# What is the difference in what AIC and c-statistic (AUC) actually measure for model fit?

Akaike Information Criterion (AIC) and the c-statistic (area under ROC curve) are two measures of model fit for logistic regression. I am having trouble explaining what is going on when the results of the two measures are not consistent. I guess they are measuring slightly different aspects of model fit, but what are those specific aspects?

I have 3 logistic regressions models. Model M0 has some standard covariates. Model M1 adds X1 to M0; model M2 adds X2 to M0 (so M1 and M2 are not nested).

The difference in AIC from M0 to both M1 and M2 is about 15, indicating X1 and X2 both improve model fit, and by about the same amount.

c-statistics are: M0, 0.70; M1, 0.73; M2 0.72. The difference in c-statistic from M0 to M1 is significant (method of DeLong et al 1988), but the difference from M0 to M2 is not significant, indicating that X1 improves model fit, but X2 does not.

X1 is not routinely collected. X2 is supposed to be routinely collected but is missing in about 40% of cases. We want to decide whether to start collecting X1, or improve collection of X2, or drop both variables.

From AIC we conclude that the variables make similar improvement to the model. It's probably easier to improve collection of X2 than start collecting a completely new variable (X1), so we would aim to improve X2 collection. But from c-statistic, X1 improves the model and X2 does not, so we should forget about X2 and start collecting X1.

As our recommendation depends on which statistic we focus on, we need to clearly understand the difference in what they are measuring.

Any advice welcome.

logistic roc aic auc

edited Mar 5 '13 at 17:33
Corone
**2,906** 1 11 40

asked Mar 4 '13 at 4:40
timbp
**593** 1 4 12

## 2 Answers

AIC and c-statistic are trying to answer different questions. (Also some issues with c-statistic have been raised in recent years, but I'll come onto that as an aside)

Roughly speaking:

- AIC is telling you how good your model fits for a *specific* mis-classification cost.
- AUC is telling you how good your model would work, on average, across all mis-classification costs.

When you calculate the AIC you treat your logistic giving a prediction of say 0.9 to be a prediction of 1 (i.e. more likely 1 than 0), however it need not be. You could take your logistic score and say "anything above 0.95 is 1, everything below is 0". Why would you do this? Well this would ensure that you only predict one when you are really really confident. Your false positive rate will be really really low, but your false negative will skyrocket. In some situations this isn't a bad thing - if you are going to accuse someone of fraud, you probably want to be really really sure first. Also, if it is very expensive to follow up the positive results, then you don't want too many of them.

This is why it relates to costs. There is a cost when you classify a 1 as a 0 and a cost when you classify a 0 as a 1. Typically (assuming you used a default setup) the AIC for logistic regression refers to the special case when both mis-classifications are equally costly. That is, logistic regression gives you the best overall number of correct predictions, without any preference for positive or negative.

The ROC curve is used because this plots the true positive against the false positive in order to show how the classifier would perform if you used it under different cost requirements. The c-statistic comes about because any ROC curve that lies strictly above another is clearly a dominating classifier. It is therefore intuitive to measure the area under the curve as a measure of how good the classifier overall.

So basically, if you know your costs when fitting the model, use AIC (or similar). If you are just constructing a score, but not specifying the diagnostic threshold, then AUC approaches are needed (with the following caveat about AUC itself).

**So what is wrong with c-statistic/AUC/Gini?**

For many years AUC was the standard approach, and is still widely used, however there are a number of problems with it. One thing that made it particularly appealing was that it corresponds to a Wilcox test on the ranks of the classifications. That is it measured the probability that the score of a randomly picked member of one class will be higher than a randomly picked member of the other class. The problem is, that is almost never a useful metric.

The most critical problems with AUC were publicized by David Hand a few years back. (See references below) The crux of the problem is that while AUC does average over all costs, because the x-axis of the ROC curve is False Positive Rate, the weight that it assigns to the different cost regimes varies between classifiers. So if you calculate the AUC on two different logitic regressions it won't be measuring "the same thing" in both cases. This means it makes little sense to compare models based on AUC.

Hand proposed an alternative calculation using a fixed cost weighting, and called this the H-measure - there is a package in R called `hmeasure` that will perform this calculation, and I believe AUC for comparison.

Some references on the problems with AUC:

- *When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?* D.J. Hand, C. Anagnostopoulos **Pattern Recognition Letters** 34 (2013) 492–495

  (I found this to be a particularly accessible and useful explanation)

edited Mar 4 '13 at 10:26

answered Mar 4 '13 at 7:31

Corone

**2,906** 1 11 40

2 And here is another paper by D.J. Hand: Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning* (2009) 77: 103–123. – chl ♦ Mar 4 '13 at 12:16

That was the one I was looking for - yes that was the key first paper on this (although I think it consequently is targeted at a more technical audience than some of the later papers). – Corone Mar 4 '13 at 12:25

2 AUC (C-index) has the advantage of measuring the concordance probability as you stated, aside from cost/utility considerations. To me the bottom line is the AUC should be used to describe discrimination of one model, not to compare 2 models. For comparison we need to use the most powerful measure: deviance and those things derived from deviance: generalized $R^2$ and AIC. – Frank Harrell Mar 4 '13 at 12:28

I am confused by Corone's answer, I thought AIC did not have anything to do with the predictive performance of a model and that it is just a measure of the likelihood of the data traded off with model complexity. – Zhubarb Oct 28 '13 at 16:15

@Berkan not sure what you mean by "nothing to do with predictive performance", unless you simply mean it is an in-sample measure not out-of-sample? (The better the likelihood the better it "predicts" those data points). The point is that AIC is for a specific, pre chosen likelihood function, whereas the AIC is an average over a set of them. If you know the likelihood (i.e. threshold, costs, prevalence...) then you can use AIC. – Corone Oct 29 '13 at 10:11

For me, the bottom line is that while the C-statistic (AUC) may be problematic when comparing models with different independent variables (analogous to what Hand refers to as "classifiers"), it is still useful in other applications. For instance, validation studies where the same model is compared across different study populations (data sets). If a model or risk index/score is shown to be highly discriminant in one population, but not in others, this could mean indicate that it is not a very good tool in general, but may be in specific instances.

answered Sep 23 '13 at 20:07

Dave

**1**

1 The C-index is too insensitive to be used to compare different models, in general. I would typically use the generalized $R^2$ or other deviance-based measures including AIC. And note that AIC is *not* related to classification/cutpoints. – Frank Harrell Sep 23 '13 at 20:27