Model Checking

Patrick Lam

Outline

Posterior Predictive Distribution

Posterior Predictive Checks
- An Example

Outline

Posterior Predictive Distribution

Posterior Predictive Checks

An Example

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

We may want to make predictions in order to:

1. Predict how a system would behave in the future (substantive implications)
2. Assess model accuracy (modeling implications)

Through simulation, we can get a **posterior predictive distribution**.

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$\begin{aligned} p(y^{\text{rep}}|y) &= \int p(y^{\text{rep}}, \theta|y) d\theta \\ &= \int p(y^{\text{rep}}|\theta, y) p(\theta|y) d\theta \end{aligned}$$

If we assume $y \perp y^{\text{rep}}|\theta$, then

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|y) p(\theta|y) d\theta$$

If y is a vector of n observations, then y^{rep} is also a vector of length n with covariates set at the observed (model checking) or hypothetical values (prediction) and $p(y^{\text{rep}}|y)$ can be thought of as an n-variate distribution.

We can simulate the posterior predictive distribution.

1. Sample m values of θ from our posterior.
2. For each posterior draw, sample a value (vector) of y^{rep} from our likelihood $p(y^{\text{rep}}|\theta)$.

The m values (vectors) of y^{rep} represent draws from the posterior predictive distribution $p(y^{\text{rep}}|y)$.

We can use the posterior predictive distribution to predict the future or assess model accuracy with posterior predictive checks.

Outline

Posterior Predictive Distribution

Posterior Predictive Checks
An Example

Much of what we have done so far is based on a model that we specify, which may or may not be accurate.

Specifically, we make many assumptions with our model which may or may not be accurate (for example, independence across observations).

We can attempt to check specific model assumptions with **posterior predictive checks**.

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\text{rep}}|y)$
4. Calculate the fraction of times $T(y^{\text{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.

The idea is that if our data violates one of our model assumptions, then our observed test statistic $T(y)$ should be significantly different than our model predicted test statistics $T(y^{\text{rep}}|y)$.

If our posterior predictive p-value is close to 0 or 1 (say 0.05 or 0.95), then it suggests that our observed data has an extreme test statistic and that something in our model may be inadequate.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).
- If the model passes posterior predictive check, it does not necessarily mean there are no problems with the model.
  - Test statistic may have low power.
  - May be testing the wrong assumption.
- It is not always clear how to correct the incorrect model assumptions.

Outline

Posterior Predictive Distribution

Posterior Predictive Checks
- An Example

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Dependent variable: binary variable on civil war onset

Independent variables: the normal set of independent variables predicting civil wars

Model: Bayesian logistic regression with binomial likelihood and multivariate Normal priors (using `MCMCpack`)

```
> library(MCMCpack)
> posterior <- MCMClogit(new.onset ~ warl + gdpenl + lpopl1 + lmtnest +
+     ncontig + Oil + nwstate + instab + polity2l + ethfrac + relfrac,
+     data = data, tune = 0.6, burnin = 1000, mcmc = 5000)

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
The Metropolis acceptance rate for beta was 0.32250
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

Posterior Predictive Distribution

1. Create model matrix of covariates X.

```
> X <- cbind(1, data[, c("warl", "gdpenl", "lpopl1", "lmtnest",
+     "ncontig", "Oil", "nwstate", "instab", "polity2l", "ethfrac",
+     "relfrac")])
```

2. Get linear predictors by multiplying X and our m draws from the posterior.

```
> Xb <- as.matrix(X) %*% t(posterior)
```

3. Convert linear predictors into probabilities with the inverse logit function.

```
> probs <- 1/(1 + exp(-Xb))
```

4. Draw m samples of y^{rep} from the binomial likelihood.

```
> n <- nrow(X)
> m <- nrow(posterior)
> y.rep <- matrix(NA, nrow = n, ncol = m)
> for (i in 1:m) {
+     y.rep[, i] <- rbinom(n, size = 1, prob = probs)
+ }
```

The resulting posterior predictive distribution is an $n \times m$ matrix.

A Bad Test Statistic

Let T = the fraction of y's that take on the value of 1

What's wrong with this test statistic?

- Unclear what assumption are we testing.
- The fraction of 1s is explicitly being modeled in the logit model.
  - The test will never show anything is wrong regardless of how bad our model is.

A Better Test Statistic

Assumption: No clustering within years

Test Statistic: T = the variance of the number of 1s in each year

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

2. Calculate T for the observed data y: $T(y)$

```
> emp.year.sum <- c()
> for (i in 1:length(unique(data$year))) {
+     emp.year.sum[i] <- sum(data$new.onset[which(data$year ==
+         unique(data$year)[i])])
+ }
> T.y <- var(emp.year.sum)
```

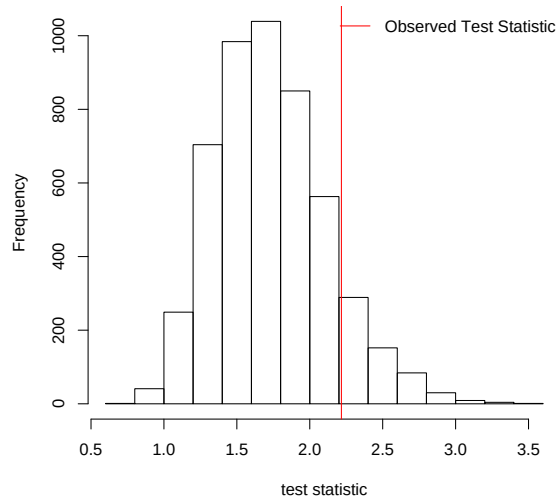
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\text{rep}}|y)$

```
> year.sum <- matrix(NA, nrow = length(unique(data$year)), ncol = ncol(y.rep))
> for (i in 1:length(unique(data$year))) {
+     year.sum[i, ] <- apply(y.rep[which(data$year == unique(data$year)[i]),
+         ], 2, sum)
+ }
> T.y.rep <- apply(year.sum, 2, var)
```

4. Calculate the fraction of times $T(y^{\text{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.

```
> mean(T.y.rep > T.y)

[1] 0.107
```



Does this mean our assumption is correct? Not necessarily (low power?)