Vector Differential Calculus in Statistics & Machine Learning

July 30, 2023

1 Motivation-Gaussian regression model

Let $(x_i, y_i), 1 \leq i \leq n$, be a set of measurements on two variables x and y, and consider the problem of fitting a line $y = \beta_0 + \beta_1 x$ to the data. The homoscedastic Gaussian regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \tag{1}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2}$$

Assume that σ^2 is known.

1.1 scalar differential calculus approach

The residual sum of squares is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \tag{3}$$

and

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} = 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right), \tag{4}$$

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} = 2 \sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i\right). \tag{5}$$

Setting this to zero, we obtain

$$
\begin{aligned}
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \\
\widehat{\beta}_1 &= \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) / \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 .
\end{aligned}
$$

1.2 Vector differential calculus approach

Recall the homoscedastic Gaussian regression model is

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \tag{6}
$$

where

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{7}
$$

where we assume that σ^2 is known.
The residual sum of squares is

$$
\ell(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{8}
$$

Note that (equation 37, The matrix cookbook):

$$
\partial(UV) = (\partial U)V + U(\partial V) \tag{9}
$$

and

$$
\partial U^T = (\partial U)^T \tag{10}
$$

Therefore

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) =& (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) & (11) \\
\Longrightarrow \quad d\ell(\boldsymbol{\beta}) =& \{d(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \{d(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} & (12) \\
=& -(\mathbf{X} d\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} d\boldsymbol{\beta} & (13) \\
=& (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} d\boldsymbol{\beta} & (14) \\
\Longrightarrow \quad \mathrm{D}\ell(\boldsymbol{\beta}) =& (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} = \mathbf{0} \text{ iff } (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} = \mathbf{0} & (15) \\
\Longrightarrow \quad \boldsymbol{\beta} =& \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. & (16)
\end{aligned}
$$

2 Derivatives

2.1 Scalar Case

For the scalar case: given a function $f : \mathbb{R} \to \mathbb{R}$, then:

$$
f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{17}
$$

$f'(x)$ tells us how much the function f changes as the input x changes by a small amount ε :

$$f(x+\varepsilon) \approx f(x)+\varepsilon f'(x) \tag{18}$$

The **chain rule** tells us how to compute the derivative of the compositon of functions. In the scalar case suppose that $f, g : \mathbb{R} \rightarrow \mathbb{R}$ and $y=f(x), z=g(y)$; then we can also write $z=(g \circ f)(x)$. The chain rule tells us that

$$\frac{\partial z}{\partial x}=\frac{\partial z}{\partial y} \frac{\partial y}{\partial x}=\frac{\partial g(f(x))}{\partial f(x)} \frac{\partial f(x)}{\partial x} \tag{19}$$

Combining these two rules lets us compute the effect of x on z : if x changes by Δx then y will change by $\frac{\partial y}{\partial x} \Delta x$, so we have $\Delta y=\frac{\partial y}{\partial x} \Delta x$. If y changes by Δy then z will change by $\frac{\partial z}{\partial y} \Delta y=\frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \Delta x$ which is exactly what the chain rule tells us.

2.2 Gradient

This same intuition carries over into the vector case. Now suppose that f : $\mathbb{R}^{N} \rightarrow \mathbb{R}$ takes a vector as input and produces a scalar. The derivative of f at the point $x \in \mathbb{R}^{N}$ is now called the gradient, and it is defined as:

$$\nabla_{x} f(x)=\lim _{h \rightarrow 0} \frac{f(x+h)-f(x)}{\|h\|} \tag{20}$$

$\nabla_{x} f(x) \in \mathbb{R}^{N}$ is a vector, where the i th coordinate of $\frac{\partial y}{\partial x}$ tells us the approximate amount by which y will change if we move x along the i th coordinate axis. We can also view the gradient $\frac{\partial y}{\partial x}$ as a vector of partial derivatives:

$$\frac{\partial y}{\partial x}=\left(\frac{\partial y}{\partial x_{1}}, \frac{\partial y}{\partial x_{2}}, \ldots, \frac{\partial y}{\partial x_{N}}\right)$$

where x_{i} is the i th coordinate of the vector x, which is a scalar, so each partial derivative $\frac{\partial y}{\partial x_{i}}$ is also a scalar.

Example 1.

$$y(x)=x_{1}^{2}+x_{2}+x_{2}^{3} \tag{21}$$

Then

$$\frac{\partial y}{\partial x}=\left(\frac{\partial y}{\partial x_{1}}, \frac{\partial y}{\partial x_{2}}\right)=\left(2 x_{1}, 1+3 x_{2}^{2}\right) \tag{22}$$

Example 2.

$$y(x)=x_{1}+x_{2}+x_{3}^{3} \tag{23}$$

Then

$$\frac{\partial y}{\partial x}=\left(\frac{\partial y}{\partial x_{1}}, \frac{\partial y}{\partial x_{2}}, \frac{\partial y}{\partial x_{3}}\right)=\left(1,1,3 x_{3}^{2}\right) \tag{24}$$

2.3 Jacobian

Suppose that $f : \mathbb{R}^N \to \mathbb{R}^M$ takes a vector as input and produces a vector as output. Then the derivative of f at a point x, also called the Jacobian, is the $M \times N$ matrix of partial derivatives. Let $y = f(x)$, then:

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_M}{\partial x} \end{pmatrix} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_M}{\partial x_1} & \cdots & \frac{\partial y_M}{\partial x_N} \end{pmatrix} \tag{25}$$

Example 1.

$$y = (y_1, y_2) = f(x) = (x_1^2 + x_2 + x_2^3, x_1 + x_2 + x_3^3) \tag{26}$$

Then

$$\frac{\partial y_1}{\partial x} = \left(\frac{\partial y_1}{\partial x_1}, \frac{\partial y_1}{\partial x_2}, \frac{\partial y_1}{\partial x_3}\right) = \left(2x_1, 1 + 3x_2^2, 0\right) \tag{27}$$

and

$$\frac{\partial y_2}{\partial x} = \left(\frac{\partial y_2}{\partial x_1}, \frac{\partial y_2}{\partial x_2}, \frac{\partial y_2}{\partial x_3}\right) = \left(1, 1, 3x_3^2\right) \tag{28}$$

Therefore,

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \end{pmatrix} = \begin{pmatrix} 2x_1 & 1 + 3x_2^2 & 0 \\ 1 & 1 & 3x_3^2 \end{pmatrix} \tag{29}$$

Multivariate chain rule. The chain rule can be extended to the vector case using Jacobian matrices. Suppose that $f : \mathbb{R}^N \to \mathbb{R}^M$ and $g : \mathbb{R}^M \to \mathbb{R}^K$. Let $x \in \mathbb{R}^N, y \in \mathbb{R}^M$, and $z \in \mathbb{R}^K$ with $y = f(x)$ and $z = g(y)$, i.e., $z = g(f(x))$. The **multivariate chain rule** says:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x} = \frac{\partial g(f(x))}{\partial f(x)}\frac{\partial f(x)}{\partial x} \tag{30}$$

Here, $\frac{\partial z}{\partial y} \in \mathbb{R}^{K \times M}$, $\frac{\partial y}{\partial x} \in \mathbb{R}^{M \times N}$, and $\frac{\partial z}{\partial x} \in \mathbb{R}^{K \times N}$.

Example 3.

$$y(x) = g(f(x)) \tag{31}$$

where $g(z) = z + 1$ and $f(x) = 2x_1 + x_2$. Then by the chain rule

$$\frac{\partial y}{\partial x} = \frac{\partial g(f(x))}{\partial f(x)} . \frac{\partial f(x)}{\partial x} \tag{32}$$

$$= 1.(2, 1) = (2, 1) \tag{33}$$

2.4 Hessian matrix

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function taking as input a vector $\mathbf{x} \in \mathbb{R}^n$ and outputting a scalar $f(\mathbf{x}) \in \mathbb{R}$. If all second-order partial derivatives of f exist, then the Hessian matrix is

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}. \tag{34}$$

That is, the entry of the ith row and the jth column is

$$(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i\,\partial x_j}. \tag{35}$$

2.5 Rules for Differentials

Let $\mathbf{u}$ and $\mathbf{v}$ be vector functions and $\mathbf{U}$ and $\mathbf{V}$ be matrix functions. A will denote a constant matrix and s a scalar function.

2.5.1 Rules for Scalar Functions

$$\begin{aligned} du^\alpha &= \alpha u^{\alpha-1} du, \\ d\log u &= u^{-1} du, \\ de^u &= e^u du. \end{aligned}$$

2.5.2 Rules Involving Linear Functions

$$\begin{aligned} d(\mathbf{AU}) &= \mathbf{A}d\mathbf{U}, \\ d(\mathbf{U}+\mathbf{V}) &= d\mathbf{U} + d\mathbf{V}, \\ d\,\mathrm{diag}(\mathbf{u}) &= \mathrm{diag}(d\mathbf{u}), \\ d\mathbf{U}^\top &= (d\mathbf{U})^\top, \\ d\,\mathrm{vec}\,\mathbf{U} &= \mathrm{vec}(d\mathbf{U}), \\ d(\mathrm{tr}\,\mathbf{U}) &= \mathrm{tr}(d\mathbf{U}), \\ d(E\mathbf{U}) &= E(d\mathbf{U}). \end{aligned}$$

2.5.3 Rules for Determinant and Matrix Inverse

$$\begin{aligned} d|\mathbf{U}| &= |\mathbf{U}|\,\mathrm{tr}\left(\mathbf{U}^{-1} d\mathbf{U}\right), \\ d\mathbf{U}^{-1} &= -\mathbf{U}^{-1}(d\mathbf{U})\mathbf{U}^{-1}. \end{aligned}$$

2.5.4 Rules Involving Quadratic Forms

$$
\begin{aligned}
&d\mathbf{u}^\top \mathbf{A}\mathbf{u} = \mathbf{u}^\top \left(\mathbf{A} + \mathbf{A}^\top\right) d\mathbf{u}, \\
&d\mathbf{u}^\top \mathbf{A}\mathbf{u} = 2\mathbf{u}^\top \mathbf{A} d\mathbf{u}, \quad \mathbf{A} \text{ symmetric. } .
\end{aligned}
$$

3 Maximum Likelihood Estimate for parameters of a multivariate Gaussian Distribution

Given a set of i.i.d. data $X = \{x_1, \ldots, x_N\}$ drawn from $\mathcal{N}(x; \mu, \Sigma)$, we want to estimate (μ, Σ) by MLE. The log-likelihood function is

$$
\ln p(X \mid \mu, \Sigma) = -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) + \text{ const} \tag{36}
$$

Taking its derivative w.r.t. μ and setting it to zero we have

$$
\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{37}
$$

Rewrite the log-likelihood using $Trace(constant) = constant$, and $Trace(AB) = Trace(BA)$,

$$
\begin{aligned}
\ln p(X \mid \mu, \Sigma) &= -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) + \text{ const} &(38)\\
&\propto -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} \operatorname{Trace}\left(\Sigma^{-1} (x_n - \mu) (x_n - \mu)^T\right) &(39)\\
&= -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \operatorname{Trace}\left(\Sigma^{-1} \sum_{n=1}^{N} \left[(x_n - \mu) (x_n - \mu)^T\right]\right) &(40)
\end{aligned}
$$

Taking the derivative w.r.t. Σ^{-1}, and using 1) $\frac{\partial}{\partial A} \log |A| = A^{-T}$; 2) $\frac{\partial}{\partial A} \operatorname{Tr}[AB] = \frac{\partial}{\partial A} \operatorname{Tr}[BA] = B^T$, we obtain

$$
\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}) (x_n - \hat{\mu})^T . \tag{41}
$$

4 Generalized Linear Models

Let $\mathbf{y}$ be a vector of responses and $\mathbf{X}$ be a corresponding design matrix. The one-parameter exponential family model, with canonical link, is characterized by the joint density

$$
f(\mathbf{y}; \boldsymbol{\beta}) = \exp\left\{\mathbf{y}^\top (\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}^\top c(\mathbf{y})\right\} \tag{42}
$$

where $\boldsymbol{\beta}$ is the vector of coefficients. For example, $b(x) = \log(1+e^x)$ corresponds to binary regression with a logit link function.

The log-likelihood of $\boldsymbol{\beta}$ is

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) &= \mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}^\top c(\mathbf{y}) && (43)\\
\implies \quad d\ell(\boldsymbol{\beta}) &= \mathbf{y}^\top \mathbf{X} d\boldsymbol{\beta} - \mathbf{1}^\top db(\mathbf{X}\boldsymbol{\beta}) && (44)\\
&= \mathbf{y}^\top \mathbf{X} d\boldsymbol{\beta} - \mathbf{1}^\top \operatorname{diag}\{b'(\mathbf{X}\boldsymbol{\beta})\}\, d(\mathbf{X}\boldsymbol{\beta}) && (45)\\
&= \mathbf{y}^\top \mathbf{X} d\boldsymbol{\beta} - b'(\mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} d\boldsymbol{\beta} && (46)\\
&= \{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top \mathbf{X} d\boldsymbol{\beta}. && (47)
\end{aligned}
$$

Hence,

$$
\mathrm{D}\ell(\boldsymbol{\beta}) = \{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top \mathbf{X}
$$

Also,

$$
\begin{aligned}
d^2\ell(\boldsymbol{\beta}) &= d\,\{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top \mathbf{X} d\boldsymbol{\beta} && (48)\\
&= -\left\{\operatorname{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\,\mathbf{X} d\boldsymbol{\beta}\right\}^\top \mathbf{X} d\boldsymbol{\beta} && (49)\\
&= (d\boldsymbol{\beta})^\top \mathbf{X}^\top \left[-\operatorname{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\right] \mathbf{X}(d\boldsymbol{\beta}) && (50)
\end{aligned}
$$

which leads to

$$
\mathrm{H}\ell(\boldsymbol{\beta}) = -\mathbf{X}^\top \operatorname{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\,\mathbf{X} \qquad (51)
$$

5 Extra examples and reading

- Vector Differential Calculus in Statistics
- Matrix handbook for multivariate statistics and Machine Learning
- The Matrix Cookbook