L.6: Simple Regression

Econometrics 1: ver. 2024 Fall Semester

Naoki Awaya

Review

Review

- Outcome variable of interest dependent variable
- Variables that determine the dependent variable explanatory variables
- Let Y denote a dependent variable and $\mathbf{X} = (X_1, \dots, X_k)$ denote a set of explanatory variables.
- The purpose of regression analysis is

> to estimate a function $g(\cdot)$ of $\mathbf{X}$ that predicts the value of Y.

The function

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

is called the regression function.

Review

Simple linear regression model

- Linear regression model with a single explanatory variable:

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

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

Multiple linear regression model

- Linear regression model with multiple explanatory variables:

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

- β_0: intercept, and $(\beta_1, \ldots, \beta_k)$: coefficients.

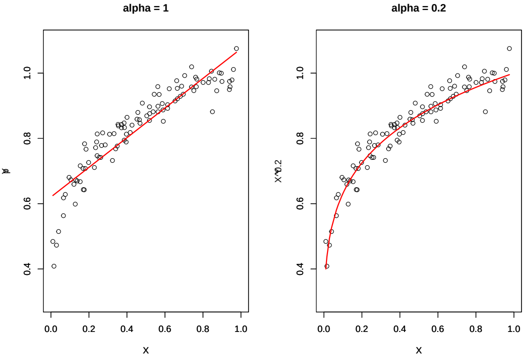
Review

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

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

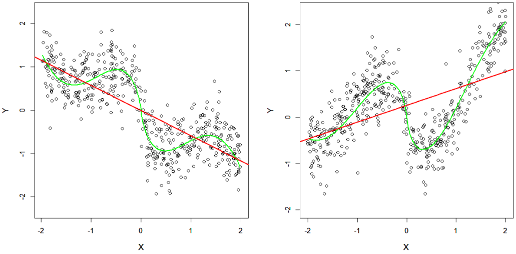
- One may consider a more general "nonlinear" regression function: e.g.,

$$g(\mathbf{X}) = (\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k)^{\alpha}.$$



Review

- In terms of MSE minimization, the best regression function is $\mathbb{E}[Y|\mathbf{X}]$; i.e., $\mathbb{E}[Y|\mathbf{X}] = \arg\min_g \mathbb{E}[(Y - g(\mathbf{X}))^2]$
- Even when $\mathbb{E}[Y|\mathbf{X}]$ is not actually a linear function of $\mathbf{X}$, a linear regression model can still give a <u>linear approximation</u> of $\mathbb{E}[Y|\mathbf{X}]$.



- Nonlinearity can be accommodated by adding polynomials of $\mathbf{X}$ as regressors.

Estimation of Simple Regression Models

OLS Estimator

- Suppose that we have data of n observations $\{(Y_i, X_i) : 1 \leq i \leq n\}$
- The data are assumed to be IID (independent and identically distributed).
  - E.g., data are obtained by random sampling from the same population.
- We consider a simple linear regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$

- For a given candidate (β_0, β_1), the prediction error is defined as

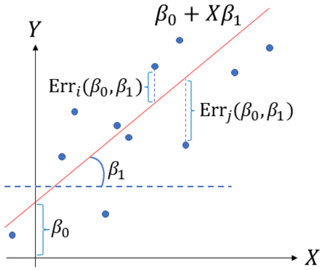
$$\mathsf{Err}_i(\beta_0, \beta_1) = Y_i - \beta_0 - X_i\beta_1.$$

OLS Estimator

Least squares method: estimate the parameter by minimizing the average squared errors (= empirical MSE).

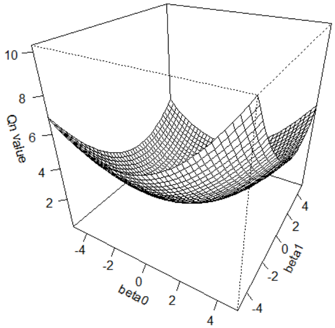
$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \arg \min_{(\beta_0, \beta_1)} Q_n(\beta_0, \beta_1)$$

where $$Q_n(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{Err}_i(\beta_0, \beta_1))^2$$



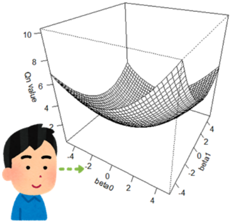
OLS Estimator

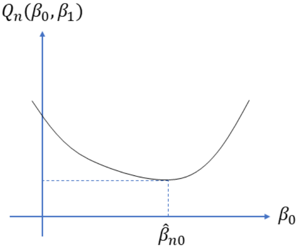
[Shape of the objective function $Q_n(\beta_0, \beta_1)$]

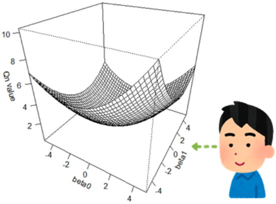


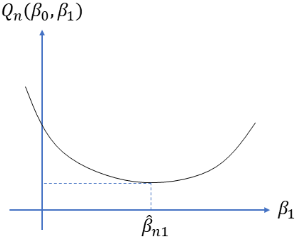
- The figure shows that the minimizer of $Q_n(\beta_0, \beta_1)$ can be uniquely found.

OLS Estimator









OLS Estimator

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

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

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

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

OLS Estimator

- First, solving (1) with respect to $\hat{\beta}_{n0}$ yields

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

- (2) can be re-written as

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

- Combining these two yields

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

OLS Estimator

Hence, we obtain

$$\text{Least squares estimator } \hat{\beta}_{n1} = \frac{\text{sample cov}(Y, X)}{\text{sample var}(X)}$$

- 3 numerically equivalent expressions:

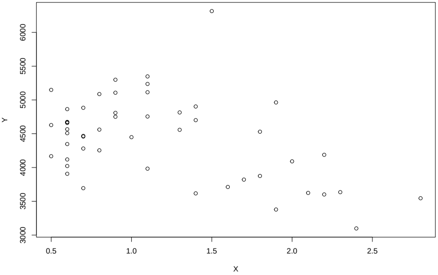
$$\hat{\beta}_{n1} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i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

- The estimator $\hat{\beta}_{n1}$ obtained by this formula is called the OLS (Ordinary Least Squares) slope estimator.
- The OLS intercept estimator is immediate from $\hat{\beta}_{n0} = \overline{Y}_n - \overline{X}_n \hat{\beta}_{n1}$.

OLS Estimator

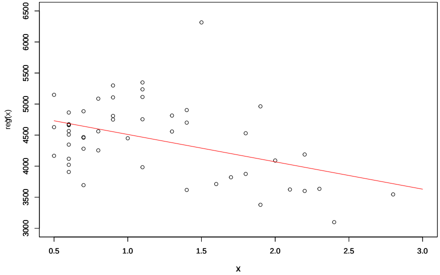
- `state.x77`: Income $= \beta_0 +$ Illiteracy$\beta_1 + \epsilon$

```
> data <- as.data.frame(state.x77)
> X <- data$Illiteracy
> Y <- data$Income
> plot(X, Y)
```



OLS Estimator

```
> beta_1 <- cov(X, Y)/var(X)
> beta_0 <- mean(Y) - mean(X)*beta_1
> reg <- function(x) beta_0 + beta_1*x
> plot(X, Y, xlim = c(0.5,3), ylim = c(3000, 6500))
> par(new = T)
> curve(reg(x), col = "red", xlim = c(0.5,3), ylim = c(3000, 6500))
```



OLS Estimator

```
> c(beta_0, beta_1)
```

```
## [1] 4951.3198 -440.6152
```

```
> lm(Y ~ X)$coef
```

```
## (Intercept)           X
##   4951.3198   -440.6152
```

Evaluation of Estimators

Evaluation of Estimators

- For the estimation of linear regression models, not only OLS but many different methods are available.
- OLS:
$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \arg \min_{(\beta_0, \beta_1)} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - X_i \beta_1)^2$$
- Weighted Least Squares (WLS):
$$(\hat{\beta}_{n0}, \hat{\beta}_{n1}) = \arg \min_{(\beta_0, \beta_1)} \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - \beta_0 - X_i \beta_1)^2$$
- Maximum Likelihood, (Generalized) Method of Moments, Least Absolute Deviations, etc.

Evaluation of Estimators

- Formally, an estimator is a "procedure" to estimate the parameter of interest; namely, the estimator is a function of the data $\{(Y_i, X_i) : 1 \leq i \leq n\}$. Thus, **the estimator is a random variable**.

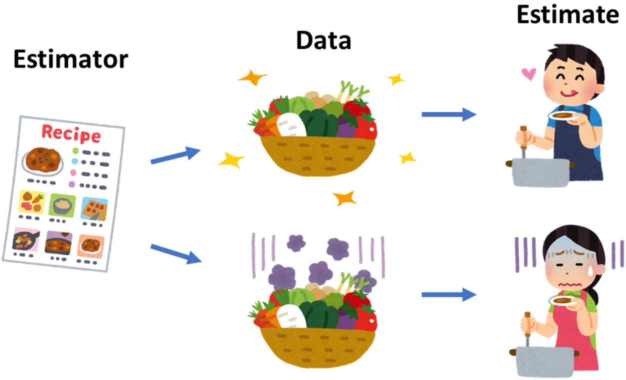
$$\text{Estimator} = T((Y_1, X_1), \dots, (Y_n, X_n))$$

- The value obtained by plugging the real data into $T(\cdot)$, is called the estimate.

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

Estimator = Estimation procedure (random variable)
Estimate = Estimation result (realized value)

Evaluation of Estimators



- The quality of the estimate is highly dependent on the quality of the data used.

Evaluation of Estimators

- Since the true parameter values are unknown, we cannot compare the precision of different estimates.
- The estimators are random variables. Thus, we can discuss which estimator is more preferable <u>probabilistically</u>, independently from the data.
- There are three major criteria used to evaluate the performance of estimators:
  - *Consistency*
  - *Unbiasedness*
  - *Efficiency*

Consistency

- Denote θ_0 as the parameter of interest to be estimated, and let $\widehat{\theta}_n$ be any estimator of θ_0.

Consistency

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

$$\widehat{\theta}_n \xrightarrow{p} \theta_0, \ (n \to \infty).$$

- For example, let $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\theta_0 = \mathbb{E}(X)$. Then, by the weak law of large numbers, $\widehat{\theta}_n$ is consistent for θ_0.
- Consistency is a minimum requirement for a good estimator.

Unbiasedness

Unbiasedness

The estimator $\widehat{\theta}_n$ is said to be an unbiased estimator of θ_0 if its expected value is θ_0:

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

- The gap between the expected value of the estimator and the true parameter $\mathbb{E}(\widehat{\theta}_n) - \theta_0$ is called bias.
  - For unbiased estimators, positive errors and negative errors balance out.
- If X_i's are drawn from the same population, the sample average is an unbiased estimator of the population mean:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathbb{E}(X)$$

Unbiasedness

- It is well known that the sample variance

$$\mathbb{V}_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

is NOT an unbiased estimator for the population variance $\mathbb{V}(X)$; see Appendix for the proof.

- An unbiased variance estimator can be obtained by

$$\mathbb{V}'_n(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Showing $\mathbb{E}[\mathbb{V}'_n(X)] = \mathbb{V}(X)$ is left as a good exercise for you.

Consistent but Biased

- There are many estimators that are consistent but are biased. For example, let
$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i + 1/n, \;\; \theta_0 = \mathbb{E}(X).$$
- This estimator is consistent
$$\hat{\theta}_n \xrightarrow{p} \mathbb{E}(X) + 0 = \theta_0$$
but is not unbiased
$$\mathbb{E}(\hat{\theta}_n) = \mathbb{E}(X) + 1/n \neq \theta_0$$
(asymptotically unbiased though).

Unbiased but Inconsistent

- Similarly, we can easily construct estimators that are unbiased but are inconsistent. For example, let

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

  Then, the estimator is clearly unbiased.
- However, noting that $\mathbb{V}(\hat{\theta}_n) = \mathbb{V}(X)/2$, by Chebyshev's inequality,

$$\begin{aligned} \Pr(|\hat{\theta}_n - \theta_0| \geq \epsilon) &\leq \frac{\mathbb{V}(\hat{\theta}_n)}{\epsilon^2} \\ &= \frac{\mathbb{V}(X)}{2\epsilon^2} \nrightarrow 0 \ \ (n \to \infty) \end{aligned}$$

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

Unbiased but Inconsistent

> Three statisticians go deer hunting with bows and arrows. They spot a big buck and take aim. One shoots and his arrow flies off three meters to the right. The second shoots and his arrow flies off three meters to the left. The third statistician jumps up and down yelling, "We got him! We got him!" (*Greg Hancock)

Lomax and Moosavi (2002) Using Humor to Teach Statistics: Must They Be Orthogonal?

Efficiency

- When multiple consistent and unbiased estimators are available, 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 that $\widehat{\theta}_{n,1}$ and $\widehat{\theta}_{n,2}$ are consistent and unbiased estimators of θ_0. We say that estimator $\widehat{\theta}_{n,1}$ is more efficient than $\widehat{\theta}_{n,2}$ if

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

- An efficient estimator is more precise in the sense that it can produce a narrower confidence interval.

Efficiency

- Let $\widehat{\theta}_{n,1}$ be the sample average of X and $\widehat{\theta}_{n,2}$ be the average of the observations with odd indices; i.e., assuming that n is even,

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

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

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

$\Longrightarrow \widehat{\theta}_{n,1}$ is more efficient than $\widehat{\theta}_{n,2}$.

Statistical Properties of OLS

Statistical Properties of the OLS Estimator

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

Assumption 1. Correct model specification:

$$\begin{aligned} Y &= \beta_0^* + X\beta_1^* + \epsilon \\ \mathbb{E}(Y|X) &= \beta_0^* + X\beta_1^* \end{aligned}$$

- Assumption 1 is generally a restrictive requirement.
- When this assumption is not satisfied, our regression result should be viewed as a linear approximation, rather than the true relationship of the variables.
- The assumption automatically implies that $\mathbb{E}(\epsilon|X) = 0$. Further, by LIE, $\mathbb{E}(\epsilon) = \mathbb{E}[\mathbb{E}(\epsilon|X)] = 0$.

Statistical Properties of the OLS Estimator

Assumption 2. The observations $\{(Y_i, X_i) : 1 \le i \le n\}$ are IID.

Assumption 3. The error term ϵ is independent of X, and its variance is given by $\mathbb{E}(\epsilon^2) = \sigma^2$.

- Assumption 2 is introduced in order to use LLN and CLT.
- Assumption 3 says that the error variance is constant over all values of X, which is called the homoskedasticity assumption.

Theorem

Under Assumptions 1–3, the OLS estimator $(\widehat{\beta}_{n0}, \widehat{\beta}_{n1})$ are unbiased and consistent for (β_0^*, β_1^*), respectively.

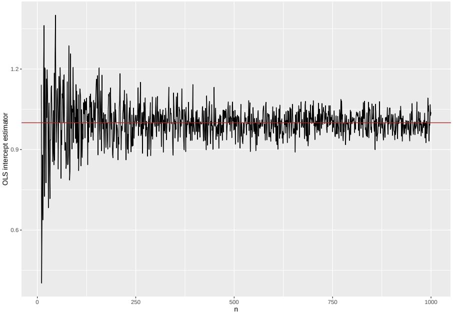
See Appendix for the proof

Numerical simulation

```
> beta0 <- 1
> beta1 <- 1
> OLS <- function(n){
+   X <- rnorm(n)
+   e <- rnorm(n)
+   Y <- beta0 + X*beta1 + e
+   bhat1 <- cov(X,Y)/var(X)
+   bhat0 <- mean(Y) - mean(X)*bhat1
+   c(bhat0, bhat1)
+   }
> N <- 10:1000 # sample size = 10, 11, 12, ...
> B0 <- numeric(length(N))
> B1 <- numeric(length(N))
> for(i in 1:length(N)){
+     est <- OLS(N[i])
+     B0[i] <- est[1]
+     B1[i] <- est[2]
+     }
```

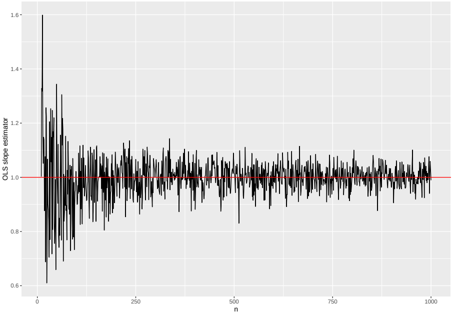
Numerical simulation

```
> library(tidyverse)
> data <- data.frame(n = N, beta_hat0 = B0)
> ggplot(data, aes(x = n, y = beta_hat0)) +
+     geom_line() + ylab("OLS intercept estimator") +
+     geom_abline(intercept = beta0, slope = 0, color = "red")
```



Numerical simulation

```
> data <- data.frame(n = N, beta_hat1 = B1)
> ggplot(data, aes(x = n, y = beta_hat1)) +
+     geom_line() + ylab("OLS slope estimator") +
+     geom_abline(intercept = beta1, slope = 0, color = "red")
```



Statistical Properties of the OLS Estimator

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

$$
\begin{aligned}
\mathbb{V}(\widehat{\beta}_{n1}) &= \mathbb{E}\left[(\widehat{\beta}_{n1} - \mathbb{E}(\widehat{\beta}_{n1}))^2\right] \\
&= \mathbb{E}\left[(\widehat{\beta}_{n1} - \beta_1^*)^2\right] \quad \text{(by the unbiasedness)} \\
&= \mathbb{E}\left\{\mathbb{E}\left[(\widehat{\beta}_{n1} - \beta_1^*)^2 \middle| X_1, \dots, X_n\right]\right\} \quad \text{(by LIE).}
\end{aligned}
$$

- Under Assumptions 1–3, we can show that

$$
\mathbb{E}\left[(\widehat{\beta}_{n1} - \beta_1^*)^2 \middle| X_1, \dots, X_n\right] = \frac{\sigma^2}{n} \frac{1}{\text{sample var}(X)}
$$

Statistical Properties of the OLS Estimator

- Thus,

$$\mathbb{V}(\hat{\beta}_{n1}) = \frac{\sigma^2}{n}\mathbb{E}\left(\frac{1}{\text{sample var}(X)}\right)$$

- The variance of the OLS estimator decreases as the sample size increases and as the variance of X increases:
  - The larger the sample size n, the smaller the estimation error.
  - The larger variation in the values of X, the easier the estimation of β_1^*.
- Equivalently, the above result can be written as

$$\mathbb{V}\left(\sqrt{n}(\hat{\beta}_{n1} - \beta_1^*)\right) = \sigma^2\mathbb{E}\left(\frac{1}{\text{sample var}(X)}\right)$$

Statistical Properties of the OLS Estimator

- Further, by taking the limit of $\mathbb{V}\left(\sqrt{n}(\widehat{\beta}_{n1} - \beta_1^*)\right)$ as $n \to \infty$,

$$\mathbb{V}_\infty\left(\sqrt{n}(\widehat{\beta}_{n1} - \beta_1^*)\right) = \sigma^2/\mathbb{V}(X)$$

  $\mathbb{V}_\infty(\cdot)$ is referred to as the asymptotic variance.

- Finally, by CLT, we have

Asymptotic distribution of $\widehat{\beta}_{n1}$

$$\sqrt{n}(\widehat{\beta}_{n1} - \beta_1^*) \xrightarrow{d} N\left(0, \sigma^2/\mathbb{V}(X)\right)$$

- That is, the distribution of $\sqrt{n}(\widehat{\beta}_{n1} - \beta_1^*)$ can be approximated by the normal distribution $N\left(0, \sigma^2/\mathbb{V}(X)\right)$ when n is sufficiently large.

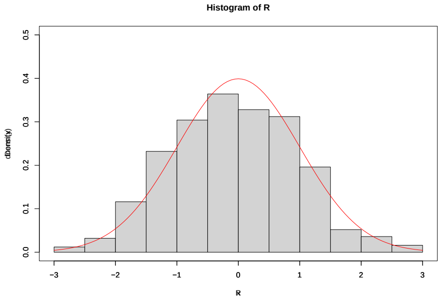
Numerical simulation

```
> beta0 <- 1
> beta1 <- 1
> n <- 1000
> r <- function(){
+   X <- rnorm(n)
+   e <- rnorm(n)
+   Y <- beta0 + X*beta1 + e
+   bhat1 <- cov(X,Y)/var(X)
+   sqrt(n)*(bhat1 - beta1)
+   }
```

- The asymptotic distribution of $\sqrt{n}(\widehat{\beta}_{n1} - \beta_1^*)$ is $N(0,1)$ ($\mathbb{V}(X) = \sigma^2 = 1$).

Numerical simulation

```
> R <- numeric(500)
> for(i in 1:500) R[i] <- r()
>
> hist(R, freq = FALSE, xlim = c(-3,3), ylim = c(0,0.5))
> par(new = T)
> # sigma^2/V(X) = 1 => the asymp dist = N(0,1)
> curve(dnorm(x), xlim = c(-3,3), ylim = c(0,0.5), col = "red")
```



Summary

- OLS estimator for a simple linear regression model:

$$\hat{\beta}_{n0} = \overline{Y}_n - \overline{X}_n \hat{\beta}_{n1}, \quad \hat{\beta}_{n1} = \frac{\text{sample cov}(Y, X)}{\text{sample var}(X)}$$

- Three criteria to evaluate the performance of estimators:
    - Consistency: $\hat{\theta}_n \xrightarrow{p} \theta_0$
    - Unbiasedness: $\mathbb{E}(\hat{\theta}_n) = \theta_0$
    - Efficiency: the variance of $\hat{\theta}_n$ is relatively small
- The OLS estimator is consistent and unbiased under certain conditions.
- The distribution of the OLS estimator can be approximated by a normal distribution in the following sense:

$$\sqrt{n}(\hat{\beta}_{n1} - \beta_1^*) \xrightarrow{d} N\left(0, \sigma^2/\mathbb{V}(X)\right)$$

Technical appendix for math lovers

* This part is optional.

Appendix

Proof of the "biasedness" of $\mathbb{V}_n(X)$

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

Noting that $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X)$ and $\mathbb{V}(\overline{X}_n) = \mathbb{V}(X)/n$,

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

From the third equality, we can see that although $\mathbb{V}_n(X)$ is biased for $\mathbb{V}(X)$ for finite n, the bias vanishes as $n \to \infty$. ■

Appendix

Unbiasedness of $\hat{\beta}_{n1}$

First, note that

$$\begin{aligned}\mathbb{E}(Y_i|X_1,\dots,X_n) &= \mathbb{E}(Y_i|X_i)\\ &= \beta_0^* + X_i\beta_1^*\end{aligned}$$

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

Recall that the OLS slope estimator is

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

Appendix

Then, taking the conditional expectation of $\widehat{\beta}_{n1}$ given $(X_1, \dots, X_n)$ gives

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_{n1}|X_1, \dots, X_n) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\mathbb{E}(Y_i|X_1, \dots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(\beta_0^* + X_i\beta_1^*)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\beta_0^*}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i} + \beta_1^*
\end{aligned}$$

Noting that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\beta_0^* = (\overline{X}_n - \overline{X}_n)\beta_0^* = 0,$$

we have $\mathbb{E}(\widehat{\beta}_{n1}|X_1, \dots, X_n) = \beta_1^*$. Finally, the result follows from LIE. ■

Appendix

Unbiasedness of $\hat{\beta}_{n0}$

Recall that $\hat{\beta}_{n0} = \overline{Y}_n - \overline{X}_n \hat{\beta}_{n1}$. The conditional expectation of $\hat{\beta}_{n0}$ given $(X_1, \dots, X_n)$ is

$$\begin{aligned}\mathbb{E}(\hat{\beta}_{n0} | X_1, \dots, X_n) &= \mathbb{E}(\overline{Y}_n | X_1, \dots, X_n) - \overline{X}_n \mathbb{E}(\hat{\beta}_{n1} | X_1, \dots, X_n) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i | X_i) - \overline{X}_n \beta_1^* \\ &= \beta_0^* + \overline{X}_n \beta_1^* - \overline{X}_n \beta_1^* = \beta_0^*\end{aligned}$$

Then, the result follows from LIE. ■

Appendix

Consistency of $\hat{\beta}_{n1}$

Recall that the OLS slope estimator is

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

Since $Y_i = \beta_0^* + X_i\beta_1^* + \epsilon_i$, we have

$$\begin{aligned}\hat{\beta}_{n1} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(\beta_0^* + X_i\beta_1^* + \epsilon_i)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i} \\ &= \beta_1^* + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i}.\end{aligned}$$

Thus, in order to show the consistency, it suffices to show that the second term on the right-hand side converges to zero in probability.

Appendix

Note that the numerator and denominator of the second term converge to $\mathbb{C}(X, \epsilon)$ and $\mathbb{V}(X)$, respectively:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i = \text{sample cov}(X, \epsilon) \xrightarrow{p} \mathbb{C}(X, \epsilon) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i = \text{sample var}(X) \xrightarrow{p} \mathbb{V}(X).$$

Then, we have

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i} \xrightarrow{p} \frac{\mathbb{C}(X, \epsilon)}{\mathbb{V}(X)} = 0.$$

This completes the proof. ■

Appendix

Consistency of $\hat{\beta}_{n0}$

Since $\overline{Y}_n = \beta_0^* + \overline{X}_n \beta_1^* + \frac{1}{n}\sum_{i=1}^n \epsilon_i$,

$$\begin{aligned}\hat{\beta}_{n0} &= \overline{Y}_n - \overline{X}_n \hat{\beta}_{n1} \\ &= \beta_0^* + \overline{X}_n(\beta_1^* - \hat{\beta}_{n1}) + \frac{1}{n}\sum_{i=1}^n \epsilon_i.\end{aligned}$$

By the consistency of $\hat{\beta}_{n1}$, $\overline{X}_n(\beta_1^* - \hat{\beta}_{n1}) \xrightarrow{p} 0$. Further, by LLN, $\frac{1}{n}\sum_{i=1}^n \epsilon_i \xrightarrow{p} 0$. Hence, $\hat{\beta}_{n0} \xrightarrow{p} \beta_0^*$. ■