Probability Distributions

1) Normal Distribution:

A random variable Y is said to have a normal distribution with mean μ and variance σ^2, which is denoted by using the notation $Y \sim \mathrm{N}(\mu, \sigma^2)$ if the probability density function (pdf) of Y is given by

$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$, where $-\infty < y < +\infty, -\infty < \mu < +\infty$, $\sigma^2 > 0$

$$\mu = \mathrm{E}(Y) = \int_{-\infty}^{+\infty} y f(y) dy = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} y e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$$

$$\begin{aligned}
\sigma^2 &= \mathrm{E}(Y-\mu)^2 \\
&= \mathrm{E}(Y^2 - 2\mu Y + \mu^2) \\
&= \mathrm{E}(Y^2) - 2\mu \mathrm{E}(Y) + \mu^2 \\
&= \mathrm{E}(Y^2) - 2\mu^2 + \mu^2 \\
&= \mathrm{E}(Y^2) - \mu^2 \\
&= \left\{ \int_{-\infty}^{+\infty} y^2 f(y) dy \right\} - \mu^2 \\
&= \left\{ \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} y^2 e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy \right\} - \mu^2
\end{aligned}$$

This is the same as $\sigma^2 = \mathrm{E}(Y-\mu)^2 = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (y-\mu)^2 e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$.

Y is said to have a **standard normal** distribution if Y has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

The **distribution function** of Y, denoted $\Phi(t)$, is given by

$$\Phi(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$$

$\Phi(t) = \mathrm{P}(Y < t)$, which is the area under the normal curve from $-\infty$ to t.

2) Chi-square Distribution

A random variable Y is said to have a chi-square distribution with ν degrees of freedom, denoted by $Y \sim \chi^2_\nu$, if

$f(y) = \frac{(\frac{1}{2})^{\frac{\nu}{2}} y^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})} e^{-\frac{1}{2}y}$, where $y > 0$ and $f(y) = 0$ otherwise; ν is a positive integer

Note: The **gamma function** is given by $\Gamma(\alpha) = \int_0^{+\infty} u^{\alpha-1} e^{-u} du$

$\mu = \mathrm{E}(Y) = \nu$

$\sigma^2 = 2\nu$

3) Student's t Distribution

Suppose we had two independent random variables Y and X. Suppose $Y \sim \mathrm{N}(0, 1)$ and $X \sim \chi^2_\nu$. Consider a new variable T which is given by

$$T = \frac{Y}{\sqrt{X/\nu}}$$

Then T is said to have the Student's t distribution with ν degrees of freedom and

$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})(1 + \frac{t^2}{\nu})^{\frac{1}{2}(\nu+1)}}$, where $-\infty < t < +\infty$, ν is a positive integer

$\mu = \mathrm{E}(T) = 0$ for $\nu > 2$

$\sigma^2 = \frac{\nu}{\nu-2}$ for $\nu > 2$

4) F Distribution

Suppose we had two independent random variables Y and X. Suppose $Y \sim \chi^2_{\nu_1}$ and $X \sim \chi^2_{\nu_2}$. Consider a new variable F which is given by

$$F = \frac{Y/\nu_1}{X/\nu_2}$$

Then F is said to have the F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom and

$$g(f) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{f^{\frac{\nu_1}{2}-1}}{(1+\frac{\nu_1 f}{\nu_2})^{\frac{\nu_1+\nu_2}{2}}}$$, where $f > 0$, ν_1 and ν_2 are positive integers.

$\mu = \mathrm{E}(F) = \frac{\nu_2}{\nu_2-2}$ for $\nu_2 > 2$

$\sigma^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$

Parameter Estimation

1) Method of Least-squares

Suppose we had a random sample $Y_1, Y_2, \ldots, Y_n$ from a variable $Y \sim \mathrm{N}(\mu, 1)$. To estimate the parameter μ by the method of least squares simply means we would like to find the point $\widehat{\mu}$ which makes μ as close as possible to the observed Y's. Hence, we would like to minimize the following function (with respect to μ)

$$g(y_1, y_2, \ldots, y_n; \mu) = \sum_{i=1}^{n} (y_i - \mu)^2$$

From calculus we know that to minimize the above function with respect to μ we must do the following steps:

1. Take the first derivative
2. Set the first derivative to zero and find the critical value
3. Take the second derivative
4. Plug the critical value into the second derivative to observe the sign of the result. If the sign is negative then the critical value is a maximum; if the sign is positive then the critical value is a minimum.

Let's do it:

Step 1:

$$\begin{aligned}
\frac{d}{d\mu} g(y_1, y_2, \ldots, y_n; \mu) &= \frac{d}{d\mu} \sum_{i=1}^{n} (y_i - \mu)^2 \\
&= \sum_{i=1}^{n} \frac{d}{d\mu} (y_i - \mu)^2 \\
&= \sum_{i=1}^{n} 2(y_i - \mu)(-1) \\
&= -2 \sum_{i=1}^{n} (y_i - \mu)
\end{aligned}$$

Step 2:

$$-2\sum_{i=1}^{n}(y_i-\mu)=0 \Rightarrow \sum_{i=1}^{n}\mu=\sum_{i=1}^{n}y_i$$
$$\Rightarrow n\mu=\sum_{i=1}^{n}y_i$$
$$\Rightarrow \mu=\frac{1}{n}\sum_{i=1}^{n}y_i$$

Hence, we have the critical value

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

Step 3:

$$\frac{d^2}{d\mu^2}g(y_1,y_2,\ldots,y_n;\mu)=\frac{d}{d\mu}\left[-2\sum_{i=1}^{n}(y_i-\mu)\right]$$
$$=-2\sum_{i=1}^{n}(-1)$$
$$=2n$$
$$>0$$

Step 4: The second derivative is a constant with a positive sign. Hence, the value of g must be a minimum at $\widehat{\mu}=\frac{1}{n}\Sigma_{i=1}^{n}y_i$. Thus, the least-sqares estimate of μ is

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\overline{Y}_n$$

Straight-line Regression:

Let $Y_i=\beta_0+\beta_1X_i+E_i$. We are interested in finding the least-squares estimates of β_0 and β_1. Since we have two parameters, we must take partial derivatives with respect to each parameter and solve the two resulting equations:

1.

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2$$
$$\text{and } \frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2$$

2. Upon taking the derivatives (and a small amount of reduction) we get

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \text{ and } \sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$$

Note: The above equations are called the **normal equations**.

3. After solving for the estimates and verifying that they minimize the functions we get the following estimates:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\text{and } \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

2) Method of Maximum Likelihood

Suppose we had a random sample $Y_1, Y_2, \ldots, Y_n$ from a variable $Y \sim \mathrm{N}(\mu, 1)$. To estimate the parameter μ by the method of maximum likelihood we first have to form the likelihood function. Each Y_i is $\mathrm{N}(\mu, 1)$ and because they represent a random sample, they are independent of each other. Hence, the joint probability density function for the Y_i's is the product of each the pdf's and this is the **likelihood function**:

$$\begin{aligned}
&L(Y_1, Y_2, \ldots, Y_n; \mu) \\
&\quad = f(y_1)f(y_2)\cdots f(y_n) \\
&\quad = \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_1-\mu)^2}\right]\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_2-\mu)^2}\right]\cdots\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2}\right] \\
&\quad = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}\right] \\
&\quad = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} \\
&\quad = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}
\end{aligned}$$

The idea is to maximize the likelihood function with respect to μ . To make the process easier, we take the logarithm (base e) of the likelihood function and maximize it. By a theorem in calculus, we know that if we take a monotonic transformation of a function (the logarithm is a monotonic function) then the value which maximizes (or minimizes) the original function will maximize (or minimize) the transformed function.

We have

$$
\begin{aligned}
l(Y_1, Y_2, \ldots, Y_n; \mu) &\equiv \log[L(Y_1, Y_2, \ldots, Y_n; \mu)] \\
&= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 \qquad \text{since } \sigma = 1
\end{aligned}
$$

It can easily be seen that maximizing the likelihood function is the same as minimizing the log-likelihood function. Indeed, minimizing the log-likelihood function is the same as what we did in the case of least-squares. Hence, under the assumption of normality, the method of least-squares is the same as the method of maximum likelihood.

EXERCISE: Show that the method of maximum likelihood is the same as the method of least-squares for the straight-line regression under the assumption of normality.

Matrix Representation of (Multiple) Regression Analysis

The (multiple) regression analysis equation is written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E$$

For the i-th subject ($i = 1, 2, \ldots, n$), the regression equation is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + E_i$$

We can write list each of the subject's regression equation:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + E_1 \\
Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + E_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + E_n
\end{aligned}$$

Let **Y** be the $n \times 1$ vector of Y_i's, i.e., let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Another more classical notation for vectors and matrices uses tildes instead of boldace letters. We could write the vector of Y_i's as

$$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

This style is commonly used when writing on paper (or chalkboard, etc.) since boldfacing letters is tedious by hand.

On word-processed and published documents, the boldface notation is more commonly used, but one should be able to adapt between the two.

Let $\boldsymbol{\beta}$ be the $(k+1) \times 1$ vector of β_i's, i.e., let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Let **X** be the $n \times (k+1)$ matrix of X_{ki}'s, i.e., let

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix}$$

Let **E** be the vector of E_i's, i.e., let

$$\mathbf{E} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}$$

Then we can write the regression equation in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

The matrix representation of the estimate of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The fundamental equation of regression analysis states that the total unexplained variation is equal to the variation due to regression plus the unexplained residual variation:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

i.e., SSY = SSR + SSE.

Derivation:

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}\left[(\widehat{Y}_i - \overline{Y})^2 + 2(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) + (Y_i - \widehat{Y}_i)^2\right] \\ &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) \end{aligned}$$

Expanding the furthest right term out we get:

$$\begin{aligned} 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) &= 2\overline{Y}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) - 2\sum_{i=1}^{n}\widehat{Y}_i(Y_i - \widehat{Y}_i) \\ &= 2\overline{Y}\sum_{i=1}^{n}\widehat{E}_i - 2\sum_{i=1}^{n}\widehat{Y}_i\widehat{E}_i \\ &= 0 \end{aligned}$$

(using properties of a fitted regression line KKMN Chapter 12, p. 217, where $\widehat{E}_i = e_i$ and using the normal equations)