Model Checking and Improvement

Statistics 220

Spring 2005

Model Checking

"All models are wrong but some models are useful"

– George E. P. Box

So far we have looked at a number of models and examined them with example data sets. Do the models used accurately describe the data used?

In standard analyses, we will often check model assumptions. For example, in standard regression we will check for

- Correct form of the regression function (e.g. linear vs quadratic)
- Constant variance of the residuals
- Independence of of the residuals
- Normality of the residuals

Basic question: How sensitive are our posterior inferences to our modelling assumptions?

Rat Example: Will the following models give significantly different answers about tumor rates in each group

1. Original model

   - Data model: $y_i =$ number of tumors in group i

$$y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \ldots, 71$$

   - Process model: $\theta_i =$ tumor rate in group i

$$\theta_i|\alpha, \beta \overset{ind}{\sim} Beta(\alpha, \beta)$$

   - Parameter model:

$$p(\alpha, \beta) \propto \frac{1}{(\alpha + \beta)^{5/2}}$$

2. Alternative model 1

- Data model: y_i = number of tumors in group i

$$y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \ldots, 71$$

- Process model: θ_i = tumor rate in group i

$$\text{logit}(\theta_i)|\mu, \sigma^2 \overset{ind}{\sim} N(\mu, \sigma^2)$$

- Parameter model:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

3. Alternative model 2

- Data model: y_i = number of tumors in group i

$$y_i|\alpha_i,\beta_i \overset{ind}{\sim} Beta-bin(n_i,\alpha_i,\beta_i) \quad i=1,\ldots,71$$

- Process model: (α_i,β_i) = tumor rate parameters in group i

$$\alpha_i,\beta_i|\gamma_\alpha,\delta_\alpha,\gamma_\beta,\delta_\beta \overset{ind}{\sim} Gamma(\alpha_i|\gamma_\alpha,\delta_\alpha)Gamma(\beta_i|\gamma_\beta,\delta_\beta)$$

  The tumor rate for group i is

$$E\left[\frac{y_i}{n_i}|\alpha_i,\beta_i\right]=\frac{\alpha_i}{\alpha_i+\beta_i}$$

- Parameter model:

$$p(\gamma_\alpha,\delta_\alpha,\gamma_\beta,\delta_\beta)\propto 1$$

Note that we will not be trying to answer the question of whether our model is correct or not. Its not (see Box). We are interested in whether the inaccuracies matter.

Examples you may have seen in the past where deviations from assumptions don't hurt much (at least in big samples):

- t-test of $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$

  Normality often isn't important, though large skewness can hurt.

- Linear Regression: $Y = X\beta + \epsilon$
  - $\hat{\beta} = (X^T X)^{-1} X^T Y$ is unbiased if $E[\epsilon] = 0$
  - $\hat{\beta}$ is minimum variance unbiased estimator if $E[\epsilon] = 0$ and constant variance. (Gauss-Markov theorem)

  Neither of these results require normality of ϵ.

There are cases where assumptions can matter. For example consider the F-test for examining $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_A : \sigma_1^2 \neq \sigma_2^2$. The results of this test can be highly dependent on the iid normal assumptions for each group.

One approach to build a super-model that contains all of our models of interest as special cases. This approach usually isn't taken as it is usually difficult to build this super-model and computation is usually infeasible, assuming you can build the model.

Instead we will base these checks on the posterior predictive distribution. Does our data look like our fitted model says it should.

This can either be done by

- *External validation*: future data is compared with the posterior predictive distribution.

- *Internal validation*: observed data is compared with the posterior predictive distribution.

Posterior Predictive Checking

Idea: If the model fits, replicated data generated under the model should look similar to the observed data.

If we see some discrepancy, is it due to model misspecification or due to chance.

Approach: Generate L datasets, $y_1^{rep}, \ldots, y_L^{rep}$ from the posterior predictive distribution $p(y^{rep}|y)$. y^{rep} corresponds to replicated data. So if there are any covariates that are conditioned on in the original data.

For example, in the rat tumor example, we need to use the same group sample sizes as in the original data set.

$\tilde{y}$ represents any future outcome whereas y^{rep} indicates a replication exactly like the observed y. $\tilde{y}$ does not need to have the same covariate structure as the original data.

The approach has a similar feel to hypothesis testing, where a test statistic $T(y, \theta)$ needs to be defined to measure the discrepancy between the data and the predictive simulations.

Note that the test statistic can depend on the data y and the parameters and hyperparameters θ, which is different from standard hypothesis testing where the test statistic only depends on the data, but not the parameters.

Tail-area probabilities

The lack of fit of the data as compared to the posterior predictive distribution can be compared by a tail-area probability (e.g. p-value) of the test statistic $T(y, \theta)$. To calculate this probability we will use the replicates sampled from $p(y^{rep}|y)$.

- Classical p-value

$$p_C = P[T(y^{rep}) \geq T(y)|\theta]$$

  where the probability is calculated over the distribution of y^{rep} given a fixed θ. In the classical testing setting θ would correspond to the null hypothesis value. It could also be a point estimate (say the MLE).

- Posterior predictive p-values

  To evaluate the fit of a Bayesian model, we need to consider what possible data sets are plausible under the model. When doing this we need to consider not only the observations y, but also the parameter values θ. Thus the p-value of interest is

$$\begin{aligned} p_B &= P[T(y^{rep}, \theta) \geq T(y, \theta)|y] \\ &= \int \int I(T(y^{rep}, \theta) \geq T(y, \theta)) p(y^{rep}|\theta) p(\theta|y) dy^{rep} d\theta \end{aligned}$$

  Usually we can't calculate the Bayesian p-value exactly, but can do it by simulation. Suppose that we have L simulations of $\theta(\theta^1, \ldots, \theta^L)$ from the posterior distribution $p(\theta|y)$. Then for each of these θ samples, generate one sample y^{repl} from $p(y^{rep}|\theta^l)$.

We want to compare each of the $T(y^{repl}, \theta^l)$ with $T(y, \theta^l)$

Then

$$\hat{p}_B = \frac{1}{L} \sum_{l=1}^{L} I(T(y^{repl}, \theta^l) \geq T(y, \theta^l))$$

(i.e. the proportion of samples where $T(y^{repl}, \theta^l) \geq T(y, \theta^l)$) is an estimate of p_B.

Note that the test statistic $T(y, \theta)$ needs to be chosen to investigate deviations of interest. This is similar to choosing a powerful test statistic when conducting a hypothesis test

For example, in the analysis of Newcomb's speed of light experiment discussed in the text, a worry was the effect of outliers. Thus $T(y, \theta)$ needs to be chosen to focus on this issue.

In the book $T(y, \theta) = \min y_i$ was used (they were worried about low outliers). Another possibility would be $T(y, \theta) = \max |y_i - \mu|$ (e.g. the biggest residual in magnitude). This would be appropriate if the worry was either big positive or big negative residuals. This might occur if

$$y_i | \mu, \sigma^2 \sim t_\nu(\mu, \sigma^2)$$

instead of

$$y_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

as used in the analysis.

While the approach is a bit more focused on test statistics, this has a similar feel to residual analysis in regression.

- Is there any pattern in the residual plot (e_i vs $\hat{y}_i$)
- Plotting e_i vs e_{i-1} or the Durbin-Watson test to examine whether residuals are correlated over time
- Normal scores plot or Anderson-Darling test for normality of residuals

For example, if we see some curvature (but constant variance) in the residual plot, it suggests we might be missing a x^2 term in the model.

If there is some curvature and non-constant variance in the residual plot, maybe we need to transform y.

Examples:

For the two random effects model examples (detergent filling and sodium level in beer), two concerns might be

1. Conditional normality of the observations (e.g. $y_{ij} \sim N(\theta_j, \sigma^2)$)

2. Constant variance of observations within each group

Note that these are probably of limited concern in both of these examples, as the total sample size is fairly large and there are equal numbers of observations in each group for both data sets.

Possible test statistics to evaluate these are

1. Normality: Let $e_{ij} = y_{ij} - \theta_j$ and $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$ be the ordered residuals. Let

$$T(y, \theta) = \text{Corr}\left(e_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$$

(e.g. Correlation of points in a normal scores plot). If the data is conditionally normal, this correlation should be close to one. Otherwise the normal scores plot will have some non-linearity, which will pull this correlation down from one.

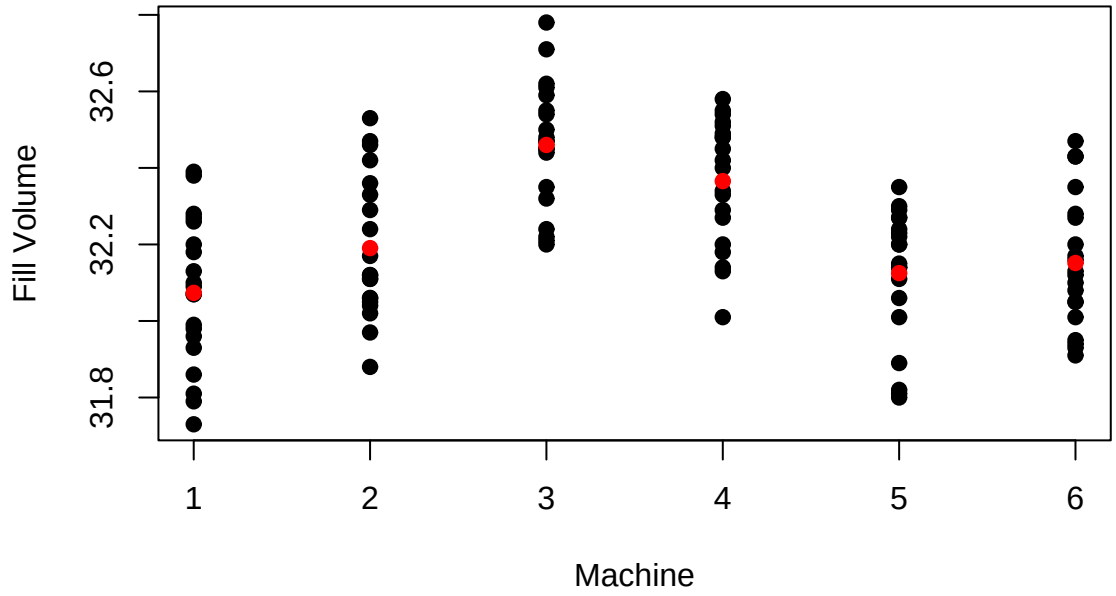
(This test statistic has the feel of the Shapiro-Wilks normality test.)

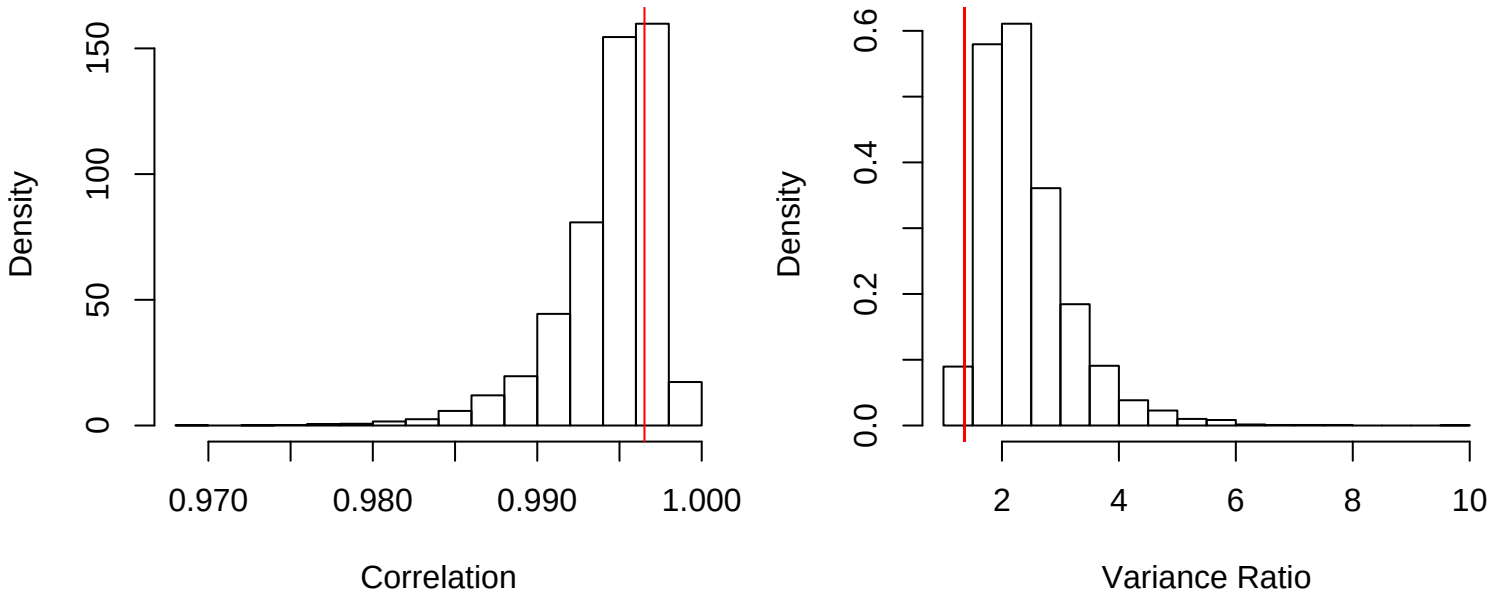
2. Equal variance: Let s_i^2 be the sample variance of the observation in group i. If the constant variance assumption is reasonable

$$T(y, \theta) = \frac{\max s_i^2}{\min s_i^2}$$

should not be much bigger than one.

Detergent example:

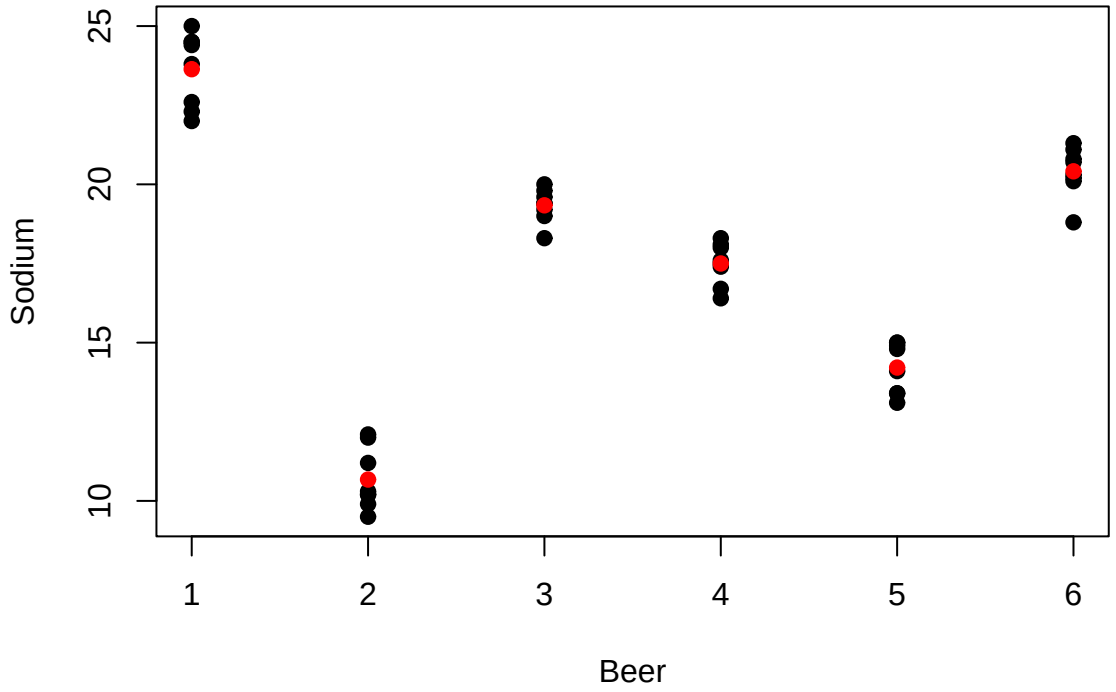


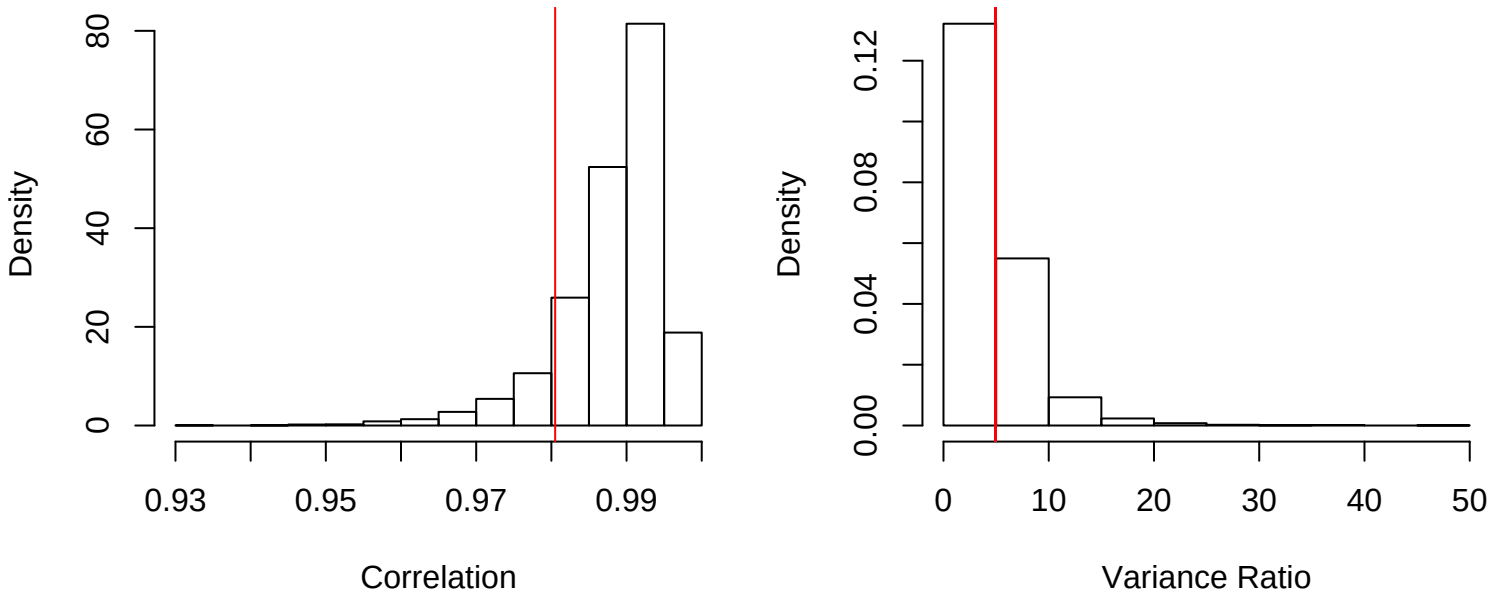


Normality test: $\hat{p}_B = 0.4886$

Equal variance test: $\hat{p}_B = 0.9866$

Beer example:





Normality test: $\hat{p}_B = 0.5270$

Equal variance test: $\hat{p}_B = 0.3520$