Least Squares Estimation

2017-01-20

In this set of notes, you will learn how the coefficients from the fitted regression equation are estimated from the data. Recall that in the previous set of notes, we used the *riverside.csv* data to examine whether education level is related to income.

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(ggplot2)

# Fit regression model
lm.1 = lm(income ~ 1 + education, data = city)
lm.1
```

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

Coefficients:
(Intercept)    education
      11321         2651
```

The fitted regression equation is

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

How does R determine the coefficient values of $\hat{\beta}_0 = 11,321$ and $\hat{\beta}_1 = 2,651$? These values are estimated from the data using a method called *Ordinary Least Squares* (OLS). To understand how OLS works, consider the following toy data set of five observations:

Table 1: Toy data set with predictor (x) and outcome (y) for $n = 5$ observations.

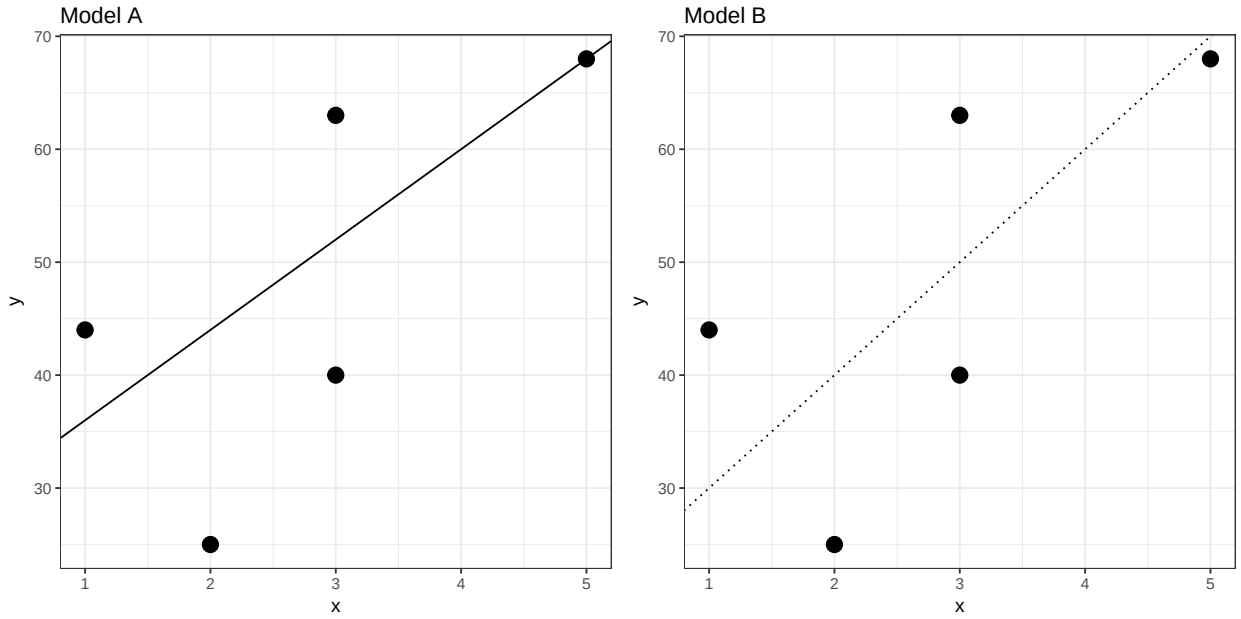
x	y
3	63
1	44
3	40

x	y
5	68
2	25

Which of the following two models fits these data better?

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

We could plot the data and both lines and try to determine which seems to fit better.



Data–Model Fit

In this case, the lines are similar and it is difficult to make a determination of which fits the data better by eyeballing the two plots. Instead of guessing which model fits better, we can actually quantify the fit to the data by computing the residuals (errors) for each model and then compare both sets of residuals; larger errors indicate a worse fitting model (i.e., more misfit to the data).

Remember, to compute the residuals, we will first need to compute the predicted value ($\hat{Y}_i$) for each of the five observations for both models.

Table 2: Predicted values and residuals for Model A

x	y	Predicted	Residual
3	63	52	11
1	44	36	8
3	40	52	-12
5	68	68	0
2	25	44	-19

Table 3: Predicted values and residuals for Model B

x	y	Predicted	Residual
3	63	50	13
1	44	30	14
3	40	50	-10
5	68	70	-2
2	25	40	-15

Eyeballing the numeric values of the residuals is also problematic. The size of the residuals is similar for both Models. Also, the eyeballing method would be impractical for larger datasets. So, we have to further quantify the model 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.

$\text{SSE}_{\text{Model A}} =$

$\text{SSE}_{\text{Model B}} =$

Once we have quantified the model misfit, we can choose the model that has the least amount of error. Since Model A has a lower SSE than Model B, we would conclude that Model A is the better fitting model to the data.

"Best" Fitting Model

In the vocabulary of statistical estimation, the process we just used to adopt Model A over Model B was composed of two parts:

- **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.)

In our example we used the SSE as the quantification of model fit, and then we optimized by selecting the model with the lower SSE. When we use `lm()` to fit a regression analysis to the data, R needs to consider not just two models like we did in our example, but all potential models (i.e., any intercept and slope). The model coefficeints that `lm()` returns are the "best" in that no other values for intercept or slope would produce a lower SSE. The model returned has the lowest SSE possible . . . thus *least squares*. 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:

Table 4: Predicted values and residuals for the 'best' fitting model.

x	y	Predicted	Residual
3	63	49.61	13.39
1	44	33.48	10.52
3	40	49.61	-9.614
5	68	65.75	2.25
2	25	41.55	-16.55

The SSE is 661.16. This is the smallest SSE possible for a linear model. Any other value for the slope or intercept would result in a higher SSE.

Optimization

Finding the the intercept and slope that give the lowest SSE is referred to as an optimization problem in the field of mathematics. Optimization is such an important (and sometimes difficult) probelm that there have been several mathematical and computational optimization methods that have been developed over the years. You can read more about mathematical optimization on Wikipedia if you are interested.

One common mathematical method to find the minimum SSE involves calculus. We would write the SSE as a function ofβ_0 and β_1, compute the partial derivatives (w.r.t. each of the coefficients), set these equal to zero, and solve to find the values of the coefficients. You can read here. The `lm()` function actually uses an optimization method called QR decomposition to obtain the regression coefficients. The actual mechanics and computation of these methods are beyond the scope of this course. We will just trust that the `lm()` function is doing things correctly in this course.

Computing the SSE

Since the regression model is based on the lowest SSE, it is often useful to compute and report the model's 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$education
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]   4920  -6099   9203  -3649 -12352   3355  -5477   7132   9347 -15804
[11]   1533  16491 -13784   2045   9734 -15151   2141   5762   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]  24206400  37197801  84695209  13315201 152571904  11256025  29997529
 [8]  50865424  87366409 249766416   2350089 271953081 189998656   4182025
[15]  94750756 229552801   4583881  33200644  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$income - (11321 + 2651 * city$education)) ^ 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 often report the SSE, but *we do not interpret the actual value.* The value of the SSE is more useful 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

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
```

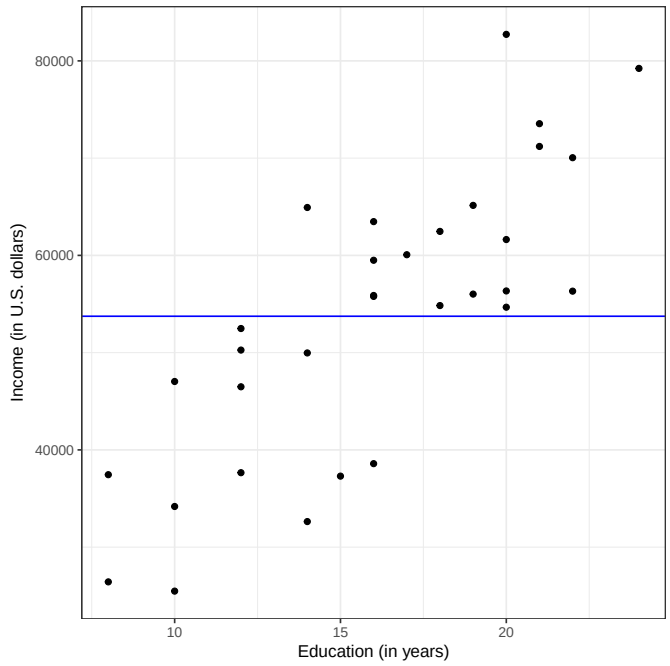
```
Call:
lm(formula = income ~ 1, data = city)

Coefficients:
(Intercept)
      53742
```

The fitted regression equation for the intercept-only model can be written as,

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

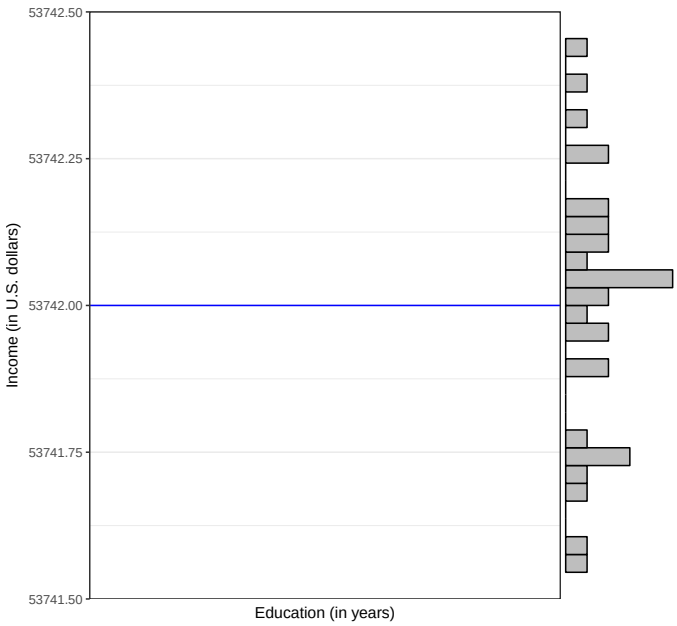
Graphically, the fitted line is a flat line crossing the y-axis at 53,742 (see plot below).



Does the estimate for β_0, 53,742, seem familiar? If not, go back to the exploration of the response variable earlier in the notes. The estimated intercept in the intercept-only model is the mean value of the response variable. 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 Baseline

Since the intercept-only model does not include any predictors, the SSE is a quantification of the total variation in the response variable. It can be used as baseline measure of the error variation in the data. 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
```

Proportion Reduction in Error

The SSE for the intercept-only model 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 accounts for 63.2% of the variation in incomes.

Since the model uses the predictor of education level, tt is also common for researchers to interpret this value using the language,

> Differences in education level account 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.

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

```
             income education
income    1.0000000 0.6316828
education 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,

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

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

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

Go back to the equation partitionuing of the sums of squares, and divide each term by SS_{Total}.

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

Re-expressing some of these terms we get,

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

Then, solving for the unexplained part, we get

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

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.

Resources

- 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/