Introduction to Simple Linear Regression

1. Regression Equation

A **simple linear regression** (also known as a **bivariate regression**) is a linear equation describing the relationship between an **explanatory variable** and an **outcome variable**.

The procedure is used to measure how the **explanatory variable** (also known as the **independent variable**) possibly influences the **outcome variable** (also known as the **dependent variable**)

Example: Let y_i denote the *income* of some individual in your sample indexed by i where $i \in \{1, 2, .., n\}$, let x_i denote the number of *years of education* of the same individual, and let n denote the sample size. A simple linear regression equation of these two variables in the sample takes the form,

$$y_i = b_0 + b_1 x_i + e_i$$

where b_1 is the sample estimate of the slope of the regression line with respect to years of education and b_0 is the sample estimate for the vertical intercept of the regression line.

The term e_i is **residual**, that is the error term in regression. Since we would not expect education to *exactly* predict income, not all data points in a sample will line up exactly on the regression line. For some individual $i \in \{1, 2, .., n\}$ in the sample, e_i is the difference between his or her actual income and the predicted level of income that is on the regression line given his or her actual education attainment equal to x_i.

The point on the regression equation line is the **predicted value** for y_i given some value for x_i. The prediced value from an estimated regression is given by,

$$\hat{y}_i = b_0 + b_1 x_i.$$

Since some actual values for y_i will be above the regression line and some will be below, some e_i will be positive and others will be negative. The *best fitting regression line* is one such that the positive values exactly offset the negative values so that the mean of the residuals equals zero:

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

To minimize the error that the regression line makes, the coefficients for the best fitting regression line are chosen to minimize the sum of the squared residuals:

$$\begin{aligned}
\{b_0, b_1\} &= \min_{b_0, b_1} \sum_{i=1}^{n} e_i^2 \\
&= \min_{b_0, b_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2
\end{aligned}$$

This standard method for estimating regression coefficients by minimizing the sum of squared residuals is called the **ordinary least squares (OLS)** method.

Since b_1 is the slope, it measures how much the y-variable changes when the x-variable increases by one unit. In this case, b_1 is the estimate for on average how much additional income one earns for each additional year of education.

Depending on the application, the vertical intercept sometimes has an intuitive meaning. It measures what value to be expected for the y-variable when the x-variable is equal to zero. In this case, b_0 is the measure for the average income to be expected for individuals with zero years of education. If your data does not include any observation with a zero value for education, or if a zero value for the x-variable is unrealistic, then this coefficient has little mearning.

The regression line in the equation above is a sample estimate of the population regression line. The population regression line is the best fitting line for all possible elements in the population and whose coefficients are generally unknown. The population regression equation is given by,

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

where b_0 above is a sample estimate of the population coefficient β_0 and b_1 above is a sample estimate of the population coefficient β_1

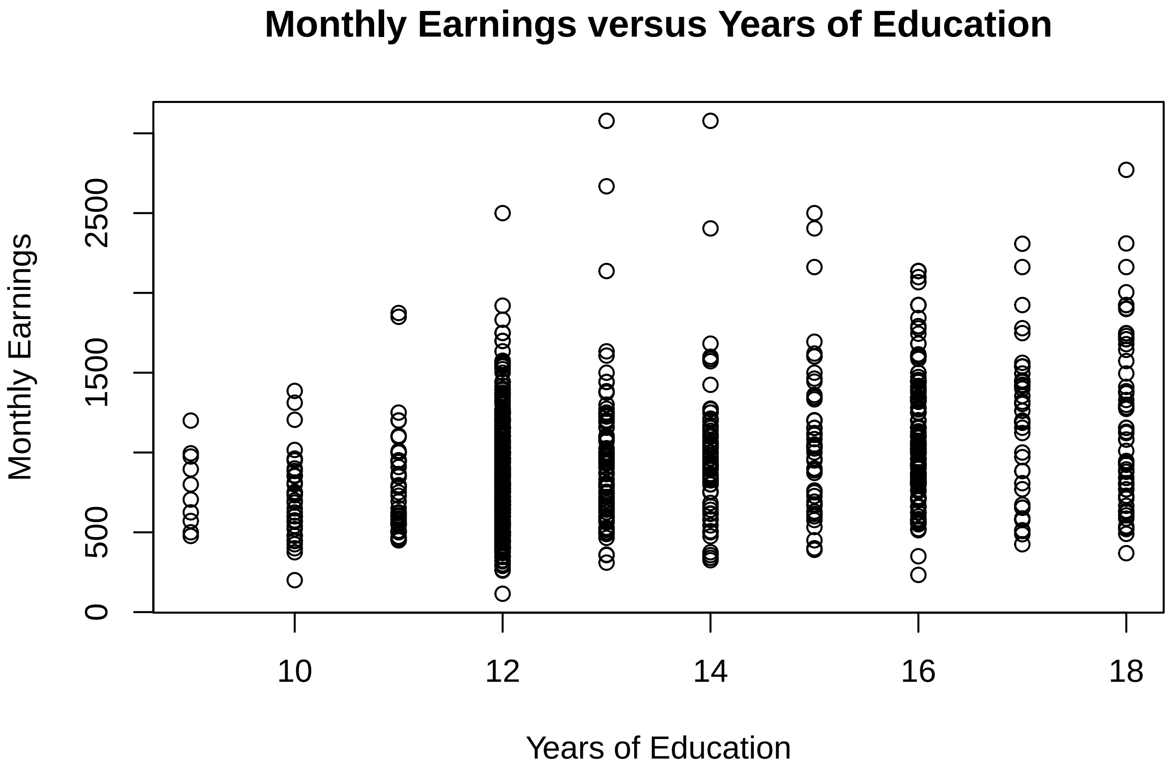
2. Download and Visualize the Dataset

The code below downloads a CSV file that includes data for 935 individuals on varaibles including their total monthly earnings (`MonthlyEarnings`) and a number of variables that could influence income, including years of education (`YearsEdu`).

```r
download.file(
  url="http://murraylax.org/datasets/wage2.csv",
  dest="wage2.csv")
wages <- read.csv("wage2.csv");
```

Let us begin by plotting the data to visually examine the relationship between years of schooling and monthly earnings. The code below produces a scatterplot with `YearsEdu` on the horizontal axis and `MonthlyEarnings` on the vertical axes.

```r
plot(x=wages$YearsEdu, y=wages$MonthlyEarnings,
     main="Monthly Earnings versus Years of Education",
     xlab="Years of Education",
     ylab="Monthly Earnings")
```



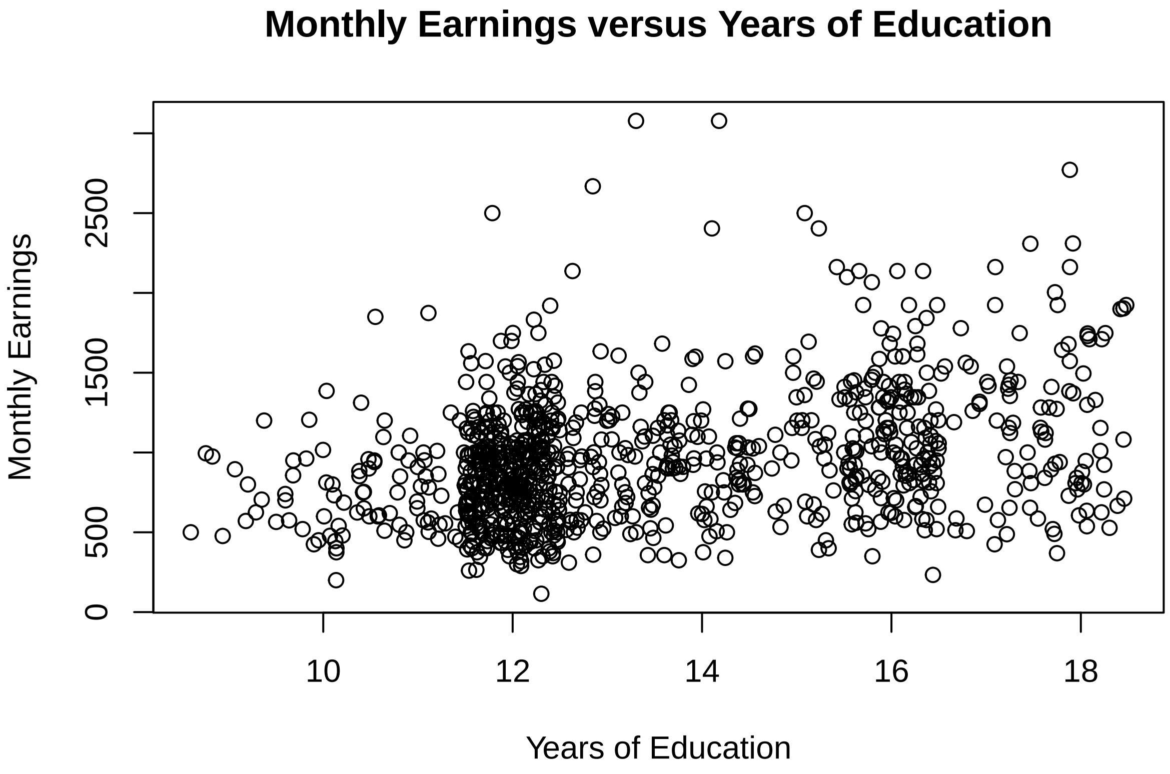
The paramters `main`, `xlab`, and `ylab` set the main title, the label for the horizontal axis, and the label for the vertical axis, respectively.

The scatterplot looks a little peculiar because the explanatory variable, years of education, consists only of a finite number of integers. When so many observations have the exact same value, they vertically line up tightly and heavily overlap, making it difficult to see the relationship in the scatterplot.

R includes a function called `jitter()` help address this visual problem. The function accepts as a parameter a numerical vector and adds a small amount of noise to it. That is, it accepts a vector of values and adds a small random number that takes negative and positive values.

Let us call the plot function again, but this time add a "jitter" to years of education. Because the distance between possible values for years of education on our plot is always equal to 1.0, we will add a jitter with a maximum magnitude of 0.5, or half of this distance. The following call to `plot` includes a call to `jitter` to accomplish this.

```
plot(x=jitter(wages$YearsEdu, amount=0.5), y=wages$MonthlyEarnings,
     main="Monthly Earnings versus Years of Education",
     xlab="Years of Education",
     ylab="Monthly Earnings")
```



3. Estimating the Regression Equation

We estimate the regression line with the R function `lm`, which stands for *linear model*. We estimate the regression line in the code that follows and assign the output to a variable we call `edulm`.

```
edulm <- lm(wages$MonthlyEarnings ~ wages$YearsEdu)
```

We passed to `lm` a single parameter that was the *formula*, `wages$MonthlyEarnings ~ wages$YearsEdu` which told `lm` to estimate a linear model for how monthly earnings depends on years of education.

The object `edulm` is a list of many other objects that include summary many statistics, hypothesis tests, and confidence intervals, regarding the equation of the best fit line. For the moment, let us examine the estimates of the coefficients. We can view these by accessing the `estimate` object within `edulm`:

```
edulm$coefficients
```

```
##    (Intercept) wages$YearsEdu
##      146.95244       60.21428
```

These results imply the equation for the best fitting line is given by,

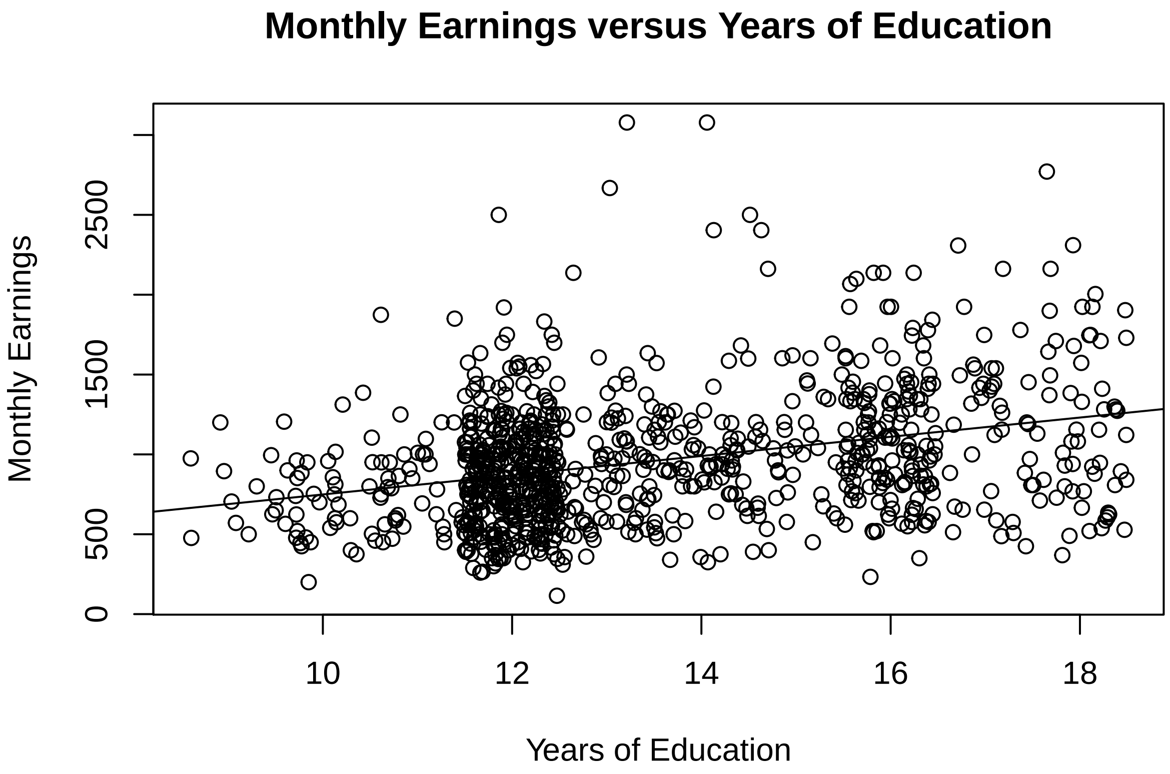
$$y_i = 146.95 + 60.21x_i + e_i,$$

and the predicted value for monthly earnings for person i with x_i years of education is given by,

$$\hat{y}_i = 146.95 + 60.21x_i.$$

We can add the regression line to our scatterplot with a call to `abline()`. In the code below, we regenerate the original scatterplot and call `abline` to draw the regression line:

```
plot(x=jitter(wages$YearsEdu, amount=0.5), y=wages$MonthlyEarnings,
     main="Monthly Earnings versus Years of Education",
     xlab="Years of Education",
     ylab="Monthly Earnings")
abline(edulm)
```


Monthly Earnings versus Years of Education
Monthly Earnings
Years of Education