Contents

[Prediction Metrics 1](#_Toc522149307)

[Mallow’s Cp 1](#_Toc522149308)

[MAE 2](#_Toc522149309)

[R² ( Coefficient of Determination ) 2](#_Toc522149310)

[RMSE 5](#_Toc522149311)

[Interview 6](#_Toc522149312)

# Prediction Metrics

Mallow’s Cp : Estimate of the size of bias introduced into the predicted response. Lower

the Cp, better the model.

Akaike information criterion (AIC) : Amount of information lost due to the predictions on the response variable. AICs’ main goal is to build model that effectively predicts the response variable.

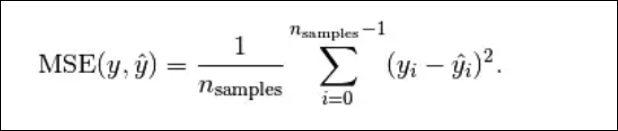
Bayesian information criterion (BIC) : BIC is similar to AIC, but BIC penalizes the

noise predictors. BICs’ main goal is to extract the features that are actually influencing

the response variable.

AIC and BIC should be less

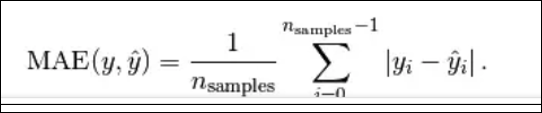
#### Mean square error:

estimates σ2, the common variance of the many subpopulations.

How does the mean square error formula differ from the sample variance formula? The similarities are more striking than the differences. The numerator again adds up, in squared units, how far each response yi is from its estimated mean. In the regression setting, though, the estimated mean is y^i. And, the denominator divides the sum by n-2, not n-1, because in using y^i to estimate  μY, we effectively estimate two parameters — the population intercept β0 and the population slope β1. That is, we lose two degrees of freedom.

# MAE

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.



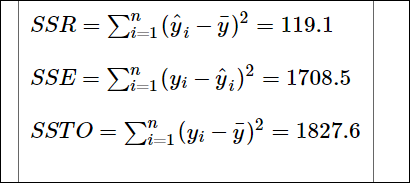
# **R² ( Coefficient of Determination )**

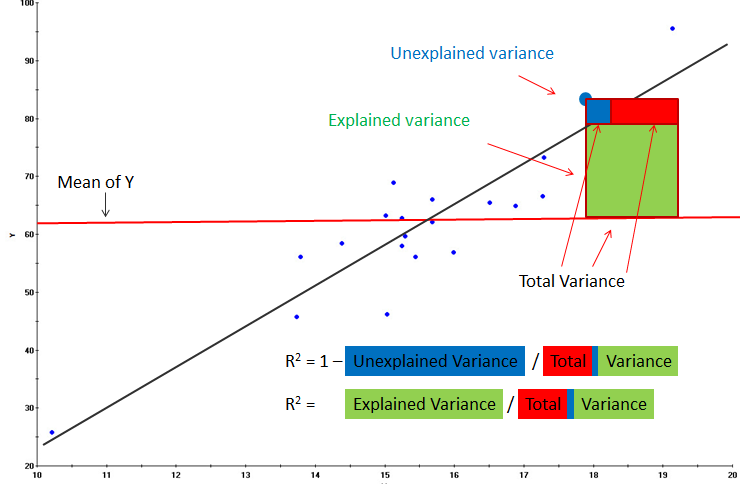
Sst = sse + ssr

Sst- total sum of squares

Sse- sum of squares unexplained

Ssr- regression sum of squares ( explained = y predicted – mean of y)^2





* SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line, y^i, is from the horizontal "no relationship line," the sample mean or y¯.
* SSE is the "error sum of squares" and quantifies how much the data points, yi, vary around the estimated regression line, y^i.

Amount if variability that is left unexplained after performing the regression

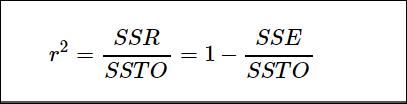
* SSTO is the "total sum of squares" and quantifies how much the data points, yi, vary around their mean, y¯.It can be thought like amount of variability inherent in the response before the regression is performed.

Note that SSTO = SSR + SSE. The sums of squares appear to tell the story pretty well. They tell us that most of the variation in the response y (SSTO = 1827.6) is just due to random variation (SSE = 1708.5), not due to the regression of y on x (SSR = 119.1).



[R-squared](http://statisticsbyjim.com/glossary/r-squared/) is a goodness-of-fit measure for linear [regression](http://statisticsbyjim.com/glossary/regression-analysis/) models. This statistic indicates the percentage of the variance in the [dependent variable](http://statisticsbyjim.com/glossary/response-variables/) that the [independent variables](http://statisticsbyjim.com/glossary/predictor-variables/) explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

. In short, the "**coefficient of determination**" or "**r-squared value**," denoted r2, is the regression sum of squares divided by the total sum of squares. Alternatively, as demonstrated in this, since SSTO = SSR + SSE, the quantity r2 also equals one minus the ratio of the error sum of squares to the total sum of squares:



R-squared is the percentage of the dependent variable variation that a linear model explains.

{\displaystyle R^2 = \frac {\text{Variance explained by the model}}{\text{Total variance}}}

R-squared is always between 0 and 100%:

* 0% represents a model that does not explain any of the variation in the response variable around its [mean](http://statisticsbyjim.com/glossary/mean/). The mean of the dependent variable predicts the dependent variable as well as the regression model.
* 100% represents a model that explains all of the variation in the response variable around its mean.

Usually, the larger the R2, the better the regression model fits your observations. However, this guideline has important caveats that I’ll discuss in both this post and the next post.

Caution with R square

Caution # 1

**The coefficient of determination r2 and the correlation coefficient r quantify the strength of a linear relationship. It is possible that r2 = 0% and r = 0, suggesting there is no linear relation between x and y, and yet a perfect curved (or "curvilinear" relationship) exists.**

**Caution # 2**

**A large r2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data**

**Caution # 3**

**The coefficient of determination r2 and the correlation coefficient r can both be greatly affected by just one data point (or a few data points).**

**Caution # 4**

**Correlation (or association) does not imply causation.**

**Caution # 7**

**A large r2 value does not necessarily mean that a useful prediction of the response ynew, or estimation of the mean response µY, can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.**

Adjusted-R²:Proportion of the response variable explained by the independent

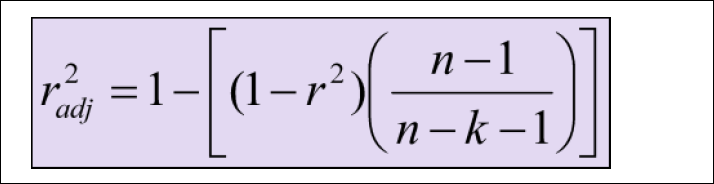
variables.

By adding more and more useless variables to a model, adjusted r-squared will decrease ;

• If you add more useful variables, adjusted r-squared will Increase

• Adjusted R 2 will always be less than or equal to R 2 ’;

• Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

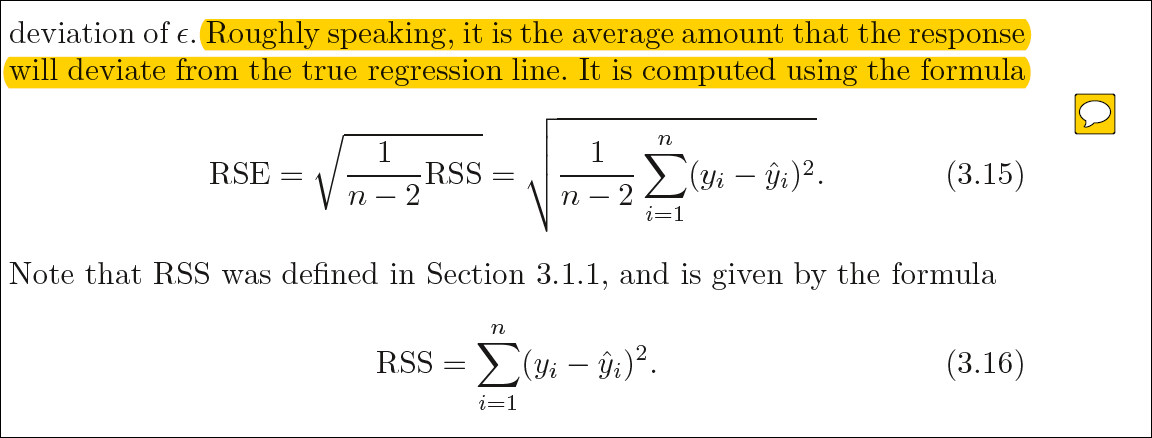


• Penalizes excessive use of unimportant independent variables • Smaller than R 2

• Useful in comparing among models

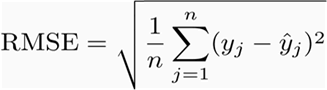
#### Residual Standard Error

RSE is an estimate of standard error.



# RMSE

RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation.

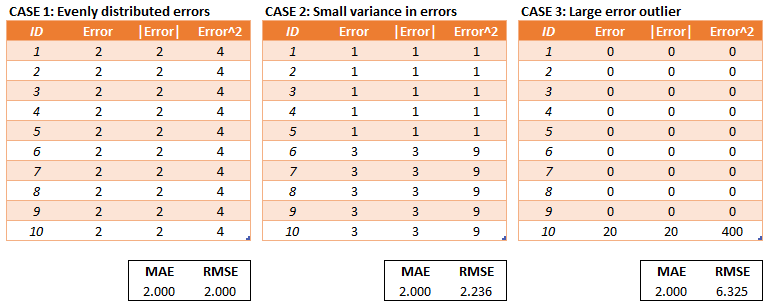


# Interview

##### In which cases would each error metric be appropriate?

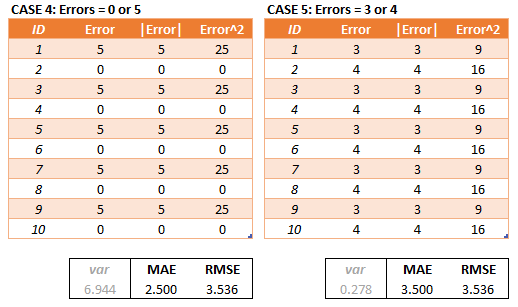
MSE is more strict to having outliers.

Differences: Taking the square root of the average squared errors has some interesting implications for RMSE. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable. The three tables below show examples where MAE is steady and RMSE increases as the variance associated with the frequency distribution of error magnitudes also increases.



*RMSE does not necessarily increase with the variance of the errors. RMSE increases with the variance of the frequency distribution of error magnitudes.*

To demonstrate, consider Case 4 and Case 5 in the tables below. Case 4 has an equal number of test errors of 0 and 5 and Case 5 has an equal number of test errors of 3 and 4. The variance of the errors is greater in Case 4 but the RMSE is the same for Case 4 and Case 5.



3,4,5 is a Pythagorean Triple

There may be cases where the variance of the frequency distribution of error magnitudes (still a mouthful) is of interest but in most cases (that I can think of) the variance of the errors is of more interest.

Another implication of the RMSE formula that is not often discussed has to do with sample size. Using MAE, we can put a lower and upper bound on RMSE.

1. **[MAE] ≤ [RMSE]**. The RMSE result will always be larger or equal to the MAE. If all of the errors have the same magnitude, then RMSE=MAE.
2. **[RMSE] ≤ [MAE \* sqrt(n)]**, where n is the number of test samples. The difference between RMSE and MAE is greatest when all of the prediction error comes from a single test sample. The squared error then equals to [MAE^2 \* n] for that single test sample and 0 for all other samples. Taking the square root, RMSE then equals to [MAE \* sqrt(n)].

*Focusing on the upper bound, this means that RMSE has a tendency to be increasingly larger than MAE as the test sample size increases.*

This can problematic when comparing RMSE results calculated on different sized test samples, which is frequently the case in real world modeling.

conclusion

RMSE has the benefit of penalizing large errors more so can be more appropriate in some cases, for example, if being off by 10 is more than twice as bad as being off by 5. But if being off by 10 is just twice as bad as being off by 5, then MAE is more appropriate.

From an interpretation standpoint, MAE is clearly the winner. RMSE does not describe average error alone and has other implications that are more difficult to tease out and understand.

On the other hand, one distinct advantage of RMSE over MAE is that RMSE avoids the use of taking the absolute value, which is undesirable in many mathematical calculations*(not discussed in this article, another time…)*.

##### How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression?

