Lecture 8-1: Sample Statistics

Lecturer: Prof. F. Jalayer, 26 February 2021
Notes prepared by: Prof. F. Jalayer and Dr. H. Ebrahimian

1. Sample mean and standard deviation

Suppose that $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ is a sample of observations. The unbiased estimators for the expected value and variance are defined as (known as sample mean and sample standard deviation):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
$$s_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}} \quad (1)$$

If *Xi*'s are i.i.d Normal (μ,σ), it can be shown that expected value of sample average is equal to μ:

$$E(\bar{X}) = E(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{\sum_{i=1}^{N} E(X_i)}{N} = \frac{NE(X_i)}{N} = \mu \quad (2)$$

And its standard deviation is equal to:

$$Var(\bar{X}) = Var(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{\sum_{i=1}^{N} Var(X_i)}{N^2} = \frac{\sigma^2}{N} \quad (3)$$

The expected value of the sample variance is equal to:

$$
\begin{aligned}
&E(s_x{}^2) = E(\frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{N-1}) = E(\frac{\sum_{i=1}^{N}\left(X_i{}^2 - 2X_i\bar{X} + \bar{X}^2\right)}{N-1}) \\
&E(X_i{}^2) = \mu^2 + \sigma^2 \\
&E(X_i\bar{X}) = E(X_i\frac{\sum X_i}{N}) = E(\frac{\left(\sum_{i\neq j, j=1}^{N} X_iX_j\right) + X_i{}^2}{N}) = \frac{(N-1)E(X_i)E(X_j) + \mu^2 + \sigma^2}{N} \\
&= \mu^2 + \frac{\sigma^2}{N} \\
&E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{N} \\
&E(s_x{}^2) = \frac{N(\mu^2+\sigma^2) - 2N(\mu^2 + \frac{\sigma^2}{N}) + N\mu^2 + N\frac{\sigma^2}{N}}{N-1} = \frac{(N-1)\sigma^2}{N-1} = \sigma^2
\end{aligned}
\tag{4}
$$

The sample covariance and correlation coefficient between the sets of data **X** and **Y** are calculated as follows:

$$
\begin{aligned}
&s_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1} \\
&r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}
\end{aligned}
\tag{2}
$$

The expected value of sample covariance is equal to:

$$E(s_{XY}{}^2) = E(\frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{N-1}) = E(\frac{\sum_{i=1}^{N}\left(X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}\right)}{N-1})$$

$$E(X_iY_i) = \rho_{XY}{}^2\sigma_y\sigma_x + \mu_y\mu_x$$

$$E(X_iY_j) = \mu_y\mu_x \quad , i \neq j$$

$$E(X_i\bar{Y}) = E(X_i\frac{\sum Y_j}{N}) = E(\frac{\left(\sum_{i\neq j, j=1}^{N} X_iY_j\right)}{N}) = \frac{(N-1)\mu_y\mu_x + (\rho_{XY}{}^2\sigma_y\sigma_x + \mu_y\mu_x)}{N} = \frac{\rho_{XY}{}^2\sigma_y\sigma_x}{N} + \mu_y\mu_x$$

$$E(\bar{X}\bar{Y}) = E(\frac{\sum Y_i}{N}\frac{\sum X_j}{N}) = E(\frac{\sum_{i\neq j, i=1}^{N}\sum_{i\neq j, j=1}^{N} Y_iX_j + N(\mu_y\mu_x + \rho_{XY}{}^2\sigma_y\sigma_x)}{N^2}) = \frac{(N^2-N)\mu_y\mu_x + N(\mu_y\mu_x + \rho^2\sigma_y\sigma_x)}{N^2}$$

$$E(s_{XY}{}^2) = \frac{N(\rho_{XY}{}^2\sigma_y\sigma_x + \mu_y\mu_x - 2\frac{\rho_{XY}{}^2\sigma_y\sigma_x}{N} - 2\mu_y\mu_x + \frac{\rho_{XY}{}^2\sigma_y\sigma_x}{N} + \mu_y\mu_x)}{N-1} = \rho_{XY}{}^2\sigma_y\sigma_x = COV(X,Y)$$

Note that it has been assumed that X_i and Y_i have a correlation expressed by ρ_{XY} and X_i and Y_j where i and j are not equal are uncorrelated.

2. Ordered Statistics

Suppose that the vector of sample data $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ is first sorted in an ascending order. The parameter p_{50} is defined as the 50th percentile of the sorted data which is equal in definition to the *median* of data, and denoted as η_X or $\widehat{X}$. It is statistically the value that 50% of the data are located below it. Similarly, p_{16} and p_{84} are the 16th and 84th percentiles of the data, where 16% and 84% of the data are below those values, respectively.

The aforementioned percentiles provide an efficient and transparent approach to estimate the main statistical parameters of a data set especially in case where there are values that are clearly distant from the other available observations; i.e. the data contains very low or high values. Accordingly,

the *counted* statistical parameters of a data set are: (1) the counted median, $\eta_X = p_{50}$, and (2) the counted standard deviation which is estimated by the average of $p_{84} - p_{50}$ and $p_{50} - p_{16}$:

$$s_X = \frac{p_{84} - p_{16}}{2} \equiv \frac{\left(p_{84} - p_{50}\right) + \left(p_{50} - p_{16}\right)}{2} \tag{3}$$

In the next section, we will discuss about the derivation of Eq. (3). The following example provides more insight into the sample statistics:

Example: A building structure is subjected to 8 different ground-motion records and the results in terms of maximum interstory drifts are as follow:

$$\mathbf{D} = \{0.02, 0.025, 0.03, 0.01, 0.04, 0.03, 0.05, 0.60\}$$

The sample mean and variance can be calculated as:

$$\begin{aligned} \overline{D} &= \frac{\sum_{i=1}^{8} D_i}{8} = 0.1006 \\ s_D &= \sqrt{\frac{\sum_{i=1}^{8} \left(D_i - \overline{D}\right)^2}{8-1}} = 0.2021 \end{aligned} \tag{4}$$

For calculating the counted statistics, one needs to first sort the data:

$$\mathbf{D}_{\text{sorted}} = \{0.01, 0.02, 0.025, 0.03, 0.03, 0.04, 0.05, 0.60\}$$

Thus,

$$\eta_X = p_{50} = 0.03$$

$$p_{16} = \mathbf{D}\left(0.16 * 8 \cong 1\right) = 0.01$$

$$p_{84} = \mathbf{D}\left(0.84 * 8 \cong 7\right) = 0.05$$

$$s_X = \frac{p_{84} - p_{16}}{2} = \frac{0.05 - 0.01}{2} = 0.02 \quad \text{or} \quad s_X = \frac{\left(0.05 - 0.03\right) + \left(0.03 - 0.01\right)}{2} = 0.02$$

Our dataset has an *outlier* which is D=0.6, and the ordered statistics can account for these data, while the sample mean and variance are affected by this high value.

In addition, the parameters of the lognormal distribution (i.e. the mean and standard deviation of ln$X$) can also be estimated according to the ordered statistics as follows:

$$\mu_{\ln X} = \ln\left(\eta_X\right) = \ln\left(x_{50}\right)$$
$$\sigma_{\ln X} = \frac{\ln x_{84} - \ln x_{16}}{2} = \frac{1}{2}\ln\left(\frac{x_{84}}{x_{16}}\right) \equiv \frac{\ln\left(x_{84}/x_{50}\right) + \ln\left(x_{50}/x_{16}\right)}{2} \tag{5}$$

In case of Lognormal distribution, the sample mean and standard deviation (see Section 1) can be estimated as:

$$\mu_{\ln X} \cong \frac{\sum_{i=1}^{n} \ln X_i}{n}$$
$$\sigma_{\ln X} \cong \sqrt{\frac{\sum_{i=1}^{n}\left(\ln X_i - \mu_{\ln X}\right)^2}{n-1}} \tag{6}$$

Lecture 8-2: Linear Regression

Lecturer: Prof. F. Jalayer, 26 February 2021
Notes prepared by: Prof. F. Jalayer and Dr. H. Ebrahimian

1. General

The most common form of linear regression is called *least squares fitting*; a mathematical procedure for finding the best-fitting line to a given set of points by minimizing the *sum of the squares* of the *residuals* (offsets) of the points from the line. The *sum of the squares* of the residuals is used instead of the offset absolute values since it allows the residuals to be treated as a continuous differentiable quantity.

2. Linear Regression Model

In practice, the vertical offsets from a line are almost always minimized (see Figure 1). This provides a fitting function in terms of the *independent variable* (or *predictor variable*) X (actually includes our observations or data) that estimates Y, which is called the *dependent variable* (or *response variable*), for a given X (i.e. $Y|X$).

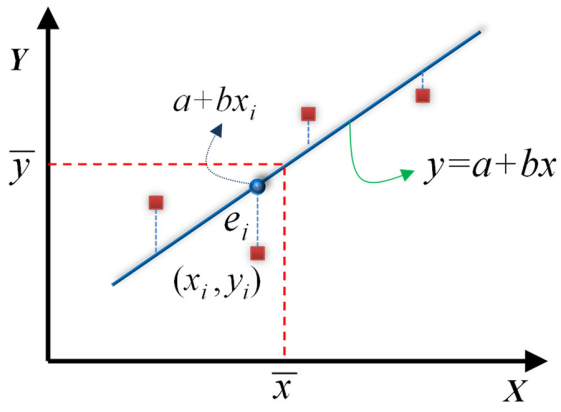

Figure 1: Linear regression

The *Residual Sum of Squares* ($\mathbb{RSS}$) of a set of *n* data points (*residual* for each data point is denoted as e_i, where *i*=1:*n*) has to be minimized. According to Fig. 1, the expression $a+b\,x_i$ is the vertical coordinate of the best-fit line for the data point x_i; thus the $\mathbb{RSS}$ of residuals can be defined as :

$$\mathbb{RSS}=\sum_{i=1}^{n}e_i^{\,2}=\sum_{i=1}^{n}\left[y_i-\left(a+b\,x_i\right)\right]^2 \tag{1}$$

According to Eq. (1), the square deviations from each point are summed, and the resulting residual is then minimized to find the best line. Therefore, this method is also called *Least Square Regression*. The condition for $\mathbb{RSS}$ to be a minimum is that:

$$\begin{aligned}\frac{\partial}{\partial a}\mathbb{RSS}&=-2\sum_{i=1}^{n}\left[y_i-\left(a+b\,x_i\right)\right]=0\\ \frac{\partial}{\partial b}\mathbb{RSS}&=-2\sum_{i=1}^{n}\left[y_i-\left(a+b\,x_i\right)\right]x_i=0\end{aligned} \tag{2}$$

The first expression yields to

$$\begin{aligned}\sum_{i=1}^{n}\left[y_i-\left(a+b\,x_i\right)\right]\triangleq\sum_{i=1}^{n}e_i=0&\Rightarrow\textstyle\sum_{i=1}^{n}y_i=na+b\sum_{i=1}^{n}x_i\\ &\Rightarrow\frac{\sum_{i=1}^{n}y_i}{n}=a+b\frac{\sum_{i=1}^{n}x_i}{n}\Rightarrow\bar{y}=a+b\,\bar{x}\end{aligned} \tag{3}$$

Thus, (1) the sum of the residuals equals zero; (2) the regression line passes through the mean of the data points (as shown in Figure 1). Moreover, Eq. (2) can be re-written in the matrix form as:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{4}$$

$$\Rightarrow b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Given the slope b according to Eq. (4), the intercept a is estimated from Eq. (3). The parameters of the regression can be re-written in a simpler form by defining the Sums of Squares (SS) as follows:

$$\begin{aligned} SS_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ SS_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\ SS_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} \end{aligned} \tag{5}$$

Therefore, Eq. (4) can be re-written as:

$$b = \frac{SS_{xy}}{SS_{xx}} \tag{6}$$

With reference to Eq. (6), SS_{xx} is related to the dispersion of X. As SS_{xx} increases (data with high dispersion), the slope b decreases. SS_{xy} is related to the correlation between X and Y.

3. Probabilistic Model of Linear Regression

The regression can also be regarded as a probabilistic model to define the probability distribution of $Y|X$, i.e. $P(Y|X)$, as follows (see also Figure 2):

$$(Y|X) \sim \mathbb{N}\left[\mathbb{E}(Y|X), \mathbb{VAR}(Y|X)\right] \tag{7}$$

where

$$\begin{aligned} Y &= (a + bX) + e \\ \mathbb{E}(Y|X) &= a + b\,x \\ \mathbb{VAR}(Y|X) &= \sigma^2_{Y|X} = \sigma^2 \end{aligned} \tag{8}$$

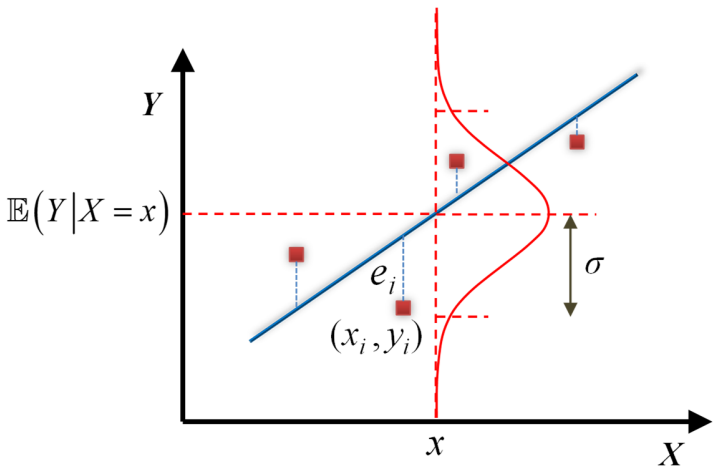

Figure 2: Probabilistic model of linear regression

Note that it has been assumed that the conditional standard deviation of Y given X does not depend on X. In other words, the variation of Y around the regression line is the same across the X values. The assumption of having *constant dispersion* is called the *homoskedasticity* assumption**.** The word comes from the Greek: *homo* (equal) and *skedasticity* (spread). The expressions in Eq. (8)

can also be written in terms of the residuals, e_i, as *i.i.d.* random variables being normally distributed with a mean equal to zero:

$$e_i \sim \mathbb{N}\left(0, \sigma^2\right) \tag{9}$$

Strictly speaking, regression can be seen as a probabilistic model for distribution of residuals (errors). The standard deviation of the regression (or the standard error of regression) can be estimated as:

$$\sigma^2 \approx s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} \left[y_i - \left(a + b\, x_i\right)\right]^2}{n-2} \triangleq \frac{\mathbb{RSS}}{n-2} \tag{10}$$

The deviator n-2 (the degrees of freedom) is used rather than n or n-1. Since the n data points are already used in order to estimate the two regression coefficients, i.e. the slope and the intercept of the regression line, n–2 degree of freedom remains. Hence,

$$\mathbb{STD}\left(Y \middle| X\right) = s = \sqrt{\frac{\sum_{i=1}^{n} \left[y_i - \left(a\, x_i + b\right)\right]^2}{n-2}} \tag{11}$$

4. Coefficient of determination (R^2)

One of the most well-known measures, that indicate how well the sample regression line fits the data, is called *coefficient of determination* or the *R*-square (R^2). The R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left[y_i - \left(a\,x_i + b\right)\right]^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{\mathbb{RSS}}{SS_{yy}} \tag{12}$$

Therefore, R^2 is defined as the proportion of the total sum of squares which is explained by the regression. It is always between zero and one. If we have a perfect fit, then $\mathbb{RSS} \cong 0$ and $R^2 \cong 1$. If $R^2 \cong 0$, we have a poor fit.

5. Logarithmic Linear Regression Model

The logarithmic regression model can be shown as:

$$\ln y_i = \left(\ln a + b \cdot \ln x_i\right) = \ln\left(a\,x_i^{\,b}\right) \Rightarrow y_i = a\,x_i^{\,b} \tag{13}$$

The slope and intercept of the logarithmic linear regression is estimated as (with reference to Eqs. 3 to 5):

$$\ln a = \overline{\ln y} - b\overline{\ln x} \triangleq \mu_{\ln y} - b\mu_{\ln x}$$

$$b = \frac{\sum_{i=1}^{n}\left(\ln x_i - \overline{\ln x}\right)\left(\ln y_i - \overline{\ln y}\right)}{\sum_{i=1}^{n}\left(\ln x_i - \overline{\ln x}\right)^2} = \frac{\sum_{i=1}^{n}\left(\ln x_i - \mu_{\ln x}\right)\left(\ln y_i - \mu_{\ln y}\right)}{\sum_{i=1}^{n}\left(\ln x_i - \mu_{\ln x}\right)^2} = \frac{SS_{\ln x \ln y}}{SS_{\ln x}} \tag{14}$$

6. Probabilistic Model of Logarithmic Linear Regression

The regression can be taken into account as a probabilistic model to define the probability distribution of $Y|X$ to be lognormal or $\ln Y|X$ to be normal as follows (see Figure 3):

$$\left(\ln Y|X\right) \sim \mathbb{N}\left[\mathbb{E}\left(\ln Y|X\right), \mathbb{VAR}\left(\ln Y|X\right)\right] \tag{15}$$

where

$$\begin{aligned} &Y = \left(a \cdot X^b\right) \cdot \varepsilon \\ &\eta_{Y|X} = a \cdot X^b \\ &\mathbb{E}\left(\ln Y|X\right) = \mu_{\ln Y|X} = \ln\left(\eta_{Y|X}\right) = \ln a + b \cdot \ln x \\ &\mathbb{VAR}\left(Y|X\right) = \sigma^2_{\ln Y|X} \triangleq \beta^2_{Y|X} = \text{const.} \end{aligned} \tag{16}$$

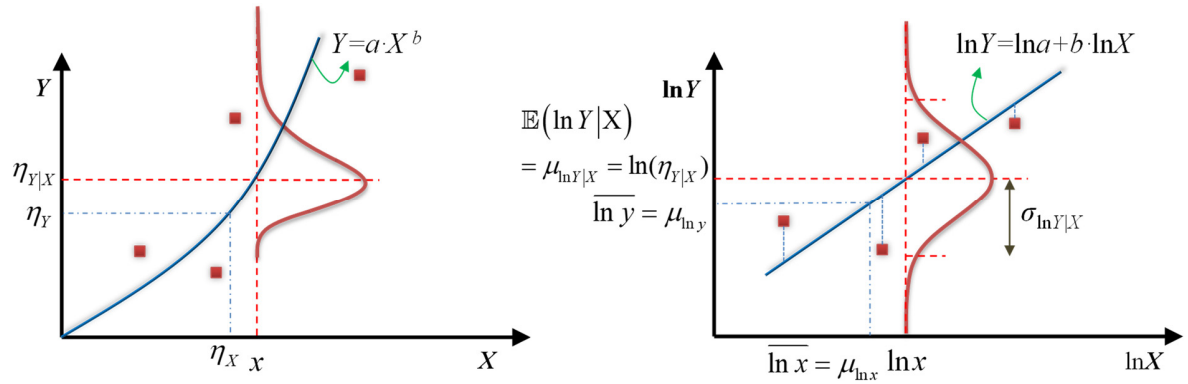

Figure 3: Probabilistic model of logarithmic linear regression: logarithmic scale (right), and arithmetic scale (left)

where ε can be represented by a lognormal distribution, in which we define its parameters, the median and variance as:

$$\begin{aligned} &\eta_\varepsilon = e^{\mu_{\ln \varepsilon}} = 1 \\ &\mathbb{VAR}\left(\ln \varepsilon\right) = \sigma^2_{\ln \varepsilon} = \beta^2_{Y|X} \end{aligned} \tag{17}$$

Note that $\beta_{Y|X}$ is called the logarithmic standard deviation of the regression, and can be estimated as follows:

$$\beta_{Y|X}^2 \approx s^2 = \frac{\sum_{i=1}^{n}\left(\ln \varepsilon_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}\left(\ln \frac{y_i}{a \cdot x_i^{\,b}}\right)^2}{n-2} \tag{18}$$