Simple Linear Regression

Frederick De Baene

2024-11-19

Contents

1 Introduction

The simple linear regression model is represented as follows:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The following restrictions are enforced:

- $E[\epsilon_i] = 0$
- $Var(\epsilon_i) = \sigma^2$
- ϵ_i is normally distributed
- $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

The responses are drawn from conditional probability distributions. These conditional probability distributions condition on the level of the predictor variable. The mean of the conditional probability distributions of the response depends on the level of the predictor. The regression function relates the expected value for the response to the predictor value:

$$E[Y_i] = E[\beta_0 + \beta_1 x_i + \epsilon_i] = E[\beta_0] + E[\beta_1 x_i] + E[\epsilon_i] = \beta_0 + \beta_1 x_i.$$

Remember that $E[\epsilon_i] = 0$.

2 Simulations

We assume the following simple linear regression model for the population:

$$Y = 10 + 2X + \epsilon_i.$$

In this example, we consider a sample comprising 10 observations. For each observation, the level for X is fixed. We repeatedly sample an outcome for each observation from their respective conditional probability distributions for the outcome.

```r
# Initialize the population parameters and the simulation size
beta_0 <- 10
beta_1 <- 2
err_var <- 4
sim_size <- 10000
n <- sim_size * 10

# Initialize a vector with varying levels for the predictor
x <- seq(from = 0, to = 90, by = 10)

# Initialize a matrix to hold the outcomes for the simulations
y <- matrix(
  data = rep(NA, times = length(x) * sim_size),
  nrow = length(x), ncol = sim_size
)
```

```r
# During each simulation, for each level of the predictor, sample from the
# conditional probability distribution of the response
for (i in 1:sim_size) y[, i] <- beta_0 + beta_1 * x + rnorm(length(x), sd = sqrt(err_var))
```

2.1 Assumptions and Restrictions

2.1.1 Independent Errors

The assumption of independent and identically distributed errors is stated as follows:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

and

$$Cov(\epsilon_i, \epsilon_j) = 0 with i \neq j.$$

Note that σ^2 is the variance of the conditional probability distributions of Y across all levels of X. Later on, we will see that one of the assumptions of linear regression states that the variance for all conditional probability distributions of Y is equal. Therefore, we do not have to write σ_i^2 but we can just write σ^2.

Simulation setup. In the simulation setup, we make use of a sample comprising 10 observations. For each observation, we keep the level of X fixed. Suppose we zoom in on observations $i = 3$ and $i = 5$. Their predictor values are:

```r
x[3]
```

```
## [1] 20
```

```r
x[5]
```

```
## [1] 40
```

Given that the population parameters are known, we can determine the mean of the conditional probability distributions for $i = 3$ and $i = 5$ given X:

$$\begin{aligned} E(Y_i|x_i) &= \beta_0 + \beta_1 \times x_i \\ &= 10 + 2 \times x_i \end{aligned}$$

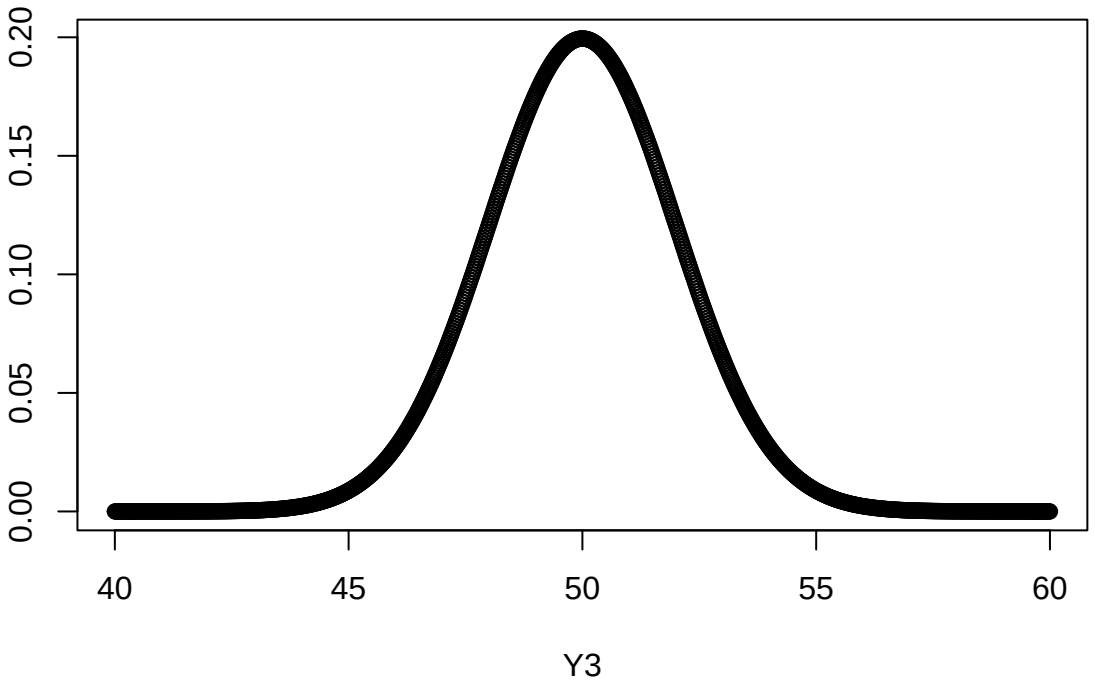
This gives us the following:

$$E(Y_3|x_3) = E(Y_3|20) = 10 + 2 \times 20$$
$$= 50$$

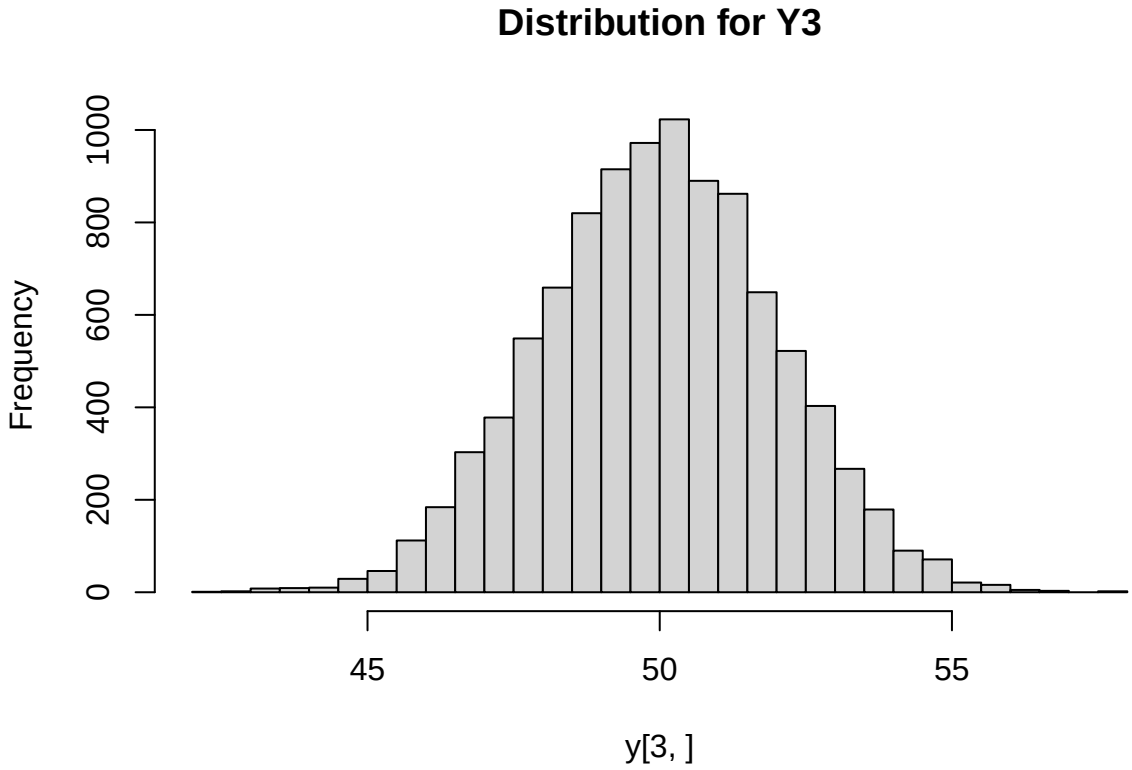
and

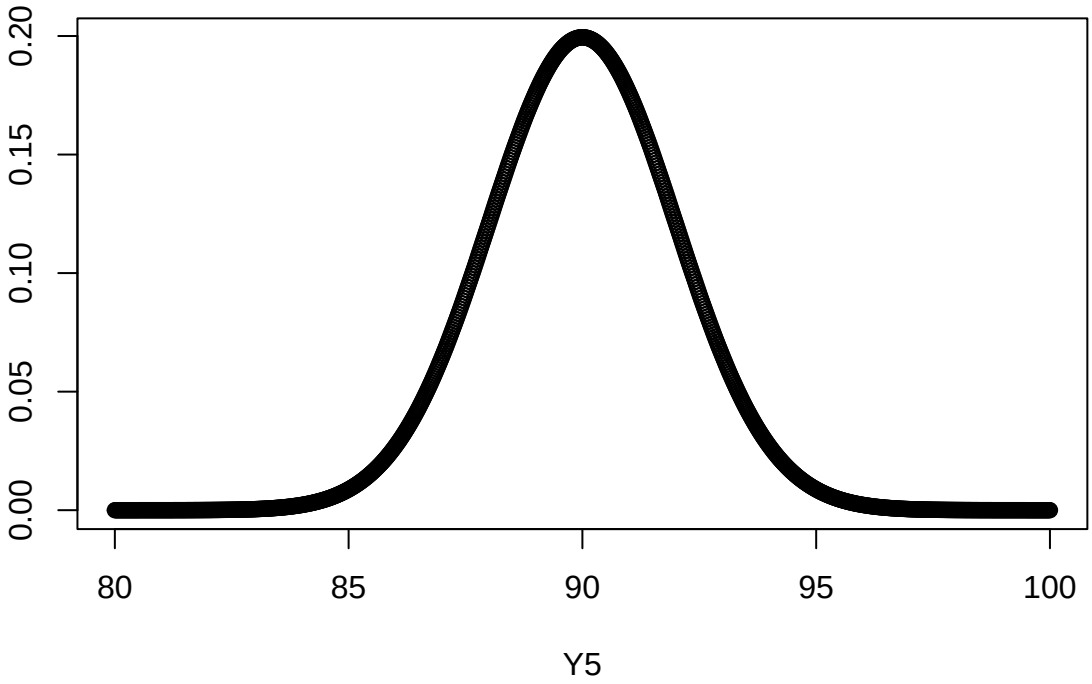
$$E(Y_5|x_5) = E(Y_5|40) = 10 + 2 \times 40$$
$$= 90$$

Given the means of the conditional probability distributions of Y for $i = 3$ and $i = 5$, we can draw the curves representing the conditional probability distributions for these two observations. From the population model we know that $\sigma^2 = 4$.

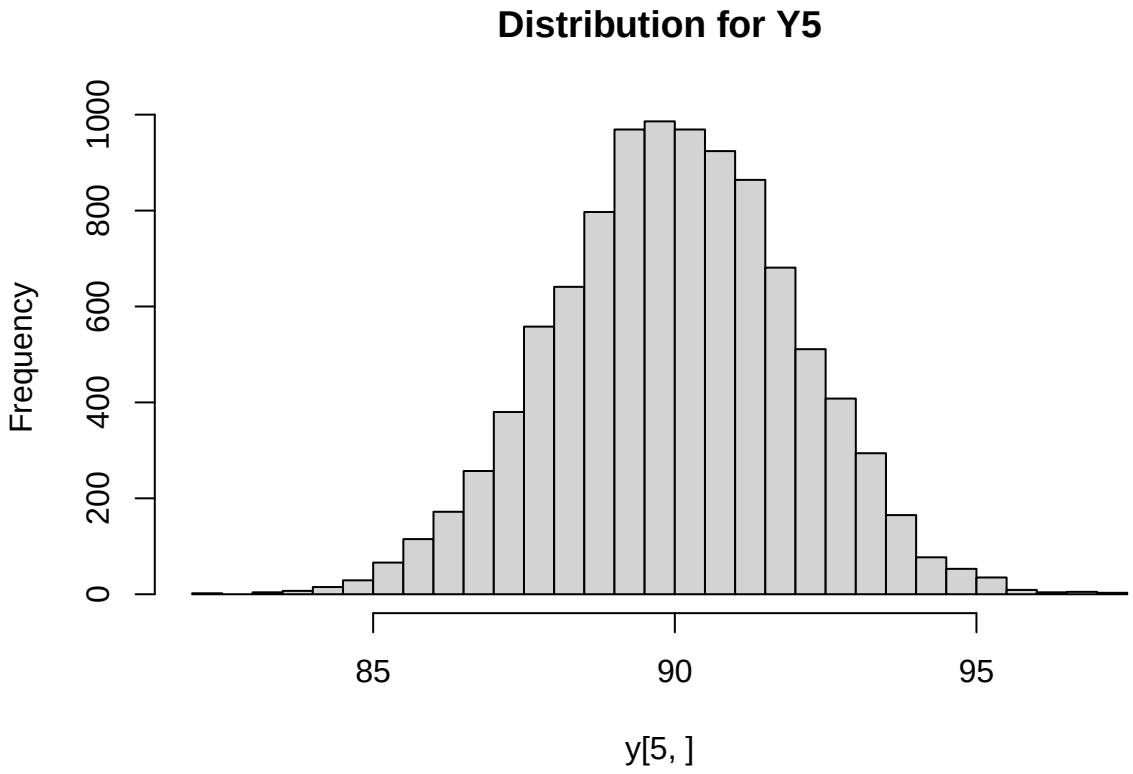


Looking at the actual outcomes, we observe the following distribution for Y_3:

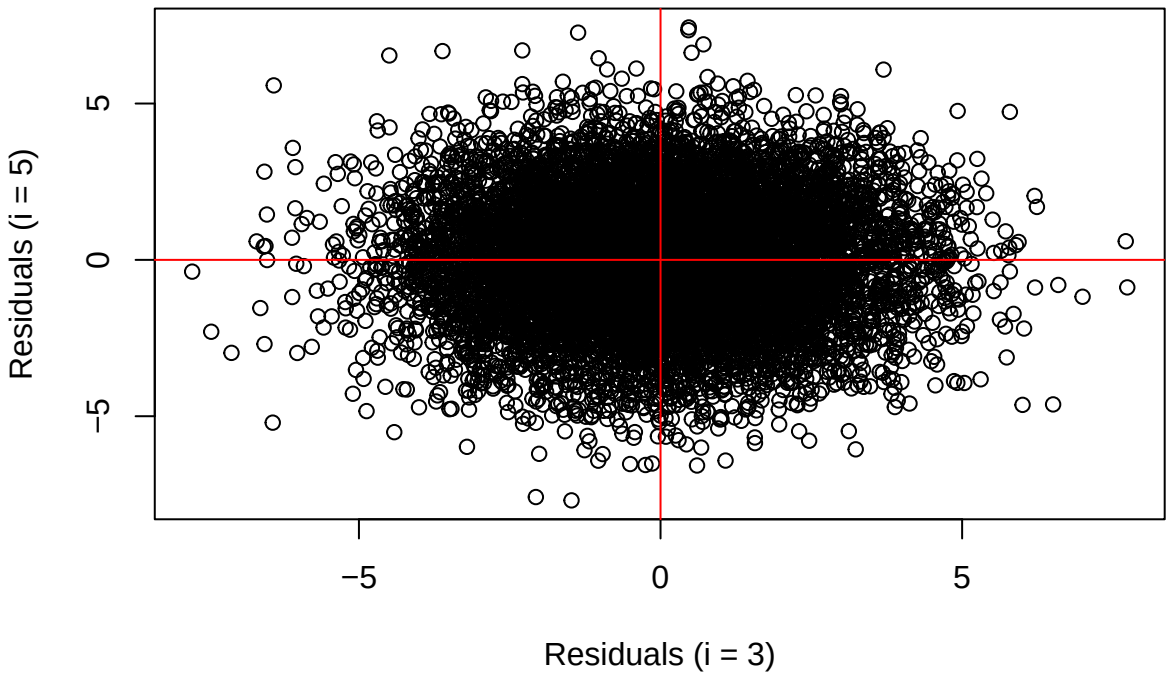




Looking at the actual outcomes, we observe the following distribution for Y_5:



The error terms are independent. This implies that the error terms for $i = 3$ are not correlated with the error terms for $i = 5$. The following scatter plot indicates that the error terms are independent:



The covariance $Cov(\epsilon_3, \epsilon_5) = -0.007$ and the scatter plot indicate no correlation between the error terms for $i = 3$ and $i = 5$. Also note the horizontal and vertical red lines, which seem to indicate that the mean for ϵ_3 and ϵ_5 is 0, in accordance with the assumption $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

A variance-covariance matrix is a square matrix. The diagonal elements equal the variance of the error terms σ^2, while the off-diagonal elements equal the covariance $Cov(\epsilon_i, \epsilon_j)$ with $i \neq j$.

```r
vars <- rep(NA, times = nrow(y))
for (i in 1:nrow(y)) vars[i] <- var(y[i, ])
```

2.2 Variance

The variance σ^2 of the conditional probability distributions of Y is the same for each level of X. A point estimator for σ^2 is the mean squared error (MSE). We must take into account that the deviations for each observation are the deviations between the observations and the mean of their respective conditional probability distribution.

For $i = 3$, the mean of the conditional probability distribution is 50. Therefore, to calculate the deviance for $i = 3$, we do:

$$Y_3 - E(Y_3|20) = Y_3 - 50$$

For $i = 5$, we have:

$$Y_5 - E(Y_5|40) = Y_5 - 90$$

To calculate the sum of squared errors taking into account the means of the conditional probability distributions, we utilize the residuals (or errors):

$$E_i = Y_i - E(Y_i|X_i) = Y_i - \hat{Y}_i$$

The sum of squared errors is calculated as follows:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

To obtain the variance, we divide the sum of squares by its associated number of degrees of freedom. For a simple linear regression model, this is $n - 2$, because we need two parameters β_0 and β_1 to calculate (or estimate) the means of the conditional probability distributions of Y. This gives us:

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}$$

The mean squared error is a point estimator for σ^2:

$$s^2 = MSE$$

We know that $\sigma^2 = 4$. Now let's verify that the mean squared error is an appropriate point estimator for σ^2:

```r
# First, calculate the mean of the conditional probability distributions of Y
estimated_means <- beta_0 + beta_1 * x

# Second, calculate the residuals
errors <- sweep(y, MARGIN = 1, STATS = estimated_means, FUN = "-")

# Third, sum the squared errors and divide by the appropriate degress of freedom
mse <- sum(errors^2) / (n - 2)
```

We see that our estimate $s^2 = 4$ approximates $\sigma^2 = 4$.

3 Inference

Throughout this chapter, we assume the normal error regression model holds:

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

where β_0 and β_1 are unknown population parameters, x_i are known constants, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

3.1 β_1

Drawing inferences on β_1 comprises either interval estimation or hypothesis testing of the form:

$$H_o : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

If $\beta_1 = 0$, there is no linear association between X and Y. The means of the conditional probability distributions of Y is the same for every level of X.

$$E[Y] = \beta_0 + 0 \times X = \beta_0$$

3.1.1 Sampling Distribution

The estimator $B1$ can be used to produce a point estimate for β_1. We will use the samples from our simulation to produce the point estimates for β_1 (and also for β_0).

```
beta_0_estimates <- rep(NA, times = sim_size)
beta_1_estimates <- rep(NA, times = sim_size)

for (i in 1:sim_size) {

  fit <- glm(y[, i] ~ x, family = gaussian)
  beta_0_estimates[i] <- fit$coefficients[1]
  beta_1_estimates[i] <- fit$coefficients[2]

}
```

The point estimates of β_1 can be used to create a sampling distribution of $B1$. The sampling distribution of $B1$ follows a normal distribution. Its mean equals the true value of the unknown population parameter β_1. The variance of the sampling distribution is defined as follows:

$$Var(B1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

We know how the variance of the estimator $B1$ is calculated. Given that we know the true variance of the errors $\sigma^2 = 4$, let's calculate the true variance of the sampling distribution:

$$
\begin{aligned}
Var(B1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
&= \frac{4}{8250} \\
&= 0.0004848485
\end{aligned}
$$

We obtain a variance for $B1$ of $Var(B1) = 0$, which gives us a standard error for $B1$ of $SE(B1) = 0.022$. Let us verify. Approximately 2.5% of the point estimates for β_1 must be less than $2 - 1.96 \times 0.022$ and 2.5% of the point estimates must be greater than $2 + 1.96 \times 0.022$

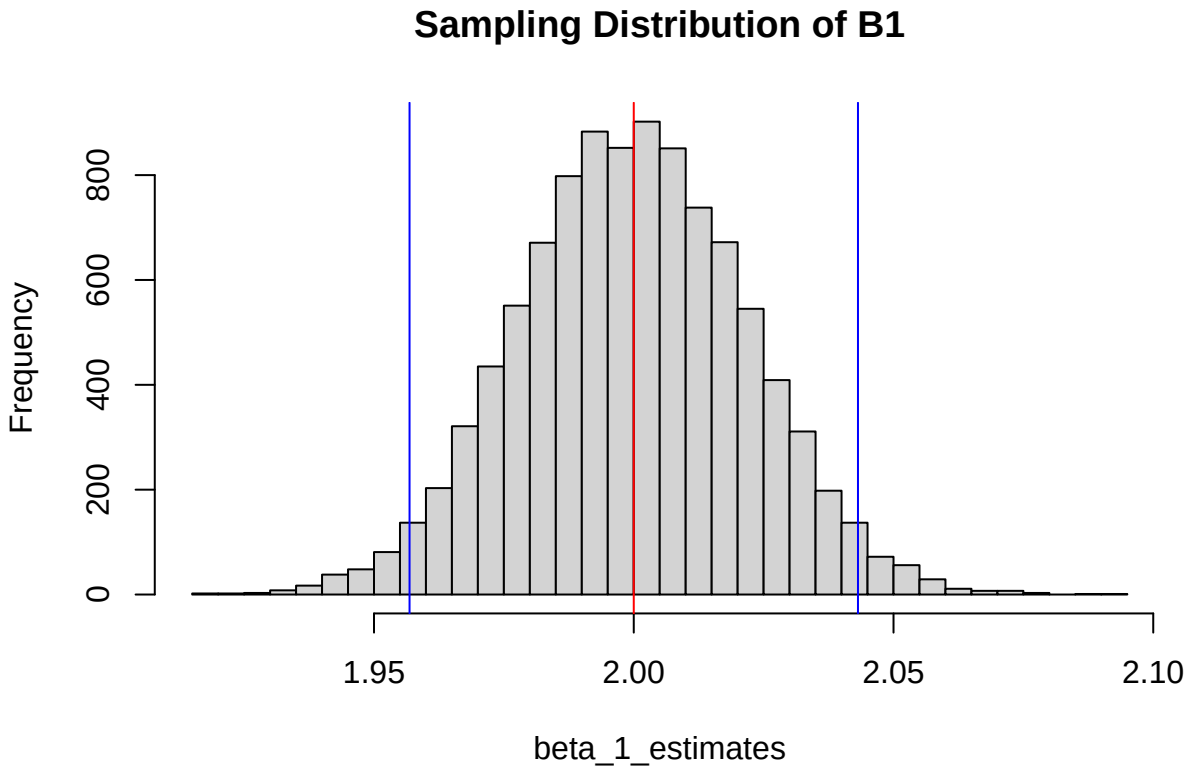
```
mean(beta_1_estimates < (beta_1 - 1.96*B1_se))
```

```
## [1] 0.0238
```

```
mean(beta_1_estimates > (beta_1 + 1.96*B1_se))
```

```
## [1] 0.0229
```

Let us have a look again at the sampling distribution of $B1$:



The normality of the sampling distribution of $B1$ follows from the fact that a linear combination of the outcomes Y_i, which are independent and normally distributed, is also normally distributed. Furthermore, because $B1$ is an unbiased estimator of β_1, this implies that $E[B1] = \beta_1$, and, thus, that the mean of the sampling distribution of $B1$ equals β_1.

Standardizing B1 gives us a standardized statistic. Because B1 follows a normal distribution, the standardization of B1 results in a standard normal distribution.

$$
\frac{B1 - \beta_1}{SE(B1)} \sim \mathcal{N}(0, 1)
$$