Odds and Ends from the Theory of Linear Regression

1 MLE Properties of Least-Squares

1.1 The Multivariate Normal Density

The univariate normal density ($n = 1$) we all remember is

$$\varphi(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

for $y \in \mathcal{R}$. We often write $Y \sim N(\mu, \sigma^2)$ to convey that the random variable Y has this density.

The multivariate normal density, for $\underline{y} \in \mathcal{R}^n$, is

$$\varphi(\underline{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1}(\underline{y} - \underline{\mu})] \tag{1}$$

The mean vector $\underline{\mu}$ replaces the univariate mean μ and the variance-covariance matrix Σ replaces the univariate variance σ^2. We often write $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, or just $\underline{Y} \sim N(\underline{\mu}, \Sigma)$ if the dimension n is clear from context. Sometimes we also say that the components of $\underline{Y}$ are "jointly normal" or "multivariate normal".

One way to build up the multivariate normal is to start with a n-dimensional vector of independent identically distributed $N(0, 1)$ variables $\underline{Z} = (Z_1, \ldots, Z_n)$. Since densities for independent random variables multiply, the density for $\underline{Z}$ must be

$$\begin{aligned}
\varphi(\underline{z}) &= \prod_1^n \varphi(z_i) \\
&= \frac{1}{(2\pi)^{n/2}} \exp[-\frac{1}{2}\sum_1^n z_i^2] \\
&= \frac{1}{(2\pi)^{n/2}} \exp[-\frac{1}{2}\underline{z}^T I^{-1} \underline{z}]
\end{aligned}$$

where I is the $n \times n$ identity matrix.

Then if $\underline{Y} = A\underline{Z} + \underline{\mu}$, it is clear that $E[\underline{Y}] = \underline{\mu}$ and $\text{Var}\,(\underline{Y}) = E[(\underline{Y} - \underline{\mu})(\underline{Y} - \underline{\mu})^T] = AA^T = \Sigma$ (so for example it is immediately obvious that Σ must by symmtric and positive definite).

The multivariate change-of-variables formula from calculus will ensure that the density for $\underline{Y}$ looks like (1). In fact, it is generally true that if $\underline{Y} = A\underline{Y} + \underline{b}$, where A is $m \times n$ and $\underline{b} \in \mathcal{R}^m$, and $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, then $\underline{Y} \sim N_m(A\underline{\mu} + \underline{b}, A\Sigma A^T)$.

1.2 Maximum Likelihood Estimation

Now let us derive maximum likelihood estimates for $\underline{\beta}$ and σ_ϵ^2 in the multiple regression model

$$\underline{y}_{n\times 1} = X_{n\times k'}\underline{\beta}_{k'\times 1} + \underline{\epsilon}_{n\times 1} \quad \underline{\epsilon} \sim N_n(\underline{0}, \sigma_\epsilon^2 I_{n\times n}).$$

Since this implies the y_i's are independent

$$\underline{y} \sim N_n(X\beta, \sigma_\epsilon^2 I_{n\times n})$$

we can immediately write down the likelihood for $\underline{\beta}$ and σ_ϵ^2 by writing down the product of the individual normals[1],

$$\begin{aligned} L(\underline{y}; \underline{\beta}, \sigma_\epsilon^2) &= \prod_1^n \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{1}{2}(y_i - \underline{x}_i^T\underline{\beta})^2/\sigma_\epsilon^2\right] \\ &= \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_1^n (y_i - \underline{x}_i^T\underline{\beta})^2/\sigma_\epsilon^2\right] \\ &= \frac{1}{(2\pi)^{n/2}|\sigma^2 I|^{1/2}} \exp\left[-\frac{1}{2}(\underline{y} - X\underline{\beta})^T(\sigma_\epsilon^2 I)^{-1}(\underline{y} - X\underline{\beta})\right], \end{aligned}$$

and hence the log-likelihood is

$$\ell(\underline{\beta}, \sigma_\epsilon^2) = \log L((\underline{y}; \underline{\beta}, \sigma_\epsilon^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$$

Now we differentiate with respect to $\underline{\beta}$ and σ_ϵ^2 to obtain

$$\begin{aligned} \frac{\partial \ell}{\partial \underline{\beta}} &= -\frac{1}{2\sigma_\epsilon^2}[-2X^T(y - X\underline{\beta})] \\ \frac{\partial \ell}{\partial \sigma_\epsilon^2} &= -\frac{n}{2}\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\epsilon^4}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) \end{aligned}$$

Setting the first set of partial derivatives equal to zero produces the normal equations again, and hence

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y},$$

the usual least-squares regression estimates. Setting the second set of partial derivatives equal to zero produces

$$\hat{\sigma_\epsilon^2} = \frac{(\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})}{n} = \frac{RSS}{n}$$

which we now know is a biased estimate of σ_ϵ^2 (the correct denominator should be $n - k'$).

1.3 Some Geometry; Eigensystems; Square Roots; Principal Components

A useful geometric property of the multivariate normal distribution is that it is "log-quadratic"; $\log\varphi(\underline{y})$ is a simple linear function of the symmetric positive-definite quadratic form $(\underline{y}-\underline{\mu})^T\mathbf{\Sigma}^{-1}(\underline{y}-\underline{\mu})$ in $(\underline{y} - \underline{\mu})$. This means that the level surfaces $\{\underline{y} : \varphi(\underline{y}) = K\}$, or equivalently the level surfaces $\{\underline{y} : \log\varphi(\underline{y}) = c\}$ will be ellipsoids.

[1] *Warning:* In this handout, $\underline{x}_i^T$ is the i^{th} row of the regression matrix X.

Finding the principal axes of the ellipsoidal level sets (contours of $\varphi(\underline{y})$) is easy with Lagrange multipliers. To find the first principal axis for example, you need a point $\underline{y}$ that is a maximum distance from $\underline{\mu}$ [maximize the squared distance $(\underline{y}-\underline{\mu})^T(\underline{y}-\underline{\mu})$], but is still on the level set [subject to the constraint $(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) = c$]. Differentiating the Lagrange-multiplier objective function

$$f(\underline{y}) = (\underline{y}-\underline{\mu})^T(\underline{y}-\underline{\mu}) - \lambda[(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) - c] \tag{2}$$

with respect to $\underline{y}$ and setting these derivatives equal to zero leads to the eigenvalue/eigenvector problem

$$(\Sigma - \lambda I)(\underline{y}-\underline{\mu}) = 0$$

So there must be multiple critical points for (2), one for one corresponding to each eigenvalue of σ. It now a matter of calculation to see that the eigenvalues will satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$$

with corresponding eigenvectors $\underline{e}_i = (\underline{y}_i - \underline{\mu})$ lying along the principal axes of the ellipsoid $(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) = c$; the half-lengths of the principal axes will be $\sqrt{\lambda c}$.

Several interesting algebraic and geometric properties come out of this analysis. To every eigenvalue λ_i there corresponds an eigenvector $\underline{e}_i$, and it can be shown that all the $\underline{e}_i$ are linearly independent. If two eigenvectors $\underline{e}_i$ and $\underline{e}_j$ correspond to different eigenvalues $\lambda_i \neq \lambda_j$ then, since $\underline{e}_j^T\Sigma\underline{e}_i$ is a scalar (hence equal to its transpose),

$$\lambda_i\underline{e}_j^T\underline{e}_i = \underline{e}_j^T\Sigma\underline{e}_i = [\underline{e}_j^T\Sigma\underline{e}_i]^T = \underline{e}_i^T\Sigma\underline{e}_j = \lambda_j\underline{e}_i^T\underline{e}_j$$

so it must be that $\underline{e}_i^T\underline{e}_i = 0$: the eigenvectors from different eigenvalues are orthogonal. If $\lambda_i = \lambda_j$, there will still be two eigenvectors, and using the Gram-Schmidt procedure we can select two eigenvectors that are orthogonal to each other and to all the other eigenvectors in the problem. If the two eigenvalues are equal, then the cross-section of the ellipse in the plane containing the two linearly independent eigenvectors $\underline{e}_i$ and $\underline{y}_j$ will be circular (it would always be an ellipse, and now the major and minor principal axes will be of the same length $2\sqrt{\lambda_i c}$); if three are equal the corresponding cross-section will be spherical, etc.

We can assume all the e_i are orthogonal; so let $\underline{a}_i = \underline{e}_i/||\underline{e}_i||$ and consider the matrix

$$A = \begin{bmatrix} \underline{a}_1 & \underline{a}_2 & \cdots & \underline{a}_p \end{bmatrix}$$

You should verify that $AA^T = I$ (the matrix is orthogonal—it has orthonormal columns); that $\Sigma A = A\,\mathit{diag}\,(\lambda_1, \ldots, \lambda_p)$ (the square $p \times p$ matrix with λ's on the main diagonal and zeros everywhere else); and therefore that $A^T\Sigma A = \mathit{diag}\,(\lambda_1, \ldots, \lambda_p)$.

Now consider the random vector $\underline{W} = A(\underline{Y} - \underline{\mu})$. This vector $\underline{W}$ will be multivariate normal, and you should verify that it has mean $\underline{0}$ (the zero vector) and variance-covariance matrix $\mathit{diag}\,(\lambda_1, \ldots, \lambda_p)$. Hence each W_i is uncorrelated with W_j (in fact, independent); and $\mathrm{Var}\,(W_i) = \lambda_i$. The W_i's are called the *principal components* of $\underline{Y}$ and they play a large role in understanding the structure of $\underline{Y}$.

2 The Gauss-Markov Theorem

In the multiple regression model

$$\underline{y}_{n\times 1} = X_{n\times k'}\underline{\beta}_{k'\times 1} + \underline{\epsilon}_{n\times 1}$$

we will only assume that the error terms ϵ_i are independent, mean 0, constant variance σ^2_ϵ.

Let us define $M = (X^TX)^{-1}X^T$, so that $\hat{\underline{\beta}} = M\underline{y}$ are the least-squares estimators of $\underline{\beta}$, and let $\tilde{\underline{\beta}} = U\underline{y}$ be any unbiased linear estimate of β. Finally, let $A = U - M$.

Lemma 1 *For any unbiased linear estimate $U\underline{y}$ of $\underline{\beta}$, $A = U - M$ is orthogonal to the rows of X; hence $AM^T = MA^T = 0$.*

Proof: Since both $\tilde{\underline{\beta}}$ and $\hat{\underline{\beta}}$ are unbiased,

$$\underline{\beta} = E[\tilde{\underline{\beta}}] = E[(M + A)\underline{y}] = E[M\underline{y}] + E[M\underline{y}] = E[\hat{\underline{\beta}}] + AE[\underline{y}] = \underline{\beta} + AX\underline{\beta},$$

and since this must be true for all $\underline{\beta}$ it follows that $AX = 0$. It immediately follows that $AM^T = AX(X^TX)^{-1} = 0$. $\square$

Theorem 1 Gauss-Markov Theorem.

$\hat{\underline{\beta}}$ is the Best (minimum-variance) Linear Unbiased Estimate (BLUE) of $\underline{\beta}$.

Proof: We will do this by minimizing the variance of the arbitrary unbiased linear estimate $\tilde{\underline{\beta}} = U\underline{y} = (M + A)\underline{y}$. Since $AM^T = MA^T = 0$, it follows that

$$\begin{aligned}
\mathrm{Var}\,(\tilde{\underline{\beta}}) &= \mathrm{Var}\,[(M + A)\underline{y}] \\
&= (M + A)\mathrm{Var}\,(\underline{y})(M + A)^T \;=\; (M + A)\sigma^2_\epsilon I(M + A)^T \\
&= \sigma^2_\epsilon(MM^T + MA^T + AM^T + AA^T) \\
&= \sigma^2_\epsilon(MM^T + AA^T)
\end{aligned}$$

and hence, reading off the diagonal elements of this variance-covariance matrix,

$$\mathrm{Var}\,(\tilde{\beta}_j) = \sigma^2_\epsilon\left(\sum_{i=1}^n m^2_{ij} + \sum_{i=1}^n a^2_{ij}\right).$$

Since the m_{ij}'s are fixed, we choose different unbiased linear estimates $\tilde{\underline{\beta}}$ by choosing different a_{ij}'s; from the above expression it is clear that the only way to minimize the variance of all the $\tilde{\beta}_j$'s is to take $a_{ij} \equiv 0$; so that $\tilde{\beta}_j = \hat{\beta}_j$ for every j. $\square$

Remark: Essentially the same technique can be used to prove the following result:

Corollary 1 *Let $\varphi = L\underline{\beta}$ be a linear function of the β's. The BLUE of φ is $L\hat{\underline{\beta}}$ (i.e. plug in the least-squares estimates).*