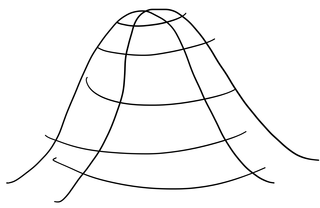
Lecture 7

Monday, June 3, 2019 2:30 PM

The Bivariate Normal Distribution

*special 2D-case of MVN $\vec{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

$$f(\vec{y}; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_1-\mu_1}{\sigma_1} \right)^2 - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y_2-\mu_2}{\sigma_2} \right)^2 \right] \right\}$$



Properties of MVN:

- Linear transformations of $\vec{y} \sim MVN(\vec{\mu}, \Sigma)$ are still MVN.
  - $\rightarrow \vec{a}\vec{y} \sim N(\vec{a}\vec{\mu}, \vec{a}\Sigma\vec{a}^T)$
  - $\rightarrow A\vec{y} \sim MVN(A\vec{\mu}, A\Sigma A^T)$
- If $\vec{y} = (y_1, y_2, \ldots, y_n)^T \sim MVN(\vec{\mu}, \Sigma)$ then $y_i \sim N(\mu_i, \Sigma_{ii})$ where μ_i is the i^{th} element of $\vec{\mu}$ and Σ_{ii} is the i^{th} diagonal element of Σ
- All joint and conditional distributions of $y_1, y_2, \ldots, y_n$ are also normal/multivariate normal
- Zero correlation $\iff$ Independence (i.e., if $Cov[y_i, y_j] = 0$ then y_i and y_j are independent).

Parameter Estimation (MLR)

Recall that in this setting, our model is:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{\mu_i} + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. Or, equivalently

$$\vec{y} = \underbrace{X\vec{\beta}}_{\vec{\mu}} + \vec{\varepsilon}$$

We assume: $\vec{\varepsilon} \sim MVN(\vec{0}_{n\times 1}, \sigma^2 I_{n\times n}) \implies \vec{y} \sim MVN(\vec{\mu} = X\vec{\beta}, \sigma^2 I)$

We'll first take a calculus-based approach to parameter estimation. As in SLR, we care about minimizing sum of squared error:

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

$$= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2$$

Take partial derivatives wrt the β's:

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})](-1) = -2\sum_{i=1}^{n} \varepsilon_i$$

$$\frac{\partial S}{\partial \beta_j} = \sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})](-x_{ij}) = -2\sum_{i=1}^{n} x_{ij}\varepsilon_i, \quad j = 1, 2, \ldots, p$$

Setting these $(p+1)$ equations equal to 0 and solving, yields:

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \mu_i) = 0$$

$$\sum_{i=1}^{n} x_{ij}\varepsilon_i = \sum_{i=1}^{n} x_{ij}(y_i - \mu_i) = 0, \quad j = 1, 2, \ldots, p$$

In vector notation, this is equivalent to:

$\vec{1}_{n\times 1}^T (\vec{y} - \vec{\mu}) = 0$ $\quad \vec{x}_j^T (\vec{y} - \vec{\mu}) = 0$ for $j = 1, 2, \ldots, p$

where $\vec{1}_{n\times 1}$ is $(n\times 1)$ vector of 1's and $\vec{x}_j$ is the $n\times 1$ vector of observations of explanatory variable $j = 1, 2, \ldots, p$.

Recognizing that $X = [\vec{1}_{n\times 1} \; \vec{x}_1 \; \vec{x}_2 \; \cdots \; \vec{x}_p]$ we can rewrite these $(p+1)$ equations as

$$X^T(\vec{y} - \vec{\mu}) = \vec{0}$$

Let's solve this:

$$X^T(\vec{y} - X\vec{\beta}) = \vec{0}$$
$$X^T\vec{y} - X^TX\vec{\beta} = \vec{0}$$
$$X^TX\vec{\beta} = X^T\vec{y}$$
$$\vec{\beta} = (X^TX)^{-1}X^T\vec{y}$$

So, if X^TX is invertible (i.e., non-singular) then the least squares estimate (LSE) of $\vec{\beta}$ is:

$$\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$$

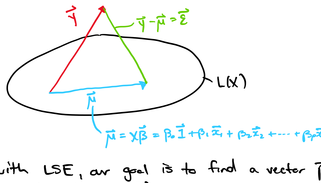
* X^TX is invertible as long as X has full rank $(p+1)$ (i.e., linearly independent columns)

This LSE can also be derived from a linear algebra perspective.

Define $L(X)$ to be span of the columns of X (i.e., all possible linear combinations of $\vec{1}_{n\times 1}, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$):

$$L(X) = \{ c_0\vec{1} + c_1\vec{x}_1 + c_2\vec{x}_2 + \cdots + c_p\vec{x}_p \mid c_0, c_1, \ldots, c_p \in \mathbb{R} \}$$ ← Subspace of $\mathbb{R}^n$

Now consider the response vector $\vec{y} \in \mathbb{R}^n$ which is not in $L(X)$



In accordance with LSE, our goal is to find a vector $\vec{\beta}$ that makes these two vectors ($\vec{y}$ and $\vec{\mu}$) as close to each other as possible. In other words we want to minimize the magnitude of error vector $\vec{\varepsilon}$.