Bivariate Linear Regression

R Tutorials for Applied Statistics

Note on required packages: The following code requires the packages in the `tidyverse`. The `tidyverse` contains many packages that allow you to organize, summarize, and plot data. If you have not already done so, download and install the libraries (needed only once per computer), and load the libraries (need to do every time you start R) with the following code:

```
# This only needs to be executed once for your machine
install.packages("tidyverse")

# This needs to be executed every time you load R
library("tidyverse")
```

---

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**, specifically with the assumption that the explanatory variable influences the outcome variable, and not vice-versa.

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

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

yi=b0+b1xi+ei

where b_1 b1 is the sample estimate of the slope of the regression line with respect to years of education and b_0 b0 is the sample estimate for the vertical intercept of the regression line.

The term e_i ei is **residual**, or 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 $i \in \{1, 2, . . , n\}$ i∈{1,2,..,n} in the sample, e_i ei is the difference between his or her actual income and the predicted level of income on line based on the person's actual education attainment.

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

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

y^i=b0+b1xi.

Since some actual values for y_i yi will be above the regression line and some will be below, some e_i ei 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:

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

1n∑i=1nei=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:

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

{b0,b1}=minb0,b1∑i=1nei2=minb0,b1∑i=1n(yi−y^i)2=minb0,b1∑i=1n(yi−b0−b1xi)2

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

Interpreting slope: Since b_1 b1 is the slope, it measures how much the y-variable changes when the x-variable increases by one unit. In this case, b_1 b1 is the estimate for on average how much additional income one earns for each additional year of education.

Interpreting intercept: Depending on the application, the vertical intercept sometimes has a useful intuitive meaning and sometimes it does not. It measures what value to be expected for the y-variable when the x-variable is equal to zero. In this case, b_0 b0 is the measure for the average income to be expected for individuals with zero years of education. If your data does not include any observations with a zero value for education, or if a zero value for the x-variable is unrealistic, then this coefficient has little meaning.

The regression line 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,

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

yi=β0+β1xi+ϵi,

where b_0b0 above is a sample estimate of the population coefficient β_0β0 and b_1b1 above is a sample estimate of the population coefficient β_1β1.

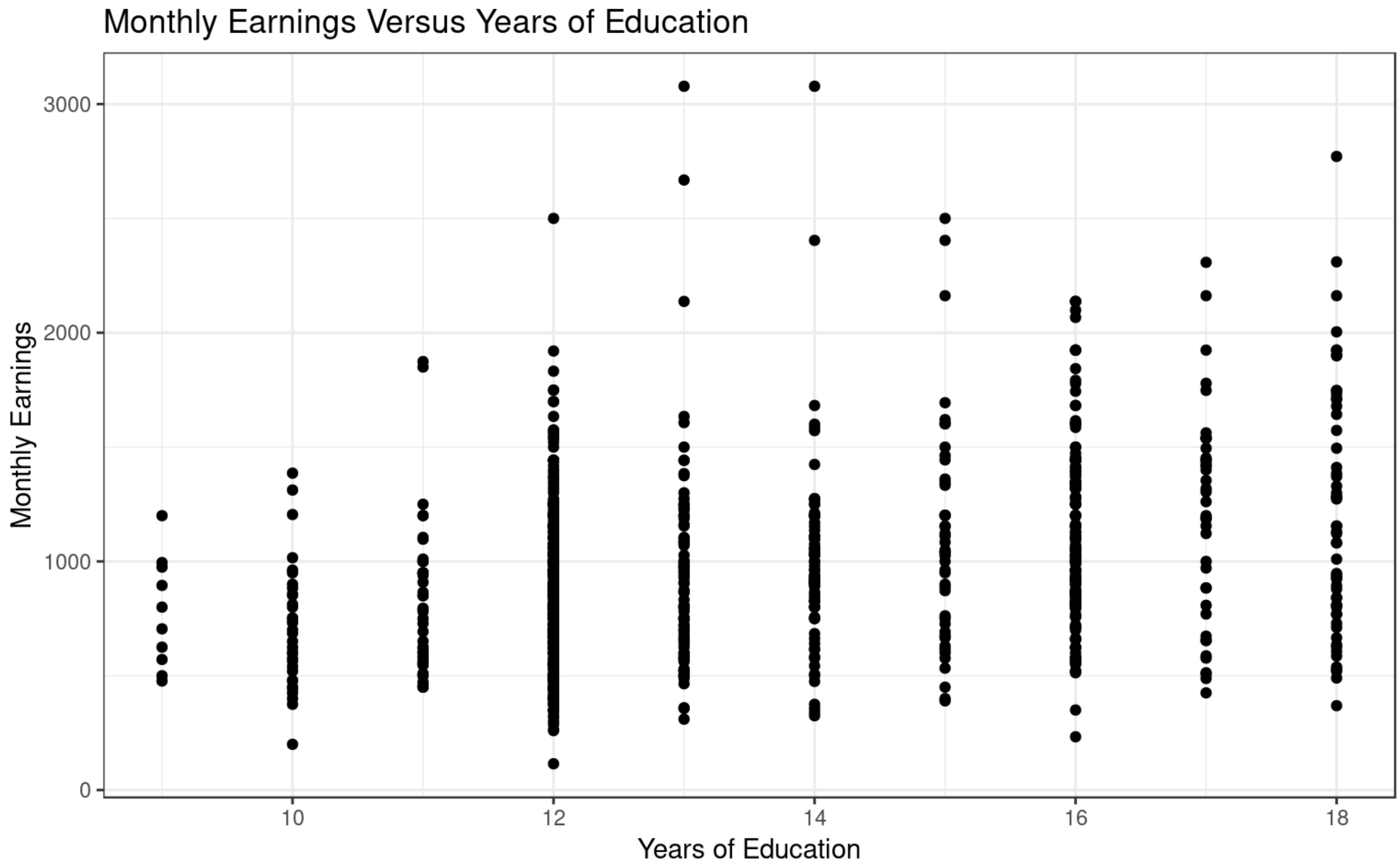
2 Download and Visualize the Data

The code below downloads a CSV file that includes data from 1980 for 935 individuals on variables including their total monthly earnings (`MonthlyEarnings`) and a number of variables that could influence income, including years of education (`YearsEdu`). The data set originally comes from textbook website for Stock and Watson's *Introduction to Econometrics*.

```
wages <- read.csv("http://murraylax.org/datasets/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 scatter plot with `YearsEdu` on the horizontal axis and `MonthlyEarnings` on the vertical axes.

```
ggplot(data=wages, mapping=aes(x=YearsEdu, y=MonthlyEarnings)) +
  geom_point() +
  labs(title="Monthly Earnings Versus Years of Education",
       x="Years of Education",
       y="Monthly Earnings") +
  theme_bw()
```



In the first line above, we call the `ggplot()` function and set the data and aesthetic layers of the graph. We define the data layer by setting the parameter `data` equal to the `wages` data frame. We define the aesthetics layer with the mapping parameter. We create a mapping with a call to the `aes()` function and map `YearsEdu` to the x-axis and `MonthlyEarnings` to the y-axis.

We add the geometry layer in the second line with a call to `geom_point()`. This creates a plot with points on it.

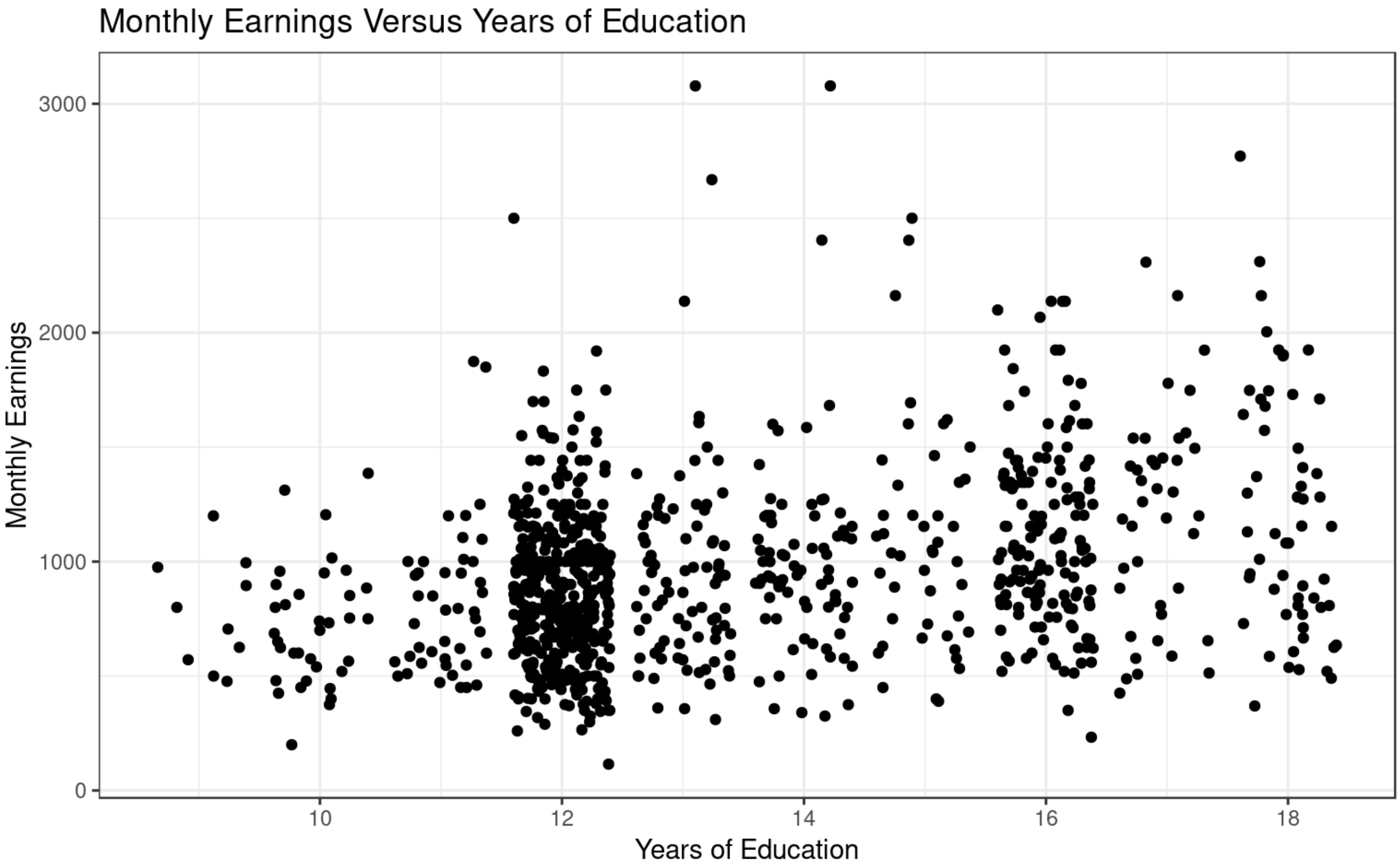
Finally, we add descriptive labels to the x and y axes and the title with the `labs()` function.

The scatter plot 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 line up vertically and tightly with heavy overlap. This makes it difficult to see the relationship in the scatter plot. This is a problem called **overplotting**.

One way to solve the overplotting problem is to alter the position of the points in the geometry layer with jittering. Jittering randomly shuffles around the points, making it easier to see all the points that are in one tight area. Generally, you want to make sure that when you jitter points that you do not shuffle them out too far, so as to end aligning with different points on the x-axis or y-axis scales.

In the code below, we create the plot again, this time setting the parameter `position=jitter` in the `geom_point()` function.

```
ggplot(data=wages, mapping=aes(x=YearsEdu, y=MonthlyEarnings)) +
  geom_point(position="jitter") +
  labs(title="Monthly Earnings Versus Years of Education",
       x="Years of Education",
       y="Monthly Earnings") +
  theme_bw()
```



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 object we call `edulm`.

```
edulm <- lm(MonthlyEarnings ~ YearsEdu, data=wages)
```

We passed to `lm()` a single parameter that was the *formula*, `MonthlyEarnings ~ 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 includes many summary 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 `coefficients` object within `edulm`:

```
edulm$coefficients
```

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

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

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

yi=146.95+60.21xi+ei,

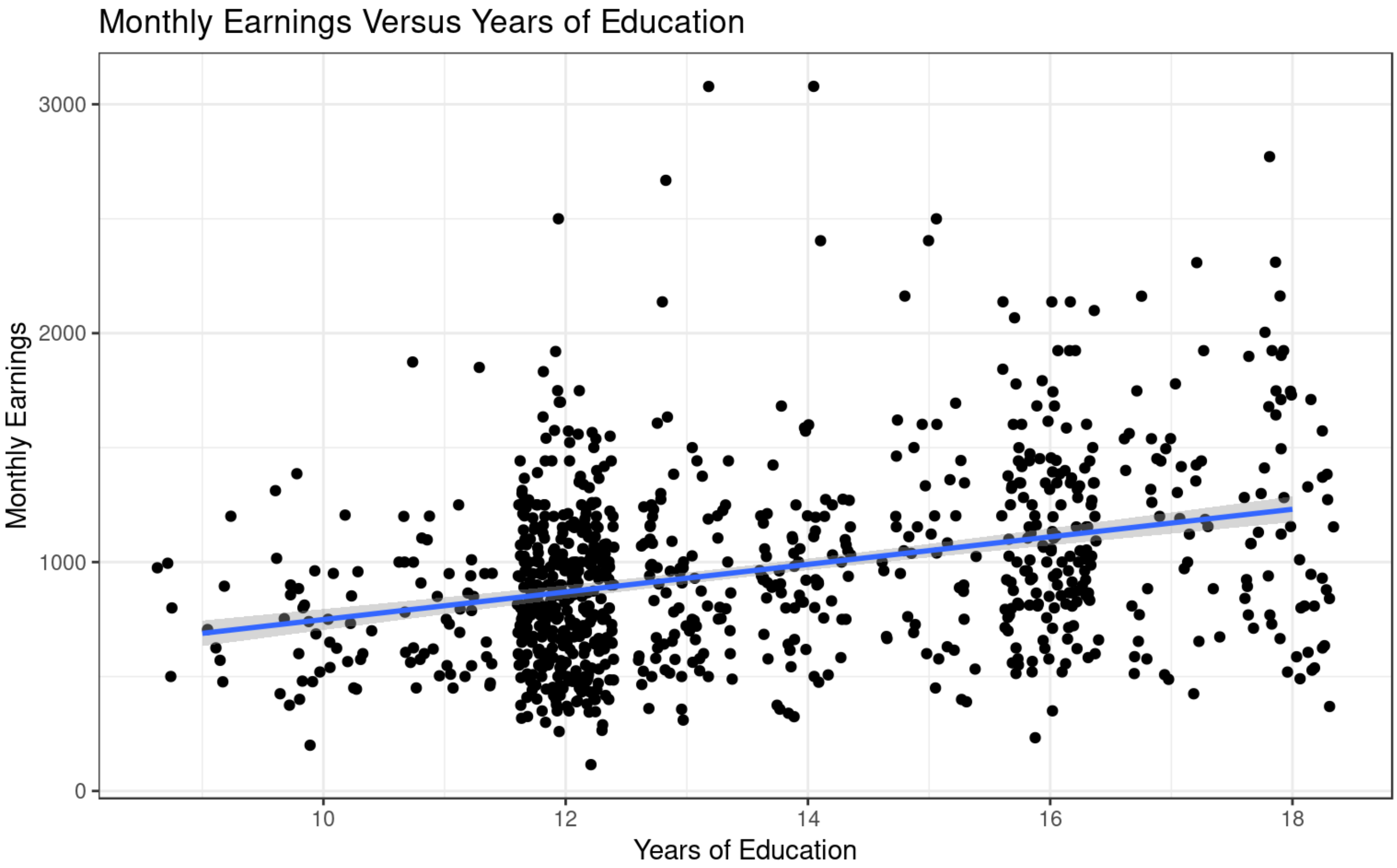
and the predicted value for monthly earnings for person ii with $x_i$xi years of education is given by,

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

y^i=146.95+60.21xi.

We can add an estimated linear regression line to our scatter plot with another geometry layer (which also creates a statistics layer). In the code that follows, we recreate the code for the scatter plot above and we add a visual of the regression line with a call to `geom_smooth()`

```
ggplot(data=wages, mapping=aes(x=YearsEdu, y=MonthlyEarnings)) +
  geom_point(position="jitter") +
  labs(title="Monthly Earnings Versus Years of Education",
       x="Years of Education",
       y="Monthly Earnings") +
  geom_smooth(method="lm") +
  theme_bw()
```



In the `geom_smooth()` function call, we set the parameter `method="lm"` to create a linear regression line. The shaded area surrounding the line represents a 95% confidence interval for the estimated regression line.