Recall the example where we looked to see whether age predicted the vocabulary of children. In that example, there were n=10 children, whose average age was 5 years. The data are copied below.

|  |  |  |
| --- | --- | --- |
| **Person #** | **Child Age** | **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 |
| Average | 5 | 1530 |

If we regress vocabulary on age, we get the following results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | -336.667 | 152.6252 | -2.20584 | 0.058459 |
| ChildAge | 373.3333 | 28.84345 | 12.94344 | 0.0000012 |

Here, the intercept is interpreted as the estimated value of vocabulary when the age of the child is 0. In other words, this model predicts an average 0-year-old to know -336.667 words (which is obviously nonsense). But the equation of the predicted regression line is:

, or

What would happen if we were to instead regress vocabulary on *centered* age? Centering the *age* variable would involve us subtracting the average age (5) from each child’s age, as is done below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Person #** | **Age** | **Centered Age** | **Vocabulary** |
| 1 | 8 | 3 | 2800 |
| 2 | 3 | -2 | 800 |
| 3 | 4 | -1 | 1200 |
| 4 | 5 | 0 | 1700 |
| 5 | 6 | 1 | 1600 |
| 6 | 6 | 1 | 1800 |
| 7 | 2 | -3 | 300 |
| 8 | 7 | 2 | 2200 |
| 9 | 4 | -1 | 1300 |
| 10 | 5 | 0 | 1600 |
| Average | 5 | 0 | 1530 |

The regression results are below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | 1530 | 49.95832 | 30.62553 | <0.0001 |
| CenteredAge | 373.3333 | 28.84345 | 12.94344 | 0.0000012 |

We can see that the slope () remains the same as in the first regression, but the y-intercept (), as well as its significance, are completely different. The Intercept term is now interpreted as the predicted value of vocabulary when *centered age* is 0 – which is what happens when the original age variable is 5. In other words, on average, 5-year-olds are estimated to have a vocabulary of 1530 words. Furthermore, 1530 is significantly different from 0, as we can see by looking at the p-value. This output makes much more sense than the predicted vocabulary of 0-year-olds. Here, the equation of the regression line is

, or

What if we wanted to use our regression models to predict the vocabulary of a 7-year old? In the original regression, where age has not been centered, we can do this by plugging in age = 7 into the regression equation to get

words.

In the second regression, where age has been centered, we first calculated centered age when age is 7, which is simply 7 - 5 = 2. We then plug in 2 into the regression equation to get

words.

So the predicted values of the dependent variable, as well as all model diagnostics (e.g., R2) remain the same regardless of centering.

We can also look at the scatter plots below. The second scatter plot simply shifts everything in the first by 5 units to the left (which is exactly what centering is).