Assessing How Well a Model Fits the Data

Brian Hunt
University of Maryland

AMSC/MATH 420, Spring 2015

“Best” Fit

- Given a model depending on some parameters, and some data, we have said that certain parameter values “best” fit the data if they minimize the error quantified by the sum of the squares of the residuals:

$$E = \sum_{j=1}^{J} R_j^2$$

- Each residual R_j is the difference between the observed value y_j (or some related value) and the model prediction for this value.
- While E quantifies the relative quality of different fits, the value of E is not so easy to interpret.

RMS Error

- The RMS error of a fit with residuals $R_1, \ldots, R_J$ is

$$\text{RMS error} = \sqrt{\frac{1}{J}\sum_{j=1}^{J} R_j^2}$$

- Minimizing the RMS error is equivalent to minimizing the sum-of-squares error E, but the RMS error has a more natural interpretation.
- The RMS error has the same units as the residuals and the data, unlike E.
- It is the root-mean-square average of the residuals, so it is not proportional to the number of data points like E is.

Does a parameter improve the model?

- Suppose we want to compare a model $f(t; \beta_1, \ldots, \beta_k, \beta_{k+1})$ to the model $f(t; \beta_1, \ldots, \beta_k, 0)$ with one fewer parameter.
- The best fit with the former model will always have an error no larger than the best fit with the latter model.
- How can we tell if the improvement is enough to make the additional parameter worth using?
- One approach is to develop a statistical model for the errors, in order to quantify the "significance" of the improvement.
- More simply (perhaps too simply), one can make a value judgment that (say) a 1% improvement is not worth the complication, but a 10% improvement is.

Does the parameter improve predictions?

- If the goal of the model is to predict data that hasn't yet or can't be measured, then we can assess whether the model with the additional parameter makes better predictions.
- Keep in mind that the more complicated model might make worse predictions than the simpler model.
- Thus, comparing the models' prediction residuals is more of a "fair fight" than comparing their residuals for the fitted data.
- How can we assess the quality of the predictions without waiting for new data to be available?

Training and Test Data

- A common way to assess predictive power of a model for data taken at a sequence of times starts by dividing the data into two time intervals.
- The data from the first time interval is called the training data set; the data from the second time interval is called the test data.
- Fit the model to the training data only, then see how well the parameters that best fit the training data are able to predict the test data.
- What proportion of the data to put in the training set depends on how much data you have and how far into the future you want the model to predict.

Interpreting the Results

- The RMS error for the test data, both by itself and in comparison with the RMS error for the training data, give some assessment of the model's predictive power. However, a simple comparison of training and test RMS errors is inadequate if the two data sets have different amounts of variability.
- Comparing the test data RMS errors for two different models is a reasonable way to assess which makes better predictions (for the time interval of the test data, at least).
- Whatever conclusions you draw, they are more convincing if tested on multiple data sets.

Model Heirarchy

- Suppose we have two models A and B. Let's write A $\prec$ B if setting a certain parameter or parameters of model B to 0 reduces model B to model A. More colloquially, this means that the model B equations consist of the model A equations plus some additional term(s). For example, SI $\prec$ SIR.
- Writing SIg and SIRg for the SI and SIR models with the growth/renewal term added, we also have SI $\prec$ SIg $\prec$ SIRg and SIR $\prec$ SIRg.
- If A $\prec$ B, then model B can fit any data set at least as well as model A.
- This doesn't necessarily mean that model B better describes the process that generated the data than model A does.

Comparing Models with Test/Training Data

- Suppose A $\prec$ B (“model B” adds an additional parameter or parameters to “model A”).
- If model B consistently predicts test data better than model A, this is a good argument in favor of model B.
- If model B predicts test data worse than model A, despite fitting the training data better, then model B may be “overfitting” the training data. But not necessarily; model B may still have some advantage over model A. For example, it may be worse at extrapolation but better at interpolation (this could be examined using test data that is interspersed in time with training data).

Fits to first 80% of San Francisco Data

Model	SI	SIR	SIg	SIRg
N	29237	28671	58503	*
λ	0.032288	0.016684	0.035714	0.061
μ	—	—	–0.071832	–0.078
ν	—	–0.74075	—	0.24
α	0.021206	0.054567	0.010270	*
rms 80%	24.015	23.168	23.234	23.173
rms 20%	14.061	35.888	69.313	63.6
rms 100%	22.826	26.212	37.319	35.2

- As on last week's slides, β was set to 0.
- For the SIRg model, minimization does not appear to converge; the rms error and some of the parameters nearly stabilize, but the error keeps getting slightly smaller as N increases well beyond reason.

Predicting Farther into the Future

- In 2013, San Francisco had 359 new HIV diagnoses (source: `http://sfaf.org/hiv-info/statistics/`).
- The models we studied, using parameters fit to the entire 1982–2001 data set, predict the following number of diagnoses for 2013:
    - SI: 3
    - SIR: 16
    - SIg: 219
    - SIRg: 555
- The SI and SIR model severely underpredict the new diagnoses 12 years later.
- The SIg and SIRg model both give reasonable ballpark predictions.