Bayesian Statistics

Model Assessment

Nan Lin
Department of Mathematics
Washington University in St. Louis

Prediction

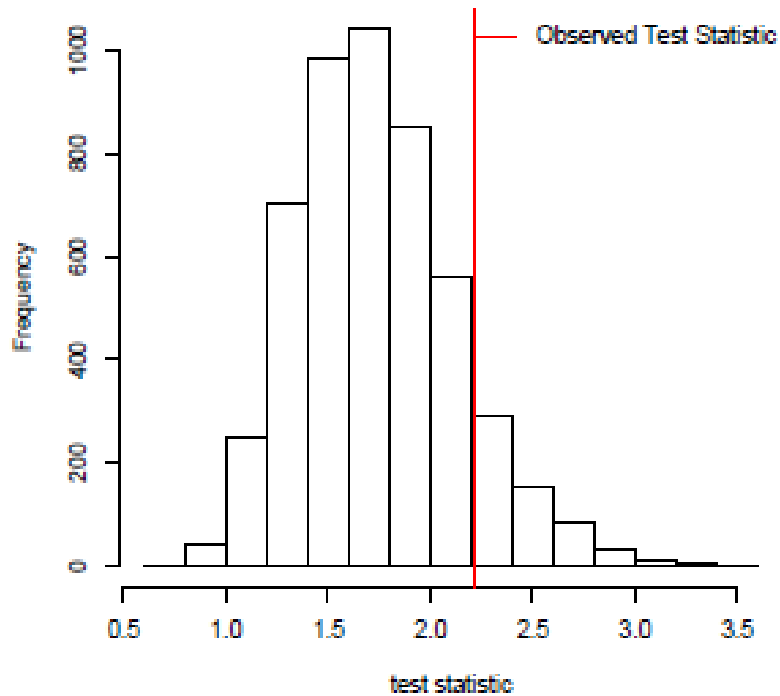
- Given a model and draws from the posterior distribution, we make prediction for future data points by simulating from the posterior predictive distribution.
- Consider making prediction for some future data point(s) $\tilde{y}$ based on the observed data y, the posterior predictive distribution is then $p(\tilde{y}|y) = \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta$
  - If we assume that $\tilde{y}$ and y are conditionally independent given θ, then $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$
- Simulating from the posterior predictive distribution
  1. Sample m values of θ from the posterior distribution $p(\theta|y)$
  2. For each simulated value of θ, sample a $\tilde{y}$ from $p(\tilde{y}|\theta)$

Model assessment by predictive checks

- Suppose that the data set contains some covariates, one may make predictions at
  - the observed values of the covariates → assess model accuracy
  - or at some hypothetical values → prediction the future

Posterior predictive checks

- Define a test statistic T that has power to diagnose violations of the assumption to be tested
  - Assume that a large value of T indicates violation
- Calculate T for the observed data y: $T(y)$
- Calculate $T(\tilde{y}|y)$ for each $\tilde{y}$ drawn from the posterior predictive distribution
- Calculate the fraction of $T(\tilde{y}|y) > T(y)$. This is an estimate of the ***posterior predictive p-value***.
  - 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.


Observed Test Statistic
Frequency
0
200
400
600
800
1000
0.5
1.0
1.5
2.0
2.5
3.0
3.5
test statistic

Possible Problems

- 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
  - Test statistics should not be based on aspects of the data that are being explicitly 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.

Example

- Model: Bayesian logistic regression with multivariate normal prior
  - $y_i | p_i \sim Bin(n_i, p_i)$ and $\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- Simulating from the posterior predictive distribution
  1. Create model matrix of covariates X.
  2. Get linear predictors by multiplying X and our m draws from the posterior.
  3. Convert linear predictors into probabilities with the inverse logit function.
  4. Draw m samples of $\tilde{y}$ from the binomial likelihood.
  - Output: a $n \times m$ matrix

Test statistic

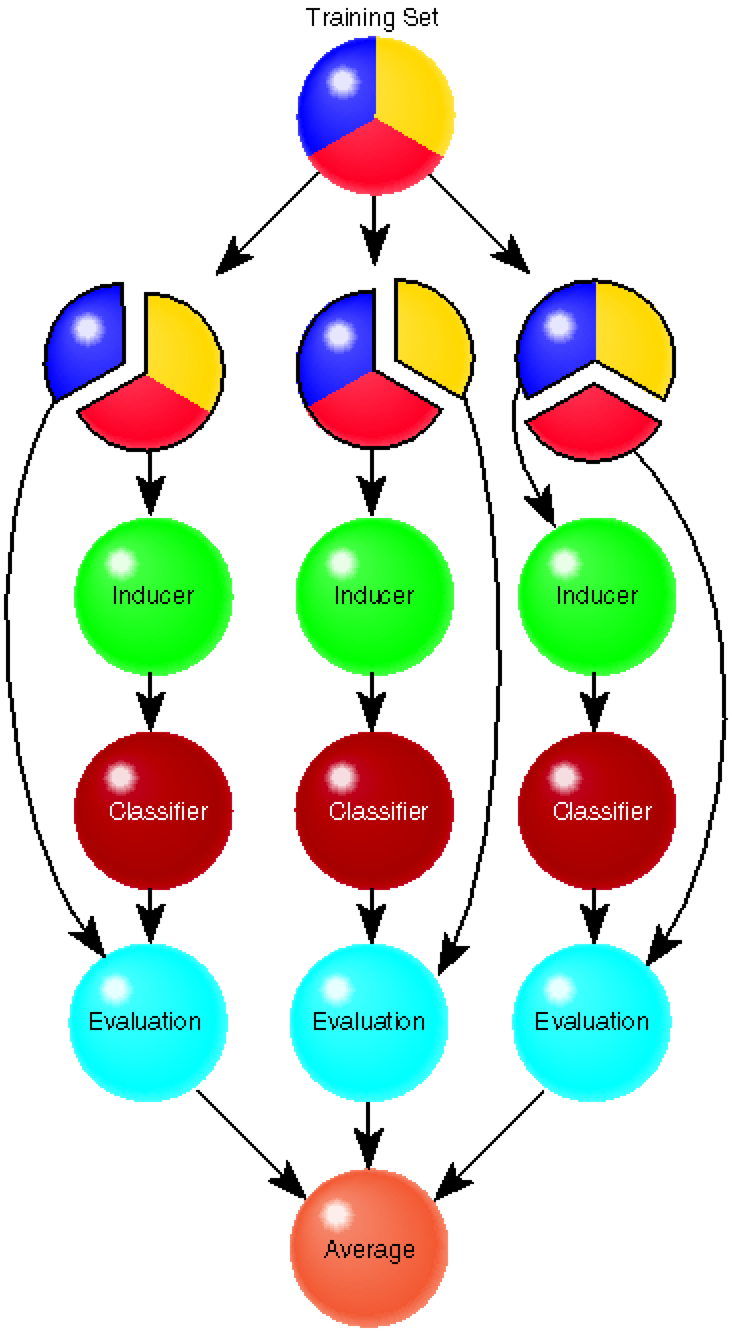
- Suppose we want to check the assumption that no clustering effect within levels of a covariate x^*
- T_1 = the fraction of y's that take on the value of 1
  - Unclear what assumption are we testing.
  - The fraction of 1s is explicitly being modeled in the logit model.
    - The test will never show anything is wrong regardless of how bad our model is.
- T_2 = the variance of the number of 1s in each level of x^*
  - When clustering effect exists, the variance within clusters tends to be small

- Assume modelling data as $\mathbf{y} = (y_1, ..., y_n) \sim N(\mu, \sigma^2)$
- Set priors on μ and σ^2
- Run WinBUGS and obtain samples: $\theta_t = \{\mu_t, \sigma_t^2\}, t = 1, \ldots, M$
- For each sampled data point θ_t, replicate n data points: $y_{rep,i}^t \sim N(\mu_t, \sigma_t^2)$, $t = 1, \ldots, M$ and $i = 1, \ldots, n$.
- For each sampled value, (μ_t, σ_t^2), we obtain M *replicated data set* $\mathbf{y}_{rep}^t = (y_{rep,1}^t, \ldots, y_{rep,n}^t)$.
- Does our model represent our data adequately? Choose a discrepancy measure, say

$$T(\mathbf{y}; \theta) = \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2}$$

Compute $T(\mathbf{y},\ \theta_t)$ and the set of of $T(\mathbf{y}_{rep}^t,\ \theta_t)$ and obtain "Bayesian p-values":

$$P(T(\mathbf{y}_{rep},\ \theta) > T(\mathbf{y},\ \theta) \mid \mathbf{y}) = \frac{1}{M} \sum_{t=1}^{M} 1[T(\mathbf{y}_{rep}^t,\ \theta_t) > T(\mathbf{y},\ \theta_t)].$$



Cross validation

- Leave-one-out cross validation
- g-fold cross validation

Basic tools for model assessment

- Cross-validation residual: $r_i = y_i - E(y_i|\boldsymbol{y}_{(i)})$, where $\boldsymbol{y}_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)^T$

- Outliers are indicated by large standardized residuals $d_i = r_i/\sqrt{Var(y_i|\boldsymbol{y}_{(i)})}$

- Conditional predictive ordinate (CPO):
- $p(y_i|\boldsymbol{y}_{(i)}) = \int p(y_i|\theta, \boldsymbol{y}_{(i)})p(\theta|\boldsymbol{y}_{(i)})d\theta$
  - Height of the conditional density at the observed value of y_i
  - Large values indicates good prediction of y_i

Approximate method

- Given Monte Carlo (MC) samples $\theta^{(g)} \sim p(\theta|y)$,

$$
\begin{aligned}
E(y_i|\mathbf{y}_{(i)}) &= \int\int y_i f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) dy_i d\boldsymbol{\theta} \\
&= \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \\
&\approx \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\
&\approx \frac{1}{G}\sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \; .
\end{aligned}
$$

  - Usually, the approximation works well unless the data set is small or y_i is an extreme outlier
  - In practice, one can use the same $\{\theta^1, \ldots, \theta^G\}$ for calculating all $E\left(y_i|\boldsymbol{y}_{(\boldsymbol{i})}\right), i = 1, \ldots, n$

Approximate method

- To obtain the standardized residual $d_i = r_i / \sqrt{Var(y_i | \boldsymbol{y}_{(i)})}$, one can use a further approximation.
  - Compute $d_i^* = \frac{y_i - E(y_i|\theta)}{\sqrt{Var(y_i|\theta)}}$
  - Then find $E(d_i^*|y)$, the posterior average of the ratio

Exact method

- Evaluate $E(y_i|\boldsymbol{y}_{(i)})$ and $Var(y_i|\boldsymbol{y}_{(i)})$ separately.
- For $Var(y_i|\boldsymbol{y}_{(i)})$, use the fact that

$$Var(y_i|\mathbf{y}_{(i)}) = E(y_i^2|\mathbf{y}_{(i)}) - [E(y_i|\mathbf{y}_{(i)})]^2,$$

$$\begin{aligned} E(y_i^2|\mathbf{y}_{(i)}) &= \int E(y_i^2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\ &= \int \{Var(y_i|\boldsymbol{\theta}) + [E(y_i|\boldsymbol{\theta})]^2\}p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta}\,. \end{aligned}$$

- For example, call the WinBUGS program n times, each time leaving one observation out
  - Use the R package BRugs

Example: stack loss data

- An oft-analyzed dataset, featuring the stack loss Y (ammonia escaping), and three covariates X_1 (air flow), X_2 (temperature), and X_3 (acid concentration).
- Linear regression with noninformative priors

$$Y_i \sim N(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} \, , \, \tau) \, ,$$

- WinBUGS code and data for approximate method: www.biostat.umn.edu/~brad/data/stacks_BUGS.txt
- BRugs code and data for exact method: www.biostat.umn.edu/~brad/software/BRugs
  - an R program that organizes the dataset, contains all the BRugs commands, and summarizes the output
  - a piece of BUGS code that is sent by R to OpenBUGS

Approximate vs Exact results

obs	sresid approx	sresid exact	CPO approx	CPO exact
1	0.948	1.098	0.178	0.124
2	–0.566	–0.628	0.224	0.188
3	1.337	1.461	0.122	0.084
4	1.672	1.851	0.078	0.047
5	–0.504	–0.477	0.251	0.244
⋮	⋮	⋮	⋮	⋮
21	–2.126	–3.012	0.046	0.005

- Approximate residuals are too small, especially for the most outlying observations
- Approximate CPOs also tend to understate lack of fit