The Model Building Process Part 3: Model Goodness Metrics

Best Practice

Authored by: Sarah Burke, PhD

24 March 2020

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

Table of Contents

Introduction

This document is the third part in a series on the steps of the (statistical) model building process. Part 1 (Burke, 2017) discusses methods to assess whether the error assumptions in a linear regression model had been satisfied while Part 2 (Burke, 2018) describes remedial measures available to use when model assumptions are violated. The model building process begins with data collection and the goal of fitting a statistical model that can be used to fully characterize the response. This best practice describes several statistical metrics that can be used to assess the quality of the model. These metrics may also be used when trying to compare two or more models for use.

Keywords: linear regression, R^2, adjusted R^2, AIC, BIC, PRESS

Motivating Example

A designed experiment was used to characterize the miss distance of the small diameter bomb (SDB) Increment II. Factors of interest included the attack mode (x_1), clutter (x_2), angle off nose (x_3), and ground range (x_4). A full factorial design (with four center points) was executed and the miss distance of the SDB from the target was measured. The notional data for this test is shown in Appendix A. An initial model was fit to the data (Equation 1) and the goal now is to evaluate the quality of this model for use by the test team.

$$\text{Miss Distance} = 62.84 + 24.6x_1 + 33.3x_2 + 3.5x_4 + 24.1x_1x_2 + 8.7x_1x_4 \quad (1)$$

Model Metrics

When evaluating the usefulness of a model, we can consider two interpretations of "usefulness." One considers the ability of the model to adequately capture the variability in the response. The other focuses on the ability of the model to predict future responses. We discuss metrics of each type throughout this document. All of the metrics discussed are readily available in statistical software packages, including JMP.

R^2

One of the most common measures of a regression model is the coefficient of determination, denoted R^2, and shown in Equation 2. This metric measures the variability in the response that is explained by the regression model. The total variability of the response (denoted SS_T) is decomposed into two components when fitting a model: the regression sum of squares and the error sum of squares. The regression sum of squares (SS_R) represents the variability of the mean response between the factor levels. Essentially, any variability in the response that can be attributed to a change in a factor level is put into the SS_R component. The error sum of squares (SS_E) represents the variability in the response that cannot be attributed to any of the factors. SS_E is the variability within a factor level and measures the discrepancy between the data and the regression model. The formula for R^2 is shown in Equation 2.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad (2)$$

R^2 ranges in value between 0 and 1. A value of 0 indicates that the regression model was completely ineffective in predicting the response; none of the variability in the response is explained by the model. A value of 1 means that the data is perfectly explained by the model. This metric is often presented as a percentage and can be interpreted as the proportion of variability in the response that is explained by the regression model. The target value of R^2 depends on the context of the problem; however, we find that in Department of Defense test and evaluation, values of 70% or higher are typically considered acceptable.

One way to visualize the goodness of the model is to graph the actual response values from the experiment versus the predicted response values from the model. The closer the actual and predicted values are, the better the model fit, and the closer R^2 is to 1. If the value of R^2 is low, then the spread of points in a plot of the actual response values versus the predicted response values are relatively large. If there was no error, then the actual values of miss distance would be equal to the predicted values of miss distance.

Figure 1 shows a scatterplot of the actual values of miss distance versus the predicted values of miss distance for the SDB example. The observations are tightly clustered around the fitted line, indicating that the model is doing a good job of capturing the variability in the miss distance. The R^2 for the model is 94%, meaning 94% of the variability in miss distance is characterized or captured by the simple regression model in Equation 1.

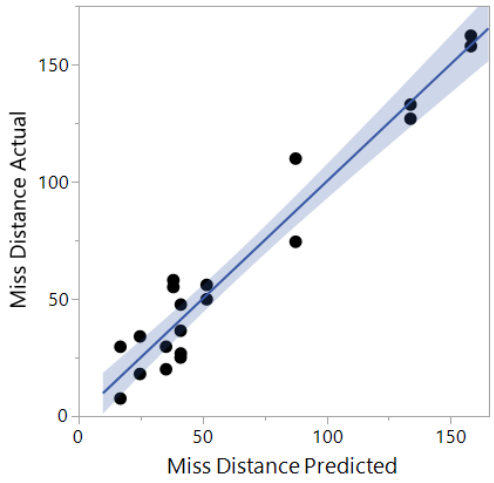

Figure 1: Actual vs Predicted Miss Distance for SDB Example

Adjusted R^2

R^2 is a nondecreasing metric and will nearly always increase as more model terms are added to the regression model. This can lead to an inflated value of the model goodness. Recall that we also want as simple a model as possible to characterize the response. Therefore, the adjusted R^2, denoted R^2_{adj}, can be used to assess a model's usefulness. This metric essentially penalizes models with a larger number of model terms and emphasizes model simplicity. Although there is a penalty term in the formula for R^2_{adj},

the interpretation is the same as that for R^2. The formula for the adjusted R^2 is shown in Equation 3, where n represents the number of observations and p represents the number of model terms.

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SS_E}{SS_T}\right) \tag{3}$$

For the SDB example, the adjusted R^2 is 92%, a difference of 2% from the R^2 metric. 92% of the variability in the response is characterized by the model. Ideally, the difference between R^2 and R^2_{adj} is relatively small. A Large difference between these two metrics indicates that there are terms in the regression model that do not contribute to explaining the response. For example, consider a more complex model for the SDB that contains all main effects and two-factor interactions. This model has an R^2 value of 95% and R^2_{adj} value of 88%. While both metrics are relatively high, the difference between the two is larger (7%) compared to the simpler model in Equation 1.

AIC and BIC

Two metrics primarily used to compare the goodness of multiple models are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both of these metrics are measures of model quality; however, on their own, they do not have the natural interpretability compared to R^2 and R^2_{adj}. They are primarily used to compare several different models; the smaller the value of AIC and BIC, the better. Because these metrics are on a log scale, differences greater than 3 in AIC or BIC between different models is considered large.

AIC, similar to R^2_{adj}, measures model goodness while balancing the tradeoffs between the model goodness and the model complexity (Silvestrini and Burke, 2018). Simpler models are generally preferred, so the criterion penalizes models with more terms. The formula for AIC is shown in Equation 4 where n represents the sample size in the data set and p is the number of terms in the model.

$$AIC = n + n\ln 2\pi + n\ln\left(\frac{SS_E}{n}\right) + 2(p+1) \tag{4}$$

The BIC is another measure of model quality relative to model complexity. The calculation and interpretation are similar to AIC. The only difference between AIC and BIC is the penalty term placed on the number of parameters. The calculation for BIC is shown in Equation 5.

$$BIC = n + n\ln 2\pi + n\ln\left(\frac{SS_E}{n}\right) + \ln n\,(p+1) \tag{5}$$

Compare the AIC and BIC for the SDB model in Equation 1 and the full model (all main effects and two-factor interactions) in Table 1, recalling that a lower value is better. The difference between the simple and complex SDB models is quite large for both AIC and BIC, indicating the simple model in Equation 1 is the better model. Note that calculating these values for a single model does not provide any insight into the goodness of a model; these two metrics are only meaningful when comparing two or more models.

Table 1: SDB Model Comparisons

Model	AIC	BIC
Simple Model (Equation 1)	177.5	175.2
Complex Model (Full model)	232.5	184.8

AIC and BIC are commonly used in distribution fitting problems; i.e., when looking to determine which statistical distribution best fits a variable. When comparing multiple distributions, we select the distribution which has the smallest AIC or BIC, as that indicates the best model fit.

Predicted Error Sum of Squares

The previous metrics are useful to assess a model's quality in capturing the variability of the response. The following metrics consider an alternative view of model goodness: how well will the model predict future (unseen) values of the response? Evaluating the prediction performance of a model is an important step when building and assessing a statistical model, particularly if the model will be used to make predictions.

One way to assess the ability of a model to predict future values is to look at the predicted sum of squares (PRESS). PRESS is a form of cross-validation, a process where we use a dataset to both fit a model and test the goodness of the model in predicting future values. In particular, PRESS is calculated using leave-one-out cross validation. To calculate the PRESS value, the regression model is fit using all but one observation. A predicted response value, denoted $\hat{y}_{(i)}$, is calculated for that withheld observation using the fitted model. The difference between the actual and this predicted value is then calculated. This method is repeated for all the observations in the data set so that n regression models are estimated, sequentially withholding one observation from each model. The PRESS value is calculated from the differences between the actual values and predicted values as shown in Equation 6.

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}\right)^2 \tag{6}$$

The smaller the value of PRESS, the better. This value, however, is not very meaningful on its own. PRESS is more often interpreted using the PRESS root mean square error (PRESS RMSE) and R^2 Prediction.

$$\text{PRESS RMSE} = \sqrt{\frac{PRESS}{n}} \tag{7}$$

$$R^2_{prediction} = 1 - \frac{PRESS}{SS_T} \tag{8}$$

PRESS RMSE can be compared to the traditional RMSE, an estimate of the noise, σ, in the response. The RMSE and PRESS RMSE are in the same measurement units as the response, leading to better interpretability. Both metrics measure the difference on average between the responses and predicted responses, respectively, just due to sampling error. Ideally, the values of the RMSE and PRESS RMSE are

similar to each other and are both relatively low. This would indicate the model does a good job at predicting future values of the response.

Similar to R^2 and R^2_{adj}, $R^2_{prediction}$ ranges in value between 0 and 1. A high value of $R^2_{prediction}$ indicates the model has good predictive capabilities. It should also ideally be close to 1 and similar in value to R^2 and R^2_{adj}. When a model includes too many insignificant terms (called model overfitting), the predicted R^2 is typically low, indicating the model does not have good predictive capabilities. For the SDB data, Table 2 compares the predictive metrics of the simple and complex models.

Table 2: SDB Predictive Metric Model Comparisons

Model	PRESS	PRESS RMSE	$R^2_{prediction}$
Simple Model (Equation 1)	4653.3	15.3	89.5%
Complex Model (Full model)	10189	22.6	76.9%

Once again, the simpler model has much better predictive capabilities compared to the more complex model. In addition, recall that the adjusted R^2 for the simple model is 92%, which is very close to the predicted R^2 value.

Conclusions

Fitting a statistical model requires several steps to ensure the model is useful. This process entails collecting data (such as via a designed experiment), fitting a model, assessing the model assumptions, applying remedial measures as necessary, assessing the model goodness, and selecting a final model for use. The metrics discussed here can be used for comparing models from data that comes from a designed experiment or an observational study. The goal is to use these metrics to ensure that the model has the evaluation and predictive capabilities needed. The metric you use to compare models may depend on the purpose of the model. If the primary goal is to characterize a response and understand which variables affect the response, using R^2, AIC, or BIC are informative. If the primary goal is to predict future values, using the PRESS metrics are valuable.

References

Burke, S.E. "The Model Building Process Part I: Checking Model Assumptions." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), July 2017.

Burke, S.E. The Model Building Process Part II: Factor Assumptions. Scientific Test and Analysis Techniques Center of Excellence (STAT COE), July 2018.

Silvestrini, R. and S. Burke. *Linear Regression Analysis with JMP and R*. ASQ Quality Press, 2018.

Appendix

Table A.1: Notional test data for SDB test

Run	Attack Mode	Clutter	Angle off Nose	Ground Range	Miss Distance (inches)
1	Normal	50	45	17.5	74
2	Normal	20	0	30	27
3	Normal	80	0	30	158
4	Laser Illuminated	20	0	30	34
5	Laser Illuminated	20	0	5	30
6	Laser Illuminated	20	90	5	20
7	Laser Illuminated	80	90	5	56
8	Normal	20	90	5	8
9	Normal	80	0	5	133
10	Laser Illuminated	50	45	17.5	58
11	Normal	80	90	30	162
12	Laser Illuminated	50	45	17.5	55
13	Laser Illuminated	80	0	30	36
14	Laser Illuminated	80	90	30	25
15	Laser Illuminated	20	90	30	18
16	Normal	80	90	5	127
17	Normal	50	45	17.5	110
18	Laser Illuminated	80	0	5	50
19	Normal	20	0	5	30
20	Normal	20	90	30	48