Some Things about Simple OLS Regression

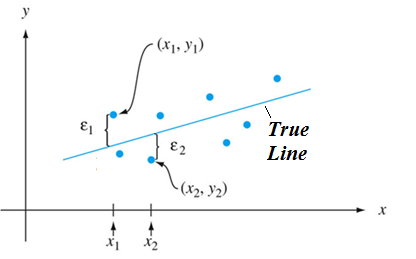
1. We assume a linear, non-deterministic (i.e., statistical) relationship between a dependent variable *y* and an independent (predictor) variable x
2. For the sake of the example, let’s say that y is the vocabulary (# words a child knows) and x is age (in years) of the child.
   1. We will assume that age and vocabulary are directly related.
3. Here is the data:

|  |  |  |
| --- | --- | --- |
| *i* | *Xi* | *Yi* |
| Person ID | Age of Child | Observed Vocabulary |
| 1 | 8 | 2800 |
| 2 | 3 | 800 |
| 3 | 4 | 1200 |
| 4 | 5 | 1700 |
| 5 | 6 | 1600 |
| 6 | 6 | 1800 |
| 7 | 2 | 300 |
| 8 | 7 | 2200 |
| 9 | 4 | 1300 |
| 10 | 5 | 1600 |

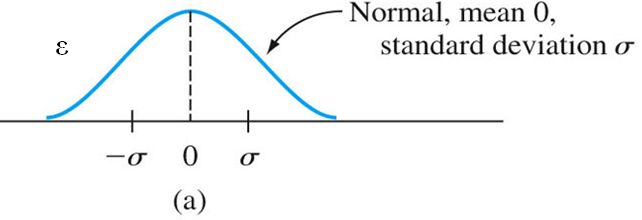
Note that y represents the observed value of the dependent variable (vocabulary) and yi represents the observed value of the dependent variable for observation i. For instance, y5 would be the vocabulary of person 5. Same notation holds for x (age).

1. Let’s plot the data:

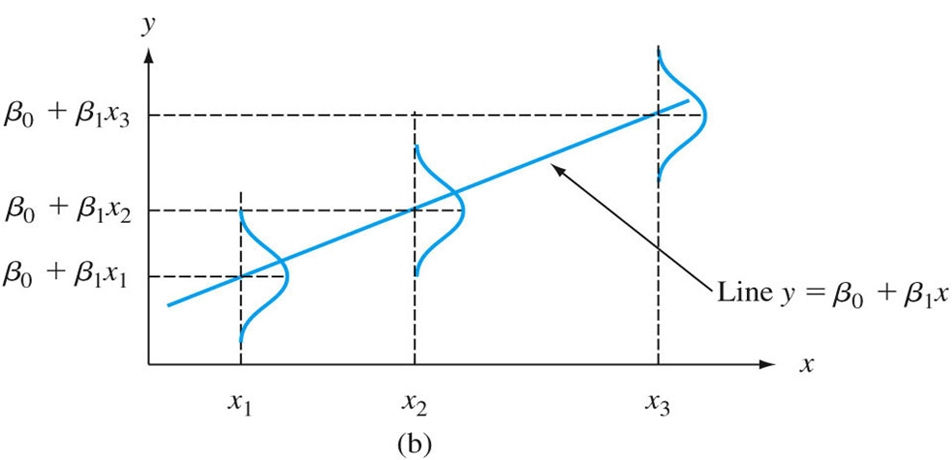
1. If we fit a line through the points, it will do a pretty good job summarizing the relationship between x(age) and vocabulary (y). We then assume that there is a *true* underlying linear relationship between x and y, and the observed data points differ from that line due to random error ().
   1. Thus, the relationship between x and y is described by a line with equation .
      1. For any x that we pick, the value of y on the true line may be thought of as the *true average* value of y for that x. For instance, if the value of x (age) is 5, the value of y (vocabulary) on that line can be thought of as the true average vocabulary of all 5 year olds.
      2. Formally speaking, for a simple linear regression model, for any fixed value of independent variable x, there are parameters , , and , such that , where .
      3. The term is known as an error term, and for each observation i, is defined as a vertical deviation (distance) between the observed value of y and the *true* value of y.
         1. These errors are something that we *cannot* predict, and are considered to be random. If the errors are not random, this becomes a problem for regression analysis.
      4. The inclusion of the random error term ε allows a point to fall either above the regression line (when ε > 0), or below the line (when ε <0)



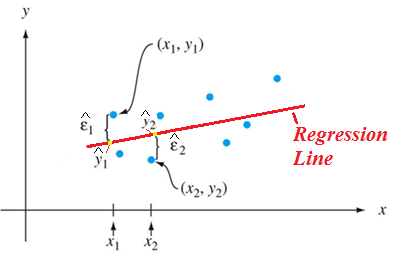
1. means that the error terms have a normal distribution with mean of 0 and variance . This holds for any given value of x (age), and implies that for any age, the average error term will be 0, and a typical deviation from the regression line will be units.



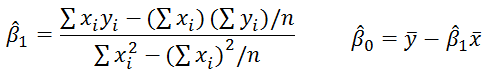
1. Again, the key thing to recall is that the above is true for ALL values of x: the distribution of is identical for any level of age. This way, for any x, the (normal) distribution of is centered at the y-value of the line (where = 0), and the variance of is always for any x. This assumption of constant variance of the error term (which is, unfortunately, often violated in practice), is something that we will discuss in more detail later, and is known as *homoscedasticity*.



1. But here’s the conundrum: we *do not* and *cannot* know the true line equation! Instead, we use our sample data to come up with an estimate of this true line. This estimate is called the regression line, and it has the equation . Here,
   1. is the predicted value of y, or the value of y that falls on the regression line.
   2. The term is a *point estimator* of the true population parameter , and the value of is a *point estimate* of the true population parameter .
   3. The term is a *point estimator* of the true population parameter , and the value of is a *point estimate* of the true population parameter .
   4. The term is known as a *residual*, and is a *point estimator* of the *error term* . Said mathematically,
   5. However, one of the difficulties is that a lot of statisticians write when they mean (that is, they omit the “hats”).
   6. Another thing is that the term residual and error are almost always used interchangeably in practice, though hard core statisticians will know that the difference is (see (d) above).



1. To summarize the above point, the goal is to use sample data to find the values and – that is, an estimate of the y-intercept and the slope of the line.
2. We are interested in values of and that minimize the sum of squared errors (SSE). That is, we want to make sure that is as small as possible. For that to happen, we perform some basic calculus and get the following:

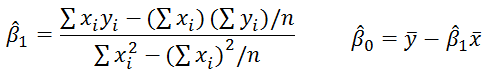


* 1. Here’s a step-by-step derivation of these using calculus, for those of you who are interested:
     1. <http://are.berkeley.edu/courses/EEP118/current/derive_ols.pdf>
  2. (As an aside, you should note that, since we’re dealing with sample data, if we look at (8) above, the theoretically correct term for SSE should be SSR (sum of squared residuals). However, some statisticians use SSR to denote regression sum of squares, which we will denote by SSM – model sum of squares – further down).

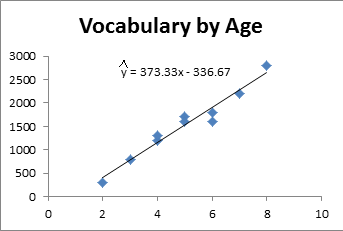
1. Let’s do it for our data. Here, n=# observations = 10.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Person # | Age of Child | Vocabulary | Age Squared | Age\*Vocabulary | Average Vocabulary | Average Age |
| *I* | *Xi* | *Yi* | *Xi2* | *Xi\*Yi* | |  | | --- | | ӯ | |  |
| 1 | 8 | 2800 | 64 | 22400 | 1530 | 5 |
| 2 | 3 | 800 | 9 | 2400 | 1530 | 5 |
| 3 | 4 | 1200 | 16 | 4800 | 1530 | 5 |
| 4 | 5 | 1700 | 25 | 8500 | 1530 | 5 |
| 5 | 6 | 1600 | 36 | 9600 | 1530 | 5 |
| 6 | 6 | 1800 | 36 | 10800 | 1530 | 5 |
| 7 | 2 | 300 | 4 | 600 | 1530 | 5 |
| 8 | 7 | 2200 | 49 | 15400 | 1530 | 5 |
| 9 | 4 | 1300 | 16 | 5200 | 1530 | 5 |
| 10 | 5 | 1600 | 25 | 8000 | 1530 | 5 |
| SUM | 50 | 15300 | 280 | 87700 | N/A | N/A |

1. Plugging everything into the formulas above, we get:



1. Let’s interpret this:
   1. means that for a 1 unit increase in x, y goes up by an average of 373.33 units. In the context of this problem, as a child’s age increases by a year, the average vocabulary increases by 373.33 words. Recall that is simply the slope of the regression line.
   2. means that when x is 0, y=-336.66. This is one of many real world examples where the y-intercept doesn’t make much sense, because it means that when age is 0, the vocabulary is -336.66. The results are nonsensical because it makes little sense to look at vocabulary of newborns. You might expect for there to be a linear relationship between vocabulary and age for kids ages 1-12, but we’re doing “extrapolation” for newborns, who are not in our dataset. In this case, we simply ignore the y-intercept.
2. Let’s plot the data, and calculate the predicted values of y for each x in our dataset, the residuals, and the SSE.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Person # | Age of Child | Observed Vocabulary | Predicted Vocabulary Formula | Predicted Vocabulary Value | Residual | Residual2 |
| *i* | *Xi* | *Yi* | |  | | --- | |  | |  |  |  |
| 1 | 8 | 2800 | -336.66+373.33 \* 8 | 2649.98 | 150.02 | 22506.0004 |
| 2 | 3 | 800 | -336.66+373.33 \* 3 | 783.33 | 16.67 | 277.8889 |
| 3 | 4 | 1200 | -336.66+373.33 \* 4 | 1156.66 | 43.34 | 1878.3556 |
| 4 | 5 | 1700 | -336.66+373.33 \* 5 | 1529.99 | 170.01 | 28903.4001 |
| 5 | 6 | 1600 | -336.66+373.33 \* 6 | 1903.32 | -303.32 | 92003.0224 |
| 6 | 6 | 1800 | -336.66+373.33 \* 6 | 1903.32 | -103.32 | 10675.0224 |
| 7 | 2 | 300 | -336.66+373.33 \* 2 | 410 | -110 | 12100 |
| 8 | 7 | 2200 | -336.66+373.33 \* 7 | 2276.65 | -76.65 | 5875.2225 |
| 9 | 4 | 1300 | -336.66+373.33 \* 4 | 1156.66 | 143.34 | 20546.3556 |
| 10 | 5 | 1600 | -336.66+373.33 \* 5 | 1529.99 | 70.01 | 4901.4001 |
| SUM | 50 | 15300 | N/A | N/A | 0 | SSE = 199667 |

1. Recall that SSE (which here is 199667) is the quantity that we’re trying to minimize (using our *least squares* method), and is basically the total variability of y that is unexplained by x (i.e., variability in vocabulary that is not explained by age).
2. When we have simple regression (regression with one predictor), , the variance of the error term ԑ, can be calculated by the formula . Here, = 199667/(10-2) = 24958.38. Then, = 157.98, meaning that the typical deviation from the regression line is 157.98 units (words). is sometimes denoted as RMSE (Root Mean Square Error) in regression outputs.
3. We can also calculate the total sum of squares (SST), which is the total variability of y’s in our dataset – that is the sum of squared deviations of y-values from (the mean of y).
   1. Mathematically,
   2. SST is a measure that is closely related to variance. In fact, variance of y can be calculated as
4. Let’s calculate SST:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Person # | Age of Child | Observed Vocabulary | Average Vocabulary | Difference between Yi and ӯ | Squared Difference between Yi and ӯ |
| *I* | *Xi* | *Yi* | ӯ | Yi-ӯ | |  | | --- | | (Yi-ӯ)2 | |
| 1 | 8 | 2800 | 1530 | 1270 | 1612900 |
| 2 | 3 | 800 | 1530 | -730 | 532900 |
| 3 | 4 | 1200 | 1530 | -330 | 108900 |
| 4 | 5 | 1700 | 1530 | 170 | 28900 |
| 5 | 6 | 1600 | 1530 | 70 | 4900 |
| 6 | 6 | 1800 | 1530 | 270 | 72900 |
| 7 | 2 | 300 | 1530 | -1230 | 1512900 |
| 8 | 7 | 2200 | 1530 | 670 | 448900 |
| 9 | 4 | 1300 | 1530 | -230 | 52900 |
| 10 | 5 | 1600 | 1530 | 70 | 4900 |
| SUM | 50 | 15300 | N/A | N/A | SST = 4381000 |

1. Now, let’s calculate the proportion of variance in vocabulary (y) that was explained by our model, which includes age (x). We know that:
   1. The total variance in y is SST = 4381000
   2. The variance in y that was *not* explained by x is SSE = 199667
   3. This means that the variance in y that was explained is SST-SSE = 4381000-199667 = 4181333. This is sometimes referred to as *model sum of squares (SSM)*.
      1. As stated earlier, sometimes this SSM is known as SSR (*regression sum of squares*)
   4. The proportion of total variance in y that *is* explained by the model can be calculated as SSM/SST = 4181333/4381000 = 0.9544.
   5. This proportion of total variance in y that *is* explained by the model is called the coefficient of determination, or R2. Again,
2. In simple regression (regression with 1 predictor), another way to calculate is to compute the value of the correlation coefficient R between x and y, and square it.
   1. Here, the correlation between age and vocabulary can be calculated in Excel using the ‘=Pearson (or ‘=Correl) function. You can see that the correlation is 0.9769. If you square 0.9769, you get 0.9544, which is our .
3. Let’s find all the stuff that we just computed manually in a regression output from R:

> fit <- lm(Vocabulary ~ ChildAge, data=simpleregdata)

> summary(fit)

Call:

lm(formula = Vocabulary ~ ChildAge, data = simpleregdata)

Residuals:

Min 1Q Median 3Q Max

-303.33 -96.67 30.00 125.00 170.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -336.67 152.63 -2.206 0.0585 .

ChildAge 373.33 28.84 12.943 0.0000012 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 158 on 8 degrees of freedom

Multiple R-squared: 0.9544, Adjusted R-squared: 0.9487

F-statistic: 167.5 on 1 and 8 DF, p-value: 0.000001202

> anova(fit)

Analysis of Variance Table

Response: Vocabulary

Df Sum Sq Mean Sq F value Pr(>F)

ChildAge 1 4181333 4181333 167.53 0.000001202 \*\*\*

Residuals 8 199667 24958

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Note: Here, SSE = 199,667 and SSR = 4,181,333. SST = SSE + SSR = 4,381,000