Chapter 6

- 6.1 The place of model checking in applied Bayesian statistics
- 6.2 Do the inferences from the model make sense?
- 6.3 Posterior predictive checking
- 6.4 Graphical posterior predictive checks (can be skipped)
- 6.5 Model checking for the educational testing example

Model checking

- demo6_1: Posterior predictive checking - light speed
- demo6_2: Posterior predictive checking - sequential dependence
- demo6_3: Posterior predictive checking - poor test statistic
- demo6_4: Posterior predictive checking - marginal predictive p-value

Model checking – overview

- Sensibility with respect to additional information not used in modeling
  - e.g., if posterior would claim that hazardous chemical decreases probability of death

Model checking – overview

- Sensibility with respect to additional information not used in modeling
    - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
    - compare predictions to completely new observations
    - cf. relativity theory predictions

Model checking – overview

- Sensibility with respect to additional information not used in modeling
  - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
  - compare predictions to completely new observations
  - cf. relativity theory predictions
- Internal validation
  - posterior predictive checking
  - cross-validation predictive checking

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$

Posterior predictive checking – example

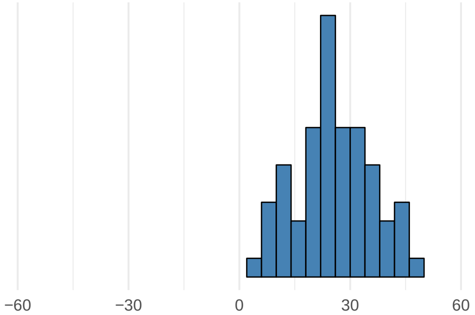
- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathrm{N}(\mu^{(s)}, \sigma^{(s)})$

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathrm{N}(\mu^{(s)}, \sigma^{(s)})$
  - repat n times to get y^{rep} with n replicates

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathrm{N}(\mu^{(s)}, \sigma^{(s)})$
  - repat n times to get y^{rep} with n replicates



Replicates vs. future observation

- Predictive $\tilde{y}$ is the next not yet observed possible observation. y^{rep} refers to replicating the whole experiment (potentially with same values of x) and obtaining as many replicated observations as in the original data.

Posterior predictive checking – example

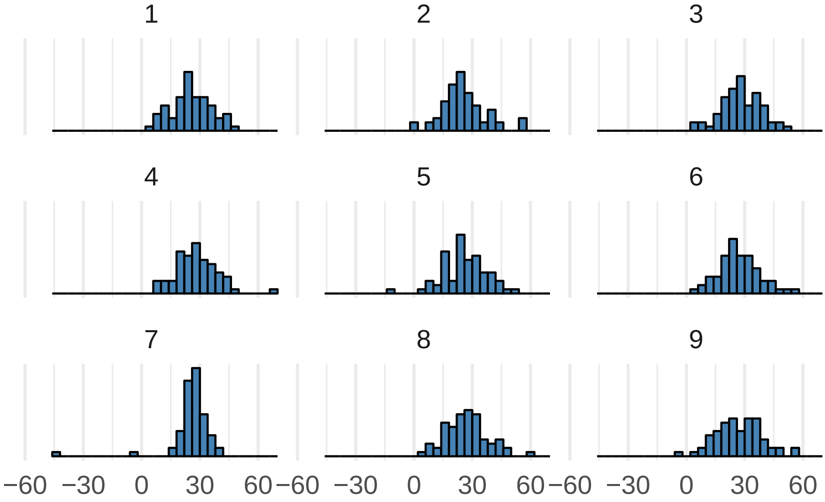
- Generate several replicated datasets y^{rep}

Posterior predictive checking – example

- Generate several replicated datasets y^{rep}
- Compare to the original dataset

Posterior predictive checking – example

- Generate several replicated datasets y^{rep}
- Compare to the original dataset



Posterior predictive checking with test statistic

- Replicated data sets y^{rep}
- Test quantity (or discrepancy measure) $T(y, \theta)$
  - summary quantity for the observed data $T(y, \theta)$
  - summary quantity for a replicated data $T(y^{\mathrm{rep}}, \theta)$
  - can be easier to compare summary quantities than data sets

Posterior predictive checking – example

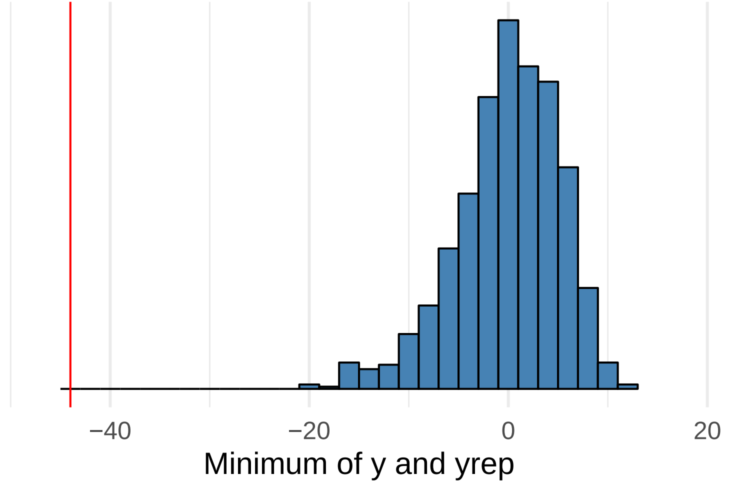
- Compute test statistic for data $T(y, \theta) = \min(y)$

Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets

Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets



Posterior predictive checking – example

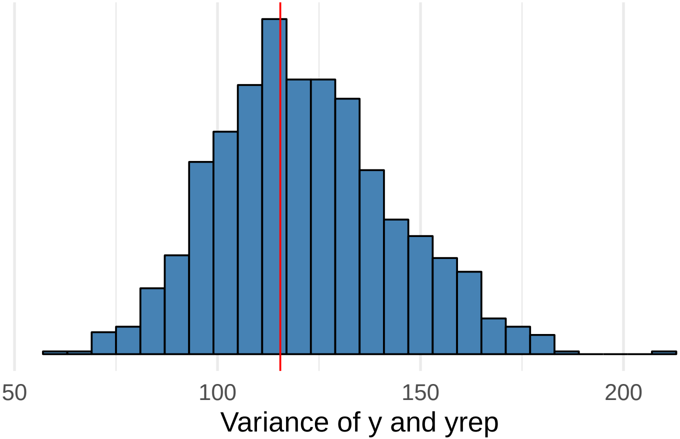
- Good test statistic is ancillary (or almost)
  - ancillary if it depends only on observed data and if its distribution is independent of the parameters of the model

Posterior predictive checking – example

- Good test statistic is ancillary (or almost)
  - ancillary if it depends only on observed data and if its distribution is independent of the parameters of the model
- Bad test statistic is highly dependent of the parameters
  - e.g. variance for normal model

Posterior predictive checking – example

- Good test statistic is ancillary (or almost)
  - ancillary if it depends only on observed data and if its distribution is independent of the parameters of the model
- Bad test statistic is highly dependent of the parameters
  - e.g. variance for normal model



Posterior predictive checking

- Posterior predictive *p*-value

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

where I is an indicator function

Posterior predictive checking

- Posterior predictive p-value

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

where I is an indicator function

  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

Posterior predictive checking

- Posterior predictive *p*-value

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

  where I is an indicator function
  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$
T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S
$$

- Posterior predictive *p*-value (ppp-value) estimated whether difference between the model and data could arise by chance

Posterior predictive checking

- Posterior predictive p-value

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

  where I is an indicator function
  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance
- Not commonly used as the distribution of test statistic often more information

Marginal and CV predictive checking

- Consider marginal predictive distributions $p(\tilde{y}_i|y)$ and each observation separately
  - marginal posterior p-values

$$p_i = \Pr(T(y_i^{\mathrm{rep}}) \leq T(y_i)|y)$$

if $T(y_i) = y_i$

$$p_i = \Pr(y_i^{\mathrm{rep}} \leq y_i|y)$$

Marginal and CV predictive checking

- Consider marginal predictive distributions $p(\tilde{y}_i|y)$ and each observation separately
  - marginal posterior p-values

$$p_i = \Pr(T(y_i^{\mathrm{rep}}) \leq T(y_i)|y)$$

if $T(y_i) = y_i$

$$p_i = \Pr(y_i^{\mathrm{rep}} \leq y_i|y)$$

- if $Pr(\tilde{y}_i|y)$ well calibrated, distribution of p_i would be uniform between 0 and 1
  - holds better for cross-validation predictive tests (cross-validation Ch 7)

Marginal predictive checking - Example

- Marginal tail area or Probability integral transform (PIT)

$$p_i = p(y_i^{\mathrm{rep}} \leq y_i | y)$$

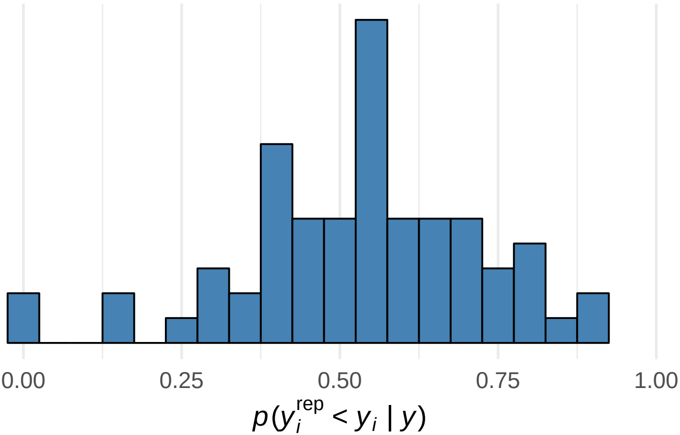
- if $p(\tilde{y}_i | y)$ is well calibrated, distribution of p_i's would be uniform between 0 and 1

Marginal predictive checking - Example

- Marginal tail area or Probability integral transform (PIT)

$$p_i = p(y_i^{\text{rep}} \leq y_i | y)$$

- if $p(\tilde{y}_i | y)$ is well calibrated, distribution of p_i's would be uniform between 0 and 1



Sensitivity analysis

- How much different choices in model structure and priors affect the results

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. t distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to “outliers”
    - e.g. t instead of Gaussian

Sensitivity analysis

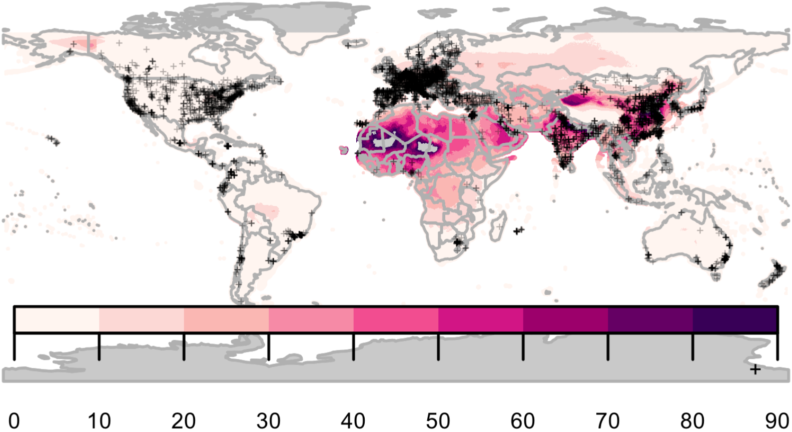
- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. *t* distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to "outliers"
    - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

Example: Exposure to air pollution

- Example from Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2017). Visualization in Bayesian workflow. https://arxiv.org/abs/1709.01449
- Estimation of human exposure to air pollution from particulate matter measuring less than 2.5 microns in diameter ($\mathrm{PM}_{2.5}$)
    - Exposure to $\mathrm{PM}_{2.5}$ is linked to a number of poor health outcomes and a recent report estimated that $\mathrm{PM}_{2.5}$ is responsible for three million deaths worldwide each year (Shaddick et al., 2017)
    - In order to estimate the public health effect of ambient $\mathrm{PM}_{2.5}$, we need a good estimate of the $\mathrm{PM}_{2.5}$ concentration at the same spatial resolution as our population estimates.

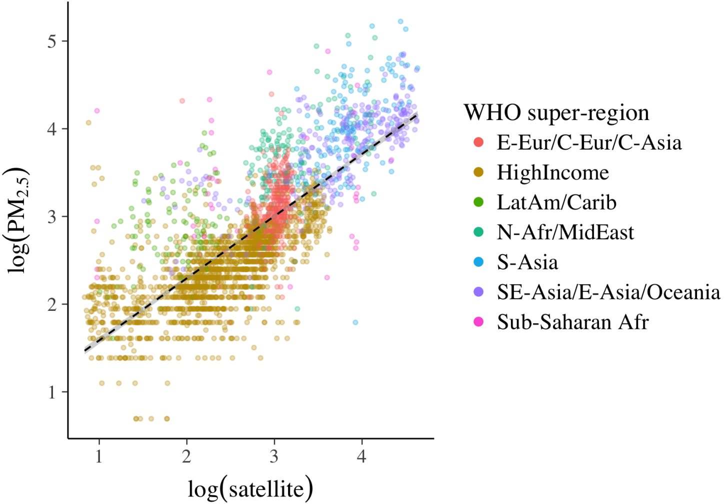
Example: Exposure to air pollution

- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



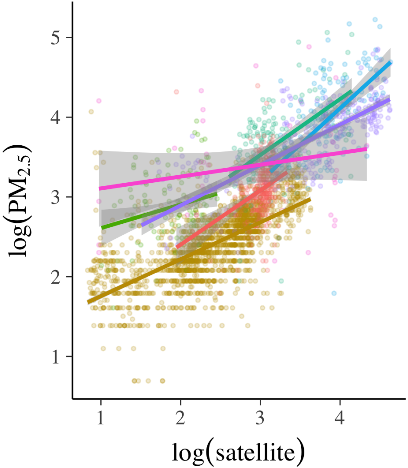
Example: Exposure to air pollution

- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



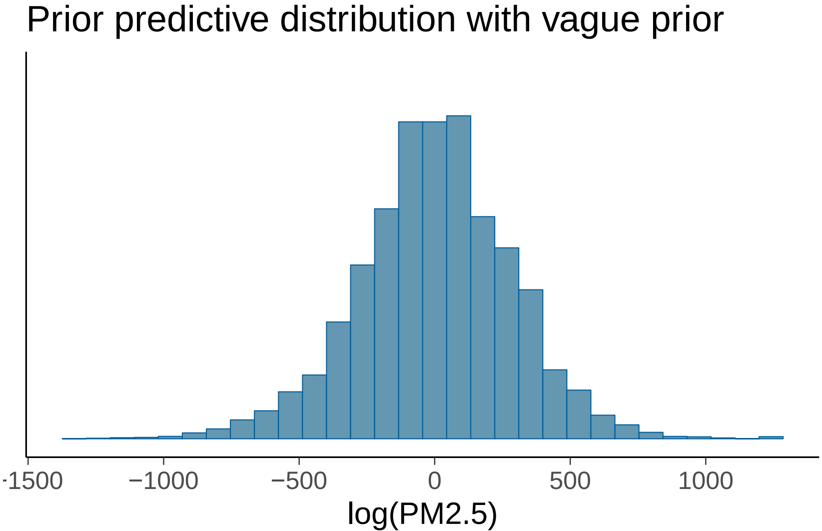
Example: Exposure to air pollution

- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



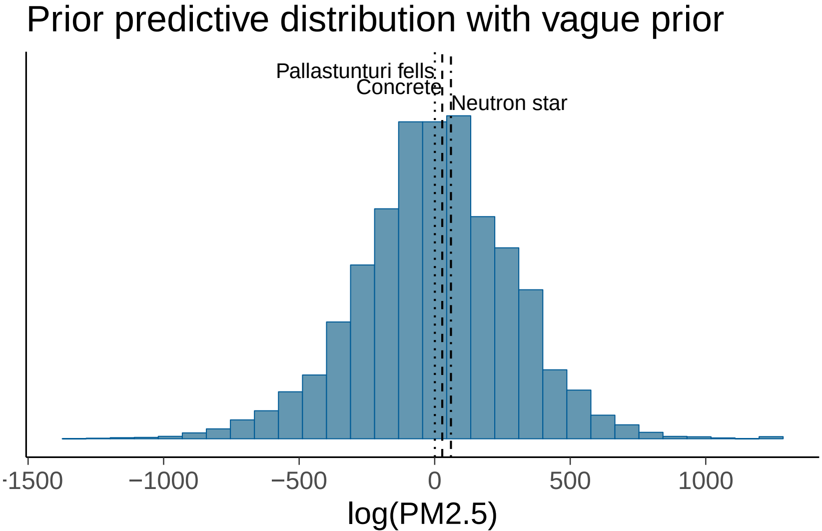
Example: Exposure to air pollution

Prior predictive checking



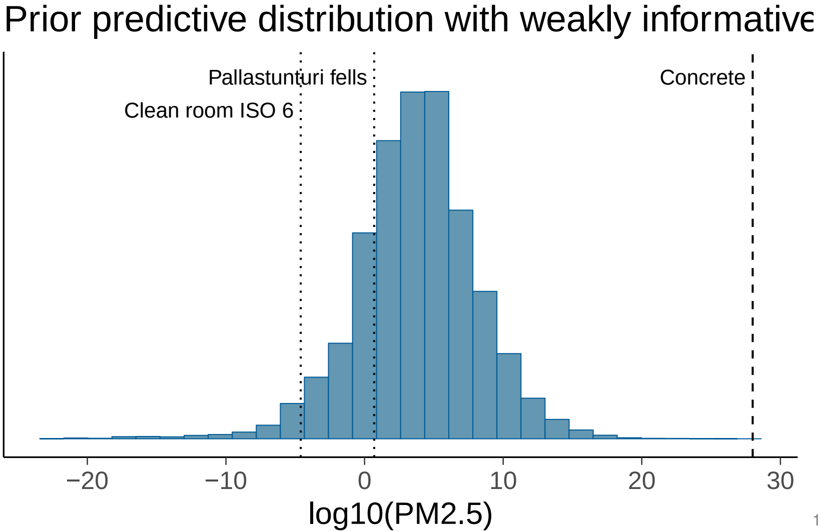
Example: Exposure to air pollution

Prior predictive checking



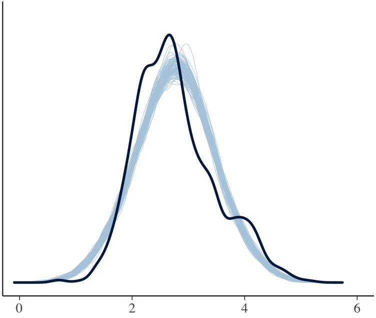
Example: Exposure to air pollution

Prior predictive checking

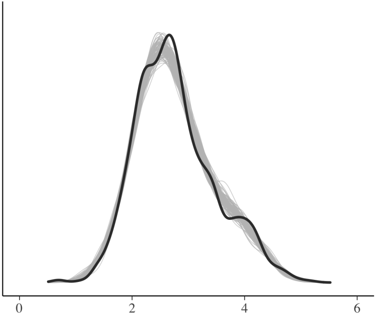


Example: Exposure to air pollution

Posterior predictive checking – marginal predictive distributions



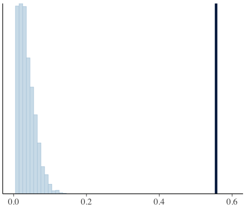
(a) Model 1



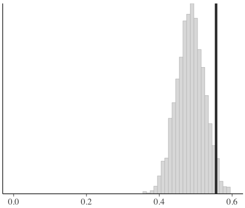
(b) Model 2

Example: Exposure to air pollution

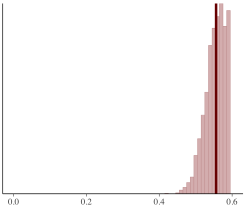
Posterior predictive checking – test statistic (skewness)



(a) Model 1



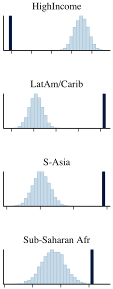
(b) Model 2



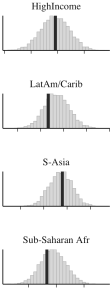
(c) Model 3

Example: Exposure to air pollution

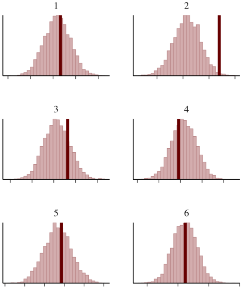
Posterior predictive checking – test statistic (median for groups)



(a) Model 1



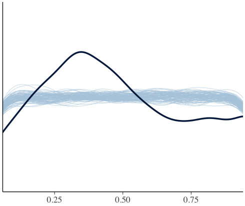
(b) Model 2



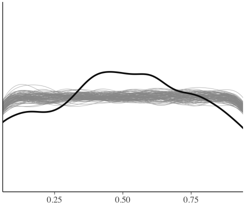
(c) Model 3

Example: Exposure to air pollution

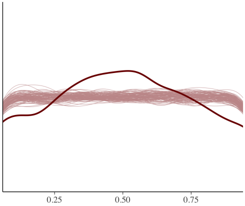
LOO predictive checking – LOO-PIT



(a) Model 1

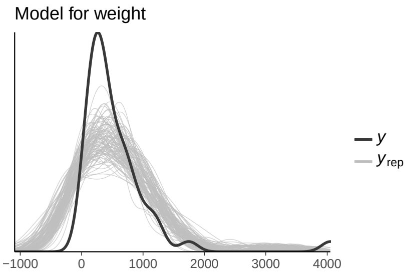


(b) Model 2



(c) Model 3

Example of posterior predictive checking

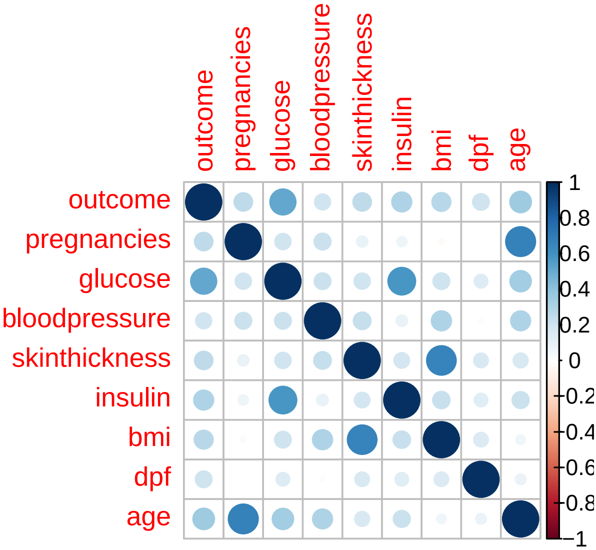


Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2019): Regression and Other Stories, Chapter 11.

Example of posterior predictive checking

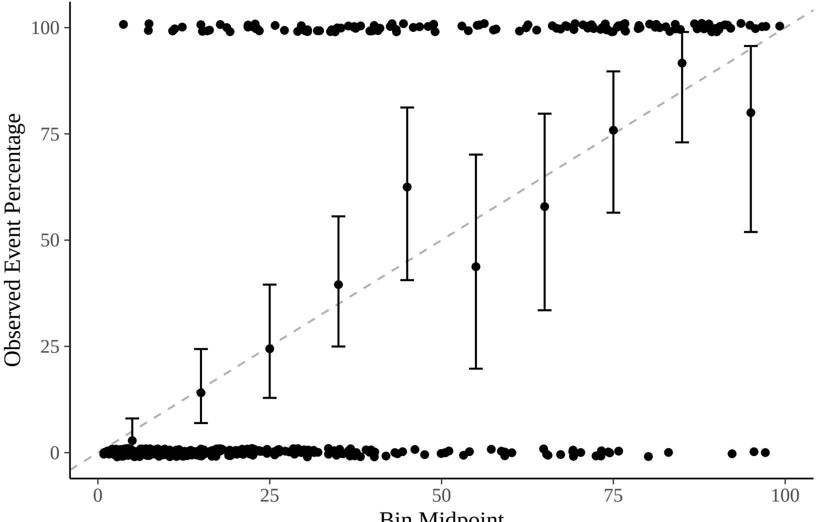
Diabetes prediction with logistic regression - diabetes demo



Example of posterior predictive checking

Diabetes prediction with logistic regression - diabetes demo

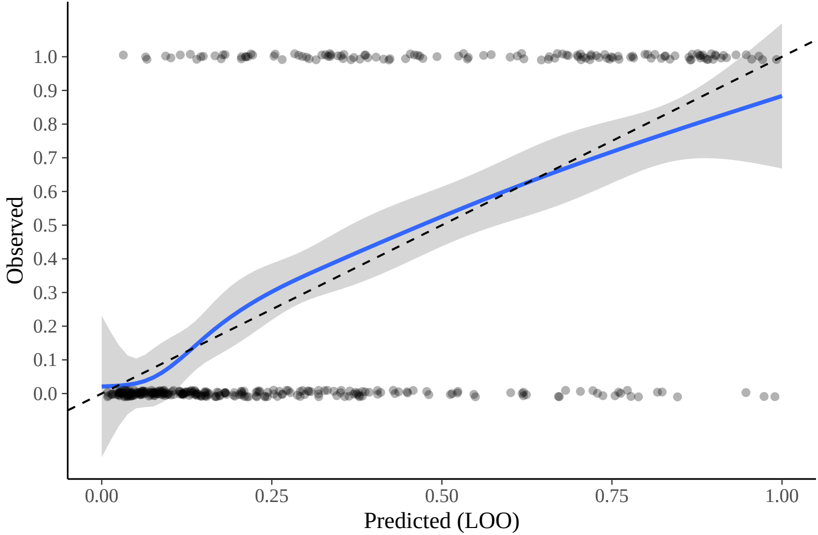
PPC with binning for binary data



Example of posterior predictive checking

Diabetes prediction with logistic regression - diabetes demo

PPC with non-linear regression for binary data



Posterior predictive checking

- demo demos_rstan/ppc/poisson-ppc.Rmd

```
data {
  int<lower=1> N;
  int<lower=0> y[N];
}
parameters {
  real<lower=0> lambda;
}
model {
  lambda ~ exponential(0.2);
  y ~ poisson(lambda);
}
generated quantities {
  real log_lik[N];
  int y_rep[N];
  for (n in 1:N) {
    y_rep[n] = poisson_rng(lambda);
    log_lik[n] = poisson_lpmf(y[n] | lambda);
    }
}
```

Further reading and examples

- Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2018). Visualization in Bayesian workflow. Journal of the Royal Statistical Society Series A, accepted for publication as discussion paper. arXiv preprint arXiv:1709.01449.
- Graphical posterior predictive checks using the bayesplot package http://mc-stan.org/bayesplot/articles/graphical-ppcs.html
- demo demos_rstan/ppc/poisson-ppc.Rmd
- Michael Betancourt's workflow case study with prior and posterior predictive checking https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html