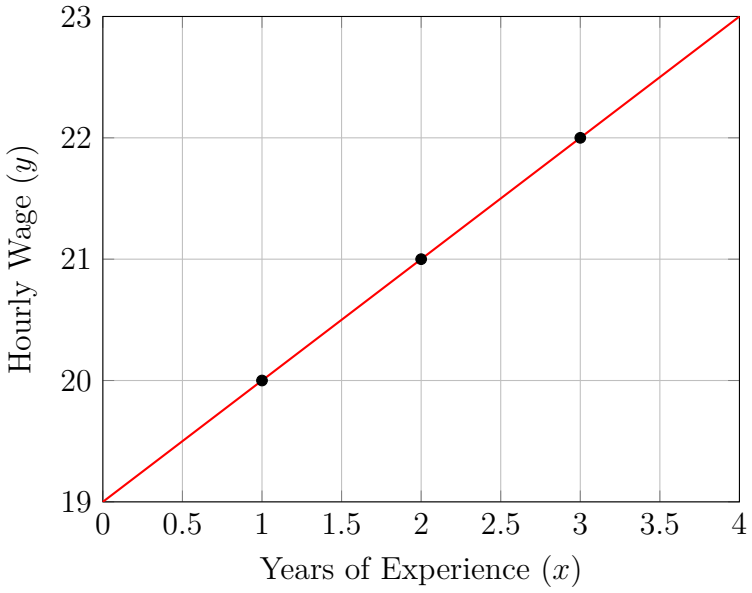
Class 03: Simple Regression Model

1 Introduction to Simple Regression

Simple Regression is one of the most fundamental techniques in econometrics and statistics, used to explore the relationship between two variables: one dependent variable (y) and one independent variable (x). The goal of simple regression is to understand how changes in the independent variable are associated with changes in the dependent variable. This relationship is typically modeled using a straight line, hence the term "linear regression."

Example: Consider a company that wants to determine how the number of years of experience of an employee affects their hourly wage. In this case, the hourly wage would be the dependent variable (y), as it depends on the years of experience, which is the independent variable (x).

Understanding this relationship allows the company to predict wages for employees based on their years of experience, which can be useful for setting salaries and making hiring decisions.



Explanation: The graph above shows a simple regression line that models the relationship between years of experience (independent variable x) and hourly wage (dependent variable y). The points on the graph represent actual data, and the red line is the regression line that best fits this data. The slope of the line indicates how much the wage is expected to increase with each additional year of experience.

2 Simple Regression Equation

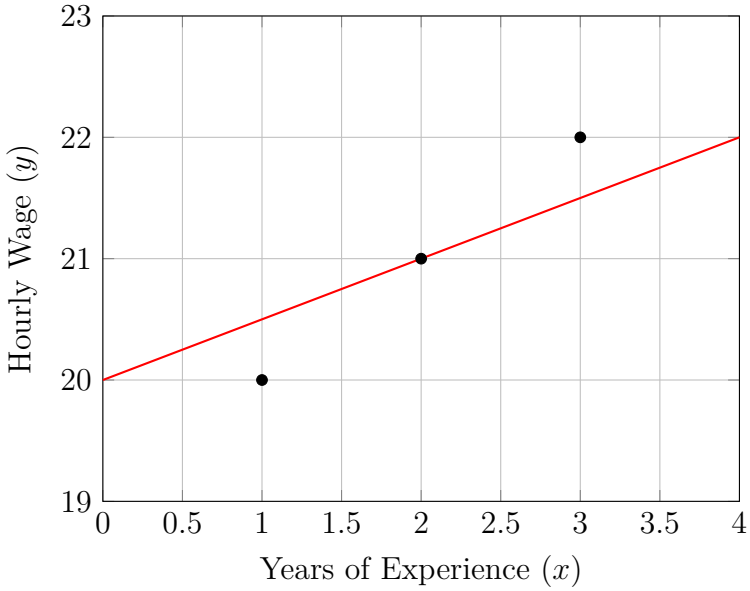
The relationship between the dependent and independent variables in simple regression is described by the following equation:

$$y = \beta_0 + \beta_1 x + u$$

Where:

- y is the **dependent variable**, which we are trying to predict or explain. In our example, this would be the hourly wage.
- x is the **independent variable**, which we use to make predictions about the dependent variable. In our example, this would be the years of experience.
- β_0 is the **intercept**, representing the expected value of y when $x = 0$. This is the point where the regression line crosses the y-axis.
- β_1 is the **slope**, which measures the change in y for each one-unit change in x. It indicates how strongly y is expected to increase or decrease as x increases.
- u is the **error term**, capturing all other factors that affect y but are not included in the model.

Interpretation: The slope (β_1) tells us how much the dependent variable y changes for a one-unit increase in the independent variable x. The intercept (β_0) gives us the starting point of y when x is zero.



Explanation: The graph visualizes the regression equation $y = 20 + 0.5x$. The intercept $\beta_0 = 20$ is where the line crosses the y-axis, indicating the expected wage when experience (x) is zero. The slope $\beta_1 = 0.5$ shows that for every additional year of experience, the wage increases by \$0.50.

3 Interpretation of Coefficients

Understanding the coefficients β_0 and β_1 is crucial for interpreting the results of a simple regression analysis:

- β_1 (Slope): This coefficient represents the average change in the dependent variable y for each one-unit increase in the independent variable x. A positive β_1 indicates that as x increases, y also increases. Conversely, a negative β_1 suggests that as x increases, y decreases.
- β_0 (Intercept): This is the expected value of y when x is zero. In some cases, the intercept might not have a meaningful interpretation (e.g., if $x = 0$ is not realistic within the context of the data).

Numerical Example: Predicting Wages Based on Experience

Let's consider a practical example. Suppose we have the following data on years of experience and hourly wage for three employees:

Years of Experience (x)	**Hourly Wage** (y)
1	20
2	21
3	22

We can use a simple regression model to estimate the relationship between years of experience and hourly wage. The estimated regression equation might look like this:

$$\text{Wage} = 20 + 0.5 \times \text{Experience}$$

Here:

- $\beta_0 = 20$ implies that an employee with zero years of experience is expected to earn a starting wage of \$20 per hour.
- $\beta_1 = 0.5$ suggests that for each additional year of experience, the hourly wage is expected to increase by \$0.50.

This equation allows us to predict the wage for any given level of experience. For instance, an employee with 4 years of experience is expected to earn:

$$\text{Wage} = 20 + 0.5 \times 4 = 20 + 2 = 22 \text{ dollars per hour.}$$

Explanation: In this graph, the red line represents the regression equation $y = 20 + 0.5x$, where $\beta_0 = 20$ is the intercept and $\beta_1 = 0.5$ is the slope. The points on the graph represent the actual data of wages and years of experience. The symbols $\hat{y}$ indicate the predicted wages at different levels of experience.

4 Residuals

After estimating the regression model, it's important to assess how well the model fits the data. One way to do this is by examining the **residuals**, which are the differences between the actual values of the dependent variable (y) and the values predicted by the model ($\hat{y}$):

$$\text{Residual} = y - \hat{y}$$

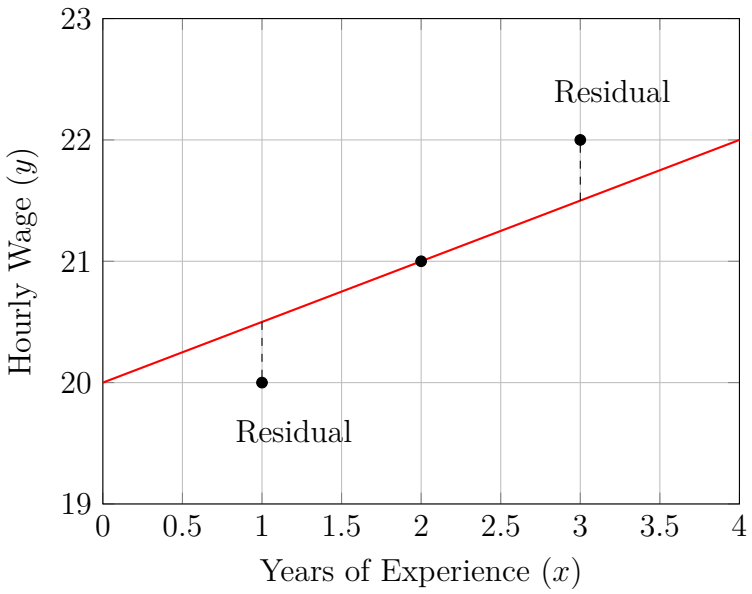
Residuals represent the portion of y that cannot be explained by the model. In other words, they reflect the "error" in the model's predictions.

Numerical Example: Residual Calculation

Let's continue with our previous example. Suppose the predicted wages based on the regression model are as follows:

Experience (x)	**Actual Wage** (y)	**Predicted Wage** ($\hat{y}$)	**Residual**
1	20	20.5	-0.5
2	21	21	0
3	22	21.5	0.5

Here, the residuals tell us how far off the model's predictions are from the actual wages. For the first employee, the model overestimates the wage by \$0.50 (since the actual wage is \$20, but the predicted wage is \$20.5). For the third employee, the model underestimates the wage by \$0.50.



Explanation: In this graph, the dashed lines represent the residuals, which are the differences between the actual data points and the predicted values on the regression line. The length of these dashed lines indicates the magnitude of the error or residual for each data point.

5 Goodness of Fit - R-Squared

The **R-Squared** (R^2) statistic is a key measure of how well the independent variable x explains the variation in the dependent variable y. It is defined as:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{SSE}}{\text{SST}}$$

Where:

- SSE (Sum of Squares for Error) is the variation in y that is explained by the regression line.
- SST (Total Sum of Squares) is the total variation in y around its mean.

An R^2 value close to 1 indicates that the regression model explains most of the variation in y, while a value close to 0 suggests that the model explains very little of the variation.

Numerical Example: Calculating SST, SSE, and R-Squared

Suppose we have the following data:

Experience (x)	**Actual Wage (y)**	**Predicted Wage ($\hat{y}$)**
1	20	20.5
2	21	21
3	22	21.5

First, calculate the mean of y:

$$\bar{y} = \frac{20 + 21 + 22}{3} = 21$$

Now, calculate the Total Sum of Squares (SST):

$$\text{SST} = (20 - 21)^2 + (21 - 21)^2 + (22 - 21)^2 = 1 + 0 + 1 = 2$$

Next, calculate the Sum of Squares for Error (SSE):

$$\text{SSE} = (20.5 - 21)^2 + (21 - 21)^2 + (21.5 - 21)^2 = 0.25 + 0 + 0.25 = 0.5$$

Finally, calculate R^2:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{0.5}{2} = 0.75$$

This means that 75% of the variation in hourly wage is explained by the years of experience, indicating that the model fits the data fairly well.

Explanation: This graph visually demonstrates the Total Sum of Squares (SST) and the Sum of Squares for Error (SSE). The SST measures the total variation of the actual data points around the mean of y, while the SSE measures the variation of the actual data points around the regression line. The closer these values are, the better the model fits the data.

6 Ordinary Least Squares (OLS)

The **Ordinary Least Squares (OLS)** method is the most common technique used to estimate the coefficients β_0 and β_1 in a simple regression model. OLS works by minimizing the sum of the squared residuals:

$$\text{Minimize } \sum(y_i - \hat{y}_i)^2$$

The goal of OLS is to find the line that best fits the data, where "best" means that the sum of squared differences between the observed values (y) and the predicted values ($\hat{y}$) is as small as possible.

Numerical Example: Applying OLS

Let's walk through an example of applying OLS to estimate β_0 and β_1.

Given the following data:

Experience (x)	**Hourly Wage** (y)
1	20
2	21
3	22

We want to fit a regression line $y = \beta_0 + \beta_1 x$. The OLS estimates of β_0 and β_1 are given by:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

First, calculate $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \frac{1+2+3}{3} = 2, \quad \bar{y} = 21$$

Next, calculate β_1:

$$\beta_1 = \frac{(1-2)(20-21) + (2-2)(21-21) + (3-2)(22-21)}{(1-2)^2 + (2-2)^2 + (3-2)^2} = \frac{1+0+1}{1+0+1} = 1$$

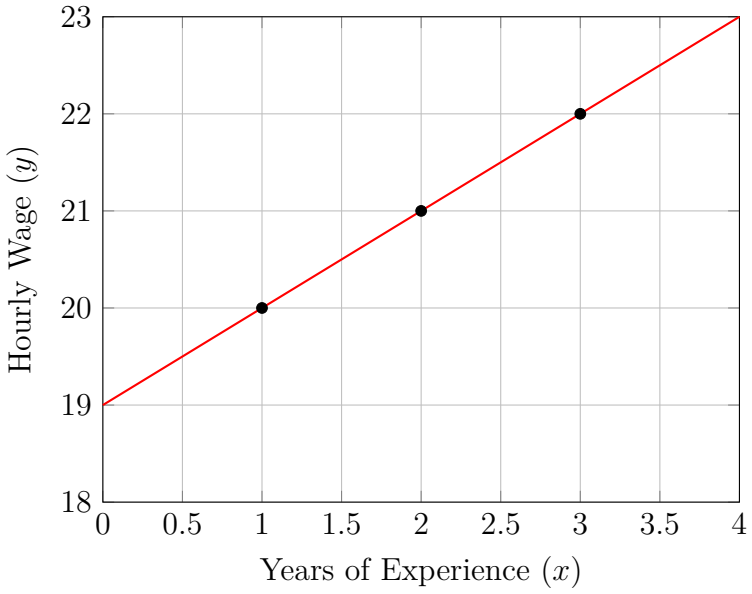
Finally, calculate β_0:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 21 - 1 \times 2 = 19$$

So, the estimated regression equation is:

$$\hat{y} = 19 + 1x$$

This equation suggests that for every additional year of experience, the hourly wage increases by \$1, and an employee with no experience is expected to earn \$19 per hour.



Explanation: The graph above shows the regression line $y = 19+1x$ estimated using the OLS method. The OLS method minimizes the sum of the squared residuals, resulting in a line that best fits the data points. Here, for every additional year of experience, the wage increases by \$1.

7 Assumptions of Simple Regression

For the OLS estimates to be reliable and valid, several key assumptions must hold:

1. **Linearity:** The relationship between the independent variable x and the dependent variable y is linear. This means that the change in y is proportional to the change in x.

2. **Independence:** The observations are independent of each other. In other words, the value of one observation does not influence or is not influenced by the value of another observation.

3. **Homoscedasticity:** The variance of the residuals (the errors) is constant across all levels of the independent variable x. This means that the spread of the residuals should be the same for all values of x.

4. **No Perfect Multicollinearity:** The independent variable must have some variation; it cannot be constant. In simple regression, this means that x must take on different values across observations.

Explanation: This graph shows the residuals from the OLS regression. The red line represents the zero residual line, and the points indicate the residuals for each data point. For OLS assumptions to hold, these residuals should have constant variance (homoscedasticity) and be randomly distributed around zero.