Chapter 12.2c Comparing Regression Models

Jim Albert and Monika Hu

Chapter 12 Bayesian Multiple Regression and Logistic Models

Introduction

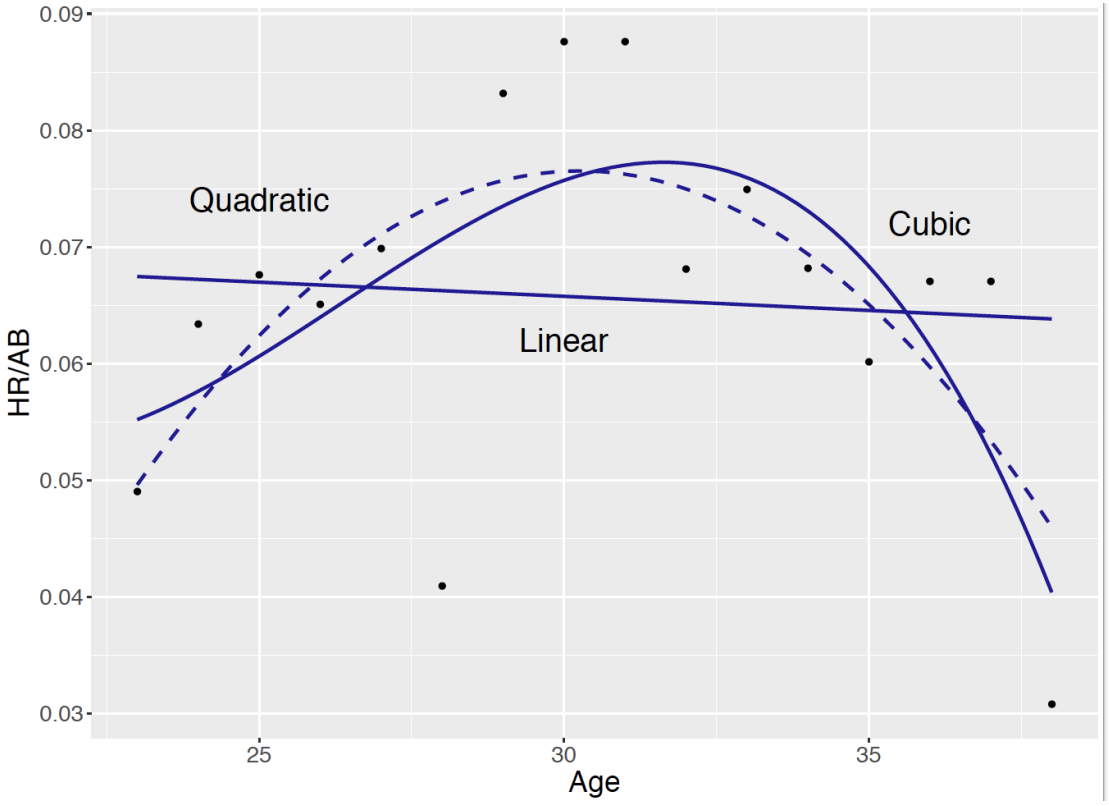
- When one fits a multiple regression model, there is a list of inputs and many possible regression models to fit.
- In the household expenditures example, there are two possible inputs, the log total income and the rural/urban status and there are 4 possible models depending on the inclusion or exclusion of each input.
- When there are many inputs, the number of possible regression models can be quite large and so there needs to be some method for choosing the "best" regression model.
- Describe a general method for selecting between models.

Learning about a career trajectory

- Consider a baseball modeling problem that will be discussed in Chapter 13.
- One is interested in seeing how a professional athlete ages during his or her career.
- One can use a regression model to explore the pattern of performance over age – this pattern is typically called the athletic's career trajectory.

Fitting models to a career trajectory

- Figure displays a scatterplot of the rate that the ballplayer Mike Schmidt hit home runs as a function of his age.
- Consider Linear, Quadratic, and Cubic model fits.



Comments

- Model 1 says that Schmidt's home run performance is a linear function of his age.
- Model 2 says that his home run performance follows a parabolic shape
- Model 3 indicates that his performance follows a cubic curve.
- The linear function of age given in Model 1 does not appear suitable but the fits of Models 2 and 3 appear to be similar in appearance.
- How can we choose between the two models?

Underfitting and overfitting

- Want the model to include all inputs helpful in explaining the variation in the response variable. Failure to include relevant inputs in the model will result in underfitting.
- Model 1 likely underfits the data.
- But one should be careful not to include too many inputs in the model.
- When one includes more inputs in our regression model than needed, one has overfitting.
- Model 3 possibly overfits the data, since it may not be necessary to represent a player's trajectory by a cubic curve.

Cross-validation

- How does one choose a suitable regression model that avoids the underfitting and overfitting problems?
- In cross-validation, one partitions the dataset into two parts – the "training" and "testing" components.
- One initially fits each regression model to the training dataset. Then one uses each fitted model to predict the response variable in the testing dataset.
- The model that is better in predicting observations in the future testing dataset is the preferred model.

Apply cross-validation here

- One randomly divides Schmidt's 8170 career at-bats into two datasets – 4085 of the at-bats are placed in a training dataset and the remaining at-bats become the testing dataset.
- For a particular model, one fits a regression model with a weakly informative prior using JAGS. Obtain a fitted regression using the posterior means of the regression coefficients.
- Use this fitted regression to predict values of the home run rate from the testing dataset. Measure the goodness of the prediction by computing the sum of squared prediction errors (SSPE).
- One uses this measure to compare predictions from alternative regression models. The best model is the model corresponding to the smallest value of SSPE.

Approximating cross-validation by DIC

- The cross validation method of assessing model performance can be generally applied in many situations.
- However, there are complications in implementing cross validation in practice.
- One issue is how the data should be divided into the training and testing components.
- That raises the question – is it necessary to perform cross validation to compare the predictive performance of two models?
- An alternative approach can be used to compare the predictive performance of models.

Approximating cross-validation by DIC

- A best regression model is the one that provides the best predictions of the response variable in an out-of-sample or future dataset.
- One can compute a measure, called the *Deviance Information Criteria* or DIC, from the simulated draws from the posterior distribution that approximates a model's out-of-sample predictive performance.
- The derivation of the DIC measure is outside of is contained in the appendix.
- But it can be applied generally and is helpful for comparing the predictive performance of several Bayesian models.

Example of model comparison using DIC

- Return to the career trajectory example. As usual practice,
- JAGS will be used to fit a specific Bayesian model such as the quadratic model M_2.
- At the sampling stage, the home run rates `y[i]` are assumed to be a quadratic function of the ages `x[i]`, and at the prior stage, the regression coefficients `beta0`, `beta1`, `beta2`, and the precision `phi` are assigned weakly informative priors. The variable `the_data` is a list containing the observed home run rates, ages, and sample size.

JAGS Script and Data Descriptions

```
modelString = "
model {
for (i in 1:N){
   y[i] ~ dnorm(mu[i], phi)
   mu[i] <- beta0 + beta1 * (x[i] - 30) +
            beta2 * pow(x[i] - 30, 2)
}
beta0 ~ dnorm(0, 0.001)
beta1 ~ dnorm(0, 0.001)
beta2 ~ dnorm(0, 0.001)
phi ~ dgamma(0.001, 0.001)
}
"
d <- filter(sluggerdata,
            Player == "Schmidt", AB >= 200)
the_data <- list(y = d$HR / d$AB,
                 x = d$Age,
```

Fitting Model

- The model is fit by the `run.jags()` function.
- To compute DIC, one needs to run multiple chains, which is indicated by the argument `n.chains = 2`.

```
post2 <- run.jags(modelString,
                  n.chains = 2,
                  data = the_data,
                  monitor = c("beta0", "beta1",
                              "beta2", "phi"))
```

Computing DIC

- To compute DIC, the `extract.runjags()` function is applied on the runjags object `post2`.
- The "Penalized deviation" output is the value of DIC computed on the simulated MCMC output.

```
extract.runjags(post2, "dic")
Mean deviance:  -88.98
penalty 4.817
Penalized deviance: -84.17
```

- The value of DIC $= -84.17$

Comparing Values of DIC

- Computing DIC for a single model is not meaningful, but one compares values of DIC for competing models.
- Suppose one wishes to compare models M_1, M_2, M_3 and a quartic regression where one represents the home run rate as a polynomial of fourth degree of age.
- For each model, a JAGS script is written with weakly informative priors. The `run.jags()` function produces a posterior sample and the `extract.runjags()` with the `dic` argument extracts the value of DIC.

Comparing Values of DIC

- Table 12.2 displays the values of DIC for the four models. The "best" model is the model with the smallest value of DIC.
- The quadratic model has the smallest value of DIC of -84.2 – this model will provide the best out-of-sample predictions.

Model	DIC
Linear	-80.4
Quadratic	-84.2
Cubic	-82.1
Quartic	-79.0