

# Simple Linear Regression—Description

2017-01-20

## Introduction and Research Question

In this set of notes, you will begin your foray into regression analysis. To do so, we will use the *riverside.csv* data to examine whether education level is related to income. The data come from C. Lewis-Beck & Lewis-Beck (2016) and contain five attributes collected from a random sample of  $n = 32$  employees working for the city of Riverview, a hypothetical midwestern city. The attributes include:

- **education:** Years of formal education
- **income:** Annual income (in U.S. dollars)
- **seniority:** Years of seniority
- **gender:** Employee's gender
- **male:** Dummy coded gender variable (0 = Female, 1 = Male)
- **party:** Political party affiliation

## Preparation

```
# Read in data
city = read.csv(file = "~/Google Drive/andy/epsy-8251/data/riverside.csv")
head(city)
```

	education	income	seniority	gender	male	party
1	8	37449	7	male	1	Democrat
2	8	26430	9	female	0	Independent
3	10	47034	14	male	1	Democrat
4	10	34182	16	female	0	Independent
5	10	25479	1	female	0	Republican
6	12	46488	11	female	0	Democrat

```
# Load libraries
library(dplyr)
library(ggplot2)
library(sm)
```

## Data Exploration

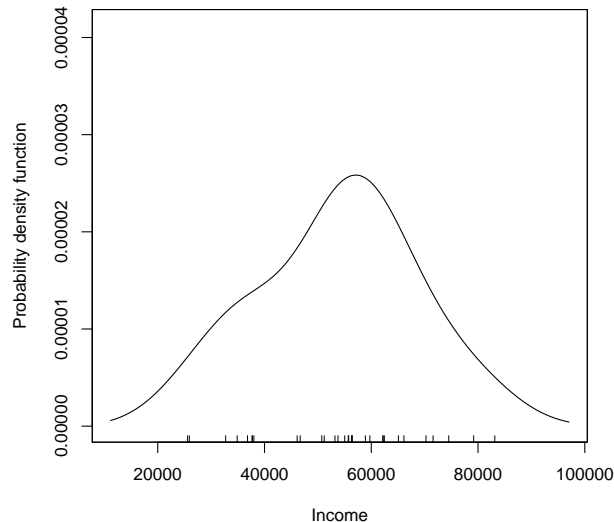
Any analysis should start with an initial exploration of the data. During this exploration, you should examine each of the variables that you will be including in the regression analysis. This will help you understand results you get in later analyses, and will also help foreshadow potential problems with the analysis. For additional detail, [this blog post](#) describes initial ideas of data exploration reasonably well. You could also refer to almost any introductory statistics text.

It is typical to begin by exploring the distribution of each variable separately. These distributions are referred to as *marginal distributions*. After that, it is appropriate to explore the relationships between the variables.

## Income

To begin this exploration, we will examine the marginal distribution of employees' incomes. We can plot a marginal distribution using the `sm.density()` function from the `sm` package.

```
sm.density(city$income, xlab = "Income")
```



This plot suggests that employees' incomes are unimodal with most incomes between roughly \$50,000 and \$70,000. (Note the scale on the  $x$ -axis uses scientific notation; `6e+04` mean  $6 \times 10^4$ . The rug at the bottom of the plot (the small vertical line segments) show the smallest income in the sample is about \$25,000 and the largest income is over \$80,000. (We could find the exact values using the `summary()` function.) This suggests there is a fair amount of variation in the data.

To further summarize the distribution, it is typical to compute and report summary statistics such as the mean and standard deviation. One way to compute these values is to use functions from the `dplyr` library.

```
city %>% summarize( M = mean(income), SD = sd(income) )
```

```
      M      SD
1 53742.12 14553.05
```

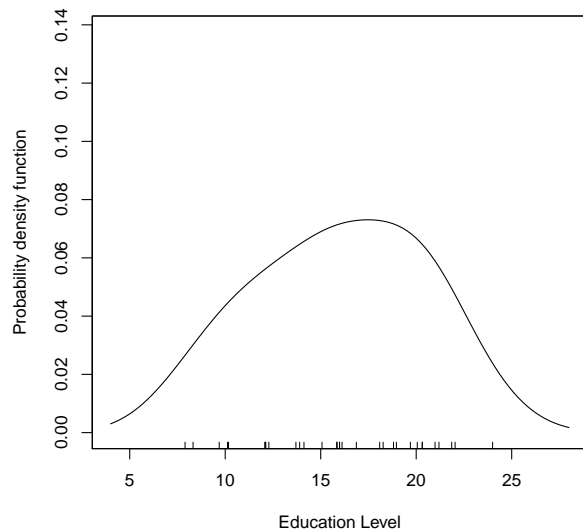
Describing this variable we might write,

The marginal distribution of income is unimodal with a mean of 53,742. There is variation in employees' salaries (SD = 14,553).

## Education Level

We will also examine the distribution of the education level variable.

```
# Plot  
sm.density(city$education, xlab = "Education Level")
```



```
# Summary statistics  
city %>% summarize( M = mean(education), SD = sd(education) )
```

```
      M      SD  
1 16 4.362598
```

Again, we might write,

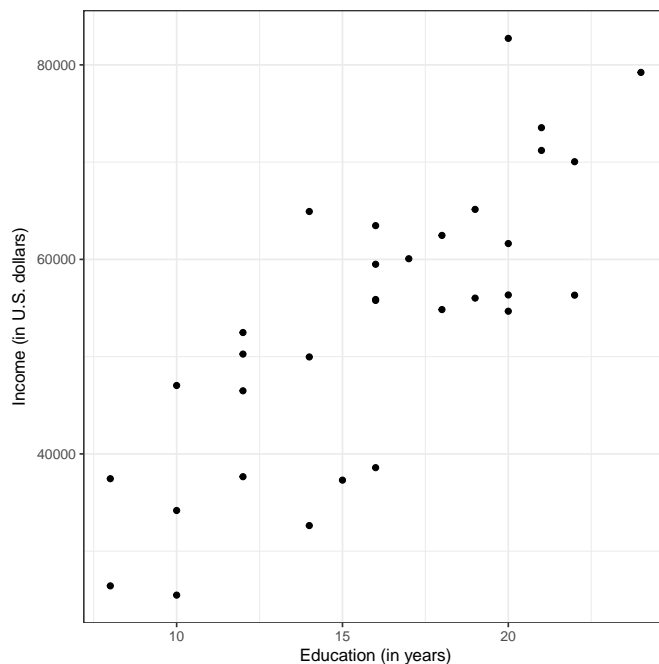
The marginal distribution of education is unimodal with a mean of 16 years. There is variation in employees' level of education ( $SD = 4.4$ ).

## Relationship Between Variables

Although examining the marginal distributions is an important first step, those descriptions do not help us directly answer our research question. To better understand any relationship between income and education level we need to explore how the distribution of income differs as a function of education. To do this, we will create a scatterplot of incomes versus education.

### Scatterplot

```
ggplot( data = city, aes(x = education, y = income) ) +  
  geom_point() +  
  theme_bw() +  
  xlab("Education (in years)") +  
  ylab("Income (in U.S. dollars)")
```



The plot suggests a relationship (at least for these employees) between level of education and income. When describing the relationship we want to touch on four characteristics of the relationship:

- Functional form of the relationship
- Direction
- Strength
- Observations that do not fit the trend (outliers)

### Correlation

To numerically summarize relationships between variables, we typically compute correlation coefficients. The correlation coefficient is a quantification of the direction and strength of the relationship. (It is important to note that the correlation coefficient is only an appropriate summarization of the relationship if the functional form of the relationship is linear.)

To compute the correlation coefficient, we use the `cor()` function. We give it an indexed data frame, `dataframe[rows, columns]`, where *dataframe* is the name of the data frame, *rows* is empty (all rows) and *columns* gives the names of the variables we want the correlation between. Since there are multiple columns, we use the `c()` function. Typically the response (or outcome) variable is the first variable provided in the columns, followed by the predictor.

```
cor( city[ , c("income", "education") ] )
```

```
           income education
income    1.0000000 0.7947847
education 0.7947847 1.0000000
```

When reporting the correlation coefficient it is conventional to use a lower-case *r* and report the value to two decimal places. Subscripts are also generally used to indicate the variables. For example,

$$r_{\text{education, income}} = 0.80$$

Combining the information culled from the scatterplot with that of the correlation analysis, we could summarize the relationship between education level and income as,

There is a strong, positive, linear relationship between education level and income ( $r = .80$ ). This suggests that city employees with lower education levels tend to have lower incomes, on average, than employees with higher education levels.

## Statistical Model

Since the relationship's functional form seems reasonably linear, we will use a *linear model* to describe the data. We can express this model mathematically as,

$$Y_i = \beta_0 + \beta_1(X_i) + \epsilon_i.$$

In this equation,

- $Y_i$  is the outcome/response value; it has an  $i$  subscript because it can vary across cases/individuals.
- $\beta_0$  is the intercept of the line that best fits the data; it does not vary across individuals.
- $\beta_1$  is the slope of the line that best fits the data; it does not vary across individuals.
- $X_i$  is the predictor value; it has an  $i$  subscript because it can vary across cases/individuals.
- $\epsilon_i$  is the error term; it has an  $i$  subscript because it can vary across cases/individuals.

## Fitted Regression Equation

The regression model can be separated into two components: a *systematic* (or fixed) component and a *random* (or stochastic) component.

$$Y_i = \underbrace{\beta_0 + \beta_1(X_i)}_{\text{Systematic (Fixed)}} + \underbrace{\epsilon_i}_{\text{Random (Stochastic)}}$$

The systematic (fixed) part of the equation gives the predicted  $Y$  given a particular  $X$ -value. The notation for the predicted  $Y$  is  $\hat{Y}$ . We express this mathematically as,

$$\hat{Y}_i = \beta_0 + \beta_1(X_i).$$

This is sometimes referred to as the *regression equation* or the *fitted equation*. The terms  $\beta_0$  and  $\beta_1$  are referred to as the regression parameters. One of the primary goals of a regression analysis is to estimate the values of the regression parameters (i.e., the intercept and slope terms). (Note that the fitted equation does not include any error terms.)

## Residuals

Now we can re-write the statistical model, substituting  $\hat{Y}$  in for the fitted part of the model.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i) + \epsilon_i \\ Y_i &= \hat{Y}_i + \epsilon_i \end{aligned}$$

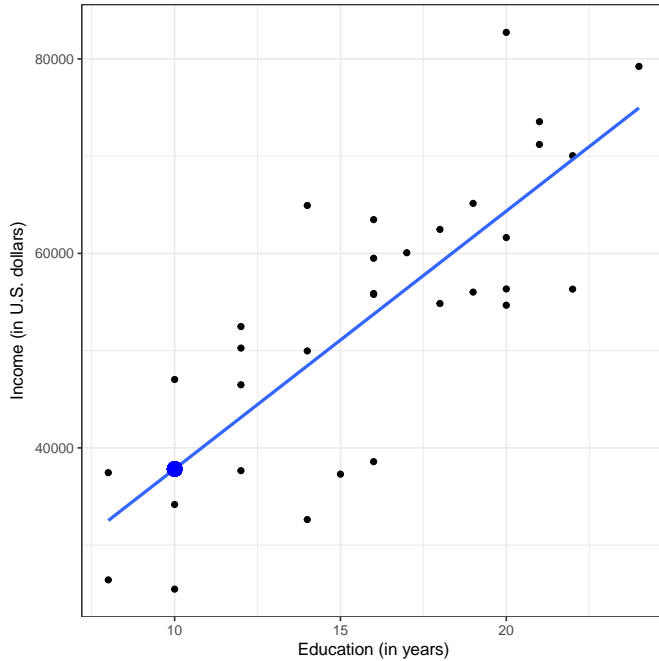
This equation implies that each observed  $Y$ -value is the sum of the predicted value of the  $Y$  (which is based on the  $X$ -value) and some residual (or error) term. Re-arranging the terms, we can mathematically express the residual term as,

$$\epsilon_i = Y_i - \hat{Y}_i$$

To compute an observation's residual, we compute the difference between the observation's  $Y$ -value and its predicted value based on the fitted equation. When the observed value of  $Y$  is larger than the predicted value of  $Y$  the residual term will be positive (underprediction). If the observed value of  $Y$  is smaller than the predicted value of  $Y$  the residual term will be negative (overprediction).

### Why is there an error term in the statistical model?

We use a single line to describe the relationship between education and income. This line is the same for all of the observations in the sample. For example, look at the figure below which shows the relationship between education and income, but this time also includes the regression line.



Consider all the employees that have an education level of 10 years. For all three of them we would predict an income of approximately \$37,800. This is denoted by the blue point on the line. The error term allows for discrepancy between the predicted  $\hat{Y}$  and the observed  $Y$ , which allows us to recover our observed value of the response variable from the model.

Graphically, the residual is represented by the vertical distance between the line and a given point on the scatterplot. Some of those points are above the line (they have a positive residual) and some are below the line (they have a negative residual). Also note that for some observations the error term is smaller than for others.

## Notation

The regression model and the fitted regression equation describe the linear relationship in the *population*. Greek letters indicate a *parameter* (a summary of the population). That is why we use the Greek letter  $\beta$  when we notate the regression model.

In most statistical analyses, you will use a *sample* of data (not the entire population). When we summarize a sample, it is referred to as a *statistic*, and we use either Roman letters or a Greek letter with a hat. This indicates that the summary measure is an estimate of the parameter. For example, the values produced for the intercept and slope from a regression analysis are estimates. Since they are estimates, we use the hat-notation on the greek letters.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(X_i)$$

The parameter estimates are also referred to as regression coefficients. Synonymously, a hat means predicted value. Some people use Roman letters when referring to sample estimates.

$$\hat{Y}_i = b_0 + b_1(X_i)$$

## Regression Estimates Using R

To fit the regression model to data using R, we will use the `lm()` function. The syntax for this function looks like this:

```
lm(outcome ~ 1 + predictor, data = dataframe)
```

where **outcome** is the name of the outcome/response variable, **predictor** is the name of the predictor variable, and **dataframe** is the name of the data frame. (The one on the right side of the tilde tells R to include the intercept in its computation.) When we fit a regression model in R, we will also assign the output to a new object in R. Below, we fit the model using education level to predict income.

```
lm.1 = lm(income ~ 1 + education, data = city)
```

Here the output is assigned to an object called `lm.1`. We can print the regression parameter estimates by typing the `lm()` object name and hitting enter.

```
lm.1
```

Call:

```
lm(formula = income ~ 1 + education, data = city)
```

Coefficients:

(Intercept)	education
11321	2651

Here the parameter estimates (or regression coefficients) are:

- $\hat{\beta}_0 = 11,321$
- $\hat{\beta}_1 = 2,651$

Remember that these are estimates and need the hats. The fitted regression equation is

$$\hat{\text{Income}} = 11,321 + 2,651(\text{Education Level}).$$

### Intercept Interpretation

The estimate for the intercept was 11,321. Graphically, this value indicates the  $y$ -value where the line passes through the  $y$ -axis (i.e.,  $y$ -intercept). As such, it gives the predicted value of  $Y$  when  $X = 0$ . Algebraically we get the same thing if we substitute 0 in for  $X_i$  in the estimated regression equation.

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1(0) \\ \hat{Y}_i &= \hat{\beta}_0\end{aligned}$$

To interpret this value, we use that same idea. Namely

The predicted income for all employees that have an education level of 0 years is \$11,321.



## Slope Interpretation

Recall from algebra that the slope of a line describes the change in  $Y$  versus the change in  $X$ . In regression, the slope describes the *predicted* change in  $\hat{Y}$  for a one-unit difference in  $X$ .

$$\hat{\beta}_1 = \frac{\Delta \hat{Y}}{\Delta X} = \frac{2651}{1}$$

In our example,

Each one-year difference in education level is associated with a \$2,651 predicted difference in income.

To better understand this, consider three city employees. The first employee has an education level of 10 years. The second has an education level of 11 years, and the third has an education level of 12 years. Now let's compute each employee's predicted income.

$$\begin{aligned}\text{Employee 1 : Income} &= 11,321 + 2,651(10) \\ &= 37,831\end{aligned}$$

$$\begin{aligned}\text{Employee 2 : Income} &= 11,321 + 2,651(11) \\ &= 40,482\end{aligned}$$

$$\begin{aligned}\text{Employee 3 : Income} &= 11,321 + 2,651(12) \\ &= 43,133\end{aligned}$$

Each of the employee's education levels differ by one year (10 to 11 to 12). The difference in predicted incomes for these employees differs by \$2,651.

## Using the Regression Equation

Consider the twelfth case in the data frame.

```
city[12, ]
```

	education	income	seniority	gender	male	party
12	14	64926	26	male	1	Independent

This employee has an education level of fourteen years ( $X_{12} = 14$ ). Her/his income is \$64,926 ( $Y_{12} = 64,926$ ). Using the fitted equation, we can compute that employee's predicted income as,

```
11321 + 2651 * 14
```

```
[1] 48435
```

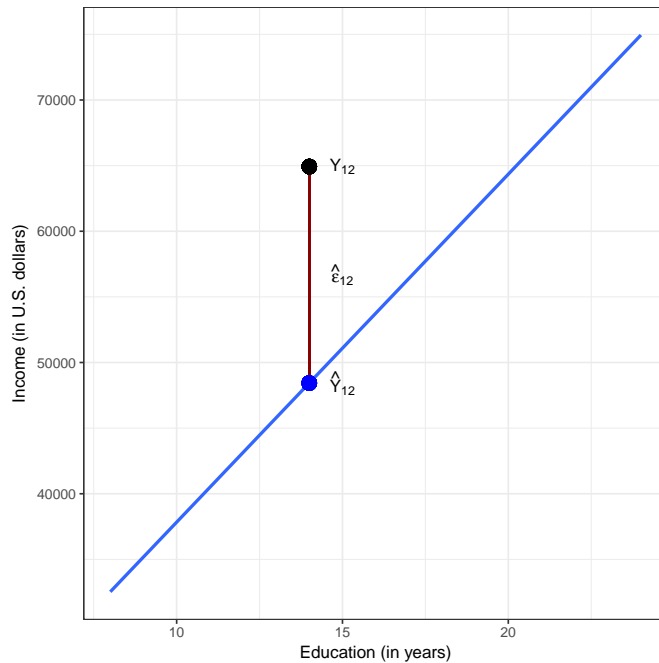
$\hat{Y}_{12} = 48,435$ .

We can also compute that employee's residual.

$$64926 - 48435$$

[1] 16491

$\hat{\epsilon}_{12} = 16,491$  The positive residual suggests that this employee earns \$16,491 more than would be expected for a city employee with 14 years of formal education. We can also represent these values graphically.

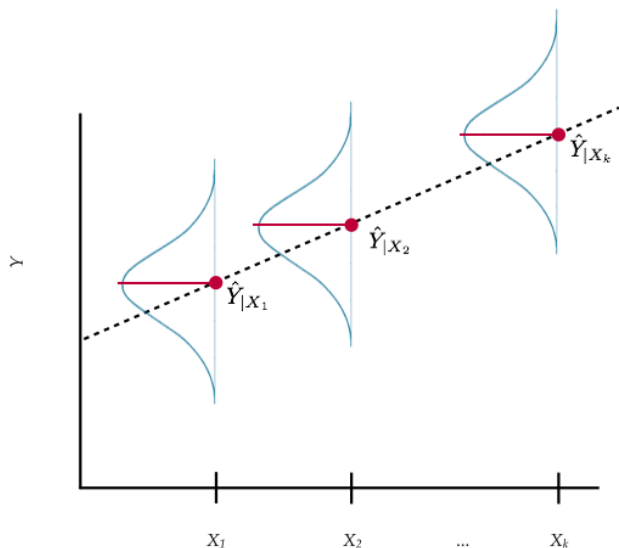


## Conditional Averages

To this point, we have used the fitted regression equation to predict for individual employees. Another way to think about the predicted value is it describes the mean value of  $Y$  for *all* cases with a particular  $X$  value. For example, using the values from the previous example we found that  $\hat{Y}_{12} = 48,435$  when  $X_{12} = 14$ . Using means, we could interpret this as,

The estimated average income for city employees having 14 years of education is \$48,435.

A statistician may refer to the predicted value of  $Y$  as a conditional mean (it is conditioned on a particular value of  $X$ ). To help better understand the idea of conditional means, consider the following plot:



At each value of  $X$  there is a distribution of  $Y$ . For example, there would be a distribution of incomes for the employees with an education level of 10 years (in the population). There would be another distribution of incomes for the employees with an education level of 11 years (in the population). And so on.

The regression equation describes the pattern of conditional means. As such, we write the fitted equation using means rather than  $\hat{Y}$ ,

$$\mu_{Y|X} = \beta_0 + \beta_1(X_i)$$

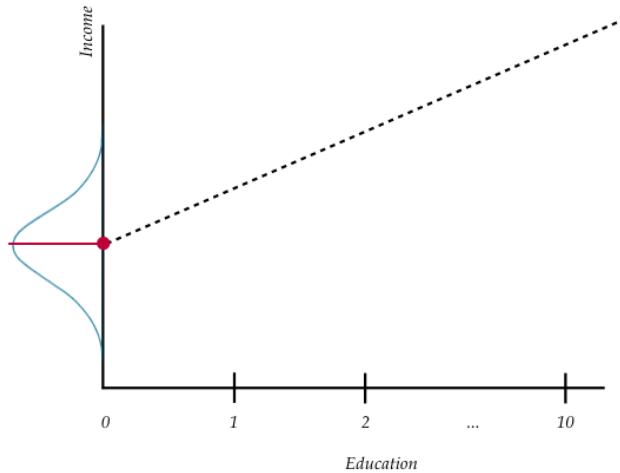
The first part is read as, “the mean of  $Y$  given  $X$ ”, or “the mean of  $Y$  conditioned on  $X$ ”. Sometimes the mean of a population is denoted as  $E(Y)$ , or the expected value of  $Y$ . Then you might see the regression equation written as,

$$E(Y|X) = \beta_0 + \beta_1(X_i)$$

When we assume a linear functional form for the model, we are saying that the mean value of  $Y$  differs by a constant amount for each one-unit difference in  $X$ . In other words, the difference between the mean income for those employees who have ten years of education and those that have 11 years of education *is the same as* the difference between the mean income for those employees who have 17 years of education and those that have 18 years of education.

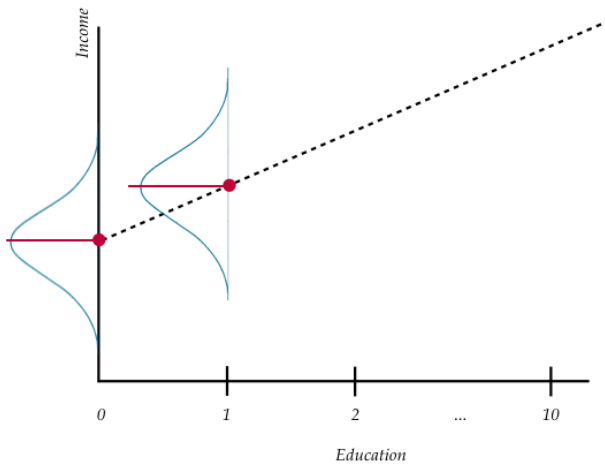
## Intercept (Re-visited)

Using the idea of conditional means, we can re-visit the interpretation of the intercept, which we had said was the predicted  $Y$  for a person with an  $X$ -value of zero. Now we can say that the intercept is the predicted mean income for all employees with zero years of formal education.



## Slope (Re-visited)

Using the idea of conditional means, we can also re-visit the interpretation of the slope, which we had said was the predicted difference in  $Y$  for employees with a one-year difference in education. Now we can say that the slope is the predicted difference in mean incomes between employees with education levels that differ by one year.



In general, when interpreting the slope and intercept, you should use the conditional mean interpretations.

## References

Lewis-Beck, C., & Lewis-Beck, M. (2016). *Applied regression: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.