Simple Linear Regression—Description

2016-07-31

Introduction and Research Question

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

- `edu`: Years of formal education
- `income`: Annual income (in U.S. dollars)
- `senior`: Years of seniority
- `gender`: Sex (0 = Female, 1 = Male)
- `party`: Political party affiliation (0 = Democrat, 1 = Independent, 2 = Republican)

Preparation

```
# Read in data
city = read.csv(file = "~/Documents/data/Applied-Regression-Lewis-Beck/riverside_final.csv")
head(city)
```

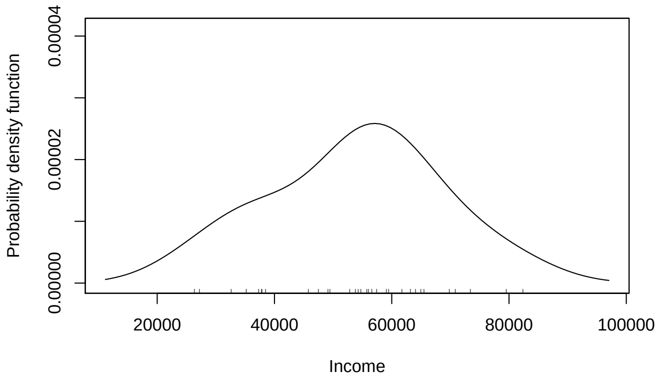
```
  edu income senior gender party
1   8  26430      9      0     1
2   8  37449      7      1     0
3  10  34182     16      0     1
4  10  25479      1      0     2
5  10  47034     14      1     0
6  12  37656     14      1     0
```

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

Examine and Describe the Outcome

Use the `sm.density()` function from the *sm* package to plot the marginal distribution of income.

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



We can also get summary measures of this variable

```r
mean(city$income)
```

```
[1] 53742.12
```

```r
sd(city$income)
```

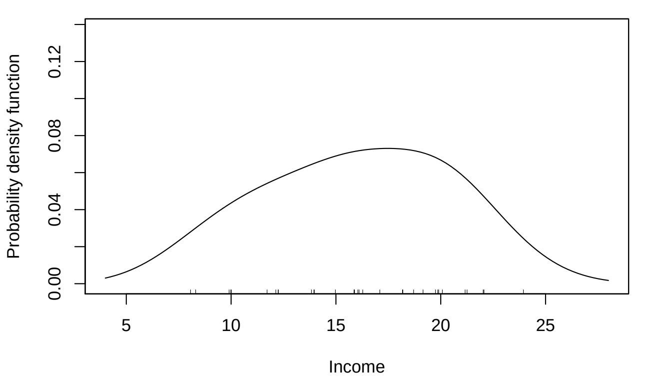
```
[1] 14553.05
```

Describing this variable we can say,

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

Examine and Describe the Predictor

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



```r
mean(city$edu)
```

```
[1] 16
```

```r
sd(city$edu)
```

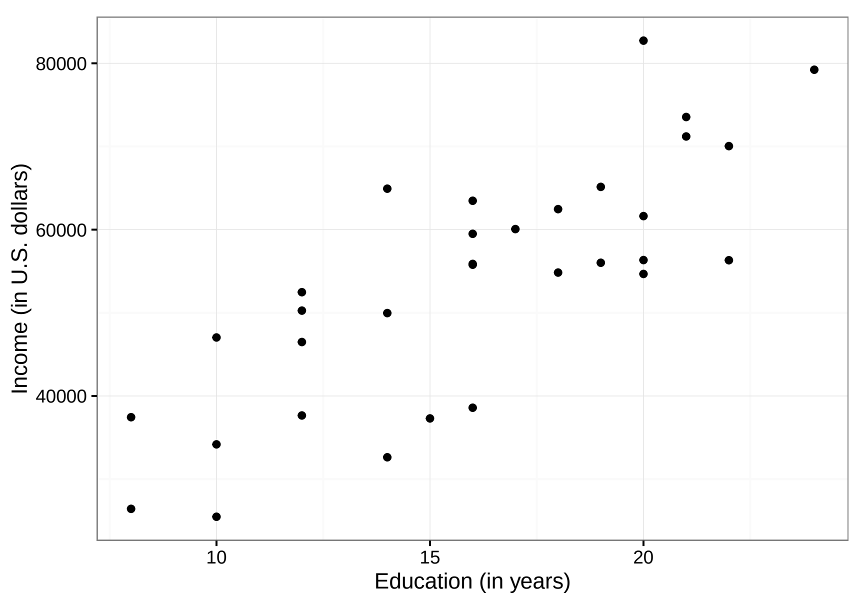
```
[1] 4.362598
```

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

Examining the Distribution of the Outcome Conditioned on the Predictor

Although examining the marginal distributions are important first steps, those descriptions do not help us directly answer our research question. To better understand how the distribution of income differs as a function of education, we need to examine a scatterplot of salary versus education.

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



The plot suggests a relationship (in the sample) between level of education and income. When describing the relationship we want to touch on four characteristics:

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

Correlation

We often also compute the correlation between the two variables. Correlation is a quantification of the direction and strength of the relationship. To compute the correlation, 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 outcome is the first variable given in the columns, followed by the predictor.

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

```
          income       edu
income 1.0000000 0.7947847
edu    0.7947847 1.0000000
```

The correlation between level of education and income, $r = 0.795$, suggests a moderate to strong positive relationship between the variables. This suggests that lower levels of education are typically associated with lower incomes, and higher levels of education are typically associated with higher incomes.

Statistical Model

Since the relationship's functional form seems reasonably linear, we will use a linear model to describe the data.

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

In this equation,

- Y_i is the outcome value; it has an i subscript because it can vary across cases/individuals
- β_0 is the intercept of the line that best fits the data
- β_1 is the slope of the line that best fits the data
- X_i is the predictor value; it has an i subscript because it can vary across cases/individuals
- ϵ_i is the error term; it has an i subscript because it can vary across cases/individuals

Regression (Fitted) Equation

The regression equation is the **systematic** part of the model that is fixed (the same) for all observations with the same predictor value.

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

The systematic (fixed) part of the equation gives the predicted Y given a particular X-value. The notation for the predicted Y is $\hat{Y}$. Note that the fitted equation does not include any error terms.

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

This is sometimes referred to as the regression equation or the fitted equation. The terms β_0 and β_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).

Error (Residual) Term

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 tells us that each observed Y-value is the sum of the predicted value of the Y (which is based on the X-value) and some error term. Re-arranging the terms, we get

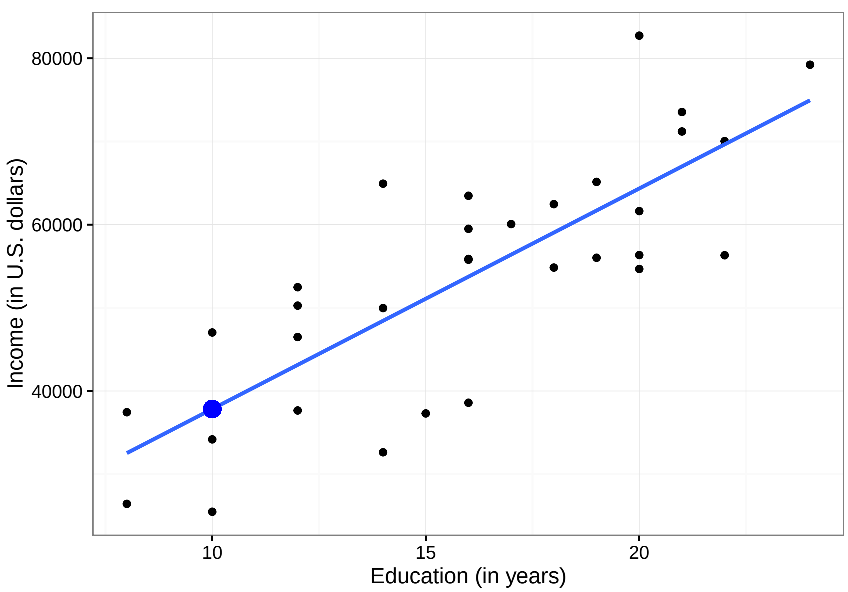
$$
\epsilon_i = Y_i - \hat{Y}_i
$$

To compute a case's/individual's error term (also called the residual), we find the difference between the observed value of the outcome and the predicted value of the outcome. 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).

Technically, the residual is also an estimate. Since the value of $\hat{Y}_i$ is based on an estimate from sample data, the difference between that estimate and the observed Y is also an estimate. Because of this, when we write thw residual term, we should either put a hat on it ($\hat{\epsilon}_i$) or use the roman notation (e_i).

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 Y and the observed Y, which allows us to recover our observed value of the outcome from the model.

Graphically the error term is the vertical distance between the line and a given point. Some of those points are above the line (have a positive error term) and some are below the line (have a negative error term). ALso note that for some observations the error term is smaller than for others.

Population Parameters versus Sample Estimates

The regression model and the fitted regression equation describe the linear relationship in the **population**. That is why we use the greek letter β when we notate the regression parameters.

In most analyses, you will use a sample of data (not the entire population) to **approximate** the parameters in this equation. The values you get for the intercept and slope are estimates. Since they are estimates, we use the hat-notation on the greek letters.

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

Hats are used to denote that the value is an estimate. Synonymously, a hat means predicted value. Some people use roman letters when referring to sample estimates.

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

The parameter estimates are also referred to as regression coefficients.

Fitting the Regression Model 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 variable, **predictor** is the name of the predictor variable, and **dataframe** is the name of the data frame. We will also typically assign the output to a new object in R. (The one on the right side of the tilde tells R to include the intercept in its computation.)

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

Here the output is assigned to `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 + edu, data = city)

Coefficients:
(Intercept)          edu
      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

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

Interpreting the Intercept

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.

Interpreting the Slope

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.

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

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

$$
\begin{aligned}
\textbf{Employee 3}:\ \hat{\text{Income}} &= 11,321 + 2,651(10) \\
&= 43,133
\end{aligned}
$$

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.

Observation, Prediction, and Error

Consider the twelfth case in the data frame.

```
city[12, ]
```

```
   edu income senior gender party
12  14  64926     26      1     1
```

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

```
11321 + 2651 * 14
```

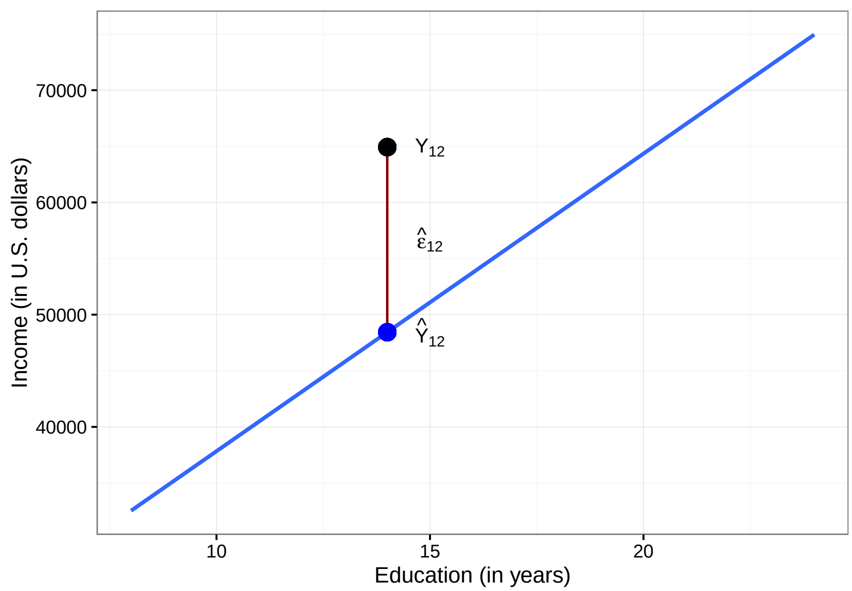
```
[1] 48435
```

$\hat{Y}_{12} = 48,435$. We can also compute that employee's residual,

```
64926 - 48435
```

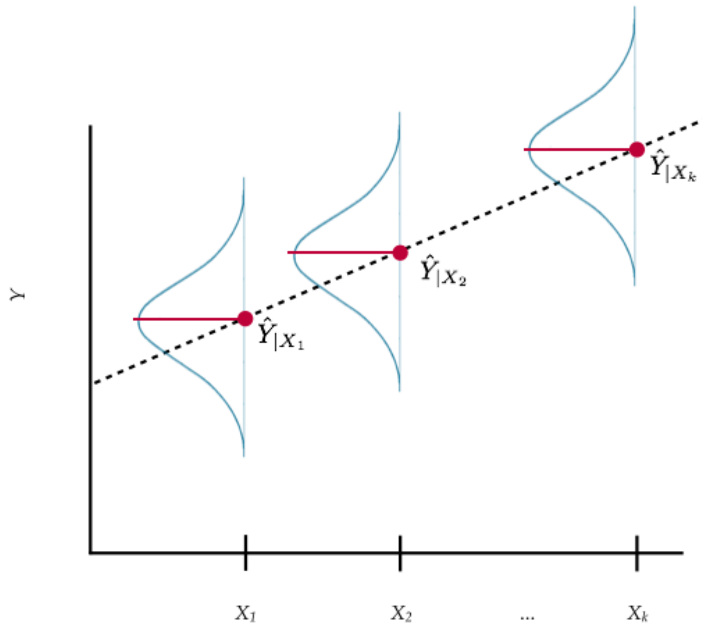
```
[1] 16491
```

$\hat{\epsilon}_{12} = 16,491$ 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 examine these values graphically.



Regression Line as Average

To this point, we have used the fitted regression equation to predict for individual cases. Another way to think about the predicted value is it describes the mean value of Y for **all** cases with a particular X value. A more scholarly way of saying this is that the predicted value of Y is a conditional mean (it is conditioned on a particular value of X). To help better understand this, 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 Y-hat,

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

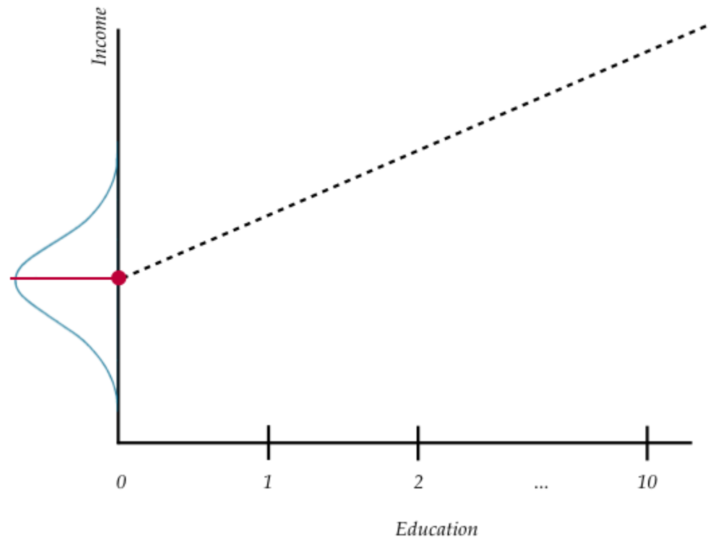
The first part is read as, "the mean of Y given X", or "the mean of Y conditioned on X". Sometimes the mean of a population is denoted as $E(Y)$, or the expected value of Y. Then you might see it written as,

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

When we assume a linear functional form, 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.

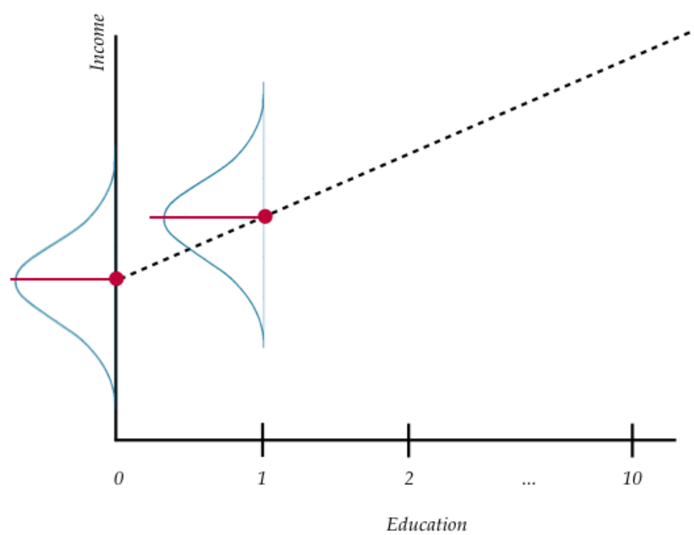
Interpretation of the Intercept (Re-visited)

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.



Interpretation of the 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.

Least Squares Estimation

How do we get the values of 11,321 and 2,651 for the intercept and slope? These values are based on an estimation method called **Least Squares**. Every estimation method requires two things:

- *Quantification of Model Fit:* We quantify how well (or not well) the estimated equation fits the data; and -*Optimization:* We find the "best" equation based on that quantification. (this boils down to finding the equation that produces the biggest or smallest measure of model fit.)

For most statistical models we quantify the model fit by computing the model's error. Error is actually a measure of model misfit (i.e., bigger errors = worse fitting model). To understand this, consider the following data set of five observations:

X	**Y**
3	63
1	44
3	64
5	68
2	25

Which of the following two models fits these data better?

- *Model A:* $\hat{Y} = 28 + 8(X)$
- *Model B:* $\hat{Y} = 20 + 10(X)$

To determine this, find the predicted values ans residuals for each of the five observations for both models.

	Model A		**Model B**	
Observation	Predicted Y	Residual	Predicted Y	Residual
1				
2				
3				
4				
5				

One issue is that eyeballing the residuals is a bit problematic for larger datasets. So, we have to further quantify the fit (or misfit). The way we do that in practice is to consider the *total* amount of error across all the observations. Unfortunately, we cannot just sum the residuals to get the total because some of our residuals are negative and some are positive. To alleviate this problem, we first square the residuals, then we sum them.

$$\begin{aligned}
\text{Total Error} &= \sum \hat{\epsilon}_i^2 \\
&= \sum \left(Y_i - \hat{Y}_i\right)^2
\end{aligned}$$

This is called a *sum of squared residuals* or *sum of squared error* (SSE; good name, isn't it). Compute the SSE for the residuals from Model A and Model B.

Now that we have quantified the model misfit, we can optimize. Ideally, the "best" model would have a low amount of error. To optimize we, in fact, find the model that has the least amount of error possible…thus *least squares*.

So the key is now to find the value for the slope and intercept that gives the smallest SSE. In mathematics, the way we optimize a function is to compute the derivative of that function, set it equal to zero, and solve it. Since there are two parameters we need to find, we would actually have to take the partial derivatives, but this type of computation is beyond the scope of this course. We have R and the `lm()` function. The values it produces for the slope and intercept are the optimized values; they produce the smallest SSE.

Here is an interactive website where you can play around with the intercept and slope of a line to visually understand the SSE: http://setosa.io/ev/ordinary-least-squares-regression/

For our toy dataset, the model that produces the smallest residuals is

$$\hat{Y} = 28.682 + 8.614(X)$$

This model gives the following residuals:

```
        1          2          3          4          5
 8.477273   6.704545   9.477273  -3.750000 -20.909091
```

The SSE is 657.89. This is the smallest SSE possible for a linear model. Any other value for the slope or intercept would result in a higher SSE.

Using R to Compute the SSE

We can use R to compute the SSE by carrying out the computations underlying the formula for SSE. Recall that the SSE is

$$\text{SSE} = \sum \left(Y_i - \hat{Y}_i\right)^2$$

We need to compute (1) the predicted values of Y, (2) the residuals, (3) the squared residuals, and finally, (4) the sum ofthe squared residuals. From the data set we have the observed X and Y values, and from the fitted `lm()` we have the intercept and slope for the regression equation.

```r
# Step 1: Compute the predicted values of Y
y_hat = 11321 + 2651 * city$edu
y_hat
```

```
 [1] 32529 32529 37831 37831 37831 43133 43133 43133 43133 48435 48435
[12] 48435 51086 53737 53737 53737 53737 53737 56388 59039 59039 61690
[23] 61690 64341 64341 64341 64341 66992 66992 69643 69643 74945
```

```r
# Step 2: Compute the residuals
errors = city$income - y_hat
errors
```

```
 [1]  -6099   4920  -3649 -12352   9203  -5477   7132   3355   9347 -15804
[11]   1533  16491 -13784 -15151   2141   5762   2045   9734   3680  -4199
[21]   3427  -5671   3452  -7998  -9669  -2712  18385   4210   6550 -13321
[31]    401   4282
```

```r
# Step 3: Compute the squared residuals
sq_errors = errors^2
sq_errors
```

```
 [1]  37197801  24206400  13315201 152571904  84695209  29997529  50865424
 [8]  11256025  87366409 249766416   2350089 271953081 189998656 229552801
[15]   4583881  33200644   4182025  94750756  13542400  17631601  11744329
[22]  32160241  11916304  63968004  93489561   7354944 338008225  17724100
[29]  42902500 177449041    160801  18335524
```

```r
# Step 4: Compute the sum of the squared residuals
sum(sq_errors)
```

```
[1] 2418197826
```

As you feel more comfortable with R and with the steps involved in computing the SSE, you can also perform these steps in a single computation.

```r
sum((city$income - (11321 + 2651 * city$edu))^2)
```

```
[1] 2418197826
```

Interpreting SSE

The SSE gives us information about the variation in Y that is left over (residual) after we fit the regression model. Since the regression model is a function of X, the SSE tells us about the variation in Y that is left over after we remove the variation associated with, or accounted for by X. In our example it tells us about the residual variation in incomes after we account for employee education level.

In practice, we report the SSE, but *we do not interpret the actual value*. The value of the SSE is used when comparing models. When researchers are considering different models, the SSEs from these models are compared to determine which model produces the least amount of misfit to the data (similar to what we did earlier).

Intercept-Only Model: A Baseline for Comparison

Consider the equation for the linear model again,

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

A simpler model (one with fewer terms) would be,

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

This model, referred to as the intercept-only model, does not include the effect of X. The value of Y is not a function of X in this model; it is not conditional on X. The fitted equation,

$$\hat{Y}_i = \hat{\beta}_0$$

indicates that the predicted Y would be the same (constant) regardless of what X is. In our example, this would be equivalent to saying that an employees' incomes would be predicted to be the same, regardless of what their education level was.

To fit the intercept-only model, we just omit the prediter term on the right-hand side of the `lm()` formula.

```
lm.0 = lm(income ~ 1, data = city)
lm.0
```

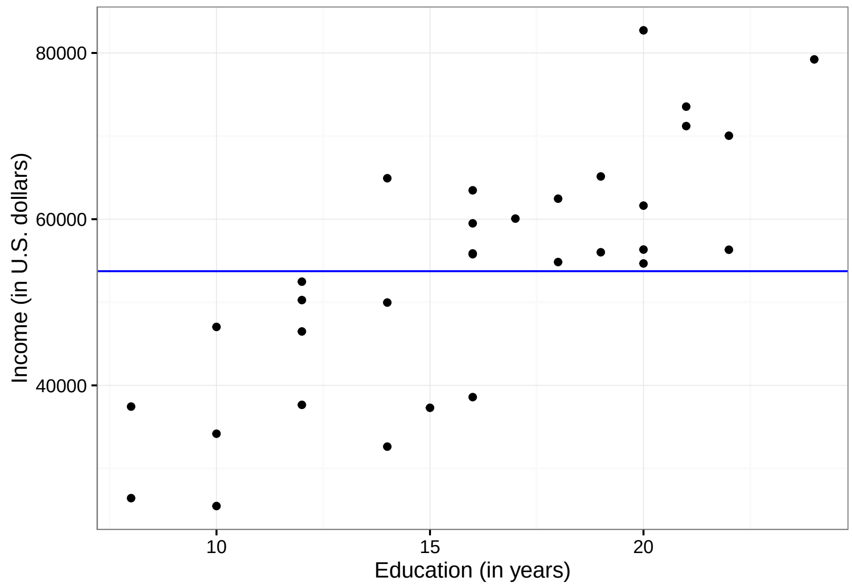
```
Call:
lm(formula = income ~ 1, data = city)

Coefficients:
(Intercept)
      53742
```

The fitted regression equation can be written as,

$$\hat{\text{Income}} = 53,742$$

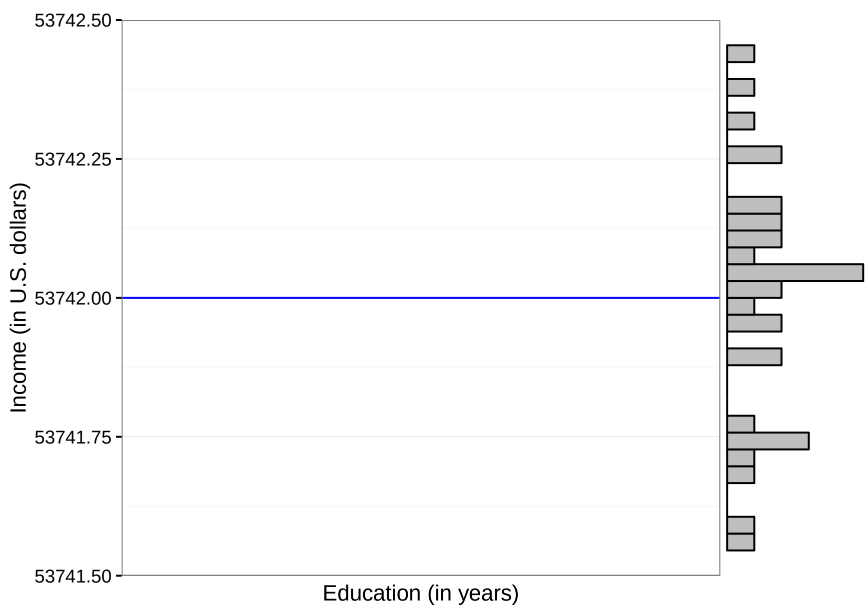
Graphically, the fitted line is a flat line crossing the y-axis at 53,742 (see plot below).



Does the estimate for β_0, 53,742, seem familiar? If not, go back to the exploration of the outcome earlier in the notes. The estimated intercept in the intercept-only model is the mean value of the outcome. This is not a coincidence. Remember that the regression model estimates the mean,

$$E(Y) = \beta_0.$$

Here, $E(Y)$ is the mean, μ_Y. The model itself does not consider any predictors, so on the plot the X variable is superfluous; we could just collapse it to its margin. This is why the mean of all the Y values is sometimes referred to as the *marginal mean.*



Yet another way to think about this is that the model is choosing a single income ($\hat{\beta}_0$) to be the predicted income for all the employees. Which value would be a good choice? Remember the `lm()` function chooses the "best" value for the parameter estimate based on minimizing the sum of squared errors. The mean is the value that minimizes the squared deviations (errors). This is one reason the mean is often used as a summary measure of a set of data.

SSE for the Intercept-Only Model

Since the intercept-only model doesn't include any predictors, the SSE is a quantification of the total variation in the outcome. Below we compute the SSE for the intercept-only model (if you need to go through the steps one-at-a-time, do so.)

```
sum((city$income - 53742)^2)
```

```
[1] 6565527426
```

This value represents the total amount of variation in the sample incomes. As such we can use it as a baseline for comparing other models that include predictors. For example,

- **SSE (Intercept-Only):** 6,565,527,426
- **SSE (w/Education Level Predictor):** 2,418,197,826

Once we account for education in the model, we reduce the SSE. This means our predictions improve (they are closer to the observed Y values). How much did they improve? They were reduced by 4,147,329,600. Is this a lot? To answer that question, we typically compute and report this reduction as a proportion of the total variation; called the *proportion of the reduction in error*, or PRE.

$$
\begin{aligned}
\text{PRE} &= \frac{6,565,527,426 - 2,418,197,826}{6,565,527,426} \\
&= \frac{4,147,329,600}{6,565,527,426} \\
&= 0.632
\end{aligned}
$$

Including education level as a predictor in the model reduced the error in the predictions by 63.2%.

Many researchers interpret this value as the percentage of *variation accounted for* by the model. They might say,

> The model which includes education level explains (or accounts for) 63.2% of the variation in incomes.

PRE's Relationship to the Correlation Coefficient

The PRE has a direct relationship to the correlation coefficient. Namely, it is the square of the correlation coefficient,

$$
\text{PRE} = r^2
$$

Try it out.

```
cor(city[c("income", "edu")])^2
```

```
          income       edu
income 1.0000000 0.6316828
edu    0.6316828 1.0000000
```

We would report this as $R^2 = .632$.

For some reason, the notation we use when reporting the correlation coefficient uses a lower-case r, while the notation for reporting the square of this value uses upper-case, R^2.

R^2, like the correlation coeffcient, is related to the strength of the linear relationship. Variables that have stronger linear relationships have a higher r value and thus higher R^2 values. Higher R^2 means more reduction in error, which implies better predictions. In a sense, it quantifies how good the model is, and because of this, R^2 is often provided as an *effect size* for regression analyses.

Partitioning Variation

Using the SSE terms we can partition the total variation in Y (the SSE value from the intercept-only model) into two parts (1) the part that is explained by the model, and (2) the part that remains unexplained. The second part is just the SSE from the regression model that includes X. Here is the partitioning of the variation in income.

$$\underbrace{6,565,527,426}_{\substack{\text{Total}\\\text{Variation}}} = \underbrace{4,147,329,600}_{\substack{\text{Explained}\\\text{Variation}}} + \underbrace{2,418,197,826}_{\substack{\text{Unexplained}\\\text{Variation}}}$$

Each of these three terms is a sum of squares (SS). The first is refereed to as the total sum of squares, as it represents the total amount of variation in Y. The second term is commmonly called the regression sum of squares or model sum of squares, as it represents the variation explained by the model. The last term is the residual sum of squares (or error sum of squares) as it represents the left-over variation that is unexplained by the model.

More generally,

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Model}} + \mathrm{SS}_{\mathrm{Error}}.$$

Since the $\mathrm{SS}_{\mathrm{Model}}$ represents the explained variation, we can express that as a proportion of the total variation by dividing by the $\mathrm{SS}_{\mathrm{Total}}$. This ratio is R^2,

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Model}}}{\mathrm{SS}_{\mathrm{Total}}}.$$

Go back to the equation partitionuing of the sums of squares, and divide each term by $\mathrm{SS}_{\mathrm{Total}}$.

$$\frac{\mathrm{SS}_{\mathrm{Total}}}{\mathrm{SS}_{\mathrm{Total}}} = \frac{\mathrm{SS}_{\mathrm{Model}}}{\mathrm{SS}_{\mathrm{Total}}} + \frac{\mathrm{SS}_{\mathrm{Error}}}{\mathrm{SS}_{\mathrm{Total}}}$$

Re-expressing some of these terms we get,

$$1 = R^2 + \frac{\mathrm{SS}_{\mathrm{Error}}}{\mathrm{SS}_{\mathrm{Total}}}$$

Then, solving for the unexplained part, we get

$$\frac{\mathrm{SS}_{\mathrm{Error}}}{\mathrm{SS}_{\mathrm{Total}}} = 1 - R^2$$

So in our example, $R^2 = 0.632$, 63.2/% of the variation in incomes was explained by the model. This implies that $1 - 0.632 = 0.368$, or 36.8% of the variation in income is unexplained by the model.

References

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