Some Theory Underlying Simple Linear Regression

2018-09-11

In these notes, we will examine the computations needed to derive the OLS simple regression coefficients, which can be computed using the following formulas:

$$
\begin{aligned}
B_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
B_0 &= \bar{y} - B_1(\bar{x})
\end{aligned}
$$

We will also mathematically develop several useful properties of the OLS simple regression model. These properties include:

- The OLS regression line passes through the means of both X and Y;
- The average value of the residual is zero;
- The residual errors around the least-squares regression are uncorrelated with the predictor variable X; and
- The residual errors around the least-squares regression are also uncorrelated with the fitted values, $\hat{Y}$.

Simple Regression Model

To begin, we will define the simple linear regression model, along with the concept of fitted values and residuals. The *population model* for simple linear regression, which uses Greek letters for the parameters, is

$$
y_i = \beta_0 + \beta_1(x_i) + \epsilon_i.
$$

We use Roman letters to denote the parameter estimates in the *sample model*, namely,

$$
y_i = B_0 + B_1(x_i) + e_i.
$$

Fitted Values and Residuals

The fitted value, $\hat{y}_i$, is

$$
\hat{y}_i = B_0 + B_1(x_i),
$$

which implies that,

$$
y_i = \hat{y}_i + e_i.
$$

Solving this for e_i, we get

$$
\begin{aligned}
e_i &= y_i - \hat{y}_i \\
&= y_i - \bigg(B_0 + B_1(x_i)\bigg) \\
&= y_i - B_0 - B_1(x_i)
\end{aligned}
$$

Empirical Example: Occupational Prestige

Here we will use the data in *duncan.csv* to illustrate the computation of the OLS simple regression coefficients and also verify several of the properties listed above. To do this, we will regress `prestige` on `income`.

```r
# Load libraries
library(broom)
library(dplyr)
library(readr)

# Read in Duncan data
duncan = read_csv("../data/duncan.csv")
head(duncan)
```

```
## # A tibble: 6 x 5
##   occupation type  income education prestige
##   <chr>      <chr>  <int>     <int>    <int>
## 1 accountant prof      62        86       82
## 2 pilot      prof      72        76       83
## 3 architect  prof      75        92       90
## 4 author     prof      55        90       76
## 5 chemist    prof      64        86       90
## 6 minister   prof      21        84       87
```

```r
# Fit regression model
lm_1 = lm(prestige ~ 1 + income, data = duncan)
coef(lm_1) # Obtain coefficient estimates
```

```
## (Intercept)      income
##    2.456574    1.080390
```

Let's first verify the computational formulas for the OLS simple regression coefficients.

```r
x = duncan$income
x_bar = mean(x)

y = duncan$prestige
y_bar = mean(y)


# Compute B_1
B_1 = sum( (x - x_bar) * (y - y_bar) ) / sum( (x - x_bar)^2 )
B_1
```

```
## [1] 1.08039
```

```r
# Compute B_0
B_0 = y_bar - B_1 * x_bar
B_0
```

```
## [1] 2.456574
```

Now we can also verify some of the properties of the OLS model.

```
y_hat = fitted(lm_1) #Obtain fitted values
e = resid(lm_1)      #Obtain residuals

# Regression lines passes through (x_bar, y_bar)
B_0 + B_1 * x_bar
```

```
## [1] 47.68889
```

```
y_bar
```

```
## [1] 47.68889
```

```
# Average residual is zero
mean(e) #Zero within rounding
```

```
## [1] 0
```

```
# Residuals are uncorrelated with X
cor(e, x) #Zero within rounding
```

```
## [1] -4.335484e-17
```

```
# Residuals are uncorrelated with y_hat
cor(e, y_hat) #Zero within rounding
```

```
## [1] -8.475441e-17
```

We can always verify certain properties and results using empirical data, but that often unsatisfactory. For example, does the property that the residuals have an average value of zero hold for all data sets? Or only for the data we verified the result on? Does the result of -4.33×10^{-17} really imply that the residuals are *uncorrelated* with X $(r = 0)$? Or do they just have a really weak relationship?

Although empirically verifying results are a good first step, we need to do more than empirical verification on a single data set. This is where mathematics, especially the ideas of mathematical proof will help. The goal in the remainder of the notes is to derive these results more generally using mathematics.

Quantifying Model-Data Fit

In general, we have good fit when the residuals are generally small. This might lead you to define a good model as having minimal residuals,

$$\sum e_i = 0,$$

After all, the sum of the residuals quantifies the total model-data misfit, and the smallest we can make this sum is zero.

One problem with this is that any line that passes through the observation $(\bar{x}, \bar{y})$ will have a sum of residuals that is equal to zero; $\sum e_i = 0$. To prove this, we begin with the regression model,

$$y_i = B_0 + B_1(x_i) + e_i$$

Then we make use of the fact that if a line passes through $(\bar{x}, \bar{y})$, it will satisfy the following fitted equation, $\bar{y} = B_0 + B_1(\bar{x})$. We can subtract one of these quantities from both sides of the equation to maintain the equality in the regression model.

$$\begin{aligned}
y_i - \bar{y} &= B_0 + B_1(x_i) + e_i - \Bigg[B_0 + B_1(\bar{x})\Bigg] \\
y_i - \bar{y} &= B_1\Bigg[x_i - \bar{x}\Bigg] + e_i \\
e_i &= y_i - \bar{y} - B_1\Bigg[x_i - \bar{x}\Bigg]
\end{aligned}$$

Since we are interested in the sum of the residuals, we want to sum the left-hand side. However, to maintain the equality, we sum both sides.

$$\begin{aligned}
\sum e_i &= \sum \Bigg(y_i - \bar{y} - B_1\Bigg[x_i - \bar{x}\Bigg]\Bigg) \\
&= \sum y_i - \sum \bar{y} - \sum B_1\Bigg[x_i - \bar{x}\Bigg] \\
&= \sum y_i - n\bar{y} - B_1\Bigg[\sum x_i - \sum \bar{x}\Bigg] \\
&= \sum y_i - n\bar{y} - B_1\Bigg[\sum x_i - n\bar{x}\Bigg]
\end{aligned}$$

But, $\bar{x} = \frac{\sum x_i}{n}$, which means $n\bar{x} = \sum x_i$. Thus,

$$\begin{aligned}
\sum e_i &= \underbrace{\sum y_i - n\bar{y}}_{0} - B_1\Bigg[\underbrace{\sum x_i - n\bar{x}}_{0}\Bigg] \\
&= 0
\end{aligned}$$

So, if we use the criteria, $\sum e_i = 0$, this leads to an infinite number of solutions, since any line passing through $(\bar{x}, \bar{y})$ will satisfy this criteria. Also, large negative residuals are just as bad as a large positive residuals when measuring misfit. Because of this, we need to be more specific about how we combine (sum) the residuals. There are two solutions that would seem to fix this problem.

- Minimize $\sum |e_i|$; Least absolute value regression
- Minimize $\sum e_i^2$; Least squares regression

Least Squares Optimization

Using the sample regression model,

$$\sum e_i^2 = \sum \Big[y_i - B_0 - B_1(x_i)\Big]^2$$

We can write the sum of squared residuals as a function of the parameter estimates,

$$f(B_0, B_1) = \sum \Big[y_i - B_0 - B_1(x_i)\Big]^2$$

The problem of least squares regression is to find the parameter estimates, B_0 and B_1, which minimize the sum of squared residuals. Mathematically, we can take the partial derivatives of $f(B_0, B_1)$ with respect to B_0 and B_1; set those partial derivatives equal to zero, and solve for B_0 and B_1. After taking the partial derivatives, we get the following:

$$\begin{aligned}\frac{\partial f(B_0, B_1)}{\partial B_0} &= \sum(-1)(2)\Big(y_i - B_0 - B_1(x_i)\Big)\\ \frac{\partial f(B_0, B_1)}{\partial B_1} &= \sum(-x_i)(2)\Big(y_i - B_0 - B_1(x_i)\Big)\end{aligned}$$

(If you have previously taken a calculus course, verify this. If not, take my word for it.) These equations are then set equal to zero.

$$\begin{aligned}0 &= \sum(-1)(2)\Big(y_i - B_0 - B_1(x_i)\Big)\\ 0 &= \sum(-x_i)(2)\Big(y_i - B_0 - B_1(x_i)\Big)\end{aligned}$$

Now we can use our summation rules to simplify these equations. Starting with the first equation:

$$\begin{aligned}0 &= \sum(-1)(2)\Big(y_i - B_0 - B_1(x_i)\Big)\\ &= -2\sum\Big(y_i - B_0 - B_1(x_i)\Big)\\ &= -2\Big(\sum y_i - \sum B_0 - \sum B_1(x_i)\Big)\\ &= -2\Big(\sum y_i - \sum B_0 - B_1\sum(x_i)\Big)\\ &= \sum y_i - \sum B_0 - B_1\sum(x_i)\\ &= \sum y_i - nB_0 - B_1\sum(x_i)\end{aligned}$$

And, re-arranging this, we get:

$$\sum y_i = nB_0 + B_1\sum(x_i)$$

Then, using summation rules on the second equation,

$$
\begin{aligned}
0 &= \sum(-x_i)(2)\Big(y_i - B_0 - B_1(x_i)\Big) \\
&= -2\sum x_i\Big(y_i - B_0 - B_1(x_i)\Big) \\
&= -2\sum\Big(x_iy_i - B_0x_i - B_1x_i^2\Big) \\
&= -2\Big(\sum x_iy_i - \sum B_0x_i - \sum B_1x_i^2\Big) \\
&= -2\Big(\sum x_iy_i - B_0\sum x_i - B_1\sum x_i^2\Big) \\
&= \sum x_iy_i - B_0\sum x_i - B_1\sum x_i^2
\end{aligned}
$$

Re-arranging this, we get:

$$\sum x_iy_i = B_0\sum x_i + B_1\sum x_i^2$$

Normal Equations and Derivation of the Coefficient Estimators

Now we have a system of two equations, referred to as the *normal equations*, with two unknowns, B_0 and B_1.

$$\sum y_i = nB_0 + B_1\sum(x_i)$$

$$\sum x_iy_i = B_0\sum x_i + B_1\sum x_i^2$$

We can use algebra to solve for B_0 and B_1. Here we will use the substitution method (see http://www.sosmath.com/soe/SE211105/SE211105.html) to solve the first normal equation for B_0 and then substitute this into the second normal equation.

$$
\begin{aligned}
\sum y_i &= nB_0 + B_1\sum(x_i) \\
\frac{\sum y_i}{n} &= \frac{nB_0 + B_1\sum(x_i)}{n} \\
\bar{y} &= B_0 + B_1(\bar{x}) \qquad \text{which means} \\
B_0 &= \bar{y} - B_1(\bar{x})
\end{aligned}
$$

Now we substitute this into the second normal equation.

$$\begin{aligned}
\sum x_i y_i &= B_0 \sum x_i + B_1 \sum x_i^2 \\
&= \Big(\bar{y} - B_1(\bar{x})\Big) \sum x_i + B_1 \sum x_i^2 \\
&= \bar{y} \sum x_i - B_1(\bar{x}) \sum x_i + B_1 \sum x_i^2 \\
\sum x_i y_i - \bar{y} \sum x_i &= B_1 \Big(\sum x_i^2 - \bar{x} \sum x_i \Big) \\
&= B_1 \Big(\sum x_i^2 - \bar{x} n \bar{x} \Big) \\
&= B_1 \Big(\sum x_i^2 - n\bar{x}^2 \Big)
\end{aligned}$$

ASIDE: Here are two results which are useful.

$$\begin{aligned}
\sum (x_i - \bar{x})^2 &= \sum \Big(x_i^2 - 2(x_i)\bar{x} + \bar{x}^2 \Big) \\
&= \sum x_i^2 - \sum 2(x_i)\bar{x} + \sum \bar{x}^2 \\
&= \sum x_i^2 - 2\bar{x} \sum (x_i) + \sum \bar{x}^2 \\
&= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\
&= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum x_i^2 - n\bar{x}^2
\end{aligned}$$

Also,

$$\begin{aligned}
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum \Big(x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y} \Big) \\
&= \sum x_i y_i - \sum x_i \bar{y} - \sum y_i \bar{x} + \sum \bar{x}\bar{y} \\
&= \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{x}\bar{y} \\
&= \sum x_i y_i - \bar{y} n \bar{x} - \bar{x} n \bar{y} + n\bar{x}\bar{y} \\
&= \sum x_i y_i - n\bar{x}\bar{y} \\
&= \sum x_i y_i - \bar{y} \sum x_i
\end{aligned}$$

We can use these results in our substitution:

$$\begin{aligned}\sum x_i y_i - \bar{y}\sum x_i &= B_1\left(\sum x_i^2 - n\bar{x}^2\right)\\ &= B_1\sum (x_i-\bar{x})^2\\ \sum (x_i-\bar{x})(y_i-\bar{y}) &= B_1\sum (x_i-\bar{x})^2\\ B_1 &= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\end{aligned}$$

The OLS coefficient estimators are then:

$$\begin{aligned}B_0 &= \bar{y} - B_1(\bar{x})\\ B_1 &= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\end{aligned}$$

These coefficients are uniquely defined, so long as $\sum (x_i-\bar{x})^2 \neq 0$. The only time this value is zero is if all the $x_i = \bar{x}$; there is no variation in the predictor. If there is no variation in X, then there are infinite solutions; any line that passes through the point $(\bar{x}, \bar{y})$ would be a potential solution.

Implication: OLS Line Passes through the Point $(\bar{x}, \bar{y})$

Through solving the first normal equation we found that $\bar{y} = B_0 + B_1(\bar{x})$. This directly implies that the least squares regression line will always pass through $(\bar{x}, \bar{y})$ since $\bar{y}$ is the predicted value for an x-value of $\bar{x}$.

Implications: Sum and Average of the Residuals is Zero

Subsequently, the fact that the line passes through the observation $(\bar{x}, \bar{y})$ also implies that the sum of the residuals from the least squares equation will be zero; $\sum e_i = 0$ (shown earlier).

Once we recognize that the sum of the residuals is zero, the average residual must be zero.

$$\begin{aligned}\bar{e} &= \frac{\sum e_i}{n}\\ &= \frac{0}{n}\\ &= 0\end{aligned}$$

Residual Standard Error

Although the OLS line has the property that its estimates minimize the sum of squared errors, that does not mean that the line fits the data well. It is worth quantifying how well the OLS line actually fits the data.

One answer to this question is to compute the *standard error of the regression*, or the *residual standard error* (aka: *root mean square error*; RMSE).

$$\text{RMSE} = s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

This represents the "average" size of the residual in the metric of the Y-variable. For example, in the corrected Duncan data, RMSE = 17.4.

```
# Obtain RMSE
n = length(duncan$prestige)

rmse = sqrt( sum(e^2) / (n-2) )
rmse
```

```
## [1] 17.40275
```

On average, when using income to predict occupational prestige, we will have an error of roughly 17.4 prestige units. (If we can believe that the residuals are normally distributed, about 70% of the errors will be in the range of ± 17.4.)

Simple Correlation

Social scientists also use the correlation coefficient as a measure of fit. It is important to realize that although the RMSE is an absolute measure of the regression fit, the correlation coefficient provides a relative measure of the regression fit. This means, we can only interpret it as a measure of regression fit in relation to another model.

The baseline we typically use is the model with no predictor; X is not explanatory. This reduces our fitted model to:

$$y_i = B_0 + e_i$$

If we minimize the sum of squared errors for this model, we find that the OLS estimate for B_0 is $\bar{y}$. Thus,

$$\text{TSS} = \sum e_i^2 = \sum (y_i - \bar{y})^2$$

We refer to this value as the *sum of squares total*; SST. In contrast, the sum of squared residuals from the regression model that includes X as a predictor is referred to as the *residual sum of squares* or *sum of squares error*; SSE.

$$\text{RSS} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

The difference between these two values is what we refer to as the *regression sum of squares*; SSReg.

$$\text{SSReg} = \text{SST} - \text{SSE}$$

The SSReg indicates the reduction in the squared error of the residuals from the model with no predictors to the model that includes X as a predictor.

The ratio of SSReg to the SST is the square of the correlation coefficient.

$$r^2 = \frac{\text{SSReg}}{\text{SST}}$$

To compute the correlation coefficient, we take the positive square root of this ratio when $B_1 > 0$ and the negative square root when $B_1 < 0$. Let's examine the relationship between the sums of squares in more detail. To start we will begin with an individual observation, i.

$$\begin{aligned}
y_i &= y_i - \hat{y}_i + \hat{y}_i \\
y_i - \bar{y} &= y_i - \hat{y}_i + \hat{y}_i - \bar{y} \\
y_i - \bar{y} &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})
\end{aligned}$$

Now we square both sides of the equation

$$\begin{aligned}
(y_i - \bar{y})^2 &= \bigg[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\bigg]^2 \\
&= (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}$$

Now we can sum both sides, distributing this over the quantitites on the right-side.

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Since this last term is zero, we have

$$\begin{aligned}
\sum (y_i - \bar{y})^2 &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \\
\text{SST} &= \text{SSE} + \sum (\hat{y}_i - \bar{y})^2 \\
\text{SST} - \text{SSE} &= \sum (\hat{y}_i - \bar{y})^2 \\
\text{SSReg} &= \sum (\hat{y}_i - \bar{y})^2
\end{aligned}$$

This process of decomposing the variation into "explained" and "unexplained" variation is referred to as *ANOVA decomposition*, or simply *ANOVA*. The `anova()` function is used to carry out an ANOVA decomposition in practice.

```
anova(lm_1)
```

```
## Analysis of Variance Table
##
## Response: prestige
##           Df Sum Sq Mean Sq F value    Pr(>F)
## income     1  30665 30664.8  101.25 7.144e-13 ***
## Residuals 43  13023   302.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
r2 = 30665 / (30665 + 13023)
r2
```

```
## [1] 0.701909
```

```r
r = sqrt(r2)
r
```

```
## [1] 0.8378001
```

Correlation Coefficient: Take 2

The correlation can also be defined as the ratio of the covariance between two random variables and the product of their standard deviations:

$$\rho = \frac{\text{Cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

We define the sample covariance as

$$s_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Thus

$$\begin{aligned} r &= \frac{\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i-\bar{x})^2}{n-1} \times \frac{\sum (y_i-\bar{y})^2}{n-1}}} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})} \end{aligned}$$

```r
# Alternate computation of r
cov(x, y) / ( sd(x) * sd(y) )
```

```
## [1] 0.8378014
```

Notice if we multiply the correlation by the ratio $\frac{\text{SD}(Y)}{\text{SD}(x)}$ we get,

$$
\begin{aligned}
r \times \frac{\text{SD}(Y)}{\text{SD}(x)} &= \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1} \times \frac{\sum(y_i-\bar{y})^2}{n-1}}} \times \frac{\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}} \\
&= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \times \frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(x_i-\bar{x})^2}} \\
&= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \\
&= B_1
\end{aligned}
$$

Drawing on this result, we can also compute B_1 using:

$$
\begin{aligned}
B_1 &= r \times \frac{\text{SD}(Y)}{\text{SD}(x)} \\
&= \frac{\text{Cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} \times \frac{\text{SD}(Y)}{\text{SD}(x)} \\
&= \frac{\text{Cov}(X,Y)}{\text{SD}(X)} \times \frac{1}{\text{SD}(X)} \\
&= \frac{\text{Cov}(X,Y)}{\text{Var}(X)}
\end{aligned}
$$

```r
# Compute B_1
cov(x, y) / var(x)
```

```
## [1] 1.08039
```

Implications: Residuals are Uncorrelated with X and the Fitted Values

Now we can use the covariance rules to show that:

- The residuals from the least squares equation are *uncorrelated* with the x-values; and
- The residuals from the least squares equation are *uncorrelated* with the fitted-values.

First we will show that the residuals from the least squares equation are *uncorrelated* with the x-values. To do this, we need to recognize that if the covariance between e_i and x_i is zero, then the correlation between them will also be zero.

$$
\begin{aligned}
\text{Cov}(e_i, x_i) &= \text{Cov}\Bigg(y_i - B_0 - B_1(x_i), x_i\Bigg) \\
&= \text{Cov}(y_i, x_i) - \text{Cov}(B_0, x_i) - \text{Cov}\Bigg(B_1(x_i), x_i\Bigg) \\
&= \text{Cov}(y_i, x_i) - 0 - B_1\text{Cov}(x_i, x_i) \\
&= \text{Cov}(y_i, x_i) - B_1\text{Var}(x_i)
\end{aligned}
$$

Since $B_1 = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)}$, we can substitute this into the last expression.

$$
\begin{aligned}
\text{Cov}(e_i, x_i) &= \text{Cov}(y_i, x_i) - B_1\text{Var}(x_i) \\
&= \text{Cov}(y_i, x_i) - \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \times \text{Var}(x_i) \\
&= \text{Cov}(y_i, x_i) - \text{Cov}(x_i, y_i) \\
&= 0
\end{aligned}
$$

We can similarly show that the residuals from the least squares equation are *uncorrelated* with the fitted-values using the covariance rules.

References

Fox, J. (2016). *Applied regression analysis & generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.