Statistics Gist

Abdullah Al Mahmud

6/23/2021

Keywords

Autocorrelation relationship of the observations between the different points in time

Baseline survey a survey which measures key conditions (indicators) before a project begins against which change and progress can be assessed.

Collinearity a linear association between two explanatory variables.

Confounder a variable that influences both the dependent variable and independent variable

Multicollinearity a situation in which more than two explanatory variables in a multiple regression model are highly linearly related

Power The chance that the study will be able to demonstrate a significant difference or effect if it is present.

Orthogonal (of an experiment) having variates which can be treated as statistically independent.

Formulae

Covariance, $Cov(X,Y) = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n}$

Regression

Let, $Y = \alpha + \beta X + \epsilon$

$E[Y|X] = \alpha + \beta X$

$\beta = E(Y|X = x+1) - E(Y|X = x)$

Assumptions of simple linear regression

- Condition of Y given X is a linear function of the parameter, i.e., $E(Y|X) = \alpha + \beta X$.
- $E(\epsilon_i) = 0 \forall i = 1, 2, 3, \cdots, n$
- $Var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2 \forall i = 1, 2, 3, \cdots, n$
- $\epsilon \sim NID(0, \sigma^2)$
- $Cov(\epsilon_i, \epsilon_j) = 0 \forall i \neq j = 1, 2, \cdots, n$ (if not, autocorrelation)
- X is non-stochastic (non-random) variable

N:B: $Cov(\epsilon_i, \epsilon_j) = E(\epsilon_i, \epsilon_j) - E(\epsilon_i)E(\epsilon_j)$

$$
\begin{aligned}
Y = \alpha + \beta X + \epsilon \\
&\Rightarrow \quad \hat{Y}_i = \hat{\alpha} + \hat{\beta} X \\
&\Rightarrow \quad \epsilon_i = Y_i - \hat{Y}_i
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
\mathrm{Cov}(x,x) &= \mathrm{E}(x,x) - [E(x)]^2 \\
&= E(x^2) - [E(x)]^2 \ \ (\#eq:covxx) \\
&= Var(x)
\end{aligned} \tag{2}
$$

See @ref(eq:covxx)

Heteroskedasticity

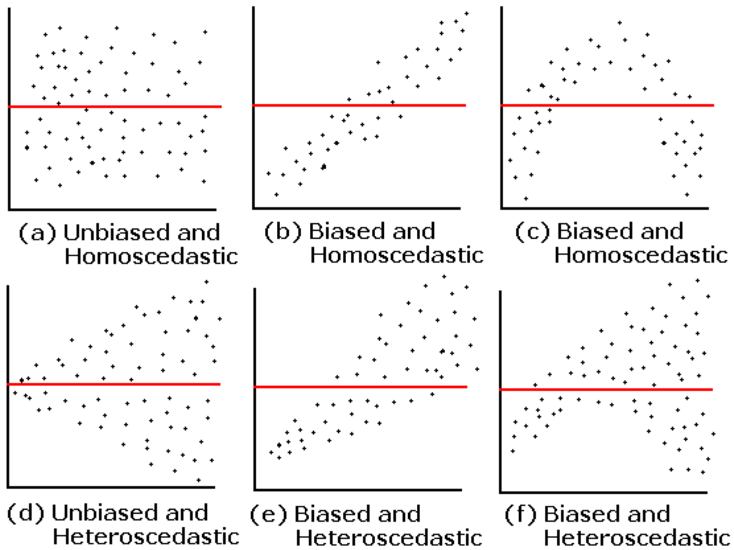

Figure 1: Bias and Heteroskedasticity

OLS estimates

$Y = \alpha + \beta X + \epsilon$

$\epsilon_i = Y_i - \hat{Y}_i$

$\sigma \sum$

$$
\begin{aligned}
&\Rightarrow \quad SSE = \Sigma\epsilon_i^2 = \sum(Y_i - \alpha - \beta X)^2 \\
&\Rightarrow \quad \frac{\delta SSE}{\delta\alpha} = 0 \\
&\Rightarrow \quad -2\Sigma(Y_i - \alpha - \beta X_i) \\
&\Rightarrow \quad \frac{\delta SSE}{\delta\beta} = 0 \\
&\Rightarrow \quad 2\Sigma(Y_i - \alpha - \beta X_i(-X_i)) = 0 \\
&\Rightarrow \quad \Sigma(Y_i - \alpha - \beta X_i)
\end{aligned} \tag{3}
$$

$$
\Sigma Y_i = n\alpha + \beta X_i (\#eq:reg1) \tag{4}
$$

$$\Sigma Y_i X_i = \alpha \sum X_i + \beta X_i^2 (\#eq:reg2) \tag{5}$$

By doing this operation: @ref(eq:reg2) $\times$ n - @ref(eq:reg1) $\times \Sigma X_i \Rightarrow$

$$\hat{\beta} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2} (\#eq:betahat) \tag{6}$$

Properties of Residual

$\Sigma e_i = 0$

$$\begin{aligned} \Sigma e_i &= \Sigma(Y_i - \hat{Y}_i) \\ &= \Sigma(Y_i - (\hat{\alpha} + \hat{\beta} X_i)) \\ &= \Sigma(Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta} X_i \quad (\#eq:res-zero) \\ &= \Sigma(Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X})) \\ &= 0 \end{aligned} \tag{7}$$

Total Sum of Squares

$$\begin{aligned} SST &= \Sigma(Y_i - \bar{Y})^2 \\ &= \Sigma[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})] \\ &= \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2 + 2\Sigma(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= SSE + SSR \end{aligned} (\#eq:sst) \tag{8}$$

Econometrics

Model Misspecification

Omission of independent variable Assume the model is $Y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon$

Estimated model: $Y_i = \beta_2^* x_{2i} + +\epsilon^*$

Now, $\beta_2^* = \frac{\Sigma x_{2i} y_i}{\Sigma x_{2i}^2}$

Finally, $E(\beta_2^*) = \beta_2 + \beta_3 \frac{Cov(x_{2i}, x_{3i})}{V(x_{2i})}$

Thus, β_2^* is biased and inconsistent.

Some other consequences

- $V(\epsilon_i)$ would be incorrectly estimated.
- $V(\hat{\beta}_2^*)$ would be biased
- CI and hypothesis testing will give misleading conclusion

Inclusion of extra variable

Epidemiology

Odds Ratio (OR)

$OR = \frac{P(D|E)}{P(\bar{D}|E)} / \frac{P(D|\bar{E})}{P(\bar{D}|\bar{E})}$

Can be calculated for

- Population-based study
- Cohort study

Cannot be calculated for case-control study

Validity and Precision

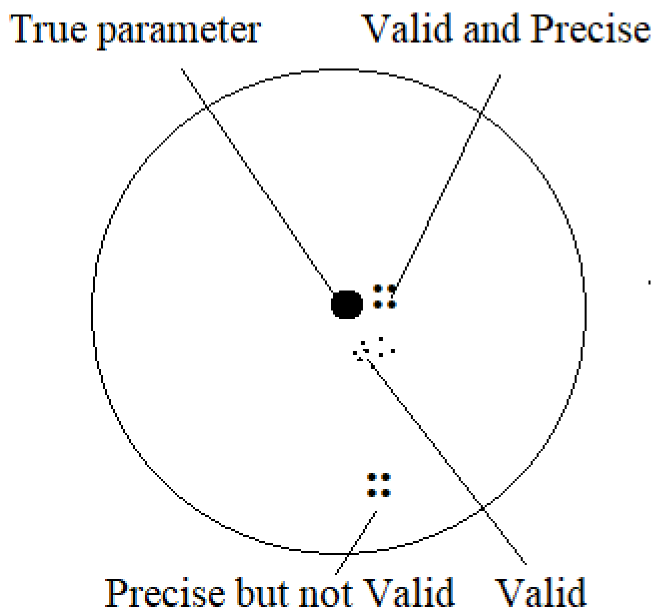

Figure 2: Validity and Precision