Lesson 13

Machine Learning – Interpreting Regression Results

Kush Kulshrestha

Interpreting p-values

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables and the response variable. Here you will learn about how to interpret the p-values and coefficients that appear in the output for 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 (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

Interpreting p-values

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.

Coefficients

Term	Coef	SE Coef	T	P
Constant	389.166	66.0937	5.8881	0.000
East	2.125	1.2145	1.7495	0.092
South	5.318	0.9629	5.5232	0.000
North	-24.132	1.8685	-12.9153	0.000

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.

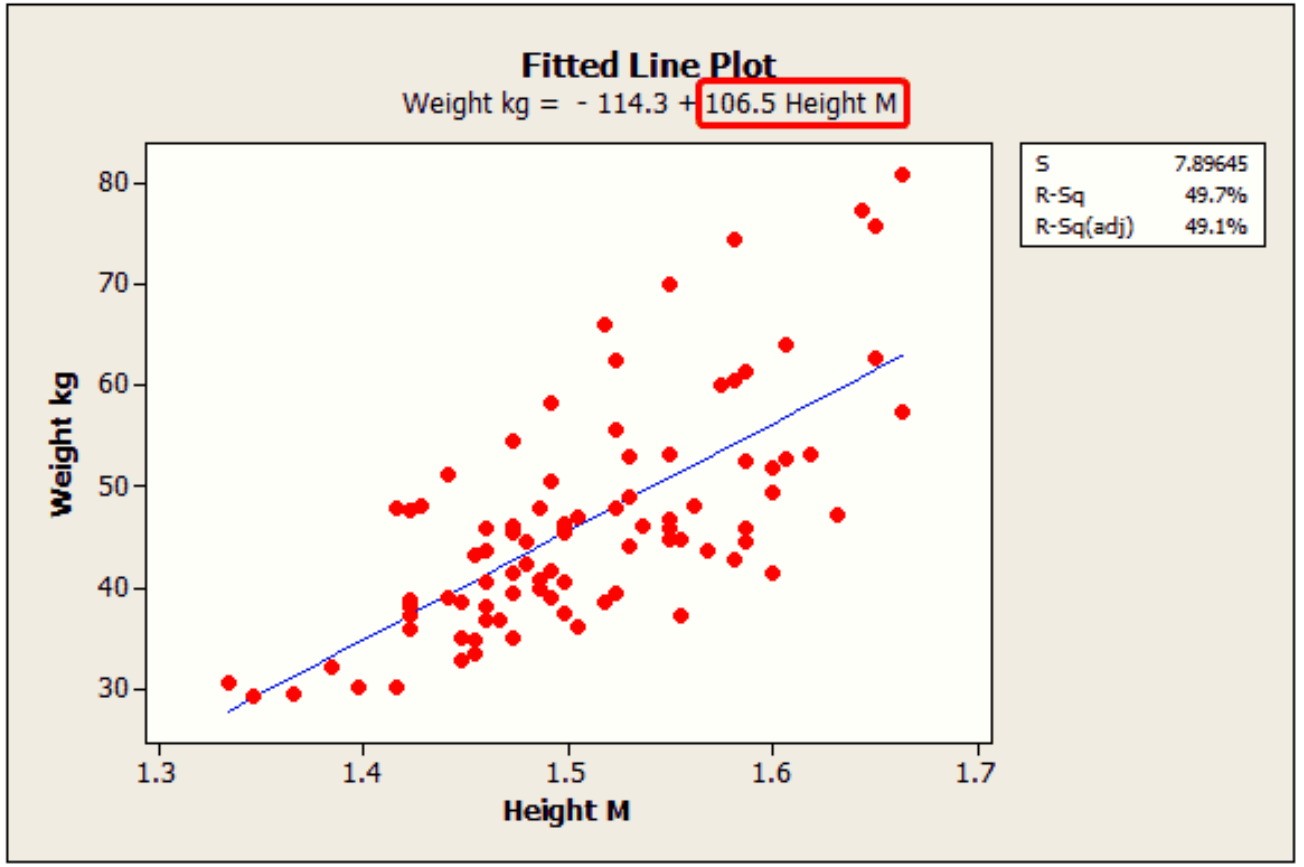
Interpreting Regression Coefficients

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 that regression provides is important because it isolates the role of one variable from all of the others in the model.

The key to understanding the coefficients is to think of them as slopes, and they're often called slope coefficients. We will this in the fitted line plot below, where we will use a person's height to model their weight.

Coefficients

Term	Coef	SE Coef	T	P
Constant	-114.326	17.4425	-6.55444	0.000
Height M	106.505	11.5500	9.22117	0.000



Interpreting Regression Coefficients

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.

The blue fitted line graphically shows the same information. If you move left or right along the x-axis by an amount that represents a one meter change in height, the fitted line rises or falls by 106.5 kilograms.

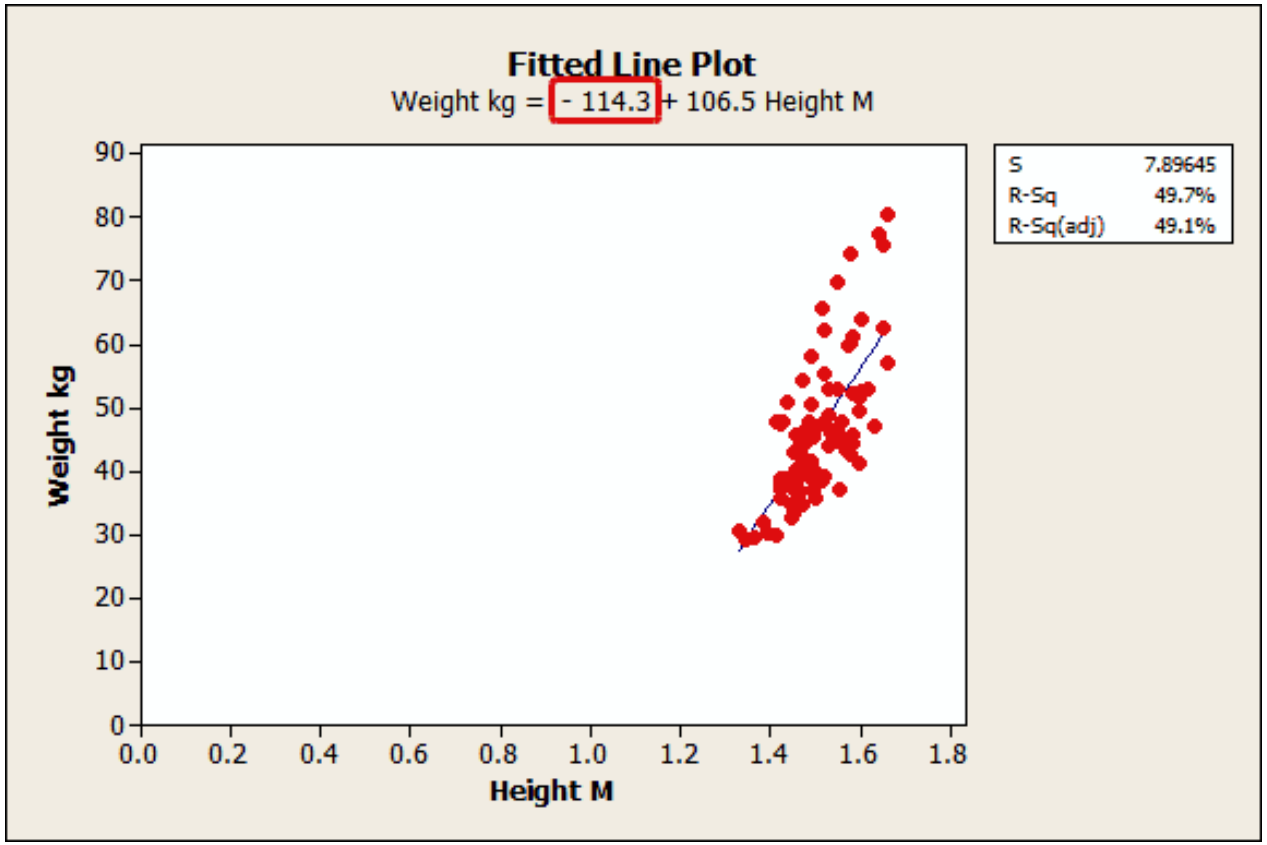
However, these heights are from middle-school aged girls and range from 1.3 m to 1.7 m. The relationship is only valid within this data range, so we would not actually shift up or down the line by a full meter in this case.

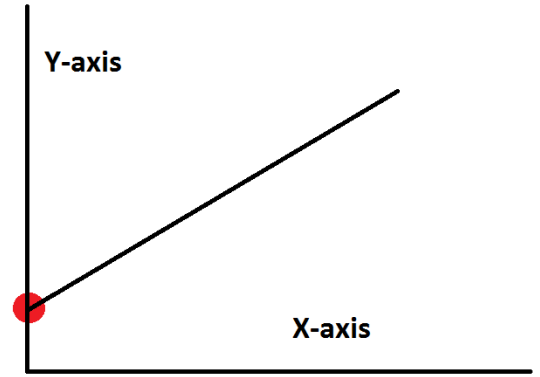
If the fitted line was flat (a slope coefficient of zero), the expected value for weight would not change no matter how far up and down the line you go. So, a low p-value suggests that the slope is not zero, which in turn suggests that changes in the predictor variable are associated with changes in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

Interpreting the constant (Y intercept)

We have 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, as it is in the following example.

In a previous example, we used a fitted line plot to illustrate a weight-by-height regression analysis. Below, I've changed the scale of the y-axis on that fitted line plot, but the regression results are the same as before.





Interpreting the constant (Y intercept)

If you follow the blue fitted line down to where it intercepts the y-axis, it is a fairly negative value. From the regression equation, we see that the intercept value is -114.3. If height is zero, the regression equation predicts that weight is -114.3 kilograms!

Clearly this constant is meaningless and you shouldn't even try to give it meaning. No human can have zero height or a negative weight!

Now imagine a multiple regression analysis with many predictors. It becomes even more unlikely that ALL of the predictors can realistically be set to zero.

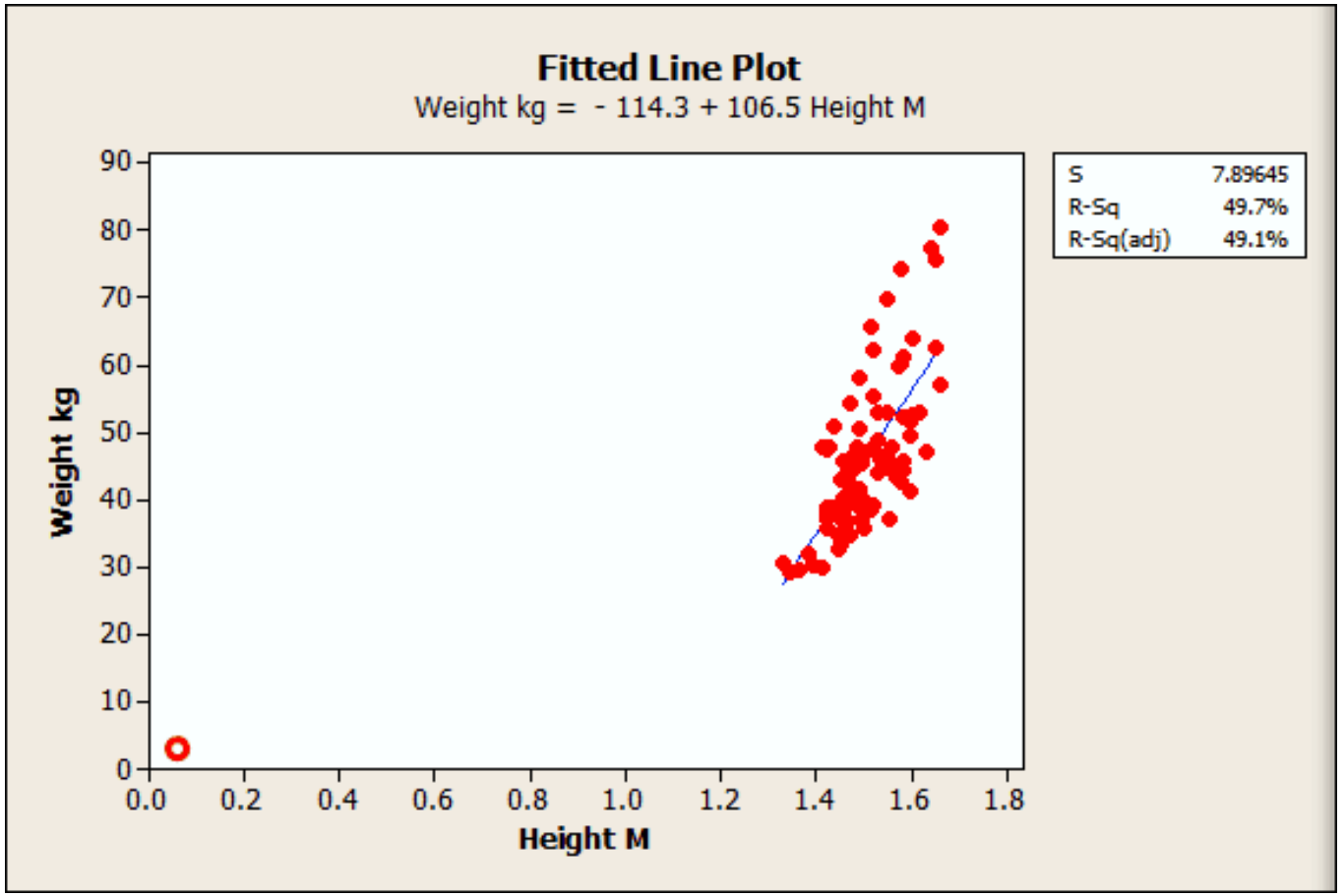
Interpreting the constant (Y intercept)

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.

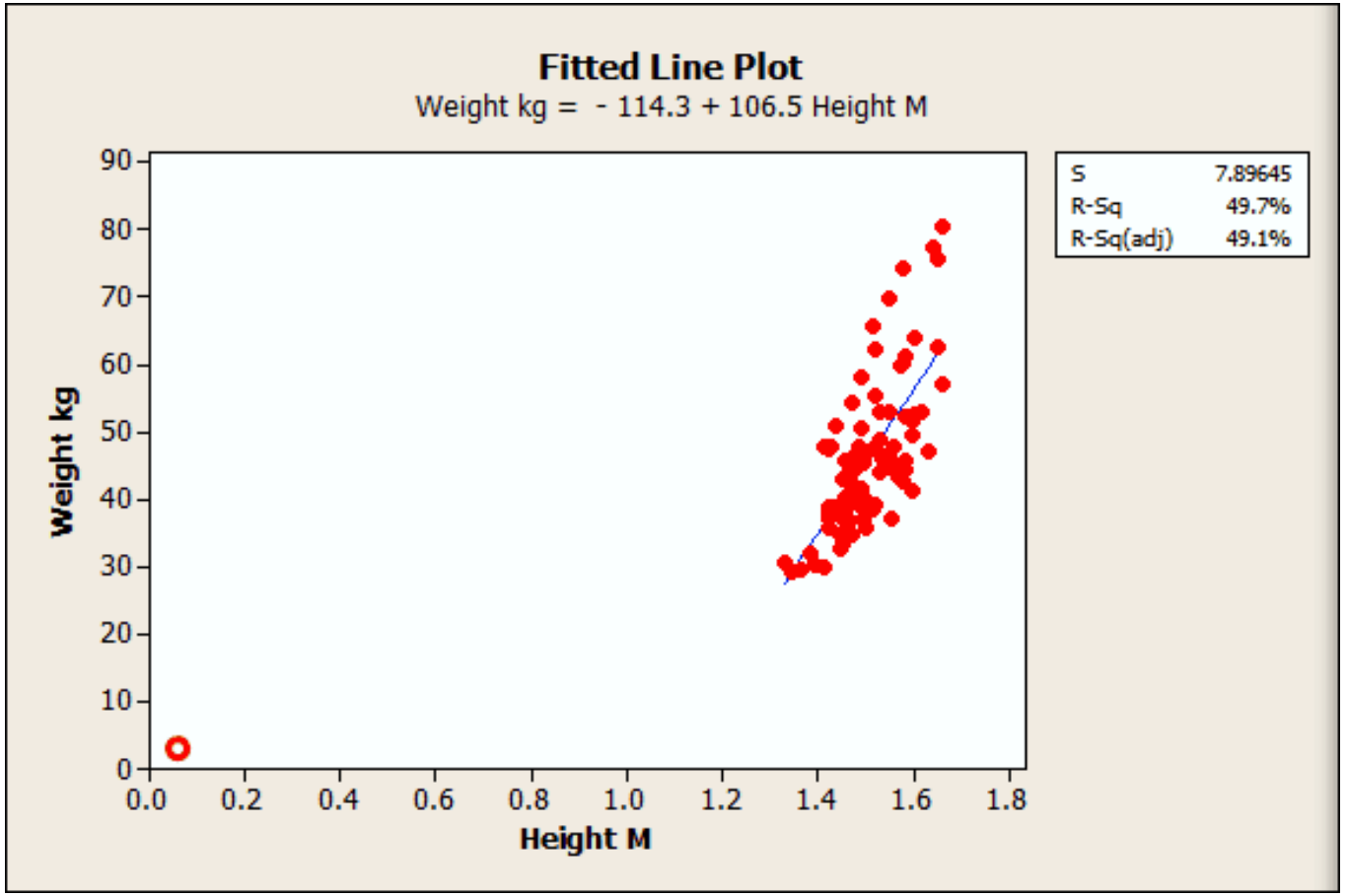
Interpreting the constant (Y intercept)

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.



Interpreting the constant (Y intercept)

There is a 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.



Goodness of fit

After you have fit a linear model using regression analysis, you need to determine how well the model fits the data.

What is Goodness of fit of a Linear Model?
Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

Before you look at the statistical measures for goodness-of-fit, we will look at checking the residual plots.
Residual plots can reveal unwanted residual patterns that indicate biased results more effectively than numbers.
When your residual plots pass muster, you can trust your numerical results and check the goodness-of-fit statistics.

Residual Plots

To start, let's breakdown and define the 2 basic components of a valid regression model:

Response = (Constant + Predictors) + Error

Another way we can say this is:

Response = Deterministic + Stochastic

The Deterministic Portion –
This is the part that is explained by the predictor variables in the model. The expected value of the response is a function of a set of predictor variables. All of the explanatory/predictive information of the model should be in this portion.

The Stochastic Error –
Stochastic is a fancy word that means random and unpredictable. Error is the difference between the expected value and the observed value. Putting this together, the differences between the expected and observed values must be unpredictable. In other words, none of the explanatory/predictive information should be in the error.

The idea is that the deterministic portion of your model is so good at explaining (or predicting) the response that only the inherent randomness of any real-world phenomenon remains leftover for the error portion. If you observe explanatory or predictive power in the error, you know that your predictors are missing some of the predictive information.
Residual plots help you check this!

Using Residual Plots

Using residual plots, you can assess whether the observed error (residuals) is consistent with stochastic error.

This process is easy to understand with a die-rolling analogy. When you roll a die, you shouldn't be able to predict which number will show on any given toss. However, you can assess a series of tosses to determine whether the displayed numbers follow a random pattern. If the number six shows up more frequently than randomness dictates, you know something is wrong with your understanding (mental model) of how the die actually behaves.

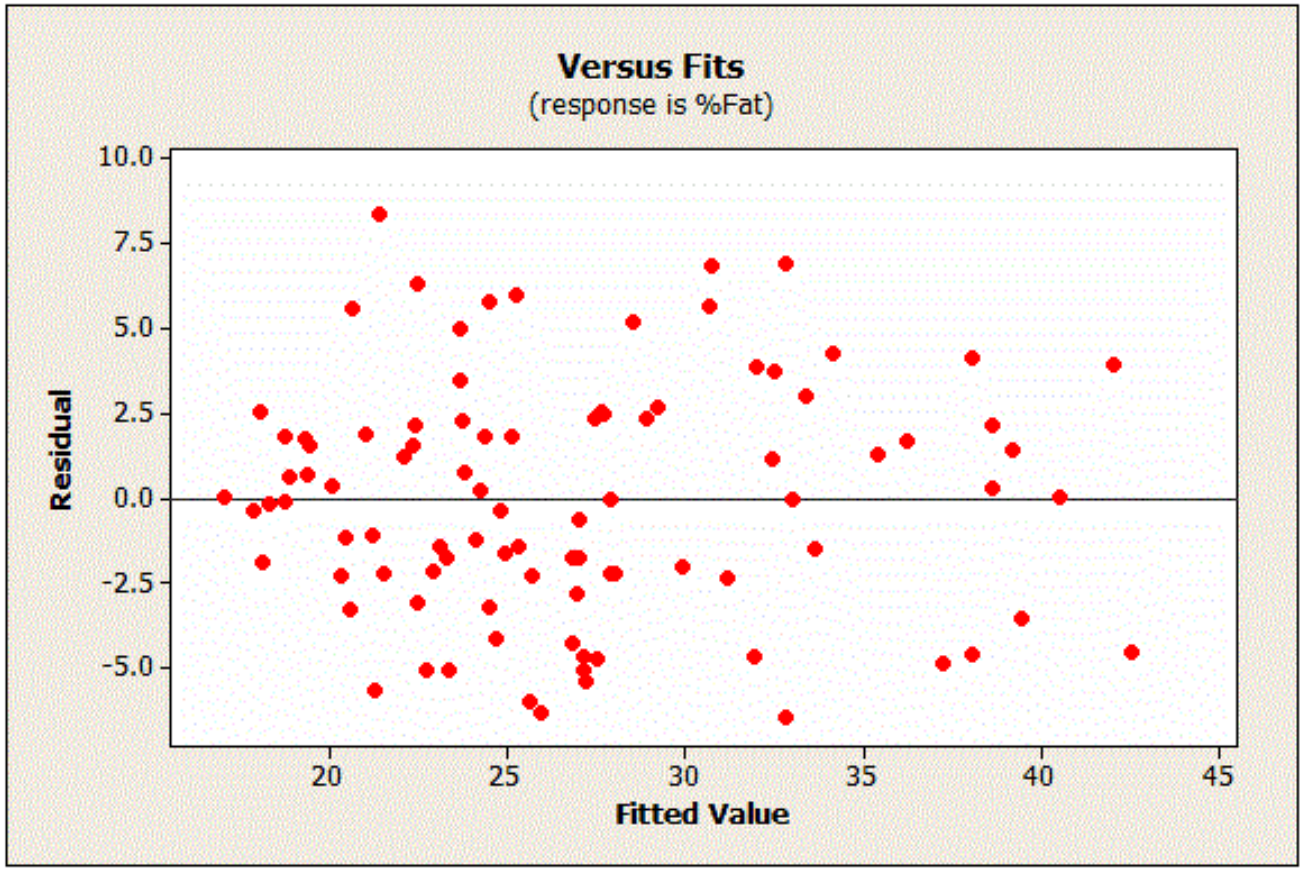
The same principle applies to regression models. You shouldn't be able to predict the error for any given observation. And, for a series of observations, you can determine whether the residuals are consistent with random error. Just like with the die, if the residuals suggest that your model is systematically incorrect, you have an opportunity to improve the model.

Using Residual Plots

The residuals should not be either systematically high or low. So, the residuals should be centered on zero throughout the range of fitted values.
In other words, the model is correct on average for all fitted values.

Further, in the OLS context, random errors are assumed to produce residuals that are normally distributed. Therefore, the residuals should fall in a symmetrical pattern and have a constant spread throughout the range.

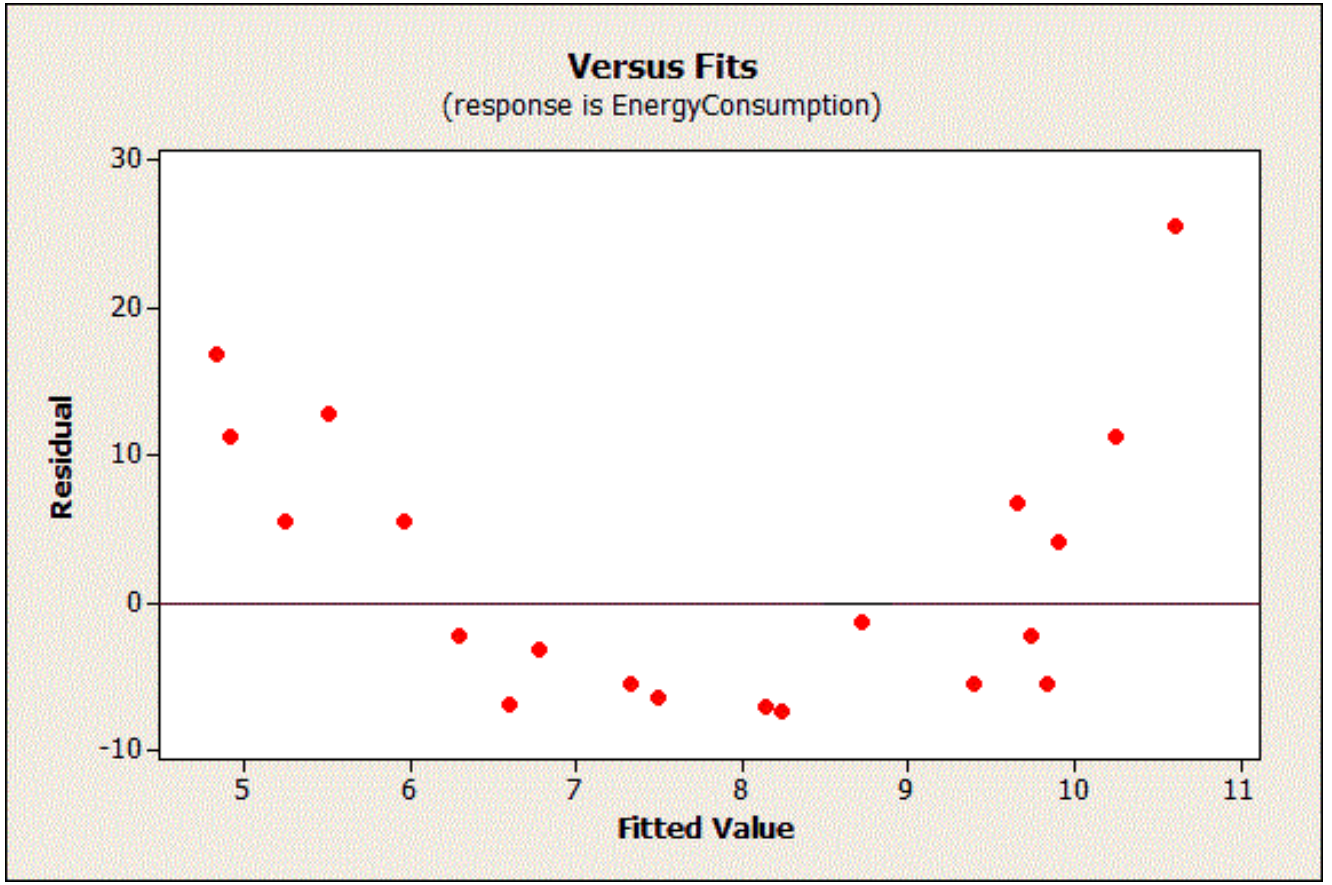
Here's how residuals should look:



Using Residual Plots

Now let's look at a problematic residual plot. Keep in mind that the residuals should not contain any predictive information.

In the graph above, you can predict non-zero values for the residuals based on the fitted value. For example, a fitted value of 8 has an expected residual that is negative. Conversely, a fitted value of 5 or 11 has an expected residual that is positive.



Using Residual Plots

The non-random pattern in the residuals indicates that the deterministic portion (predictor variables) of the model is not capturing some explanatory information that is “leaking” into the residuals.

The graph could represent several ways in which the model is not explaining all that is possible. Possibilities include:

- A missing variable
- A missing higher-order term of a variable in the model to explain the curvature
- A missing interaction between terms already in the model

In addition to the above, here are two more specific ways that predictive information can sneak into the residuals:

- **The residuals should not be correlated with another variable.** If you can predict the residuals with another variable, that variable should be included in the model.
- **Adjacent residuals should not be correlated with each other (autocorrelation).** If you can use one residual to predict the next residual, there is some predictive information present that is not captured by the predictors. Typically, this situation involves time-ordered observations.

For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated.

R-Squared

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

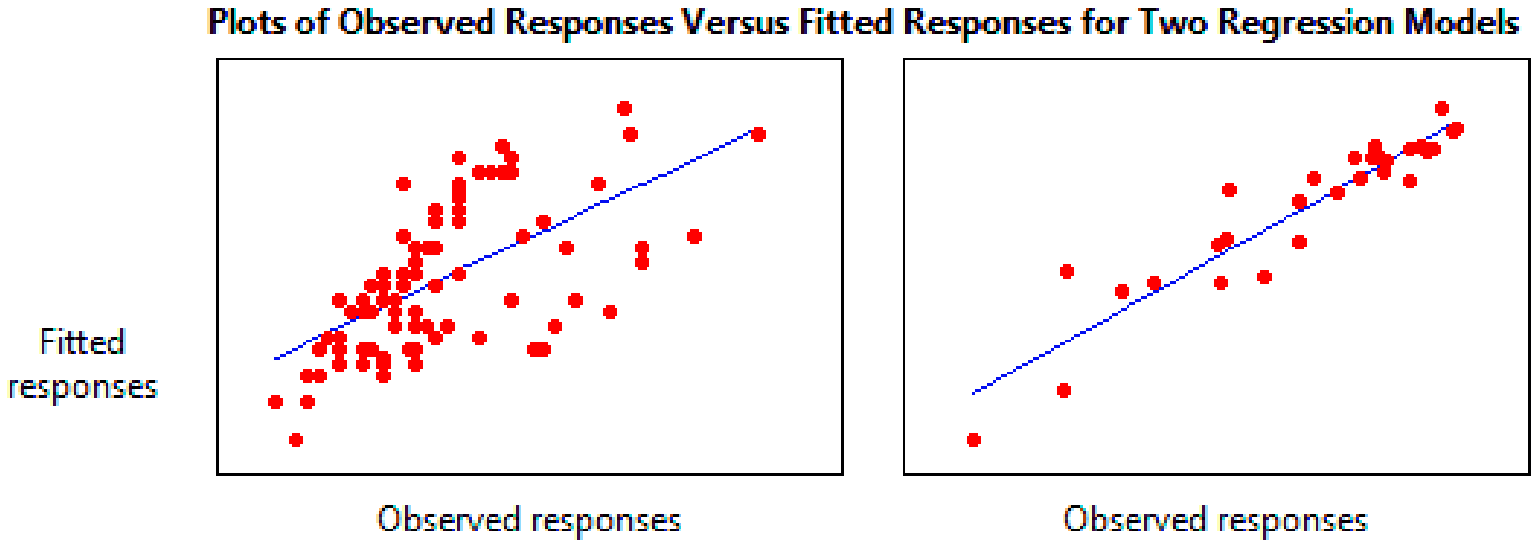
R-squared is always between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data.

Graphical Representation of R-Squared

Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.



The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

Limitations of R-Squared

1. R-squared *cannot* determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.

2. R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

Are low R-Squared values bad?

No! There are two major reasons why it can be just fine to have low R-squared values.

In some fields, it is entirely expected that your R-squared values will be low. We get a R-squared value of 37~40 % in finance and we are happy.

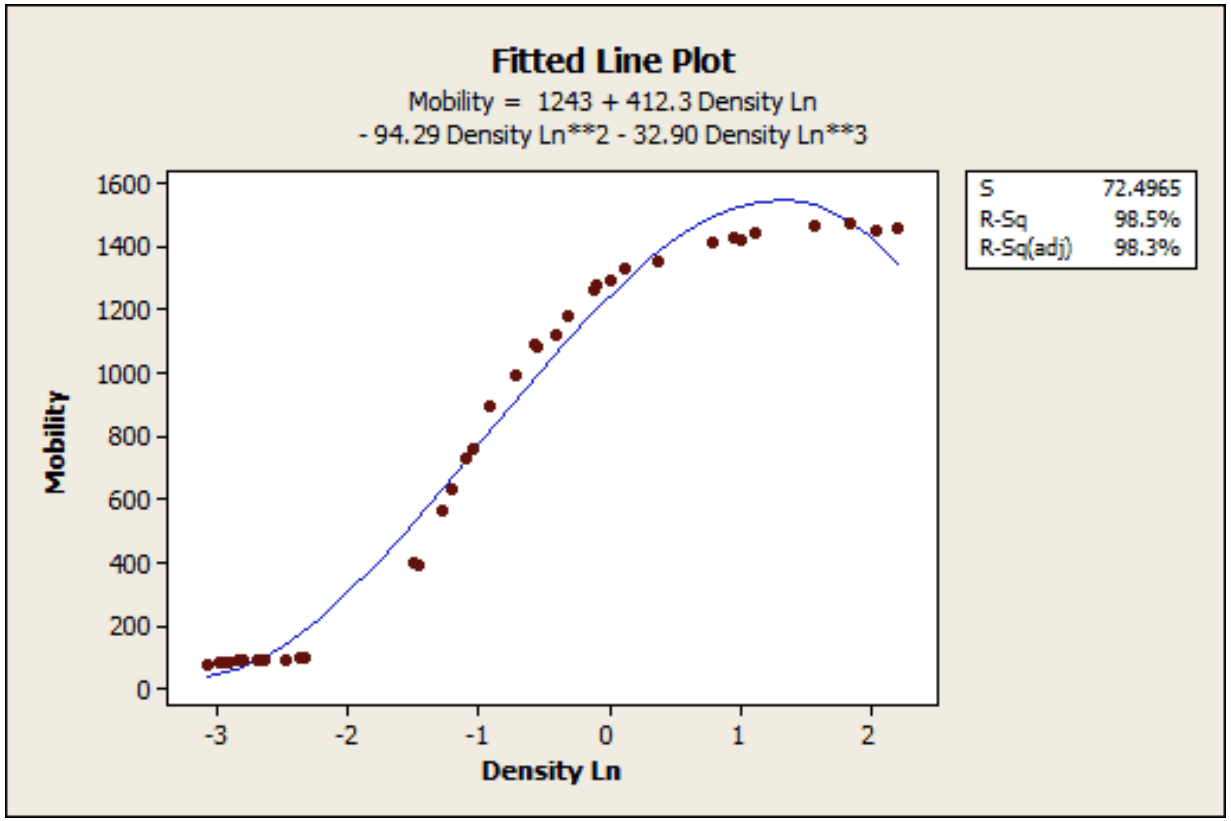
Furthermore, if your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.

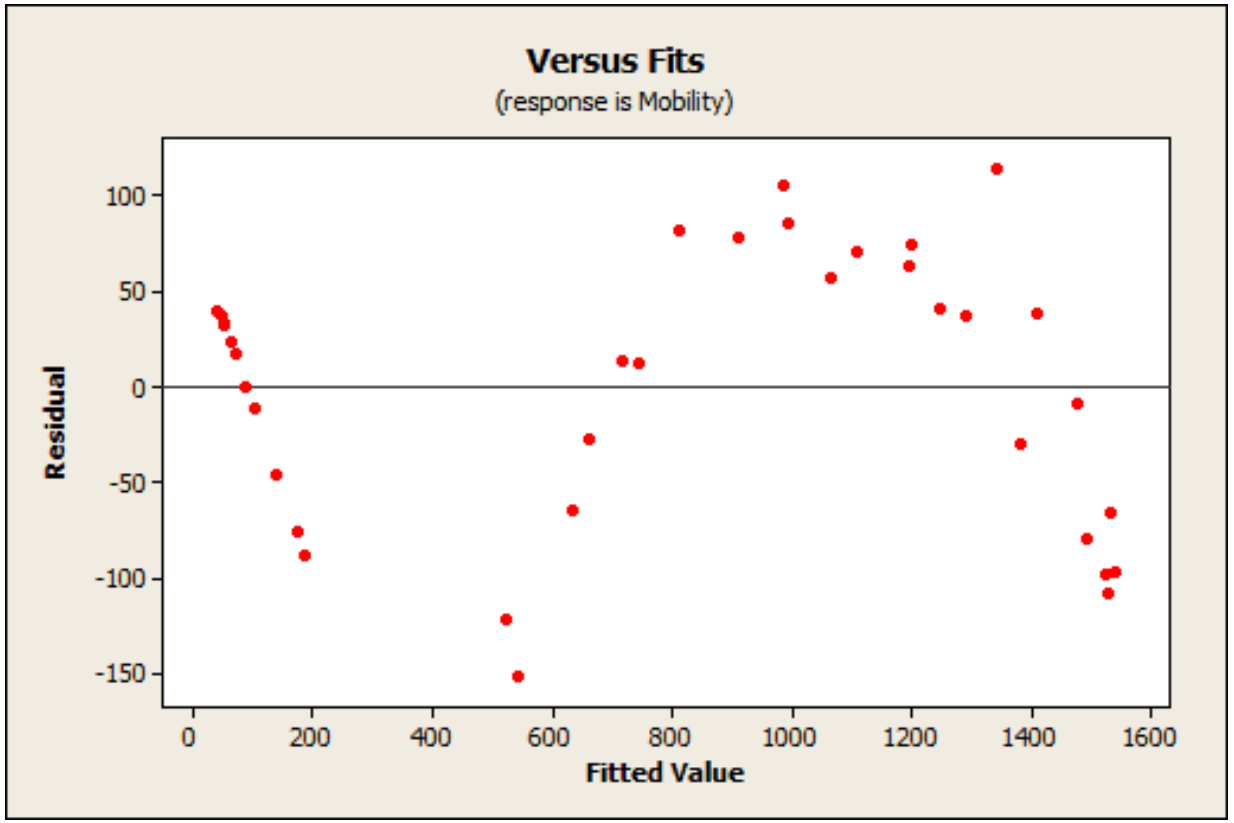
Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. Obviously, this type of information can be extremely valuable.

A low R-squared is most problematic when you want to produce predictions that are reasonably precise (have a small enough prediction interval).

Are high R-Squared values good?

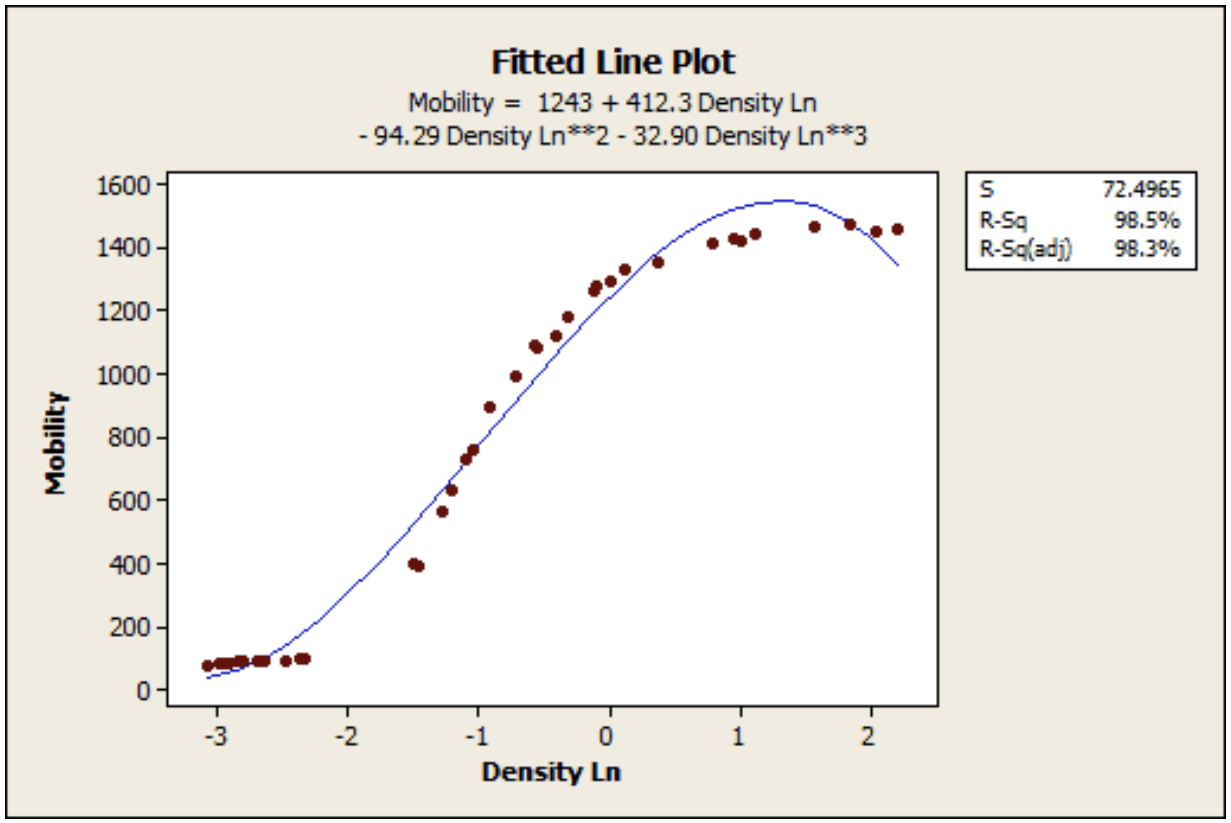
No! A high R-squared does not necessarily indicate that the model has a good fit. That might be a surprise, but look at the fitted line plot and residual plot below. The fitted line plot displays the relationship between semiconductor electron mobility and the natural log of the density for real experimental data.

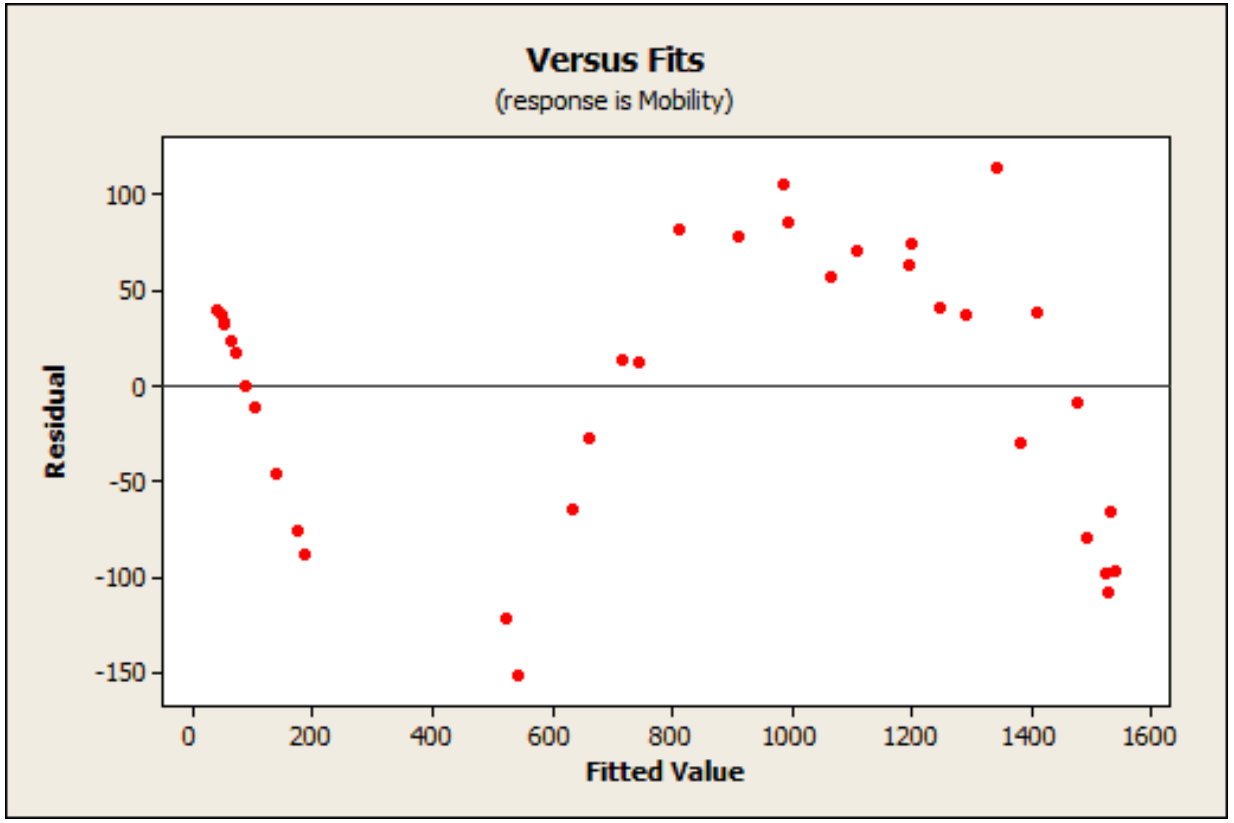




Are high R-Squared values good?

The fitted line plot shows that these data follow a nice tight function and the R-squared is 98.5%, which sounds great. However, look closer to see how the regression line systematically over and under-predicts the data (bias) at different points along the curve. You can also see patterns in the Residuals versus Fits plot, rather than the randomness that you want to see. This indicates a bad fit, and serves as a reminder as to why you should always check the residual plots.





Conclusion on R-Squared

Similar biases can occur when your linear model is missing important predictors, polynomial terms, and interaction terms.

R-squared is a handy, seemingly intuitive measure of how well your linear model fits a set of observations. However, as we saw, R-squared doesn't tell us the entire story. You should evaluate R-squared values in conjunction with residual plots.

Interpretation on Standard Error of Regression

Similar biases can occur when your linear model is missing important predictors, polynomial terms, and interaction terms.

R-squared is a handy, seemingly intuitive measure of how well your linear model fits a set of observations. However, as we saw, R-squared doesn't tell us the entire story. You should evaluate R-squared values in conjunction with residual plots.