I’ve used regression extensively and love it for all of its flexibility. You can use:

* multiple predictor variables
* continuous and categorical variables
* higher-order terms to model curvature
* interaction terms to see if the effect of one predictor depends upon the value of another
* Problems where the predictors seem enmeshed together like spaghetti.

Suppose you’re a researcher and you are studying a question that involves intertwined predictors. For example, you want to determine:

* how exercise habits and diet effect weight
* how drinking coffee and smoking cigarettes are related to heart disease

These are all research questions where the predictors are likely to be correlated with each other and they could all influence the response variable. How do you separate out the effects? How do you determine which variables are significant and how large of a role does each one play? Regression comes to the rescue!

**You Must Control Everything! (Or at least the important variables)**

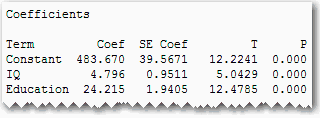
Multiple regression estimates how the changes in each predictor variable relate to changes in the response variable. Importantly, regression automatically controls for every variable that you include in the model. It means that when you look at the effect of one variable in the model, you are holding constant all of the other predictors in the model. You explain the effect that changes in one predictor have on the response without having to worry about the effects of the other predictors. In other words, you can isolate the role of one variable from all of the others in the model. And, you do this simply by including the variables in your model. It's beautiful!

For instance, a recent study assessed how coffee consumption affects mortality. Initially, the results showed that higher coffee consumption is correlated with a higher risk of death. However, many coffee drinkers also smoke. After the researchers included a variable for smoking habits in their model, they found that coffee consumption *lowered*the risk of death while smoking increased it. So, by including coffee consumption, smoking habits, and other important variables, the researchers held everything that is important constant and were able to focus on the role of coffee consumption.

Take note, this study also illustrates how not including an important variable (leaving it uncontrolled) can completely mess up your results.

**What to Look For in the Regression Output**

After you fit and verify that you have a good model, all you need to do is look at the p-value and coefficient for each predictor. If the p-value is low (usually < 0.05), the predictor is significant. Coefficients represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. For example, if your response variable is income and your predictors include IQ and education (among other relevant predictors), you might see output like this:



The p-values indicate that both IQ and education are significant. The IQ coefficient shows that an increase of one IQ point increases your earnings by an average of around $4.80, holding everything else in the model constant. Further, an increase in 1 unit of education increases your earnings by $24.22.

**How To Get Results That You Can Trust**

* [Include all of the important variables in your model.](https://blog.minitab.com/blog/adventures-in-statistics/collecting-good-data-its-a-messy-world-confound-it)  Leaving out important variables leaves them uncontrolled and can bias your coefficients (i.e., they’re probably wrong). E.g. In a regression with activity was the predictor and bone density was the response, I found that activity was **not** significant. It turns out that we didn’t include an important variable: the subject’s weight. The results for the variables you include can be biased by the significant variables that you don’t include.
* You should have good measures for the included variables, or at least include [proxy variables](https://blog.minitab.com/blog/adventures-in-statistics/proxy-variables-the-good-twin-of-confounding-variables) for those that are hard to measure.
* Check your [residual plots](https://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis) to make sure that your model fits your data.

**Correlated Predictors**

As we’ve seen, regression analysis can handle predictors that are correlated, also known as mullticollinearity. Moderate multicollinearity may not be a problem. However, severe multicollinearity is problematic because it can increase the variance of the regression coefficients, making them unstable and difficult to interpret.

For example, IQ and education are probably correlated, as is drinking coffee and smoking. As long as they aren’t excessively correlated, it’s not a problem. How do you know? VIFs are your friend! Variance inflation factor (VIF) is an easy to use measure of multicollinearity. VIF values greater than 10 may indicate that multicollinearity is unduly influencing your regression results. If you see high VIF values, you may want to remove some of the correlated predictors from your model.

**Closing Thoughts and the Bonus Tip!**

Regression gives you the power to separate out the effects of even tricky research questions. You can unravel the intertwined spaghetti noodles by holding all relevant variables constant and seeing the role that each plays.

Imagine you’re in the process of finding the proper regression model for your data. You have many variables, and you’ve included the terms for curvature and interactions. You’re reducing your model down to just the significant terms and checking the residual plots along the way. The result is a lot of output in the session window and many graphs. It can be difficult to find the specific plots for any given regression model.

Proxy variable - For example, my colleague, Kevin, does an excellent job using regression analysis to assist those who play fantasy football. Recently, he used a model that included one predictor variable -- each player’s fantasy football points from the prior season -- to predict his points for the subsequent season. Clearly, the points from one season are not causing the points for the next season. Rather, the points are a proxy variable for a host of other variables such as each player’s skills and capabilities, those of their team, the teams they play against, etc. It’s impossible to measure all of these, so a proxy variable is essential. His model for choosing quarterbacks has an r-squared of 73.68%. In this case, there is enough of a correlation from one year to the next that he can use the model for prediction, even though we don’t know or measure the exact causal variables.

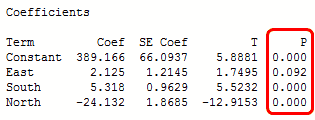
<https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients>

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables and the response variable.

## How Do I Interpret the P-Values in Linear Regression Analysis?

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger p-value suggests that changes in the predictor are not associated with changes in the response.

In the output below, we can see that the predictor variables of South and North are significant because both of their p-values are 0.000. However, the p-value for East (0.092) is greater than the common alpha level of 0.05, which indicates that it is not statistically significant.



Typically, you use the coefficient p-values to determine which terms to keep in the regression model. In the model above, we should consider removing East.

## How Do I Interpret the Regression Coefficients for Linear Relationships?

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This [statistical control](https://blog.minitab.com/blog/adventures-in-statistics/a-tribute-to-regression-analysis) that regression provides is important because it isolates the role of one variable from all of the others in the model.

The equation shows that the coefficient for height in meters is 106.5 kilograms. The coefficient indicates that for every additional meter in height you can expect weight to increase by an average of 106.5 kilograms. However, these heights are from middle-school aged girls and range from 1.3 m to 1.7 m. The relationship is only valid within this data range, so we would not actually shift up or down the line by a full meter in this case.

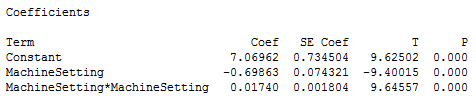
If the fitted line was flat (a slope coefficient of zero), the expected value for weight would not change no matter how far up and down the line you go. So, a low p-value suggests that the slope is not zero, which in turn suggests that changes in the predictor variable are associated with changes in the response variable.

## How Do I Interpret the Regression Coefficients for Curvilinear Relationships and Interaction Terms?

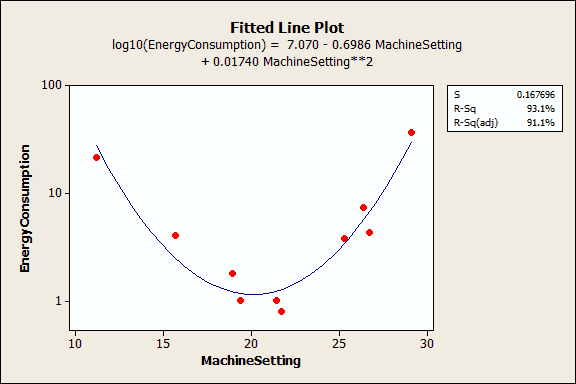
In the above example, height is a linear effect; the slope is constant, which indicates that the effect is also constant along the entire fitted line. However, if your model requires polynomial or interaction terms, the interpretation is a bit less intuitive.

As a refresher, polynomial terms [model curvature in the data](https://blog.minitab.com/blog/adventures-in-statistics/curve-fitting-with-linear-and-nonlinear-regression), while interaction terms indicate that the effect of one predictor depends on the value of another predictor.

The next example uses a data set that requires a quadratic (squared) term to model the curvature. In the output below, we see that the p-values for both the linear and quadratic terms are significant.



The residual plots (not shown) indicate a good fit, so we can proceed with the interpretation. But, how do we interpret these coefficients? It really helps to graph it in a fitted line plot.



You can see how the relationship between the machine setting and energy consumption varies depending on where you start on the fitted line. For example, if you start at a machine setting of 12 and increase the setting by 1, you’d expect energy consumption to decrease. However, if you start at 25, an increase of 1 should increase energy consumption. And if you’re around 20, energy consumption shouldn’t change much at all.

A significant polynomial term can make the interpretation less intuitive because the effect of changing the predictor varies depending on the value of that predictor. Similarly, a significant interaction term indicates that the effect of the predictor varies depending on the value of a different predictor.

Take extra care when you interpret a regression model that contains these types of terms. You can’t just look at the main effect (linear term) and understand what is happening! Unfortunately, if you are performing multiple regression analysis, you won't be able to use a fitted line plot to graphically interpret the results. This is where subject area knowledge is extra valuable!

<https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-to-interpret-the-constant-y-intercept>

The constant term in linear regression analysis also known as the y intercept, it is simply the value at which the fitted line crosses the y-axis. The value of the constant term is almost always meaningless! Paradoxically, it is crucial to include the constant term in most regression models!

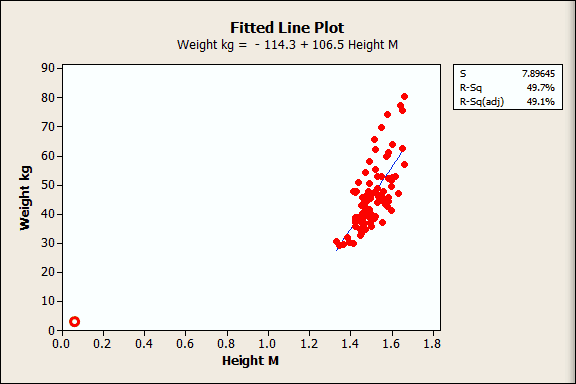
## Zero Settings for All of the Predictor Variables Is Often Impossible

I’ve often seen the constant described as the mean response value when all predictor variables are set to zero. Mathematically, that’s correct. However, a zero setting for all predictors in a model is often an impossible/nonsensical combination. If all of the predictors can’t be zero, it is impossible to interpret the value of the constant. Don't even try!

## Zero Settings for All of the Predictor Variables Can Be Outside the Data Range

Even if it’s possible for all of the predictor variables to equal zero, that data point might be outside the range of the observed data. You should never use a regression model to make a prediction for a point that is outside the range of your data because the relationship between the variables might change. The value of the constant is a prediction for the response value when all predictors equal zero. If you didn't collect data in this all-zero range, you can't trust the value of the constant.

The height-by-weight example illustrates this concept. These data are from middle school girls and we can’t estimate the relationship between the variables outside of the observed weight and height range. However, we can get a sense that the relationship changes by marking the average weight and height for a newborn baby on the graph. That’s not quite zero height, but it's as close as we can get.



I drew the red circle near the origin to approximate the newborn's average height and weight. You can clearly see that the relationship must change as you extend the data range!

So the relationship we see for the observed data is locally linear, but it changes beyond that. That’s why you shouldn’t predict outside the range of your data...and another reason why the regression constant can be meaningless.

## The Constant Is the Garbage Collector for the Regression Model

Even if a zero setting for all predictors is a plausible scenario, and even if you collect data within that all-zero range, the constant might still be meaningless!

The constant term is in part estimated by the omission of predictors from a regression analysis. In essence, it serves as a garbage bin for any bias that is not accounted for by the terms in the model. You can picture this by imagining that the regression line floats up and down (by adjusting the constant) to a point where the mean of the residuals is zero, which is a key assumption for [residual analysis](https://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis). This floating is not based on what makes sense for the constant, but rather what works mathematically to produce that zero mean.

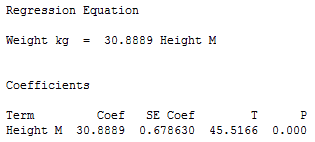
The constant guarantees that the residuals don’t have an overall positive or negative bias, but also makes it harder to interpret the value of the constant because it absorbs the bias.

## Why Is it Crucial to Include the Constant in a Regression Model?

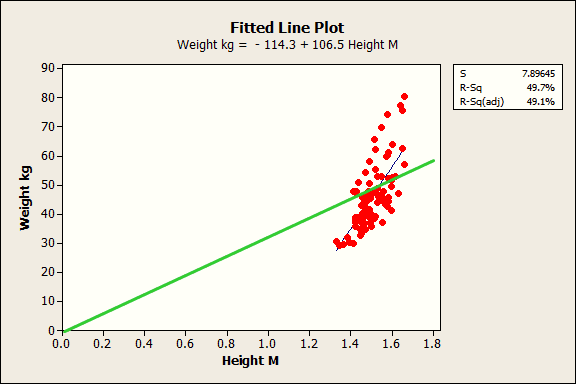
Immediately above, we saw a key reason why you should include the constant in your regression model. It guarantees that your residuals have a mean of zero.

Additionally, if you don’t include the constant, the regression line is forced to go through the origin. This means that all of the predictors and the response variable must equal zero at that point. If your fitted line doesn’t naturally go through the origin, your regression coefficients and predictions will be biased if don't include the constant.

I’ll use the height and weight regression example to illustrate this concept. First, I’ll use General Regression in Minitab [statistical software](http://www.minitab.com/en-us/products/minitab/)to fit the model without the constant. In the output below, you can see that there is no constant, just a coefficient for height.



Next, I’ll overlay the line for this equation on the previous fitted line plot so we can compare the model with and without the constant.



The blue line is the fitted line for the regression model with the constant while the green line is for the model without the constant. Clearly, the green line just doesn’t fit. The slope is way off and the predicted values are biased. For the model without the constant, the weight predictions tend to be too high for shorter subjects and too low for taller subjects.

In closing, the regression constant is generally not worth interpreting. Despite this, it is almost always a good idea to include the constant in your regression analysis. In the end, the real value of a regression model is the ability to understand how the response variable changes when you change the values of the predictor variables. Don't worry too much about the constant!