optimizingViaSSE

September 16, 2019

0.1 Minimizing SSE

0.1.1 optimal within a model

We can take the idea that a smaller SSE suggests a better model fit further. Instead of using SSE to compare different models, we can use the SSE to evaluate different parameter values inside the same model.

Consider the same dataset as above and suppose we're fitting a simple linear regression. Then our SSE becomes

$$\text{SSE}(y, \hat{y}_i) = \sum_{i=1}^{N} [y_i - \hat{y}_i]^2 \tag{1}$$

$$\text{SSE}(y, x, \beta_0, \beta_1) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2}$$

$$\tag{3}$$

where y and x are vectors of data. Step two in above equation replaced the predicted value $\hat{y}$ with the linear model used to make this prediction $\beta_0 + \beta_1 x$.

Now our SSE is a function of the data, that cannot be changed, and the parameters of our model β_0 and β_1. Changing β_0 or β_1 will change the value of the SSE. One way to find a best fit model is to find those parameters value that make the SSE as small as possible.

0.1.2 derivative

SSE is a function of β_0 and β_1, and can be optimized by taking the derivative with respect to both parameters and finding the point where the derivative of these two equations equals zero simultaneously.

We take the derivative with respect to β_0

$$\frac{d\text{SSE}(\beta_0, \beta_1)}{d\beta_0} = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{4}$$

$$\frac{d\text{SSE}(\beta_0, \beta_1)}{d\beta_0} = \sum_{i=1}^{N} \frac{d}{d\beta_0} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{5}$$

$$\frac{d\text{SSE}(\beta_0, \beta_1)}{d\beta_0} = \sum_{i=1}^{N} -2[y_i - (\beta_0 + \beta_1 x_i)] \tag{6}$$

$$\tag{7}$$

The above derivative can be set to zero and solved for β_0, our variable.

$$\sum_{i=1}^{N} -2[y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{8}$$

$$\sum_{i=1}^{N} y_i - N\beta_0 - \beta_1 \sum_{i=1}^{N} x_i = 0 \tag{9}$$

$$N\beta_0 = \sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i \tag{10}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{11}$$

$$\tag{12}$$

The value for β_0 that optimizes the SSE is the average of our y values minus the optimal β_1 times the average of our x values.

We must also take the derivative with respect to β_1.

$$\frac{d\text{SSE}(\beta_0, \beta_1)}{d\beta_1} = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{13}$$

$$\frac{d\text{SSE}(\beta_0, \beta_1)}{d\beta_1} = \sum_{i=1}^{N} -2x_i[y_i - (\beta_0 + \beta_1 x_i)] \tag{14}$$

$$\tag{15}$$

The above equation can also be set to zero and solved for β_1.

$$\sum_{i=1}^{N} -2x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{16}$$

$$\sum_{i=1}^{N} x_i y_i - x_i \beta_0 - \beta_1 x_i^2 = 0 \tag{17}$$

$$\tag{18}$$

At this point we can substitute the optimal value for β_0 we derived.

$$\sum_{i=1}^{N} x_i y_i - x_i(\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i^2 = 0 \tag{19}$$

$$\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \bar{y} + \sum_{i=1}^{N} x_i \beta_1 \bar{x} - \beta_1 x_i^2 = 0 \tag{20}$$

$$\beta_1 \left(x_i^2 - \sum_{i=1}^{N} x_i \bar{x} \right) = \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \bar{y} \tag{21}$$

$$\beta_1 = \frac{\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \bar{y}}{\left(x_i^2 - \sum_{i=1}^{N} x_i \bar{x} \right)} \tag{22}$$

This equation for β_1 doesn't look like anything we can recognize, but we can change the SSE we optimized to make this equation look more familiar. The equation we optimized was a function of β_0 and β_1, and so adding a constant value that does not include β_0 or β_1 would not change the optimal β.

From each data point, lets subtract $\bar{x}$ and $\bar{y}$, called centering our data. Then the above equation becomes

$$\beta_1 = \frac{\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \bar{y}}{\left(x_i^2 - \sum_{i=1}^{N} x_i \bar{x} \right)} \tag{23}$$

$$= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^{N} (x_i - \bar{x}) \bar{y}}{\sum_{i=1}^{N} (x_i - \bar{x})^2 - \bar{x} \sum_{i=1}^{N} (x_i - \bar{x})} \tag{24}$$

$$= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{25}$$

$$= \frac{Cov(X, Y)}{Var(X)} \tag{26}$$

Centering our data, we see the optimal β_1 is the covariance between y and x divided by the variance of x.

We can also right the above in matrix form. The covariance between X and Y is written

$$Cov(X, Y) = X'y$$

where $X = x - \bar{x}$ and $Y = y - \bar{y}$, and the variance of X is written

$$Var(X) = X'X.$$

Then the expression for β_1 is

$$\beta_1 = (X'X)^{-1}(X'y)$$

But by adding a column of 1s to X, we can see that the above expression works for both β_1 and β_0. In fact, this expression will work for any design matrix X.

So we can write

$$\beta = (X'X)^{-1}(X'y)$$

To see this more clearly, let's generalize our derivations of β_0 and β_1 to multiple βs.

We first form our SSE for multiple linear regression

$$\text{SSE}(y, X, \beta_0, \beta_1) = \sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in})]^2$$

where x_{ij} is observation i for variable j. Taking the derivative for every β and setting equal to 0 we have

For β_0

$$\sum_{i=1}^{N} 1[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in})] = 0 \tag{27}$$

$$\sum_{i=1}^{N}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in}) = \sum_{i=1}^{N} 1 y_i \tag{28}$$

$$\tag{29}$$

For β_1

$$\sum_{i=1}^{N} x_{i1}[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in})] = 0 \tag{30}$$

$$\sum_{i=1}^{N} x_{i1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in}) = \sum_{i=1}^{N} x_{i1} y_i \tag{31}$$

$$\tag{32}$$

For β_2

$$\sum_{i=1}^{N} x_{i2}[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in})] = 0 \tag{33}$$

$$\sum_{i=1}^{N} x_{i2}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in}) = \sum_{i=1}^{N} x_{i2} y_i \tag{34}$$

$$\tag{35}$$

For β_n

$$\sum_{i=1}^{N} x_{in}[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in})] = 0 \tag{36}$$

$$\sum_{i=1}^{N} x_{in}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_n x_{in}) = \sum_{i=1}^{N} x_{in} y_i \tag{37}$$

$$\tag{38}$$

The right hand side of this system of equations can be rewritten as

$$X'y,$$

and the left hand side can be rewritten as

$$(X'X)\beta.$$

Our system of equations is then

$$(X'X)\beta = X'y.$$

We can solve for β by left multiplying each side by $(X'X)^{-1}$

$$\beta = (X'X)^{-1}X'y$$

and arriving at the same solution for multiple linear regression that we found with simple linear regression.

We can verify that the above equation $(X'X)^{-1}X'y$ recovers the optimal βs using R.

[4]:
```
data <- read.csv('polynomialData.csv')
head(data)

cubicRegression <- lm(y~x+I(x^2)+I(x^3),data=data)

print("CUBIC REGRESSION")
print(cubicRegression)

y <- data$y
ones = rep(1,length(data$x))
x    = data$x
x2   = data$x^2
x3   = data$x^3

print("DESIGN MATRIX")
```

```
X = cbind(ones,x,x2,x3)
print(head(X))

print("OPTIMAL BETAS")
optimalBetas <- solve(t(X)%*%X,t(X)%*%y)  # this is the same as solving our␣
  ↪system of equatons.
print(round(optimalBetas,4))
```

A data.frame: 6 ÃŮ 2

x <dbl>	y <dbl>
0.9958723	8.2420054
-0.6556163	2.3114202
-0.9176787	4.0842076
0.1963727	-5.5386897
1.0309346	2.5166174
1.2610719	-0.5388713

```
[1] "CUBIC REGRESSION"

Call:
lm(formula = y ~ x + I(x^2) + I(x^3), data = data)

Coefficients:
(Intercept)            x       I(x^2)       I(x^3)
     0.2994       2.2877       1.0962      -1.4120

[1] "DESIGN MATRIX"
     ones          x         x2           x3
[1,]    1  0.9958723 0.99176166  0.987667975
[2,]    1 -0.6556163 0.42983269 -0.281805304
[3,]    1 -0.9176787 0.84213426 -0.772808705
[4,]    1  0.1963727 0.03856225  0.007572574
[5,]    1  1.0309346 1.06282620  1.095704329
[6,]    1  1.2610719 1.59030227  2.005485460
[1] "OPTIMAL BETAS"
        [,1]
ones  0.2994
x     2.2877
x2    1.0962
x3   -1.4120
```

The beta coefficients we found from running a cubic regression match the beta coefficients from solving the system of equations above that minimize the SSE.