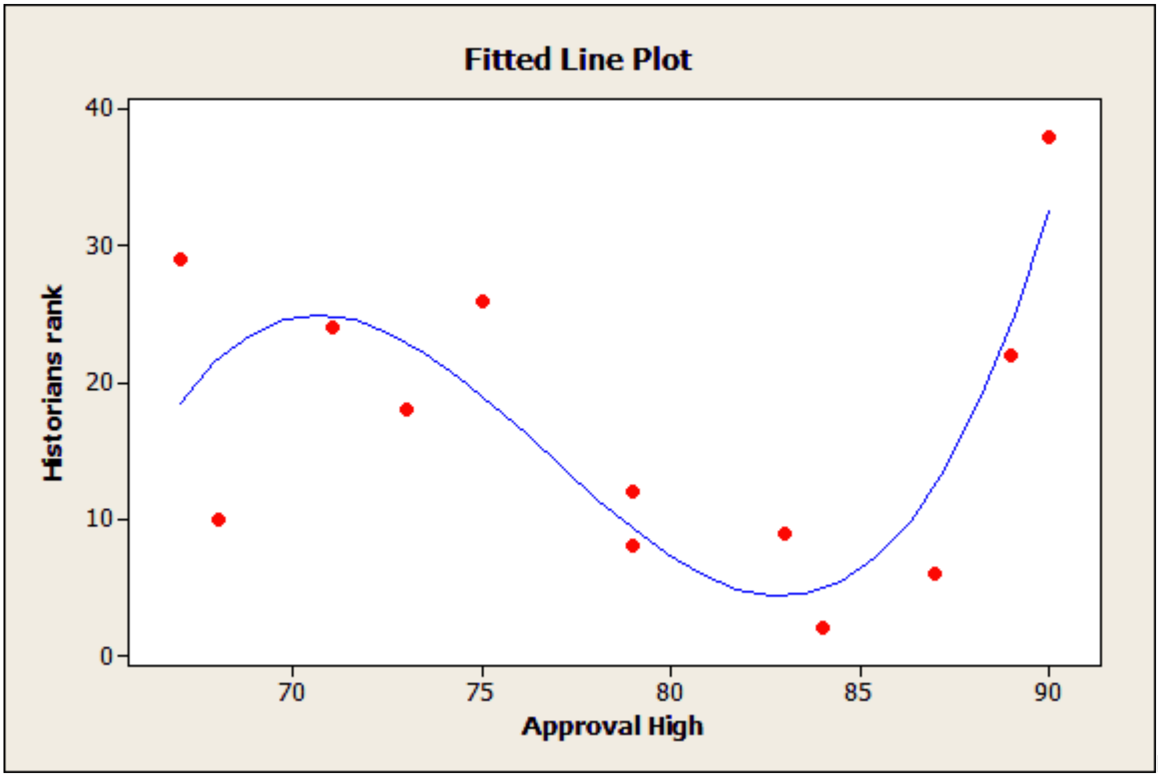
The Danger of Overfitting Regression Models

Jim Frost 3 September, 2015



In regression analysis, overfitting a model is a real problem. An overfit model can cause the regression coefficients, p-values, and R-squared to be misleading. In this post, I explain what an overfit model is and how to detect and avoid this problem.

An overfit model is one that is too complicated for your data set. When this happens, the regression model becomes tailored to fit the quirks and random noise in your specific sample rather than reflecting the overall population. If you drew another sample, it would have its own quirks, and your original overfit model would not likely fit the new data.

Instead, we want our model to approximate the true model for the entire population. Our model should not only fit the current sample, but new samples too.

The fitted line plot illustrates the dangers of overfitting regression models. This model appears to explain a lot of variation in the response variable. However, the model is too complex for the sample data. In the overall population, there is no real relationship between the predictor and the response. You can read about the model here.

Fundamentals of Inferential Statistics

To understand how overfitting causes these problems, we need to go back to the basics for inferential statistics.

The overall goal of inferential statistics is to draw conclusions about a larger population from a random sample. Inferential statistics uses the sample data to provide the following:

- Unbiased estimates of properties and relationships within the population.
- Hypothesis tests that assess statements about the entire population.

An important concept in inferential statistics is that the amount of information you can learn about a population is limited by the sample size. The more you want to learn, the larger your sample size must be.

You probably understand this concept intuitively, but here's an example. If you have a sample size of 20 and want to estimate a single population mean, you're probably in good shape. However, if you want to estimate two population means using the same total sample size, it suddenly looks iffier. If you increase it to three population means and more, it starts to look pretty bad.

The quality of the results worsens when you try to learn too much from a sample. As the number of observations per parameter decreases in the example above (20, 10, 6.7, etc), the estimates become more erratic and a new sample is less likely to reproduce them.

Applying These Concepts to Overfitting Regression Models

In a similar fashion, overfitting a regression model occurs when you attempt to estimate too many parameters from a sample that is too small. Regression analysis uses one sample to estimate the values of the coefficients for *all* of the terms in the equation. The sample size limits the number of terms that you can safely include before you begin to overfit the model. The number of terms in the model includes all of the predictors, interaction effects, and polynomials terms (to model curvature).

Larger sample sizes allow you to specify more complex models. For trustworthy results, your sample size must be large enough to support the level of complexity that is required by your research question. If your sample size isn't large enough, you won't be able to fit a model that adequately approximates the true model for your response variable. You won't be able to trust the results.

Just like the example with multiple means, you must have a sufficient number of observations for each term in a regression model. Simulation studies show that a good rule of thumb is to have 10-15 observations per term in multiple linear regression.

For example, if your model contains two predictors and the interaction term, you'll need 30-45 observations. However, if the effect size is small or there is high multicollinearity, you may need more observations per term.

How to Detect and Avoid Overfit Models

Cross-validation can detect overfit models by determining how well your model generalizes to other data sets by partitioning your data. This process helps you assess how well the model fits new observations that weren't used in the model estimation process.

Minitab statistical software provides a great cross-validation solution for linear models by calculating predicted R-squared. This statistic is a form of cross-validation that doesn't require you to collect a separate sample. Instead, Minitab calculates predicted R-squared by systematically removing each observation from the data set, estimating the regression equation, and determining how well the model predicts the removed observation.

If the model does a poor job at predicting the removed observations, this indicates that the model is probably tailored to the specific data points that are included in the sample and not generalizable outside the sample.

To avoid overfitting your model in the first place, collect a sample that is large enough so you can safely include all of the predictors, interaction effects, and polynomial terms that your response variable requires. The scientific process involves plenty of research before you even begin to collect data. You should identify the important variables, the model that you are likely to specify, and use that information to estimate a good sample size.

For more about the model selection process, read my blog post, How to Choose the Best Regression Model. Also, check out my post about overfitting regression models by using too many phantom degrees of freedom. The methods described above won't necessarily detect this problem.

How to Choose the Best Regression Model

Choosing the correct linear regression model can be difficult. After all, the world and how it works is complex. Trying to model it with only a sample doesn’t make it any easier. In this post, I'll review some common statistical methods for selecting models, complications you may face, and provide some practical advice for choosing the best regression model.

It starts when a researcher wants to mathematically describe the relationship between some predictors and the response variable. The research team tasked to investigate typically measures many variables but includes only some of them in the model. The analysts try to eliminate the variables that are not related and include only those with a true relationship. Along the way, the analysts consider many possible models.

They strive to achieve a Goldilocks balance with the number of predictors they include.

- **Too few**: An underspecified model tends to produce biased estimates.
- **Too many**: An overspecified model tends to have less precise estimates.
- **Just right**: A model with the correct terms has no bias and the most precise estimates.

Statistical Methods for Finding the Best Regression Model

For a good regression model, you want to include the variables that you are specifically testing along with other variables that affect the response in order to avoid biased results. Minitab statistical software offers statistical measures and procedures that help you specify your regression model. I'll review the common methods, but please do follow the links to read my more detailed posts about each.

Adjusted R-squared and Predicted R-squared: Generally, you choose the models that have higher adjusted and predicted R-squared values. These statistics are designed to avoid a key problem with regular R-squared—it increases *every* time you add a predictor and can trick you into specifying an overly complex model.

- The adjusted R squared increases only if the new term improves the model more than would be expected by chance and it can also decrease with poor quality predictors.
- The predicted R-squared is a form of cross-validation and it can also decrease. Cross-validation determines how well your model generalizes to other data sets by partitioning your data.

P-values for the predictors: In regression, low p-values indicate terms that are statistically significant. "Reducing the model" refers to the practice of including all candidate predictors in the model, and then systematically removing the term with the highest p-value one-by-one until you are left with only significant predictors.

Stepwise regression and Best subsets regression: These are two automated procedures that can identify useful predictors during the exploratory stages of model building. With best subsets regression, Minitab provides Mallows' Cp, which is a statistic specifically designed to help you manage the tradeoff between precision and bias.

Real World Complications

Great, there are a variety of statistical methods to help us choose the best model. Unfortunately, there also are a number of potential complications. Don't worry, I'll provide some practical advice!

- The best model can be only as good as the variables measured by the study. The results for the variables you include in the analysis can be biased by the significant variables that you don't include. Read about an example of omitted variable bias.
- Your sample might be unusual, either by chance or by data collection methodology. False positives and false negatives are part of the game when working with samples.

- P-values can change based on the specific terms in the model. In particular, multicollinearity can sap significance and make it difficult to determine the role of each predictor.
- If you assess enough models, you *will* find variables that appear to be significant but are only correlated by chance. This form of data mining can make random data appear significant. A low predicted R-squared is a good way to check for this problem.
- P-values, predicted and adjusted R-squared, and Mallows' Cp can suggest different models.
- Stepwise regression and best subsets regression are great tools and can get you close to the correct model. However, studies have found that they generally don't pick the correct model.

Recommendations for Finding the Best Regression Model

Choosing the correct regression model is as much a science as it is an art. Statistical methods can help point you in the right direction but ultimately you'll need to incorporate other considerations.

Theory

Research what others have done and incorporate those findings into constructing your model. Before beginning the regression analysis, develop an idea of what the important variables are along with their relationships, coefficient signs, and effect magnitudes. Building on the results of others makes it easier both to collect the correct data and to specify the best regression model without the need for data mining.

Theoretical considerations should not be discarded based solely on statistical measures. After you fit your model, determine whether it aligns with theory and possibly make adjustments. For example, based on theory, you might include a predictor in the model even if its p-value is not significant. If any of the coefficient signs contradict theory, investigate and either change your model or explain the inconsistency.

Complexity

You might think that complex problems require complex models, but many studies show that simpler models generally produce more precise predictions. Given several models with similar explanatory ability, the simplest is most likely to be the best choice. Start simple, and only make the model more complex as needed. The more complex you make your model, the more likely it is that you are tailoring the model to your dataset specifically, and generalizability suffers.

Verify that added complexity actually produces narrower prediction intervals. Check the predicted R-squared and don't mindlessly chase a high regular R-squared!