Modeling (Notes)

Mark DiBattista

July 12, 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*
SN	*Statistics 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 true relationship between predictors and response is a linear function. For any real problem the relationship encoded in the data points is noisy, however, and the assumption of linearity under common conditions for optimality is consistent with joint Gaussian distributions of response and predictor variables.

In fact given a set of measurements, there are two basic approaches commonly 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 model parameters for which the loss function in the engineering approach is quadratic, also knows as 'least squares', and for which the entropy measure is maximum likelihood are identical, and each can be shown to generate sample estimators for the mean vector and covariance matrix of the underlying joint Gaussian distribution. Finally, the estimated parameters of the Gaussian distribution can be used to generate interval tests and other measures of quality and stability of the model.

2.1 Gaussian Joint and Conditional Probability Function

A multivariate Gaussian distribution is completely defined by the specification of a mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, with the probability density function

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

The joint distribution can be expressed as the product of marginal and conditional distributions (*cf PN*, §7.2.1.6) upon partitioning the variables into two distinct sets,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}, \end{pmatrix} \end{cases} \tag{2}$$

and the distribution of $\mathbf{W}_1$ conditioned on the realizaed variables, $\mathbf{W}_2 = \mathbf{w}_2$ is given by

$$\mathbf{W}_1 | (\mathbf{W}_2 = \mathbf{w}_2) \sim \mathrm{N}(\boldsymbol{\mu}_1 - (\mathbf{w}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}). \tag{3}$$

If the vector of random variables is partitioned into a single response variable, Y, conditioned on the remainder predictor variables, $X_1, \cdots, X_n$,

$$\mathbf{W} = \begin{pmatrix} Y \\ X_1 \\ \vdots \\ X_m \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \boldsymbol{\mu}_{\mathbf{x}} = \begin{pmatrix} \mu_{x_1} \\ \vdots \\ \mu_{x_m} \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{y\mathbf{x}}^\top \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{pmatrix}, \begin{cases} \boldsymbol{\sigma}_{y\mathbf{x}} = \begin{pmatrix} \sigma_{yx_1} & \cdots & \sigma_{yx_m} \end{pmatrix}^\top \\ \boldsymbol{\Sigma}_{\mathbf{xx}} = \begin{pmatrix} \sigma_{x_1}^2 & \cdots & \sigma_{x_1x_m} \\ \vdots & \ddots & \vdots \\ \sigma_{x_1x_m} & \cdots & \sigma_{x_m}^2 \end{pmatrix} \end{cases} \end{cases} \tag{4}$$

then the single-variate conditional distribution is also Gaussian, and takes the form,

$$Y|(\mathbf{X} = \mathbf{x}) \sim \mathrm{N}\left(\mu_y + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}, \sigma_y^2 - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}\right). \tag{5}$$

The parameters of the Gaussian distribution can be simplified – and the linear form of the mean stressed – by introducing the following constants,

$$\left.\begin{aligned} \mathbf{a} &= \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \\ \boldsymbol{\alpha} &= \begin{pmatrix} \mu_y - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix} \end{aligned}\right\} \Rightarrow \mu_{y|\mathbf{x}} = \mu_y + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}} \equiv \boldsymbol{\alpha}^\top \mathbf{a} \tag{6}$$

$$R^2 = \frac{\boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}}{\sigma_y^2} \qquad \Rightarrow \sigma_{y|\mathbf{x}}^2 = \sigma_y^2 \left(1 - R^2\right) \tag{7}$$

which leads to the expression,

$$Y|(\mathbf{X} = \mathbf{x}) \sim \mathrm{N}(\mu(\mathbf{x}), \sigma^2) \equiv \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}, \sigma_y^2 \left(1 - R^2\right)\right). \tag{8}$$

Here, the parameter, R^2, which appears in the variance of the conditional distribution, is the familiar coefficient of determination. It is clear from the derivation provided above, however, that the parameter describes the strength of the linear relationship between predictor and response variables, and is not an independent property of the measurement data.

2.2 Estimation of Linear Relationship from Data

The linear relationship between response and predictor variables is derived from information contained within the matrix, W, which contains a list of indexed measurements,

$$W = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{y} & X \end{pmatrix}. \tag{9}$$

Here, the measurements are organized by row, in which the first column contains the response variable, $\mathbf{y}$, and the remaining columns contain the predictors, X, and is called the **data matrix**. The predictor data points may be selected by design, or by chance.

We construct the **extended data matrix** by replacing the response vector with a unit vector,

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & X \end{pmatrix}, \tag{10}$$

which contains all information from which the linear relationship is to be derived. Each row in the extended data matrix in (10) matches the form of the vector of arbitrary predictor measurements in the conditional mean show above in (6).

$$\mathbf{a}_i^\top = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{im} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_i^\top \end{pmatrix} \Rightarrow A = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix} \tag{11}$$

2.3 Engineering Approach: Least Squares

The engineering approach to linear modeling is quite direct: assert a linear relationship between the vector of response variables, $\mathbf{y}$, and the extended data matrix, A, in which the linear relationship is mediated through a vector of coefficients, $\boldsymbol{\alpha}$,

$$\mathbf{y} - A\boldsymbol{\alpha} = 0. \tag{12}$$

Typically, the number of data points, n, exceeds the number of predictors, m, and the linear relation in (12) is overdetermined. The properties of solutions to linear equations are covered in *LAN* §7.

Although there is generally no *exact* solution to the equation in (12), we can determine the *best* solution given a condition of optimality. For a quadratic penalty function the optimization problem leads to a gradient operator applied to an inner product:

$$\hat{\boldsymbol{\alpha}} \leftarrow \min_{\boldsymbol{\alpha}} ||\mathbf{y} - A\boldsymbol{\alpha}||_2^2 \Rightarrow \nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) = 0. \tag{13}$$

This problem, covered also in *LAA* §3, is solved by expanding the inner product and applying the gradient operator term by term:

$$\begin{aligned}
\nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) &= \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \left(A\boldsymbol{\alpha}\right) - \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\
&= \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top A\boldsymbol{\alpha} - \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\
&= 2A^\top A\hat{\boldsymbol{\alpha}} - 2A^\top \mathbf{y} = 0 \\
\Rightarrow \hat{\boldsymbol{\alpha}} &= \left(A^\top A\right)^{-1} A^\top \mathbf{y}.
\end{aligned} \tag{14}$$

Notice that the least-squares solution is achieved by applying a linear operator derived from the extended data matrix, $(A^\top A)^{-1} A^\top$, to the predictor vector, $\mathbf{y}$. The solution is simply a linear combination of all measurements.

2.4 Probabilistic Approach: Maximum Entropy

The probabilistic approach to linear modeling asserts that each measurement is drawn from the same underlying distribution,

$$Y|(\mathbf{X} = \mathbf{x}) \sim \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}, \sigma^2\right) \Rightarrow p_{Y|\mathbf{X}=\mathbf{x}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(y - \boldsymbol{\alpha}^\top \mathbf{a}\right)^2} \tag{15}$$

for which the conditional distribution is taken as Gaussian, the linear relation is encoded in the mean, $\boldsymbol{\alpha}^\top \mathbf{a}$, and the strength of the relationship is encoded in the variance, σ^2. The determination of the coefficients to the linear model, ***alpha***, is therefore point-estimation problem given sampled data, which is covered in *SN* §3, while maximum-likelihood estimators are addressed specifically in §3.3.

If each sample is governed by the one-dimensional Gaussian distribution in (15), the full set of samples, assuming each is independent of the other, is governed by the multivariate Gaussian,

$$\left.\begin{aligned} \mathbf{Y} &= \begin{pmatrix} Y_1 & \cdots & Y_n \end{pmatrix}^\top \\ \mathbf{X} &= \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_n \end{pmatrix}^\top \end{aligned}\right\} \Rightarrow \mathbf{Y}|(\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}^\top) = \prod_{i=1}^{n} \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}_i, \sigma^2\right) = \mathrm{N}(A\boldsymbol{\alpha}, \sigma^2 I_n). \tag{16}$$

Point estimators are derived from operations performed on the joint distribution; point estimators for the coefficients in the linear model, and the variance of the distribution are presented in the next few sections.

2.4.1 Conditional Mean

2.4.1.1 Maximum-Likelihood Estimator

The likelihood function for the sample distribution, which is also Shannon information, is simply the logarithm of the joint probability density function (*cf. PN,* §10.4),

$$
\begin{aligned}
p_{\mathbf{Y}|\mathbf{X}} &= \frac{1}{\sigma^n\sqrt{(2\pi)^n}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha})} \\
&\Rightarrow \ln p_{\mathbf{Y}|\mathbf{X}} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha}) \qquad (17)
\end{aligned}
$$

Since the linear model coefficients enter the likelihood function as an inner product, the maximum-likelihood estimator is generated by the extrema of the gradient operator,

$$\nabla_{\boldsymbol{\alpha}}\left(\mathbf{y}-A\boldsymbol{\alpha}\right)^\top\left(\mathbf{y}-A\boldsymbol{\alpha}\right) = 0 \equiv \hat{\boldsymbol{\alpha}} \leftarrow \min_{\boldsymbol{\alpha}} ||\mathbf{y}-A\boldsymbol{\alpha}||_2^2 \Rightarrow \hat{\boldsymbol{\alpha}} = (A^\top A)^{-1}A^\top\mathbf{y}. \qquad (18)$$

The optimization condition is identical to the least-squares problem shown above in §2.3, and so the solutions exactly coincide.

The probabilistic approach holds an advantage over the engineering approach, however, since the statistical parameters of the underlying joint distribution define a distribution for the parameters of the linear model. Indeed, given the mean and variance of the sample conditional distribution,

$$\mathbb{E}_{\mathbf{X}}\mathbf{Y} = A\boldsymbol{\alpha} \qquad (19)$$

$$\mathbb{V}_{\mathbf{X}}\mathbf{Y} = \sigma^2 I \qquad (20)$$

the mean and variance of the estimated coefficients, $\boldsymbol{\alpha}$, can be calculated directly,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} &= \mathbb{E}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top\mathbb{E}_{\mathbf{X}}\mathbf{Y} = (A^\top A)^{-1}A^\top A\boldsymbol{\alpha} = \boldsymbol{\alpha}, && (21)\\
\mathbb{V}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} &= \mathbb{V}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top(\mathbb{V}_{\mathbf{X}}\mathbf{Y})A(A^\top A)^{-1} = (A^\top A)^{-1}A^\top(\sigma^2 I)A(A^\top A)^{-1} \\
&= \sigma^2(A^\top A)^{-1}. && (22)
\end{aligned}
$$

Finally, since the underlying conditional distribution is Gaussian, and all arithmetic operations to generate the coefficients are linear, the *distribution* of the coefficients must also be Gaussian (*cf. PN*, §7.2.1.4), the distribution of estimated coefficients is given by

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\boldsymbol{\alpha}, \sigma^2(A^\top A)^{-1}\right). \qquad (23)$$

2.4.1.2 Interval Tests for Mean Linear Coefficients

Based on Hotelling's T^2 test, which is the multidimensional version of the Student T-test.

2.4.2 Conditional Variance

2.4.2.1 Maximum-Likelihood Estimator

The maximum-likelihood estimator for the variance of the conditional distribution, whose likelihood function is provided in (17), is the extremum of the derivative with respect to variance,

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2} \left(-\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) \right) &= 0 \\
\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \left(\mathbf{y} - A\hat{\boldsymbol{\alpha}}\right)^\top \left(\mathbf{y} - A\hat{\boldsymbol{\alpha}}\right) &= \frac{1}{n} \left(\mathbf{y} - A(A^\top A)^{-1} A^\top \mathbf{y}\right)^\top \left(\mathbf{y} - A(A^\top A)^{-1} A^\top \mathbf{y}\right) \\
= \frac{1}{n} \left(\mathbf{y} - P_A \mathbf{y}\right)^\top \left(\mathbf{y} - P_A \mathbf{y}\right) &= \frac{1}{n} \mathbf{y}^\top \left(I - P_A\right) \mathbf{y} \qquad (24)
\end{aligned}
$$

The maximum-likelihood variance estimator is a quadratic form with the matrix operator, $I - P_A$, that projects vectors into the orthogonal complement of the column space of the extended data matrix, A.

2.4.2.2 Unbiased Estimator

The quadratic form in (24) shows that information in the variance is contained within the orthogonal complement of the column space of the data matrix, which is an $n - (m + 1)$-dimensional subspace. Coupled with the observation that the mean is projected into the column space,

$$
A\hat{\boldsymbol{\alpha}} = A(A^\top A)^{-1} A^\top \mathbf{y} = P_A \mathbf{y}, \qquad (25)
$$

which is an $(m + 1)$-dimensional subspace, it is clear that the maximum-entropy estimators for the conditional Gaussian distribution generalize Fisher's Theorem (*cf. PN*, §7.2.1.7) to a multidimensional setting. The arguments that hold in the 1-dimensional case carry over to the multidimensional case, since

$$
\left.
\begin{aligned}
\mathbb{E}_{\mathbf{X}} \left(\mathbf{Y}^\top \mathbf{Y}\right) &= n\sigma^2 \\
\mathbb{E}_{\mathbf{X}} \left(\mathbf{Y}^\top P_A \mathbf{Y}\right) &= (m + 1)\sigma^2
\end{aligned}
\right\} \Rightarrow \mathbb{E}_{\mathbf{X}} \left(\frac{1}{n - (m + 1)} \mathbf{Y}^\top \left(I - P_A\right) \mathbf{Y} \right) = \sigma^2. \qquad (26)
$$

We can therefore adjust the normalizing factor in the maximum-likelihood variance estimator, which minimizes the overall average discrepancy between the estimated and actual variances, to derive a least-unbiased estimator,

$$
\hat{s}^2 = \frac{1}{n - (m + 1)} \mathbf{y}^\top \left(I - P_A\right) \mathbf{y}. \qquad (27)
$$

The distribution of the scaled unbiased estimator is then a chi-squared distribution with the appropriate degrees of freedom,

$$
(n - (m + 1)) \frac{\hat{s}^2}{\sigma^2} \sim \chi^2_{n-(m+1)}. \qquad (28)
$$

2.4.3 Linear Operations of Least Squares and Maximum-Entropy Sample Statistics

The least-squares and maximum-entropy methods generate identical linear-algebraic operations and solutions. We show here that the arithmetic operations carried out in the matrix formulae encode detailed sample estimators for every mean, variance and covariance in the underlying joint distribution. The solutions are both equivalent to replacing all statistical quantities with sample estimates, and the matrix operations shown above are exactly those necessary to achieve this.

Given the extended data matrix, A, the **scatter matrix**, $A^\top A$, can be expressed as

$$
A^\top A = \begin{pmatrix} \mathbf{1}^\top \\ X^\top \end{pmatrix} \begin{pmatrix} \mathbf{1} & X \end{pmatrix} = \begin{pmatrix} \mathbf{1}^\top \mathbf{1} & \mathbf{1}^\top X \\ X^\top \mathbf{1} & X^\top X. \end{pmatrix} \qquad (29)
$$

Each of these operations encode sample estimators for mean and covariance statistics (*cf. PN*, §3.3),

$$\mathbf{1}^\top\mathbf{1} = n, \tag{30}$$
$$\mathbf{1}^\top X = n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top, \tag{31}$$
$$X^\top\mathbf{1} = n\hat{\boldsymbol{\mu}}_{\mathbf{x}}, \tag{32}$$
$$X^\top X = n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right). \tag{33}$$

The partitioned scatter matrix in (29) can be inverted by use of the lower Schur block matrix inversion formula (*cf. LAN*, §6.2), repeated here as (note the unrelated use of the variable, A),

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1} & -A^{-1}B\left(D - CA^{-1}B\right)^{-1} \\ -\left(D - CA^{-1}B\right)^{-1}CA^{-1} & \left(D - CA^{-1}B\right)^{-1} \end{pmatrix}. \tag{34}$$

Upon assigning the sample statistics in the scatter matrix to the blocks in the Schur formula, we directly construct the inverse of the scatter matrix,

$$\left.\begin{aligned} A &= n \\ B &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \\ C &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}} \\ D &= n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right) \end{aligned}\right\} \Rightarrow \left(A^\top A\right)^{-1} = \frac{1}{n}\begin{pmatrix} 1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & -\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \\ -\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \end{pmatrix} \tag{35}$$

Similarly, the means and covarances of the response variables are carried out by linear operation with the data matrix,

$$A^\top\mathbf{y} = \begin{pmatrix} \mathbf{1}^\top\mathbf{y} \\ X^\top\mathbf{y} \end{pmatrix} = n\begin{pmatrix} \hat{\mu}_y \\ \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} - \hat{\mu}_y\hat{\boldsymbol{\mu}}_{\mathbf{x}} \end{pmatrix}. \tag{36}$$

Applying these results to the operations for calculating estimates of the conditional mean and variance

$$\hat{\boldsymbol{\alpha}} = (A^\top A)^{-1}A^\top\mathbf{y} = \begin{pmatrix} \hat{\mu}_y - \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \end{pmatrix} \tag{37}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top A(A^\top A)^{-1}A^\top\mathbf{y}\right) = \hat{\sigma}_y^2 - \hat{\boldsymbol{\sigma}}_{y\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} = \hat{\sigma}_y^2\left(1 - \hat{R}^2\right) \tag{38}$$

we generate exactly the results achieved by replacing the statistical parameters in the conditional distribution in (8) with sample estimates.

Finally, the distribution of the coefficients of the linear model can be expressed as

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y - \boldsymbol{\sigma}_{y\mathbf{x}}^\top\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \frac{\sigma^2}{n}\begin{pmatrix} 1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & -\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \\ -\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \end{pmatrix}\right). \tag{39}$$