Modeling (Notes)

Mark DiBattista

July 9, 2019

Abstract

There are two parts to the modeling problem:

- Given a complex phenomenon or process, generate a finite set of structured datapoints, sampled at random or by design, that covers its range of properties or behavior;
- Given a finite set of structured datapoints, extract information that, given a new, partially complete datapoint, affords imputation of missing values from those present.

Usually, the information is incomplete and contradictory, and imputation provisional.

Models, and modeling techniques, are distinguished by the type of datapoints, the relationship among datapoints within a single sample, and linkage of datapoints across samples. The relationships within datapoints may be constrained in form, while linkage across samples maybe facilitated by sequencing or by mapping samples to an increasing variable such as time.

The focus of these notes is on the quantitative aspect of the modeling problem for which uncertainties in the approximating relationships and linkages are sufficiently regular to support the expression and interpretation of imputed values as estimates and confidence intervals.

1 Suggested Resource Materials

Useful source texts:

- Probability/Statistics, intermediate (probability sections are better than statistics): *Statistical Inference*, Casella & Berger
- Probability, advanced: *Probability and Measure*, Billingsley

Throughout the text the acronyms refer to companion writeups,

LAN	*Linear Algebra (Notes)*
LAA	*Linear Algebra (Applications)*
PN	*Probability Notes*

within which information is referenced by chapter and/or numbered equation.

2 Linear Models

The linear model covers the case for which the data points are real values, partitioned into a single response variable and remainder predictor variables, and the relationship between predictors and response is a linear function. There are two basic approaches taken to solve the problem.

- The *engineering approach:* generate model coefficients for a prior linear relation between predictors and response variables that minimizes a loss function;
- The *probabilistic approach:* generate model parameters for a prior Gaussian conditional probability distribution, response variable conditioned on the predictors, that maximize an entropy measure.

The cases for which the loss function in the engineering approach is quadratic, also knows as 'least squares', and the entropy measure is maximum likelihood lead to identical results, since the underlying arithmetic calculations are identical.

$$W = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{y} & X \end{pmatrix} \tag{1}$$

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \tag{2}$$

$$\mathbf{a}_i^\top = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{im} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_i^\top \end{pmatrix} \Rightarrow A = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix} \tag{3}$$

$$\boldsymbol{\alpha}(W) \Rightarrow y_{n+1} = f(\mathbf{x}_{n+1}; \boldsymbol{\alpha}(W)) \tag{4}$$

2.1 Engineering Approach to Linear Modeling

$$\mathbf{y} - A\boldsymbol{\alpha} = 0, \tag{5}$$

$$\min_{\boldsymbol{\alpha}} ||\mathbf{y} - A\boldsymbol{\alpha}||_2^2 \Rightarrow \nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) = 0, \tag{6}$$

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) &= \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \left(A\boldsymbol{\alpha}\right) - \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\ &= \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top A\boldsymbol{\alpha} - \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\ &= 2A^\top A\boldsymbol{\alpha} - 2A^\top \mathbf{y} = 0 \\ \Rightarrow \boldsymbol{\alpha}_{LS} &= \left(A^\top A\right)^{-1} A^\top \mathbf{y}. \end{aligned} \tag{7}$$

2.2 Probabilistic Approach to Linear Modeling

2.2.1 Maximum Likelihood given Linear Form

$$Y|(\mathbf{X}=\mathbf{x}) \sim \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}, \sigma^2\right) \Rightarrow p_{Y|\mathbf{X}=\mathbf{x}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(y-\boldsymbol{\alpha}^\top \mathbf{a}\right)^2} \tag{8}$$

$$\left.\begin{array}{l} \mathbf{Y} = \begin{pmatrix} Y_1 & \cdots & Y_n \end{pmatrix}^\top \\ \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_n \end{pmatrix}^\top \end{array}\right\} \Rightarrow \mathbf{Y}|(\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}^\top) = \prod_{i=1}^{n} \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}_i, \sigma^2\right) = \mathrm{N}(A\boldsymbol{\alpha}, \sigma^2 I_n) \tag{9}$$

$$\begin{aligned} p_{\mathbf{Y}|\mathbf{X}} = \frac{1}{\sigma^n\sqrt{(2\pi)^n}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha})} & \\ \Rightarrow \ln p_{\mathbf{Y}|\mathbf{X}} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha}) & \end{aligned} \tag{10}$$

$$\begin{aligned} \text{maximum entropy solution: } \nabla_{\boldsymbol{\alpha}}\left(\mathbf{y}-A\boldsymbol{\alpha}\right)^\top\left(\mathbf{y}-A\boldsymbol{\alpha}\right) = 0 \equiv \min_{\boldsymbol{\alpha}} ||\mathbf{y}-A\boldsymbol{\alpha}||_2^2 & \\ \Rightarrow \hat{\boldsymbol{\alpha}} = (A^\top A)^{-1}A^\top\mathbf{y} & \end{aligned} \tag{11}$$

$$\mathbb{E}_{\mathbf{X}}\mathbf{Y} = A\boldsymbol{\alpha} \tag{12}$$

$$\mathbb{V}_{\mathbf{X}}\mathbf{Y} = \sigma^2 I \tag{13}$$

$$\mathbb{E}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} = \mathbb{E}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top\mathbb{E}_{\mathbf{X}}\mathbf{Y} = (A^\top A)^{-1}A^\top A\boldsymbol{\alpha} = \boldsymbol{\alpha} \tag{14}$$

$$\begin{aligned} \mathbb{V}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} = \mathbb{V}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top(\mathbb{V}_{\mathbf{X}}\mathbf{Y})A(A^\top A)^{-1} &= (A^\top A)^{-1}A^\top(\sigma^2 I)A(A^\top A)^{-1} \\ &= \sigma^2(A^\top A)^{-1} \end{aligned} \tag{15}$$

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\boldsymbol{\alpha}, \sigma^2(A^\top A)^{-1}\right) \tag{16}$$

2.2.2 Joint and Conditional Gaussian Models

$$\mathbf{W} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})} \tag{17}$$

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{cases} \tag{18}$$

$$\mathbf{W}_1|(\mathbf{W}_2 = \mathbf{w}_2) \sim \mathrm{N}(\boldsymbol{\mu}_1 - (\mathbf{w}_2 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{19}$$

$$\mathbf{W}=\begin{pmatrix}Y\\X_1\\\vdots\\X_m\end{pmatrix}\Rightarrow\begin{cases}\boldsymbol{\mu}=\begin{pmatrix}\mu_y\\\boldsymbol{\mu}_{\mathbf{x}}\end{pmatrix},\boldsymbol{\mu}_{\mathbf{x}}=\begin{pmatrix}\mu_{x_1}\\\vdots\\\mu_{x_m}\end{pmatrix}\\[2em]\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_y^2&\boldsymbol{\sigma}_{y\mathbf{x}}^\top\\\boldsymbol{\sigma}_{y\mathbf{x}}&\boldsymbol{\Sigma}_{\mathbf{xx}}\end{pmatrix},\begin{cases}\boldsymbol{\sigma}_{y\mathbf{x}}=\begin{pmatrix}\sigma_{yx_1}&\cdots&\sigma_{yx_m}\end{pmatrix}^\top\\\boldsymbol{\Sigma}_{\mathbf{xx}}=\begin{pmatrix}\sigma_{x_1}^2&\cdots&\sigma_{x_1x_m}\\\vdots&\ddots&\vdots\\\sigma_{x_1x_m}&\cdots&\sigma_{x_m}^2\end{pmatrix}\end{cases}\end{cases}\tag{20}$$

$$Y|(\mathbf{X}=\mathbf{x})\sim\mathrm{N}\left(\mu_y+(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{y\mathbf{x}},\sigma_y^2-\boldsymbol{\sigma}_{y\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{y\mathbf{x}}\right)\equiv\mathrm{N}\left(\mu(\mathbf{x}),\sigma^2\right)\tag{21}$$

$$\left.\begin{aligned}\mathbf{a}&=\begin{pmatrix}1\\\mathbf{x}\end{pmatrix}\\\boldsymbol{\alpha}&=\begin{pmatrix}\mu_y-\boldsymbol{\sigma}_{y\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\\\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\end{pmatrix}\end{aligned}\right\}\Rightarrow\mu_{y|\mathbf{x}}=\mu_y+(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{y\mathbf{x}}\equiv\boldsymbol{\alpha}^\top\mathbf{a}\tag{22}$$

$$R^2=\frac{\boldsymbol{\sigma}_{y\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{y\mathbf{x}}}{\sigma_y^2}\Rightarrow\sigma_{y|\mathbf{x}}^2=\sigma_y^2\left(1-R^2\right)\tag{23}$$

2.2.3 Formal Equivalence between Least Squares and Sample Statistics

$$A=\begin{pmatrix}\uparrow&&\uparrow&\\\mathbf{1}&\leftarrow&X&\rightarrow\\\downarrow&&\downarrow&\end{pmatrix}\tag{24}$$

$$A^\top=\begin{pmatrix}\leftarrow&\mathbf{1}^\top&\rightarrow\\&\uparrow&\\\leftarrow&X^\top&\rightarrow\\&\downarrow&\end{pmatrix}\tag{25}$$

$$A^\top A=\begin{pmatrix}\mathbf{1}^\top\mathbf{1}&\mathbf{1}^\top X\\X^\top\mathbf{1}&X^\top X\end{pmatrix}\tag{26}$$

$$\begin{aligned}\mathbf{1}^\top\mathbf{1}&=n,&(27)\\\mathbf{1}^\top X&=n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top,&(28)\\X^\top\mathbf{1}&=n\hat{\boldsymbol{\mu}}_{\mathbf{x}},&(29)\\X^\top X&=n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}-\hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right)&(30)\end{aligned}$$

$$\left.\begin{aligned}A&=n\\B&=n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\\C&=n\hat{\boldsymbol{\mu}}_{\mathbf{x}}\\D&=n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}-\hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right)\end{aligned}\right\}\Rightarrow\left(A^\top A\right)^{-1}=\begin{pmatrix}1+\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}}&-\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\\-\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}}&\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\end{pmatrix}\tag{31}$$

$$A^\top \mathbf{y} = \begin{pmatrix} \mathbf{1}^\top \mathbf{y} \\ X^\top \mathbf{y} \end{pmatrix} = n \begin{pmatrix} \hat{\mu}_y \\ \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} - \hat{\mu}_y \hat{\boldsymbol{\mu}}_{\mathbf{x}} \end{pmatrix} \tag{32}$$

$$\hat{\boldsymbol{\alpha}} = (A^\top A)^{-1} A^\top \mathbf{y} = \begin{pmatrix} \hat{\mu}_y - \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \end{pmatrix} \tag{33}$$

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \frac{\sigma^2}{n} \begin{pmatrix} 1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\mu}}_{\mathbf{x}} & -\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \\ -\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\mu}}_{\mathbf{x}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \end{pmatrix} \right) \tag{34}$$