}$$

yielding

$$l(\boldsymbol{\beta}) = \mathbf{r}'\boldsymbol{\beta} - \sum_{i=1}^{n} \ln(1 + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}),$$

where

$$\mathbf{r} = \sum_{i=1}^{n} Y_i \mathbf{x}_i = \sum_{Y_i=1} \mathbf{x}_i$$

is considered fixed when maximizing the log-likelihood function. The score equation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{r} - \sum_{i=1}^{n} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}\mathbf{x}_i = \mathbf{0}.$$

MLE $\widehat{\boldsymbol{\beta}}_{ML}$ solves the score equation and

$$l(\widehat{\boldsymbol{\beta}}_{ML}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}).$$

The Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = -E\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \sum_{i=1}^{n} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{(1+\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i})^2}\mathbf{x}_i\mathbf{x}_i'. \tag{3}$$

and the asymptotic covariance matrix are given by

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \mathcal{I}^{-1} = \left(\sum_{i=1}^{n} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{(1+\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i})^2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}.$$

MLE theory

$$\widehat{\boldsymbol{\beta}}_{ML} \simeq \mathcal{N}\left(\boldsymbol{\beta}, \left(\sum_{i=1}^{n} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i}}{(1+\mathrm{e}^{\boldsymbol{\beta}'\mathbf{x}_i})^2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right), \quad n \to \infty.$$

Therefore, the Fisher information matrix is

$$\begin{aligned} & \sum_{i=1}^{n} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{(1+\exp(\boldsymbol{\beta}'\mathbf{x}_i))^2}\mathbf{x}_i\mathbf{x}_i' \\ = \; & \mathbf{X}'\mathbf{D}\mathbf{X} \end{aligned}$$

where

$$\mathbf{D}^{n\times n} = \mathbf{D}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{e^{\boldsymbol{\beta}'\mathbf{x}_1}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_1})^2} & 0 & 0 & 0 \\ 0 & \frac{e^{\boldsymbol{\beta}'\mathbf{x}_2}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_2})^2} & 0 & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \frac{e^{\boldsymbol{\beta}'\mathbf{x}_n}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_n})^2} \end{bmatrix}.$$

How to run logistic regression in R

$$\mathrm{glm}(\mathrm{y\tilde{}x1 + x2}, \mathrm{family = binomial})$$

Statistical inference

Wald statistical inference

MLE is asymptotically normal

$$\widehat{\boldsymbol{\beta}}_{ML} \simeq \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1})$$

for large n. In particular, under the null hypothesis, $H_0 : \beta_j = 0$ we have

$$Z_j = \frac{\widehat{\boldsymbol{\beta}}_{ML,j}}{\sqrt{(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}_{jj}}} \simeq \mathcal{N}(0,1)$$

with the **Wald CI**

$$\widehat{\boldsymbol{\beta}}_{ML,j} \pm Z_{1-\alpha/2}\sqrt{(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}_{jj}}$$

This interval covers the true β_j approximately with probability $\lambda = 1-\alpha$, confidence level.

The hypothesis (statistical testing of regression coefficients) $H_0 : \beta_j = 0$ is rejected if the Z-score

$$Z_j = \frac{\widehat{\boldsymbol{\beta}}_{ML,j}}{\sqrt{(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}_{jj}}}$$

is outside the interval $(-Z_{1-\alpha/2}, Z_{1-\alpha/2})$ or $(-1.96, 1.96)$ for $\alpha = 0.05$. The p-value = `2*pnorm(-abs(Zj))`.

Statistical testing of regression coefficients is the backbone method to bulding parsimonous models!

The likelihood ratio test

Let $l(\boldsymbol{\theta})$ be the log-likelihood and $\widehat{\boldsymbol{\theta}}_{ML}$ be MLE, that is,

$$l(\widehat{\boldsymbol{\theta}}_{ML}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}).$$

Test the null hypothesis

$$H_0 : \mathbf{A}^{k\times m}\boldsymbol{\theta}^{m\times 1} = \mathbf{r}^{k\times 1}$$

where $\mathbf{A}$ and $\mathbf{r}$ are given. Then if this hypothesis holds then $(n \to \infty)$

$$-2[\max_{\mathbf{A}\boldsymbol{\theta}=\mathbf{r}} l(\boldsymbol{\theta}) - l(\widehat{\boldsymbol{\theta}}_{ML})] \simeq \chi^2(k).$$

Statistical testing of regression coefficients: under $H_0 : \beta_j = 0$ we have

$$-2(l_{\max,-j} - l_{\max}) \simeq \chi^2(1)$$

or $D_{-j} - D \simeq \chi^2(1)$, where $D = -2l$ is called deviance (a part of `glm`).

Homework

1. (15 points). A bank wants to develop a rule for granting or denying a credit card application. Over the years, the bank collected the data on 1996 applicants – some of them failed a minimum payment and some did not, see the file `creditpr.csv`. The first column indicates failed = 1, not failed = 0; the second column contains monthly paycheck; and the third column contains months at work. (a) Develop and formulate the rule to grant credit card applications in the form M<a+bP, where M=months at works and P=paycheck. (b) Plot the data points using red for those who failed the minimum payment and green otherwise (use `legend`) and display the discrimination line. (c) Compute and display the total empirical and theoretical misclassification probability. (d) Compute the area under the respective binormal ROC curve and compare it with the total misclassification error from LDA. Hint: adopt the R function `mah`.

2. (15 points). Logistic regression has two predictors x_1 and x_2. Develop algorithms for testing the null hypothesis that the two slope coefficients are the same via the Wald and likelihood ratio test. Formulate algorithms as step-by-step computations with the rule for accepting or rejecting the null hypothesis.