STAT 576 Bayesian Analysis

Lecture 6: Model Checking

Chencheng Cai

Washington State University

Model Checking Methods

Goal:

- Assess the fit of the model to the data.
- Assess the fit of the model to our substantive knowledge.
- Assess the adequacy/robusteness of the model.

Model Checking Methods

Goal:

- Assess the fit of the model to the data.
- Assess the fit of the model to our substantive knowledge.
- Assess the adequacy/robusteness of the model.

Methods:

- Sensitivity Analysis.
  - Check whether other models generate a similar posterior.
- External Validation.
  - Posterior predictive checking.
- Internal Validation.
  - Cross-validation predictive checking.

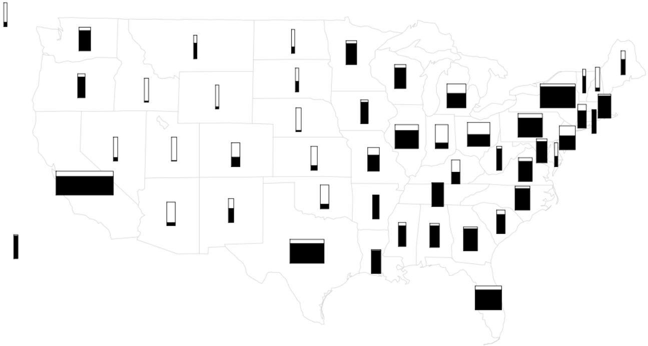
Sensitivity Analysis

- How the results are affected by different choices of the model structure?
  - different models (binomial v.s. Poisson, normal v.s. t)
  - different priors
  - different structures (hierarchical v.s. separate)
  - different distribution families (Gaussian v.s. mixed Gaussian)

Sensitivity Analysis

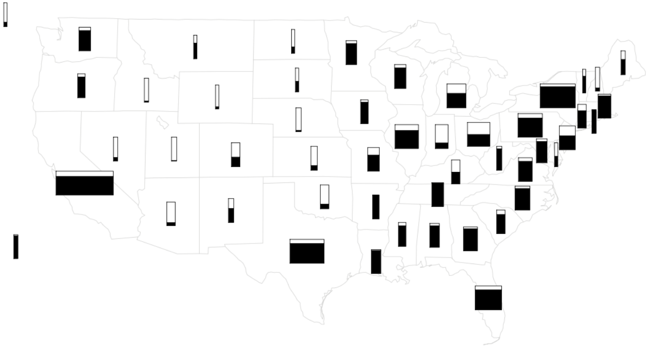
- How the results are affected by different choices of the model structure?
  - different models (binomial v.s. Poisson, normal v.s. t)
  - different priors
  - different structures (hierarchical v.s. separate)
  - different distribution families (Gaussian v.s. mixed Gaussian)
- Compare the sensitivity of essential inference quantities.
  - extreme quantities v.s. mean/median.
  - extrapolation v.s. interpolation.

Example: Election Prediction



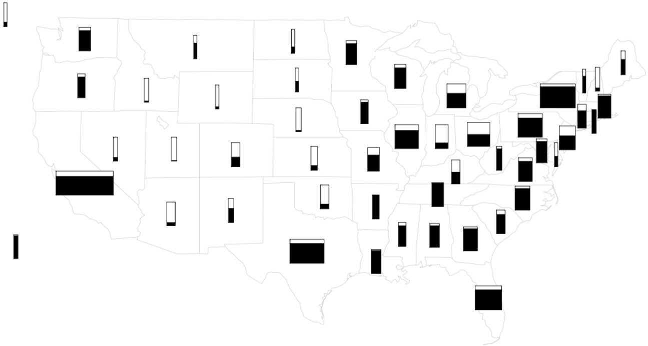
- Posterior winning probability of Bill Clinton at each state in Oct. 1992.
- Hierarchical linear regression model.

Example: Election Prediction



- Posterior winning probability of Bill Clinton at each state in Oct. 1992.
- Hierarchical linear regression model.
- The model seems wrong at Texas and Florida.

Example: Election Prediction



- Posterior winning probability of Bill Clinton at each state in Oct. 1992.
- Hierarchical linear regression model.
- The model seems wrong at Texas and Florida.
- It is much easier to evaluate the performance afterwards.

Posterior Predictive Checking

- Idea: check the discrepancy between the predicted values and the observed values.

Posterior Predictive Checking

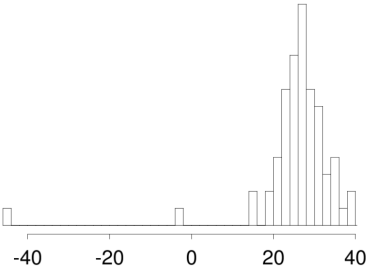
- Idea: check the discrepancy between the predicted values and the observed values.
- Procedure:
  - Generate simulated samples from the **joint posterior predictive distribution**
  - Compare the samples with the observed data.
  - Systematic differences imply the failings of the model.

Example: Light Speed Experiment

- Simon Newcomb set up an experiment in 1882 to measure the light speed.
- The travel time of light was recorded for the round-trip between
  - his lab on the Potomac river
  - a mirror at the base of the Washington Monument
- The total travel distance is 7422 meters.

Example: Light Speed Experiment

- Simon Newcomb set up an experiment in 1882 to measure the light speed.
- The travel time of light was recorded for the round-trip between
  - his lab on the Potomac river
  - a mirror at the base of the Washington Monument
- The total travel distance is 7422 meters.
- The measurement was repeated $n = 66$ times.



Histogram for deviations from 24800 ns

Example: Light Speed Experiment

- We model the travel time by a normal distribution:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

Example: Light Speed Experiment

- We model the travel time by a normal distribution:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- We can choose a noninformative prior for μ and σ^2:

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

Example: Light Speed Experiment

- We model the travel time by a normal distribution:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- We can choose a noninformative prior for μ and σ^2:

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

- Recall our previous results for multiparameter Bayesian inference. The marginal posterior for μ is

$$\mu \mid y \sim t_{66}\left(\bar{y}, \frac{65}{66^2} s^2\right)$$

Example: Light Speed Experiment

- We model the travel time by a normal distribution:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- We can choose a noninformative prior for μ and σ^2:

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

- Recall our previous results for multiparameter Bayesian inference. The marginal posterior for μ is

$$\mu \mid y \sim t_{66}\left(\bar{y}, \frac{65}{66^2} s^2\right)$$

- A 95% credible interval is $[23.6, 28.8]$.
- We know the true value should be around 33.0.

Example: Light Speed Experiment

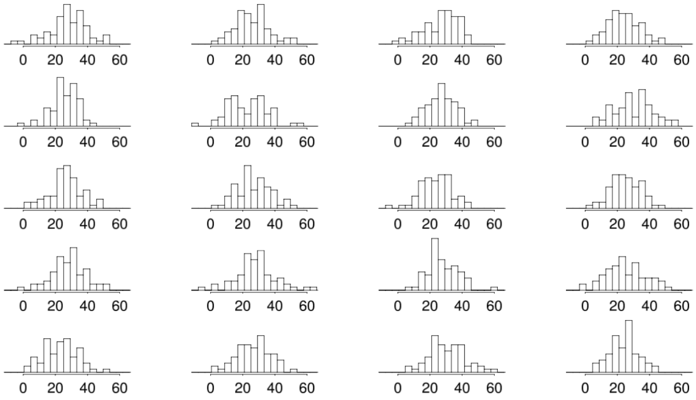
Generate posterior predictive replicates y^{rep}

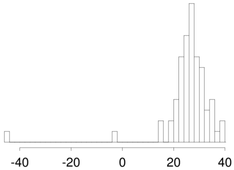
- Draw $\mu^{(s)}, \sigma^{2(s)}$ from the joint posterior distribution $p(\mu, \sigma^2 \mid y)$.
- Draw $y^{rep(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{2(s)})$.
- Repeat the drawing to get n replicates of y^{rep}.

Example: Light Speed Experiment

Generate posterior predictive replicates y^{rep}

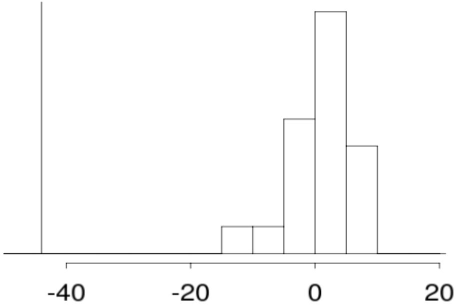
- Draw $\mu^{(s)}, \sigma^{2(s)}$ from the joint posterior distribution $p(\mu, \sigma^2 \mid y)$.
- Draw $y^{rep(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{2(s)})$.
- Repeat the drawing to get n replicates of y^{rep}.





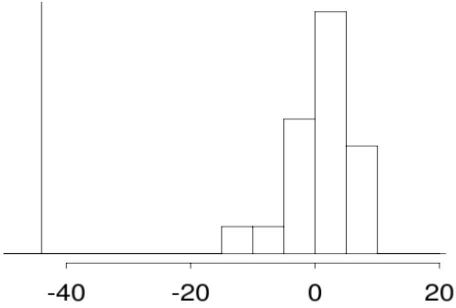
Example: Light Speed Experiment

We get the histogram of the **smallest** travel time for all replicates.



Example: Light Speed Experiment

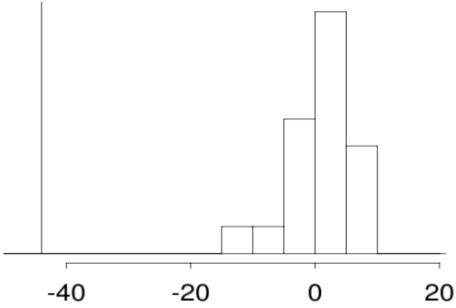
We get the histogram of the **smallest** travel time for all replicates.



- Can hardly observe an occurrence that is less than -20.

Example: Light Speed Experiment

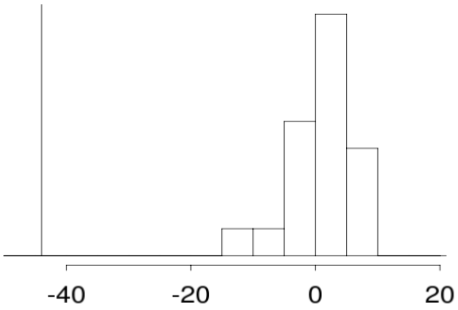
We get the histogram of the **smallest** travel time for all replicates.



- Can hardly observe an occurrence that is less than -20.
- Decide: whether the **data** was wrong or the **model** was wrong?

Example: Light Speed Experiment

We get the histogram of the **smallest** travel time for all replicates.



- Can hardly observe an occurrence that is less than -20.
- Decide: whether the **data** was wrong or the **model** was wrong?
- The model was wrong: should use heavy-tailed distribution or contaminated normal (mixed Gaussian).

Posterior Predictive Checking

- Replicated datasets:

$$p(y^{rep} \mid y) = \int \underbrace{p(y^{rep} \mid \theta)}_{obs.\ model} \underbrace{p(\theta \mid y)}_{posterior} d\mu(\theta)$$

Posterior Predictive Checking

- Replicated datasets:

$$p(y^{rep} \mid y) = \int \underbrace{p(y^{rep} \mid \theta)}_{obs.\ model} \underbrace{p(\theta \mid y)}_{posterior} d\mu(\theta)$$

- **Test quantity** (or discrepancy measure) $T(y, \theta)$
  - Summary quantity for the observed data $T(y, \theta)$
  - Summary quantity for a replicated data $T(y^{rep}, \theta)$.

Posterior Predictive Checking

- Replicated datasets:

$$p(y^{rep} \mid y) = \int \underbrace{p(y^{rep} \mid \theta)}_{obs.\ model} \underbrace{p(\theta \mid y)}_{posterior} d\mu(\theta)$$

- **Test quantity** (or discrepancy measure) $T(y, \theta)$
  - Summary quantity for the observed data $T(y, \theta)$
  - Summary quantity for a replicated data $T(y^{rep}, \theta)$.
- The frequentist counter-part is known as **test statistics** $T(y)$, which only depends on the data.

Posterior Predictive Checking

- Replicated datasets:

$$p(y^{rep} \mid y) = \int \underbrace{p(y^{rep} \mid \theta)}_{obs.\ model} \underbrace{p(\theta \mid y)}_{posterior} d\mu(\theta)$$

- **Test quantity** (or discrepancy measure) $T(y, \theta)$
  - Summary quantity for the observed data $T(y, \theta)$
  - Summary quantity for a replicated data $T(y^{rep}, \theta)$.
- The frequentist counter-part is known as **test statistics** $T(y)$, which only depends on the data.
- In the light speed example, we choose $T(y, \theta) = \min(y)$ (also a test statistic).

Posterior Predictive Checking

Classical p-values:

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

Posterior Predictive Checking

Classical p-values:

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

Posterior predictive p-values:

$$p_B = \mathbb{P}[T(y^{rep}, \theta) \geq T(y, \theta) \mid y]$$

Posterior Predictive Checking

Classical p-values:

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

Posterior predictive p-values:

$$p_B = \mathbb{P}[T(y^{rep}, \theta) \geq T(y, \theta) \mid y]$$

- The classical p-values measure how likely the data is coming from the null model.
- The posterior predictive p-values measure how likely the data is similar to the postetior predictive replicates.

Posterior Predictive Checking

Classical p-values:

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

Posterior predictive p-values:

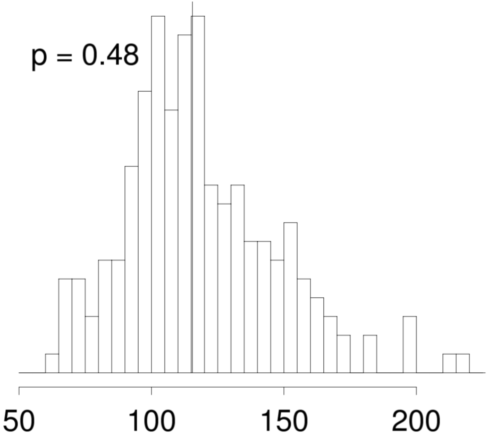
$$p_B = \mathbb{P}[T(y^{rep}, \theta) \geq T(y, \theta) \mid y]$$

- The classical p-values measure how likely the data is coming from the null model.
- The posterior predictive p-values measure how likely the data is similar to the postetior predictive replicates.
- In Bayesian, θ is also random. p_B can be estimated by joint samples of (y^{rep}, θ).

$$\begin{aligned} p_B &= \iint \mathbb{I}\{T(y^{rep}, \theta) \geq T(y, \theta)\} p(y^{rep} \mid \theta) p(\theta \mid y) d\mu(\theta) d\mu(y^{rep}) \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\{T(y^{rep(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)})\} \end{aligned}$$

Example: Light Speed Experiment

If we use the sample variance as the test quantity:



Cannot tell the discrepancy — because the sample variance is a sufficient statistics.

Posterior Predictive Checking

- A **good** test statistic is ancilliary
  - ancilliary: depends on the observed data but independent of the parameters.
- A **bad** test statsistic is highly dependent of the parameters.
  - i.e. sufficient statistics.

Posterior Predictive Checking

- A **good** test statistic is ancilliary
  - ancilliary: depends on the observed data but independent of the parameters.
- A **bad** test statsistic is highly dependent of the parameters.
  - i.e. sufficient statistics.
- If we have multiple test statistics, we do not conduct p-value justification.
  - See the smoking example in the textbook.

Posterior Predictive Checking

- A **good** test statistic is ancilliary
  - ancilliary: depends on the observed data but independent of the parameters.
- A **bad** test statsistic is highly dependent of the parameters.
  - i.e. sufficient statistics.
- If we have multiple test statistics, we do not conduct p-value justification.
  - See the smoking example in the textbook.
- An extreme p-value often suggests the weakness of the current model. The next step is to revise the model.

Example: Educational Testing

Data: the effects of coaching programs for the SAT-V scores for students in 8 schools.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

Example: Educational Testing

Separate estimation:

- Some schools have moderate effects (18-28).
- Most schools have small effects (0-12).
- Two have negative effects.
- Difficult to distinguish because of large variance.

Example: Educational Testing

Separate estimation:

- Some schools have moderate effects (18-28).
- Most schools have small effects (0-12).
- Two have negative effects.
- Difficult to distinguish because of large variance.

Pooled estimation:

- All schools have identical effect θ.
- Use noninformative prior.
- Posterior mean: 7.7 with s.e. 4.1

Example: Educational Testing

Separate estimation:

- Some schools have moderate effects (18-28).
- Most schools have small effects (0-12).
- Two have negative effects.
- Difficult to distinguish because of large variance.

Pooled estimation:

- All schools have identical effect θ.
- Use noninformative prior.
- Posterior mean: 7.7 with s.e. 4.1

Hierarchical model:

- $\theta_1, \ldots, \theta_8 \sim \mathcal{N}(\mu, \tau^2)$ i.i.d.
- $y_j \mid \theta_j \sim (\theta_j, \sigma_j^2)$ independent.
- choose flat prior $p(\mu, \tau) \propto 1$.

Example: Educational Testing

Hierarchical model:

- By drawing posterior samples:
  - draw $\mu^{(s)}, \tau^{(s)}$ from $p(\mu, \tau \mid y)$
  - draw $\theta_1^{(s)}, \ldots, \theta_8^{(s)}$ from $p(\theta_1, \ldots, \theta_8 \mid \mu^{(s)}, \tau^{(s)}, y)$

Example: Educational Testing

Hierarchical model:

- By drawing posterior samples:
  - draw $\mu^{(s)}, \tau^{(s)}$ from $p(\mu, \tau \mid y)$
  - draw $\theta_1^{(s)}, \ldots, \theta_8^{(s)}$ from $p(\theta_1, \ldots, \theta_8 \mid \mu^{(s)}, \tau^{(s)}, y)$
- we have the posterior quantiles for each school:

School	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
A	-2	7	10	16	31
B	-5	3	8	12	23
C	-11	2	7	11	19
D	-7	4	8	11	21
E	-9	1	5	10	18
F	-7	2	6	10	28
G	-1	7	10	15	26
H	-6	3	8	13	33

Example: Educational Testing — Model Checking

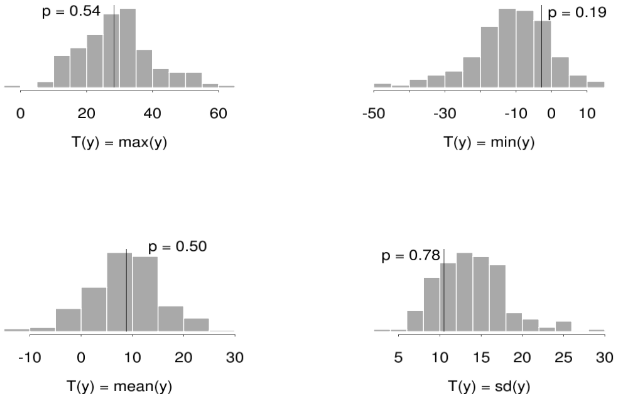
- Assumptions:
  - normality of y_j.
  - exchangeability of the priors for θ_j's.
  - normality of prior of θ_j.
  - flat hyperprior.

Example: Educational Testing — Model Checking

- Assumptions:
  - normality of y_j.
  - exchangeability of the priors for θ_j's.
  - normality of prior of θ_j.
  - flat hyperprior.
- Comparing posterior inferences to substantive knowledge:
  - Individual effects between 5 and 10 seems reasonable.
  - Some lower bounds go to negative.

Example: Educational Testing — Model Checking

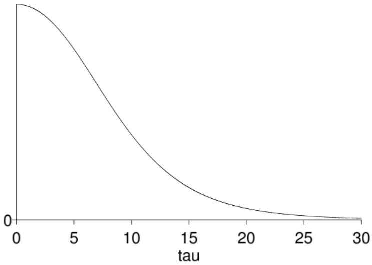
- Posterior predictive checking.
  - $y^{rep} = (y_1^{rep}, \ldots, y_8^{rep})$
  - Test statistics: max, min, mean, s.d.



Example: Educational Testing — Model Checking

Sensitivity Analysis:

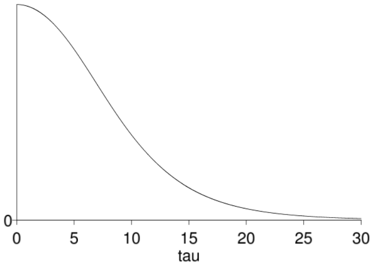
- Uniform prior for τ: the marginal posterior for τ
  — no significant change if we multiply it by another prior



Example: Educational Testing — Model Checking

Sensitivity Analysis:

- Uniform prior for τ: the marginal posterior for τ
  — no significant change if we multiply it by another prior

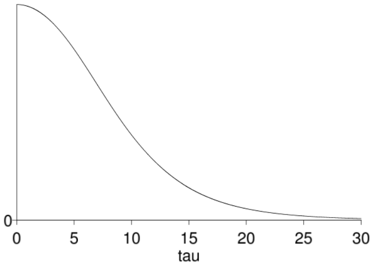


- normality of $y_j \mid \theta_j, \sigma_j$: ensured by experimental designa and CLT.

Example: Educational Testing — Model Checking

Sensitivity Analysis:

- Uniform prior for τ: the marginal posterior for τ
  — no significant change if we multiply it by another prior



- normality of $y_j \mid \theta_j, \sigma_j$: ensured by experimental designa and CLT.
- normality of the prior for θ_j's:
  One may consider other heavy-tailed distributions. But needs advanced sampling techniques.