Linear Regression

Tadao Hoshino （星野匡郎）

ver. 2018 Fall Semester

Regression Analysis

Regression Analysis: Introduction

- Outcome variable of interest : dependent variable
- Variables explaining the variation of the dependent variable : explanatory variables (also referred to as "independent variables" or simply "regressors")

Simple linear regression model

- Linear regression model with a single explanatory variable:

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

- Y: dependent variable, X: explanatory variable, and ε: error term (containing all unobserved determinants of Y).
- β_0: intercept, and β_1: regression coefficient of X. These are parameters of interest to be estimated.

Regression Analysis: Introduction

Multiple linear regression model

- Linear regression model with multiple explanatory variables:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon$$

- The model can be written in a vector form as

$$Y = \beta_0 + \mathbf{X}^\top \beta + \varepsilon$$

where $\mathbf{X} = (X_1, ..., X_k)^\top$, and $\beta = (\beta_1, ..., \beta_k)^\top$.

Example: Determinants of annual income

- A linear regression model of annual income:

$$\text{Income} = \beta_0 + \text{Experience}\beta_1 + \text{Hours}\beta_2 + \text{Education}\beta_3 + \varepsilon$$

- For example, coefficient β_1 tells us

How much an additional year of working experience affects income

$\beta_1 =$ marginal effect of Experience variable

Regression Analysis: Introduction

- "Regressing Y on $\mathbf{X} = (X_1, ..., X_k)^\top$" means

Estimating a function $g(\mathbf{x})$ that predicts the value of Y when $\mathbf{X} = \mathbf{x}$

The function

$$g(\cdot) : \mathbf{X} \rightarrow \text{predicted value of } Y$$

is called the regression function.

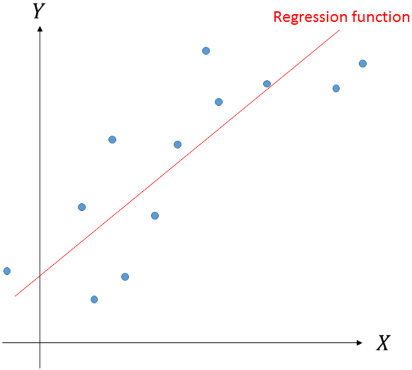
- "Linear" regression is a regression analysis based on the linear regression function:

$$g(\mathbf{x}) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$$

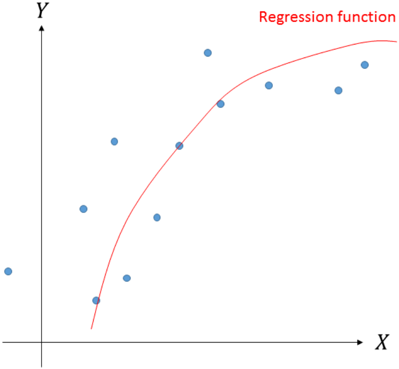
- There are many alternatives for the shape of the regression function. Examples of "nonlinear" regression model:
  - $g(\mathbf{x}) = (\beta_0 + x_1\beta_1 + \cdots + x_k\beta_k)^\alpha$
  - $g(\mathbf{x}) = \exp(\beta_0 + x_1\beta_1 + \cdots + x_k\beta_k)$, etc

Regression Analysis: Introduction

Linear Regression Model



Nonlinear Regression Model



Regression Analysis: Introduction

- What is theoretically the best choice for the shape of a regression function?
  $\Rightarrow$ conditional expectation function
- Linear regression is not the best, but is not too bad and reasonably accurate in practice.

Conditional Expectation Function

Conditional Expectation Function

- Let Y be a dependent variable, and $\mathbf{X} = (X_1, ..., X_k)^\top$ be a vector of explanatory variables. Both are random variables.
- Further, let $f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x})$ be the conditional density function of Y given $\mathbf{X} = \mathbf{x}$.
- Then, the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ is

$$E(Y|\mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x}) dy$$

- Example: $Y =$ height, $X_1 =$ gender, $X_2 =$ age
  - $E(Y|X_1 = \text{male}, X_2 = 12) \approx 152.6$
  - $E(Y|X_1 = \text{female}, X_2 = 18) \approx 157.9$（Ref. 学校保健統計調査）

Conditional Expectation Function

- Thus, the value of the conditional expectation $E(Y|\mathbf{X} = \mathbf{x})$ can vary with the value of $\mathbf{X}$.
- That is, we can view $E(Y|\mathbf{X})$ as a function of $\mathbf{X}$:

$E(Y|\mathbf{X} = \mathbf{x})$ is the value obtained by plugging $\mathbf{x}$ into $E(Y|\mathbf{X})$

Conditional Expectation Function

The function $m(\cdot)$ defined as follows

$$m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$$

is called the conditional expectation function (or conditional mean function).

Law of Iterated Expectations

Here, since $E(Y|\mathbf{X})$ is a function of the random variable $\mathbf{X}$, $E(Y|\mathbf{X})$ is also a random variable. (Recall: a function of a random variable is a random variable.)

Law of Iterated Expectations

The following result is called the law of iterated expectations

$$E(Y) = E(E(Y|\mathbf{X}))$$

NOTE : For the right-hand side, the "inner expectation" is the conditional expectation of Y given $\mathbf{X}$, and the "outer expectation" is the expectation with respect to $\mathbf{X}$.

Law of Iterated Expectations

- When $\mathbf{X}$ is a vector of discrete random variables, the LIE says

$$E(Y) = \sum_{\ell=1}^{L} E(Y|\mathbf{X} = \mathbf{x}_\ell)\,\Pr(\mathbf{X} = \mathbf{x}_\ell)$$

  That is, $E(Y)$ is a weighted average of the group means $E(Y|\mathbf{X} = \mathbf{x}_\ell)$, where each weight is the ratio (probability) of the group $\mathbf{X} = \mathbf{x}_\ell$.

- In the case of continuous $\mathbf{X}$,

$$E(Y) = \int_{-\infty}^{\infty} E(Y|\mathbf{X} = \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Law of Iterated Expectations

Numerical example

Two lotteries: $X = 1, 2$, prize amount: Y

	Lottery 1			Lottery 2	
Probability	0.5	0.5	Probability	0.9	0.1
Y	100	200	Y	100	500

Suppose that the two lotteries are mixed together with proportion $6 : 4$; that is, $\Pr(X = 1) = 0.6$ and $\Pr(X = 2) = 0.4$.

		Lottery 1 + 2	
Probability	0.66	0.3	0.04
Y	100	200	500

Law of Iterated Expectations

Numerical example (cont')

- $E[Y] = 0.66 \cdot 100 + 0.3 \cdot 200 + 0.04 \cdot 500 = 146$
- $E[Y|X=1] = 0.5 \cdot 100 + 0.5 \cdot 200 = 150$
- $E[Y|X=2] = 0.9 \cdot 100 + 0.1 \cdot 500 = 140$
- The law of iterated expectations:

$$\begin{aligned} E[E[Y|X]] &= E[Y|X=1]\Pr(X=1) + E[Y|X=2]\Pr(X=2) \\ &= 150 \cdot 0.6 + 140 \cdot 0.4 = 90 + 56 \\ &= 146 \\ &= E[Y] \end{aligned}$$

Law of Iterated Expectations

Proof. We only prove the continuous case.

$$\begin{aligned}
E(E(Y|\mathbf{X})) &= \int_{-\infty}^{\infty} E(Y|\mathbf{X}=\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) dy \right) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right) dy \\
&\overset{\text{(i)}}{=} \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{Y,\mathbf{X}}(y,\mathbf{x}) d\mathbf{x} \right) dy \\
&\overset{\text{(ii)}}{=} \int_{-\infty}^{\infty} y f_Y(y) dy = E(Y) \quad \blacksquare
\end{aligned}$$

NOTE: (i) $f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) = \frac{f_{Y,\mathbf{X}}(y,\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}$; (ii) $\int_{-\infty}^{\infty} f_{Y,\mathbf{X}}(y,\mathbf{x}) d\mathbf{x} = f_Y(y)$.

Exercise

Exercise 1

Let X and U be random variables, and $Y = U + 0.5X$. Suppose that X is normally distributed as normal with mean 0 and variance 3, $N(0,3)$, U is independent of X, and $E(U) = 1$. Calculate the following quantities:

1. $E(Y)$,
2. $E(Y|X)$,
3. $E(g(X)Y)$, where $g(X) = 4X$.

Best Regression Function

Best Regression Function

- Let $g(\cdot)$ be a candidate regression function, and $e(\mathbf{X})$ be the corresponding prediction error of Y at $\mathbf{X} = \mathbf{x}$

$$e(\mathbf{x}) = Y - g(\mathbf{x}).$$

- It is natural to think that for an ideal regression function the expectation of the prediction error should be zero:

$$E(e(\mathbf{X})) = 0.$$

- When we set $g(\cdot)$ to the conditional expectation function $m(\cdot)$, by the law of iterated expectations,

$$\begin{aligned} E(e(\mathbf{X})) &= E(Y - m(\mathbf{X})) \\ &= E(Y) - E(E(Y|\mathbf{X})) \\ &= E(Y) - E(Y) = 0. \end{aligned}$$

Thus, the conditional expectation function meets this criterion.

Best Regression Function

- However, the conditional expectation function is not the only function that satisfies $E(e(\mathbf{X})) = 0$. Note that

$$\begin{aligned} E(e(\mathbf{X})) &= E(Y - g(\mathbf{X})) \\ &= E(Y - m(\mathbf{X})) + E(m(\mathbf{X}) - g(\mathbf{X})) \\ &= E(m(\mathbf{X}) - g(\mathbf{X})). \end{aligned}$$

  There are many functions $g(\cdot)$ satisfying $E(m(\mathbf{X}) - g(\mathbf{X})) = 0$.

- For an extreme example, consider a constant regression function $g(\mathbf{x}) = E(Y)$ for all $\mathbf{x}$:

$$E(m(\mathbf{X}) - E(Y)) = E(Y) - E(Y) = 0$$

- However, using a constant regression function is not appropriate for the purpose of regression analysis (because the predicted value of Y is independent of $\mathbf{X}$, in this case).

Best Regression Function

- The requirement $E(e(\mathbf{X})) = 0$ corresponds the "unbiasedness" of the regression function.
- The "best" regression function should be not only unbiased but also have the smallest variance (variance = the risk of wrong prediction).
- That is, we consider the minimization of $E(e^2(\mathbf{X}))$, the so-called MSE (mean squared error).

Best Regression Function

(cont')

- For a candidate regression function $g(\cdot)$ and the conditional expectation function $m(\cdot)$, let

$$\begin{aligned}
\text{MSE } E(e^2(\mathbf{X})) &= E[(Y - g(\mathbf{X}))^2] \\
&= E[(Y - m(\mathbf{X}) + m(\mathbf{X}) - g(\mathbf{X}))^2] \\
&= E[(Y - m(\mathbf{X}))^2] + 2E[(Y - m(\mathbf{X}))(m(\mathbf{X}) - g(\mathbf{X}))] \\
&\quad + E[(m(\mathbf{X}) - g(\mathbf{X}))^2] \\
&= E(I_1) + 2E(I_2) + E(I_3)
\end{aligned}$$

- $E(I_1)$ is not dependent on the choice of $g(\cdot)$, and thus can be ignored.

Best Regression Function

(cont')

- Note that $E(I_2) = 0$, because

$$\begin{aligned} E(I_2|\mathbf{X}) &= E[(Y - m(\mathbf{X}))(m(\mathbf{X}) - g(\mathbf{X}))|\mathbf{X}] \\ &= \underbrace{(E(Y|\mathbf{X}) - m(\mathbf{X}))}_{=0}(m(\mathbf{X}) - g(\mathbf{X})) = 0 \end{aligned}$$

  and by the law of iterated expectations,

$$E(I_2) = E(E(I_2|\mathbf{X})) = E(0) = 0$$

- Thus, the first two parts $E(I_1)$ and $2E(I_2)$ of the MSE cannot be made smaller by manipulating the form of $g(\cdot)$.

Best Regression Function

(cont')

- Consequently, the minimizer of the MSE can be characterized by a function $g(\cdot)$ minimizing

$$E(I_3) = E[(m(\mathbf{X}) - g(\mathbf{X}))^2]$$

- Clearly from the above, it is only when we set the regression function $g(\cdot)$ equal to the conditional expectation function $m(\cdot)$ that $E(I_3)$ (and thus the whole MSE) is minimized.

$\Rightarrow$ The best regression function is given by the conditional expectation.

Summary of the Discussion So Far

- “Regressing Y on $\mathbf{X}$” means

  Finding a function that predicts Y using the information on $\mathbf{X}$

- The function of $\mathbf{X}$ that gives the predicted value of Y is called the regression function.
- Based on the MSE criterion, the best regression function coincides with the conditional expectation function of Y given $\mathbf{X}$.
- Therefore, mathematically speaking, the regression analysis is a problem of estimating the conditional expectation function.

Linear Regression Function

Linear Regression Function

- It is often very difficult to directly estimate the conditional expectation function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ without imposing any functional form assumptions. (Such an approach is called nonparametric regression.)
- In practice, we need some simplifications on the shape of the regression function:

Linear regression model	$g(\mathbf{x}) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$
Nonlinear regression model	e.g., $g(\mathbf{x}) = (\beta_0 + x_1\beta_1 + \cdots + x_k\beta_k)^\alpha$

- If the chosen regression function is "perfect" such that $g(\mathbf{x}) = m(\mathbf{x})$, we do not need to directly estimate the conditional expectation.
- Instead, it suffices to estimate only the "parameters" $\beta_0, ..., \beta_k$ and α. (This type of approach is called parametric regression.)

Linear Regression Function

- Among many parametric regression models, the linear regression model is the most often employed in both theoretical and applied researches.
- The assumption that the conditional expectation function is linear:

  $$E(Y|\mathbf{X}=\mathbf{x}) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$$

  seems a very restrictive condition.
- For example, for simplicity, consider a single explanatory variable X_1. Assuming that $E(Y|X_1)$ is linear in X_1 is indeed often very restrictive.
- However, if we use the "vector" of explanatory variables $\mathbf{X}$, including polynomials of X_1,

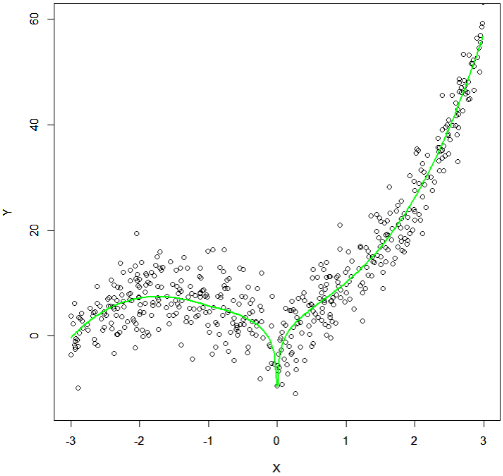
  $$\mathbf{X} = (X_1, X_1^2, X_1^3, ...)^\top$$

  a linear function of $\mathbf{X}$ can well approximate the conditional expectation function even when the function is highly nonlinear and complicated. [1]

[1] Note that $E(Y|X_1) = E(Y|\mathbf{X})$ because $X_1 = x_1 \iff X_1^2 = x_1^2, X_1^3 = x_1^3, ...$

Linear Regression Function

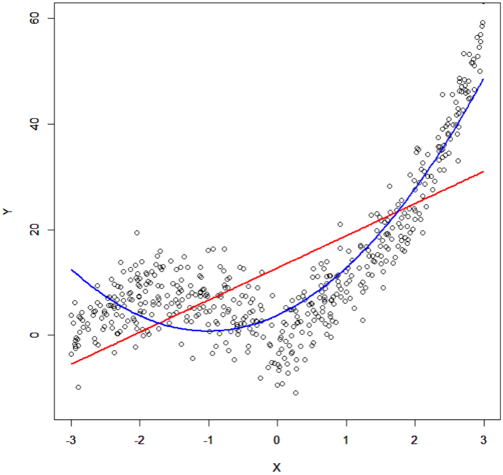
—: True conditional expectation function $m(\mathbf{x})$



Linear Regression Function

—: $g(\mathbf{x}) = \beta_0 + \mathbf{x}\beta_1$

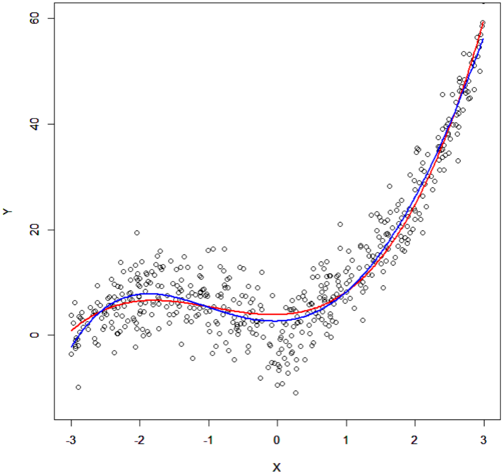
—: $g(\mathbf{x}) = \beta_0 + \mathbf{x}\beta_1 + \mathbf{x}^2\beta_2$



Linear Regression Function

—: $g(\mathbf{x}) = \beta_0 + \mathbf{x}\beta_1 + \mathbf{x}^2\beta_2 + \mathbf{x}^3\beta_3$

—: $g(\mathbf{x}) = \beta_0 + \mathbf{x}\beta_1 + \mathbf{x}^2\beta_2 + \mathbf{x}^3\beta_3 + \mathbf{x}^4\beta_4$



Linear Regression Function

- Thus, with the inclusion of polynomials, the linear regression function can handle a relatively wide range of functional forms.[2]
- In addition, compared with nonlinear regression, linear regression model is much easier to implement and analyze.

[2] The Weierstrass Approximation Theorem shows that any continuous functions on a closed interval can be approximated by polynomials with arbitrary accuracy.

Estimation of Linear Regression Model

OLS Estimator

- Suppose that we have data of n independent observations of $\{(Y_1, X_1), ..., (Y_n, X_n)\}$.
- For simplicity, we first focus on a simple linear regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i, \;\; i = 1, ..., n$$

- The most popular estimation method for the parameters (β_0, β_1) is the least squares method:

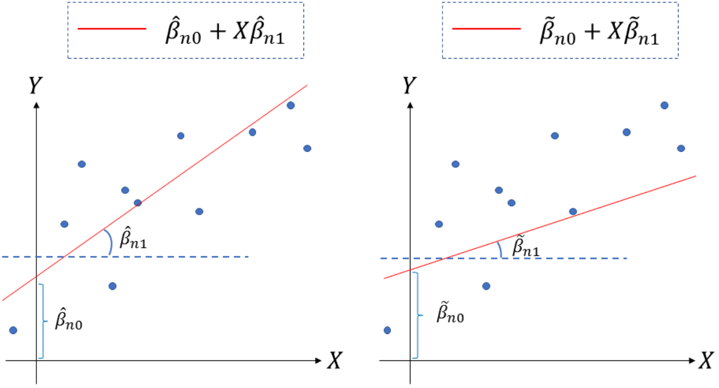
$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}} Q_n(\beta_0, \beta_1)$$

$$Q_n(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1)^2$$

where "argmin" stands for the argument that minimizes the following objective function.

OLS Estimator

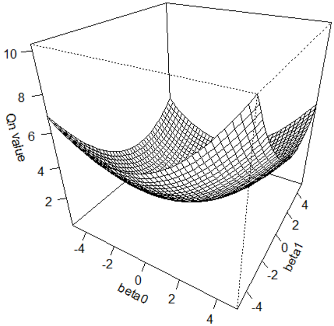
The estimated least squares regression function is characterized by the line that best fits the data.



$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \text{argmin} Q_n(\beta_0, \beta_1)$ $\qquad$ $(\tilde{\beta}_{n0}, \tilde{\beta}_{n1}) \neq \text{argmin} Q_n(\beta_0, \beta_1)$

OLS Estimator

[Shape of the objective function $Q_n(\beta_0, \beta_1)$]



∗ The figure suggests that the minimizer of $Q_n(\beta_0, \beta_1)$ is unique.

OLS Estimator

- By the first order condition of the least-squares problem, $(\hat{\beta}_{n0}, \hat{\beta}_{n1})$ satisfies the following:

$$(1) \quad 0 = \frac{Q_n(\hat{\beta}_{n0}, \hat{\beta}_{n1})}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_{n0} - X_i \hat{\beta}_{n1})$$

$$(2) \quad 0 = \frac{Q_n(\hat{\beta}_{n0}, \hat{\beta}_{n1})}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_{n0} - X_i \hat{\beta}_{n1})$$

- That is, the estimator $(\hat{\beta}_{n0}, \hat{\beta}_{n1})$ is the solution of the **system of equations** $\{$(1), (2)$\}$.

OLS Estimator

- First, solving (1) and (2) yield

$$\hat{\beta}_{n0} = \bar{Y}_n - \bar{X}_n \hat{\beta}_{n1}$$

and

$$0 = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \hat{\beta}_{n0} - \frac{1}{n}\sum_{i=1}^{n} X_i^2 \hat{\beta}_{n1},$$

respectively, where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Further, combining these two yields

$$0 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n}_{\text{sample covariance of } (Y,X)} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right)}_{\text{sample variance of } X} \hat{\beta}_{n1}$$

OLS Estimator

- Hence, we obtain

$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2}$$

  or numerically equivalently

$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2},$$

  where the numerator is the sample covariance of (Y, X) and the denominator is the sample variance of X.
- The estimator $\hat{\beta}_{n1}$ obtained by this formula is called the OLS (Ordinary Least Squares) estimator.

OLS Estimator with Multiple Regressors

- Consider a multiple regression model:

$$Y_i = \mathbf{X}_i^\top \beta + \varepsilon_i, \;\; i = 1, ..., n$$

  where $\mathbf{X}$ is a $k \times 1$ vector of explanatory variables including a constant term, and β is the associated vector of regression coefficients.

- The least squares estimator of β is similarly given by

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmin}}\, Q_n(\beta)$$

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^\top \beta)^2$$

OLS Estimator with Multiple Regressors

- By the first order condition of the least-squares problem,

$$\begin{aligned}\underbrace{\mathbf{0}}_{k\times 1 \text{ vector of zeros}} &= \frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta} \\ &= -\frac{2}{n}\sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i^\top \hat{\beta}_n)\end{aligned}$$

- Dividing both sides by 2 and rearranging, we get

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i$$

OLS Estimator with Multiple Regressors

- Then, if the inverse matrix of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top$ exists,

$$\hat{\beta}_n = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}}_{k\times k}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i}_{k\times 1}$$

- This is the formula of the OLS estimator for the multiple linear regression.

Statistical Properties of the OLS Estimator

Evaluation of Estimators

- For the estimation of the regression coefficients, many techniques are available, not limited to the least squares (OLS).
- Least Squares (OLS):

$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - X_i\beta_1)^2$$

- Least Absolute Deviations (LAD):

$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \beta_0 - X_i\beta_1|$$

- Maximum Likelihood, Method of Moments, and many others.

Evaluation of Estimators

- Formally, the estimator is a "procedure" to obtain the value of a parameter of interest; namely, the estimator is a function of the data $\{(Y_1, X_1), ..., (Y_n, X_n)\}$. Thus, **the estimator is itself a random variable**.

$$\text{Estimator } \hat{\beta}_{n1} = T((Y_1, X_1), ..., (Y_n, X_n))$$

- A particular realization of the estimator, the value obtained by plugging the real data into $T(\cdot)$, is called the estimate.

$$\text{Estimate} = T((y_1, x_1), ..., (y_n, x_n))$$

$$\begin{aligned} \text{Estimator} &= \text{ Estimation procedure (random variable)} \\ \text{Estimate} &= \text{ Estimation result (realized value)} \end{aligned}$$

Evaluation of Estimators

- When different estimates are obtained from different estimators, because the true value of the parameter to be estimated is unknown, it is in general impossible to conclude which estimate is more accurate than the other.
- On the other hand, since the estimator is a random variable, we can discuss which estimator is more preferable from the perspective of probability theory.
- There are three major criteria used to evaluate the performance of estimators:
  - Consistency
  - Unbiasedness
  - Efficiency

Consistency

- Denote θ_0 as the parameter of interest in the population, i.e., θ_0 is the true value of θ. Also, let $\hat{\theta}_n$ be any estimator of θ_0.

Consistency

The estimator $\hat{\theta}_n$ is said to be a consistent estimator of θ_0, if $\hat{\theta}_n$ converges to θ_0 in probability as n increases:

$$\hat{\theta}_n \xrightarrow{P} \theta_0, \ (n \to \infty).$$

- For example, let $\hat{\theta}_n$ be the sample average and θ_0 be the population mean of X:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \ \ \theta_0 = E(X)$$

  Then, by the weak law of large numbers, $\hat{\theta}_n$ is consistent for θ_0.

Unbiasedness

Unbiasedness

For a given estimator $\hat{\theta}_n$ of θ_0, $\hat{\theta}_n$ is said to be an unbiased estimator of θ_0, if its expectation is equal to θ_0:

$$E(\hat{\theta}_n) = \theta_0.$$

- The gap between the expected value of the estimator and the true parameter $E(\hat{\theta}_n) - \theta_0$ is called bias. As suggested by its name, an unbiased estimator is an estimator with zero bias.
- Clearly, the sample average is an unbiased estimator of the population mean:
$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = E(X)$$

Unbiasedness

- On the other hand, it is well known that the sample variance $V_n(X) = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is *not* an unbiased estimator for the population variance $V(X)$.

Proof. Observe that

$$
\begin{aligned}
V_n(X) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X) - [\bar{X}_n - E(X)])^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2 - \frac{1}{n}\sum_{i=1}^{n}(\bar{X}_n - E(X))^2
\end{aligned}
$$

Unbiasedness

(cont') Noting that $E(\bar{X}_n) = E(X)$ and $V(\bar{X}_n) = V(X)/n$,

$$\begin{aligned} E[V_n(X)] &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - E(X))^2] - \frac{1}{n}\sum_{i=1}^{n} E[(\bar{X}_n - E(X))^2] \\ &= V(X) - V(\bar{X}_n) \\ &= (1 - 1/n)V(X) \neq V(X) \quad \blacksquare \end{aligned}$$

- From the third equality, we can see that although $V_n(X)$ is biased for $V(X)$ for finite n, the bias vanishes as $n \to \infty$.
- An unbiased variance estimator can be obtained by the following formula
$$V_n'(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$
In a similar way to the above, one can show that $E[V_n'(X)] = V(X)$ even for finite n. (The proof is omitted.)

Consistent but Biased

- A minimum requirement for an estimator is consistency, which ensures that the estimate well approximates the true parameter as long as the sample size is sufficiently large.
- There are many estimators that are consistent but are biased. For example, letting

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i + 1/n, \;\; \theta_0 = E(X)$$

the estimator has consistency

$$\hat{\theta}_n \xrightarrow{P} E(X) + 0 = \theta_0$$

but is not unbiased

$$E(\hat{\theta}_n) = E(X) + 1/n \neq \theta_0.$$

Unbiased but Inconsistent

- Similarly, we can easily construct estimators that are unbiased but are inconsistent. Let

$$\hat{\theta}_n = \frac{1}{2}(X_1 + X_n), \;\; \theta_0 = E(X).$$

  Then, the estimator is clearly unbiased.

- However, noting that $V(\hat{\theta}_n) = V(X)/2$, by Chebyshev's inequality,

$$\begin{aligned}\Pr(|\hat{\theta}_n - \theta_0| \geq \epsilon) &\leq \frac{V(\hat{\theta}_n)}{\epsilon^2} \\ &= \frac{V(X)}{2\epsilon^2} \not\to 0 \;\; (n \to \infty)\end{aligned}$$

  which implies the inconsistency of $\hat{\theta}_n$.

Efficiency

- When various consistent estimators are available, we should choose one which has a smaller stochastic error. Namely, an estimator with smaller variance is more preferable because

> the smaller the variance, the higher the probability of obtaining a value close to the true parameter.

- Suppose $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ to be two consistent estimators of θ_0. We say the estimator $\hat{\theta}_{n,1}$ is more efficient than $\hat{\theta}_{n,2}$ if

$$V(\hat{\theta}_{n,1}) < V(\hat{\theta}_{n,2})$$

- When both $V(\hat{\theta}_{n,1})$ and $V(\hat{\theta}_{n,2})$ decrease to zero as n increases, the efficiency does not matter in the limit $n \to \infty$.
- For finite n, an efficient estimator is more precise, in the sense that it produces a narrower confidence interval.

Efficiency

Example

- Let $\hat{\theta}_{n,1}$ be the sample average of X, and $\hat{\theta}_{n,2}$ be the average over the observations with odd indices (assuming that n is a even number):

$$\hat{\theta}_{n,1} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\theta}_{n,2} = \frac{1}{n/2}\sum_{i=1}^{(n/2)} X_{2i-1}$$

- Clearly, both $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ are consistent and unbiased for $E(X)$.
- However, their variances differ:

$$V(\hat{\theta}_{n,1}) = \frac{V(X)}{n} < V(\hat{\theta}_{n,2}) = \frac{V(X)}{n/2}$$

This implies that $\hat{\theta}_{n,1}$ is more efficient than $\hat{\theta}_{n,2}$.

Statistical Properties of the OLS Estimator

In order to investigate the statistical properties of the OLS estimator, we introduce the following assumptions:

Assumption 1. The conditional expectation of Y given X is given by a linear function $\beta_{00} + X\beta_{01}$. That is,

$$\begin{aligned} Y &= \beta_{00} + X\beta_{01} + \varepsilon \\ E(Y|X) &= \beta_{00} + X\beta_{01} \ \ (\iff E(\varepsilon|X) = 0) \end{aligned}$$

Assumption 2. $E(X^2)$ is finite.

Assumption 3. The observations $\{(Y_1, X_1), ..., (Y_n, X_n)\}$ are independent and sampled from the same population.

Assumption 4. The error term ε is independent of X, and its variance is given by $E(\varepsilon^2) = \sigma^2$.[3]

[3] Assumption 4 is called homoskedasticity (constant variance).

Statistical Properties of the OLS Estimator

Unbiasedness of the OLS estimator

The OLS slope estimator

$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

is an unbiased estimator of β_{01}

Proof. First, note that

$$E(Y_i|X_1, ..., X_n) = E(Y_i|X_i) = \beta_{00} + X_i\beta_{01}$$

where the first equality follows by that the X_j's $(j \neq i)$ are irrelevant to Y_i, and the second equality follows by Assumption 1.

Statistical Properties of the OLS Estimator

(cont') Also note that the OLS slope estimator can be re-written as

$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)Y_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i}$$

Then, the conditional expectation of $\hat{\beta}_{n1}$ given $(X_1, ..., X_n)$ can be calculated as

$$\begin{aligned} E\left(\hat{\beta}_{n1}\middle| X_1, ..., X_n\right) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)E(Y_i|X_1, ..., X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(\beta_{00} + X_i\beta_{01})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\beta_{00}}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i} + \beta_{01} \end{aligned}$$

Statistical Properties of the OLS Estimator

(cont') Because

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\beta_{00} = (\bar{X}_n - \bar{X}_n)\beta_{00} = 0,$$

we have $E\left(\hat{\beta}_{n1}\middle| X_1, ..., X_n\right) = \beta_{01}$. Further, by the law of iterated expectations, we have

$$\begin{aligned} E(\hat{\beta}_{n1}) &= E\left[E\left(\hat{\beta}_{n1}\middle| X_1, ..., X_n\right)\right] \\ &= E(\beta_{01}) = \beta_{01} \end{aligned}$$

which implies the desired result. ■

Statistical Properties of the OLS Estimator

Unbiasedness of the OLS estimator

The OLS intercept estimator

$$\hat{\beta}_{n0} = \bar{Y}_n - \bar{X}_n \hat{\beta}_{n1}$$

is an unbiased estimator of β_{00}.

Proof. The conditional expectation of $\hat{\beta}_{n0}$ given $(X_1, ..., X_n)$ is

$$\begin{aligned} E(\hat{\beta}_{n0} | X_1, ..., X_n) &= E(\bar{Y}_n | X_1, ..., X_n) - \bar{X}_n E(\hat{\beta}_{n1} | X_1, ..., X_n) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(Y_i | X_i) - \frac{1}{n} \sum_{i=1}^{n} X_i \beta_{01} \\ &= \beta_{00} + \frac{1}{n} \sum_{i=1}^{n} X_i \beta_{01} - \frac{1}{n} \sum_{i=1}^{n} X_i \beta_{01} = \beta_{00}. \end{aligned}$$

Then, the result follows from the law of iterated expectations. ■

Statistical Properties of the OLS Estimator

Consistency of the OLS estimator

The OLS slope estimator

$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

is a consistent estimator of β_{01}.

Sketch of the proof. Substituting $Y_i = \beta_{00} + X_i\beta_{01} + \varepsilon_i$ in the definition of the OLS estimator, we have

$$\begin{aligned}\hat{\beta}_{n1} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(\beta_{00} + X_i\beta_{01} + \varepsilon_i)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i} \\ &= \beta_{01} + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i}\end{aligned}$$

Statistical Properties of the OLS Estimator

(cont') Thus, in order to show the consistency, we need to show that the second term on the right-hand side converges to zero in probability. Letting

$$S_i = \frac{(X_i - \bar{X}_n)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i},$$

we can write

$$\hat{\beta}_{n1} = \beta_{01} + \frac{1}{n}\sum_{i=1}^{n} S_i,$$

where $E(S) = 0$ holds (because $E(S_i|X_1, ..., X_n) = 0$). Applying the weak law of large numbers to $\frac{1}{n}\sum_{i=1}^{n} S_i$, we have

$$\frac{1}{n}\sum_{i=1}^{n} S_i \xrightarrow{P} E(S) = 0,$$

which implies $\hat{\beta}_{n1} \xrightarrow{P} \beta_{01}$. ■

Statistical Properties of the OLS Estimator

Consistency of the OLS estimator

The OLS intercept estimator

$$\hat{\beta}_{n0} = \bar{Y}_n - \bar{X}_n \hat{\beta}_{n1}$$

is a consistent estimator of β_{00}.

Sketch of the proof. Substituting $Y_i = \beta_{00} + X_i \beta_{01} + \varepsilon_i$ in the definition of the OLS estimator, we have

$$\hat{\beta}_{n0} = \beta_{00} + \frac{1}{n} \sum_{i=1}^{n} X_i (\beta_{01} - \hat{\beta}_{n1}) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i.$$

As shown above, $\beta_{01} - \hat{\beta}_{n1} \xrightarrow{P} 0$. In addition, by the weak law of large numbers, we have $\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \xrightarrow{P} 0$. Hence, we obtain $\hat{\beta}_{n0} \xrightarrow{P} \beta_{00}$. ■

Statistical Properties of the OLS Estimator

- Next, we derive the variance of the OLS slope estimator $\hat{\beta}_{n1}$.
- By definition, the variance of $\hat{\beta}_{n1}$, $V(\hat{\beta}_{n1})$, is given by

$$V(\hat{\beta}_{n1}) = E\left(\{\hat{\beta}_{n1} - E(\hat{\beta}_{n1})\}^2\right)$$

By the unbiasedness $E(\hat{\beta}_{n1}) = \beta_{01}$ and the law of iterated expectaions, we have

$$V(\hat{\beta}_{n1}) = E\left(E\left(\{\hat{\beta}_{n1} - \beta_{01}\}^2 | X_1, ..., X_n\right)\right)$$

- Recall that

$$\hat{\beta}_{n1} - \beta_{01} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

where

$$S_i = \frac{(X_i - \bar{X}_n)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i}.$$

Statistical Properties of the OLS Estimator

(cont') Thus,

$$V(\hat{\beta}_{n1}) = \frac{1}{n^2} E\left(\sum_{j=1}^{n}\sum_{i=1}^{n} E(S_j S_i | X_1, ..., X_n)\right)$$

Since ε_i's are independent, for $j \neq i$,

$$E(S_j S_i | X_1, ..., X_n) = E(S_j | X_1, ..., X_n) E(S_i | X_1, ..., X_n) = 0,$$

and therefore

$$\sum_{j=1}^{n}\sum_{i=1}^{n} E(S_j S_i | X_1, ..., X_n) = \sum_{i=1}^{n} E(S_i^2 | X_1, ..., X_n).$$

Statistical Properties of the OLS Estimator

(cont') Consequently, noting that $E(\varepsilon_i^2|X_i) = E(\varepsilon_i^2) = \sigma^2$ by Assumption 4,

$$
\begin{aligned}
V(\hat{\beta}_{n1}) &= \frac{1}{n^2} E\left(\sum_{i=1}^{n} E(S_i^2|X_1, ..., X_n) \right) \\
&= \frac{1}{n} E\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i} \right) \sigma^2
\end{aligned}
$$

- This formula suggests that the variance (i.e., estimation error) of the OLS estimator decreases as the sample size increases and as the variance of X increases:
    - The larger the sample size, the more information it contains.
    - The more variation of the values of X, the easier the estimation of β_{01}.

Statistical Properties of the OLS Estimator

- Equivalently, the above result can be written as

$$V(\sqrt{n}(\hat{\beta}_{n1} - \beta_{01})) = E\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)X_i}\right)\sigma^2$$

- Further, by the law of large numbers, in the limit $n \to \infty$ the right-hand side converges to

$$V_\infty(\sqrt{n}(\hat{\beta}_{n1} - \beta_{01})) = \frac{\sigma^2}{V(X)}$$

This variance in the limit is referred to as the asymptotic variance.[4]

[4] In statistics, the term "asymptotic" means that the result holds "as the sample size increases to infinity".

Statistical Properties of the OLS Estimator

- Finally, applying the central limit theorem to the OLS estimator, we can show that

$$\sqrt{n}(\hat{\beta}_{n1} - \beta_{01}) \xrightarrow{d} N\left(0, \frac{\sigma^2}{V(X)}\right)$$

- In other words, if the sample size is sufficiently large, the distribution of $\sqrt{n}(\hat{\beta}_{n1} - \beta_{01})$ can be approximated by the normal distribution $N\left(0, V(X)^{-1}\sigma^2\right)$.[5]
- Derivation of the variance and the distribution of the intercept estimator $\hat{\beta}_{n0}$ are omitted.

[5] This result plays a fundamental role in the theory of hypothesis testing of the OLS estimator.

What If the Linear Regression Function is Misspecified?

Model Misspecification

- Without loss of generality, a regression model can be expressed as

$$Y = m(\mathbf{X}) + \varepsilon$$

  where $m(\mathbf{X}) = E(Y|\mathbf{X})$, and ε is an error term, such that $E(\varepsilon|\mathbf{X}) = 0$.
- The linear regression model is based on a model specification assumption that $m(\mathbf{X})$ is a linear function in $\mathbf{X}$:

$$m(\mathbf{X}) = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$$

- Suppose that the linear model specification is not true

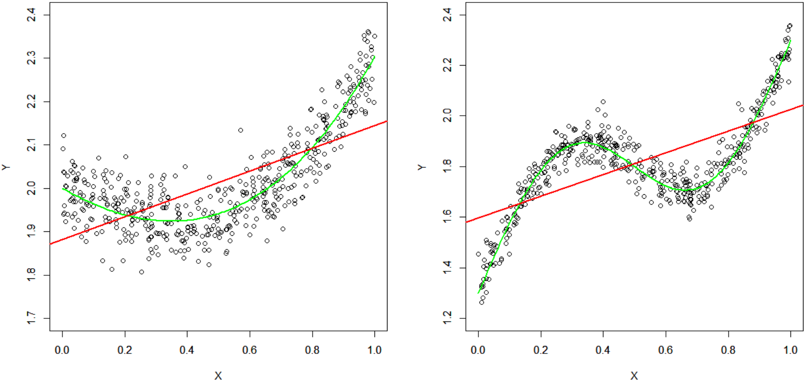
$$m(\mathbf{X}) \neq \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$$

  and we fit an incorrect linear regression function to the data. How can we interpret the resulting OLS estimates in this situation?

Model Misspecification

🟩: True conditional expectation function $m(x)$

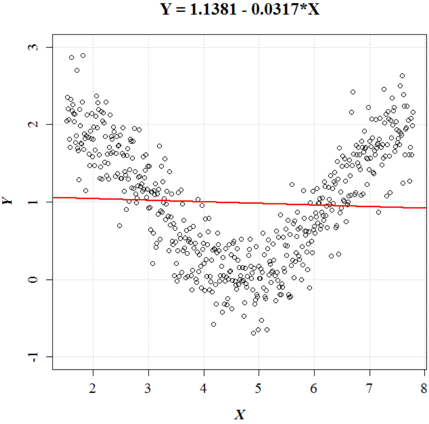
🟥: Estimated linear regression function $\hat{\beta}_{n0} + x\hat{\beta}_{n1}$



Model Misspecification

- When the linear regression function is misspecified, the estimated regression function can be interpreted as a "linear approximation" of the true regression function $m(\mathbf{X})$.
- In reality, the assumption that the data perfectly follows a linear functional form is rarely (or never) met. The OLS estimates should be interpreted as just approximates.
- In addition, it is important to note that the linear approximation is not necessarily always informative. There are situations even when the magnitude of $\hat{\beta}_{n1}$ is very small, X is significantly related to Y (as shown in the figure below).

Model Misspecification



∗ Although the estimated regression coefficient of X is almost zero, there is a clear "nonlinear" relationship between X and Y.

Glossary I

Glossary II