ECON5120 Bayesian Data Analysis

Week 5: Model Evaluation and Empirical Example

Santiago Montoya-Blandón

University of Glasgow
Adam Smith Business School

February 18, 2026

Outline

Model evaluation

- Sensibility with respect to additional information not used in modeling
- External (out-of-sample) validation: compare predictions to completely new observations
- Internal (in-sample) validation
  - Posterior predictive checking
  - Cross-validation predictive checking

Posterior predictive checking

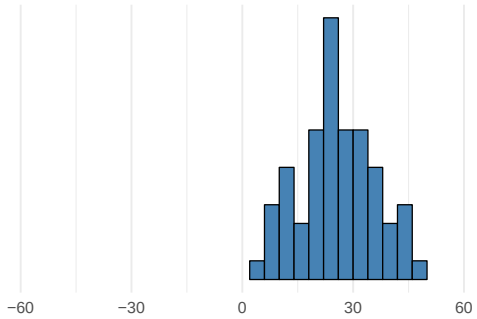
- Idea: if the model is a good fit for the data, then *simulated data* from the posterior should be close to original data
- *Statistics* of the original and simulated data should also be close
- Requires sampling from the data generating process (DGP) using posterior simulations
- Use standard sampling or MCMC methods for simulation
- Simulated (replicated) data is different to predicted data:
  - predictive $\tilde{y}$ are new and not yet observed data points
  - replicates y_{rep} follow from full model likelihood and have same size as original data

Posterior predictive checking: example

- Model $y_i \sim \mathcal{N}(\mu, \sigma^2)$ with prior $(\mu, \sigma^2) \propto 1/\sigma^2$
- Posterior $p(\mu, \sigma^2 \mid y_1, \ldots, y_n)$ is of known form: Normal-Inverse Gamma
- Obtain posterior predictive replicates y_{rep} as follows
- Draw $\mu^{(s)}, \sigma^{2(s)}$ from the posterior
- Draw $y_i^{(s)}$ from the likelihood $\mathcal{N}(\mu^{(s)}, \sigma^{2(s)})$ at the posterior draws
- Repeat over $i = 1, \ldots, n$ times to get a simulated dataset $y_{\text{rep}}^{(s)}$ of size n
- Repeat over $s = 1, \ldots, S$ to obtain S simulated datasets

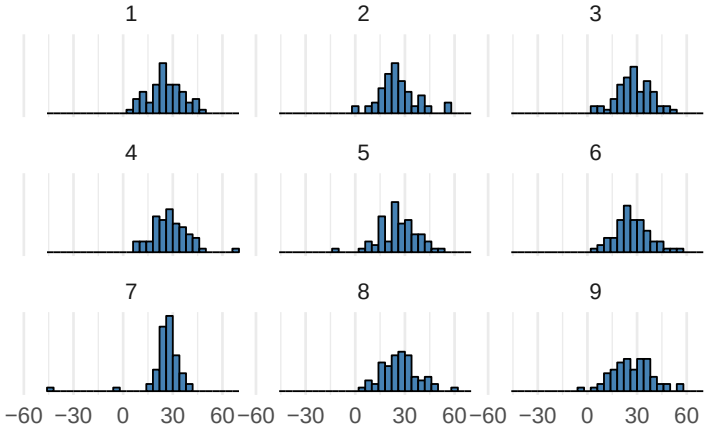
Posterior predictive checking: example

- Model $y_i \sim \mathcal{N}(\mu, \sigma^2)$ with prior $(\mu, \sigma^2) \propto 1/\sigma^2$
- Posterior $p(\mu, \sigma^2 \mid y_1, \ldots, y_n)$ is of known form: Normal-Inverse Gamma
- Obtain posterior predictive replicates y_{rep} as follows
- Draw $\mu^{(s)}, \sigma^{2(s)}$ from the posterior
- Draw $y_i^{(s)}$ from the likelihood $\mathcal{N}(\mu^{(s)}, \sigma^{2(s)})$ at the posterior draws
- Repeat over $i = 1, \ldots, n$ times to get a simulated dataset $y_{\text{rep}}^{(s)}$ of size n
- Repeat over $s = 1, \ldots, S$ to obtain S simulated datasets



Posterior predictive checking: example

Posterior predictive checking: compare y_{rep} to actual y

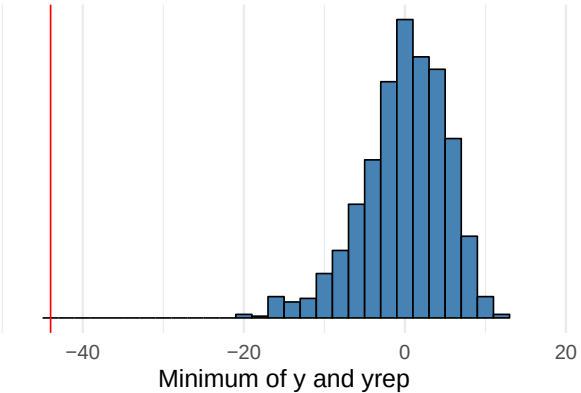


Posterior predictive checking: Test statistics

- Replicated data sets: $y_{\text{rep}}^{(1)}, \ldots, y_{\text{rep}}^{(S)}$
- It is usually easier to compare summary quantities than data sets
- Specially true for high-dimensional data
- Test quantity (or discrepancy measure): $T(y, \theta)$
- Summary quantity for the observed data: $T(y, \theta)$
- Summary quantity for a replicated data: $T(y_{\text{rep}}, \theta)$

Posterior predictive checking: Test statistics example

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

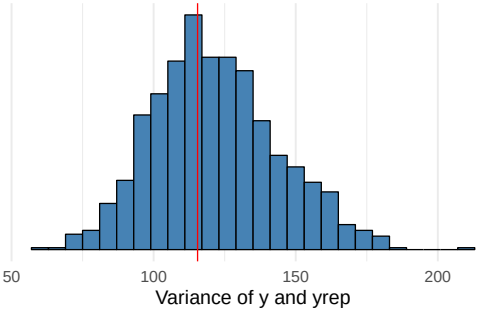


Posterior predictive checking: Test statistics

- Good test statistic is *ancillary* (or almost): $T(y, \theta) \approx T(y)$
  - Ancillary test statistics only depend on observed data
  - Their distributions are independent of model parameters

Posterior predictive checking: Test statistics

- Good test statistic is *ancillary* (or almost): $T(y, \theta) \approx T(y)$
  - Ancillary test statistics only depend on observed data
  - Their distributions are independent of model parameters
- Bad test statistics are highly dependent of the parameters
  - e.g. variance for normal model



Sensitivity analysis

- Sensitivity analyses (or robustness checks) aim to understand how different choices in model structure and priors affect the results
  - Test different models and priors
  - Alternatively, use nested models that generalize particular features
    - 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 (out-of-sample prediction) is more sensitive than interpolation (in-sampe prediction)

UofG

Example: Exposure to air pollution

- Example from Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2019). Visualization in Bayesian Workflow. https://doi.org/10.1111/rssa.12378
- 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 worldwide

Example: Exposure to air pollution

Two sources of information:

1. Direct measurements of PM 2.5 from ground monitors at 2980 locations
   - These are high-quality measurements that can be taken as population quantities
   - They are only available at select locations
   - Locations themselves are subject to sample bias: measurement stations are more likely to be present in high-income countries
2. High-resolution satellite data of aerosol optical depth
   - Available worldwide
   - Our objective will be to extrapolate the high-quality ground measurements to all areas
   - We will then be able to calculate human exposure to the particles

Example: Exposure to air pollution

- Goal: predict ambient concentration of $\mathrm{PM}_{2.5}$ using satellite measures
- Simple method for prediction: linear regression

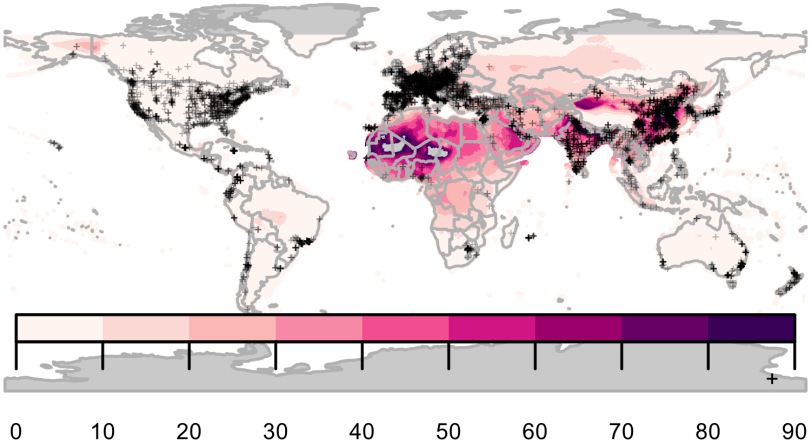
$$\log(\mathrm{PM}_{2.5}) = \beta_0 + \beta_1 \log(\text{Satellite}) + u_i$$

- Take the measured satellite concentration at a new location: $\widetilde{\text{Satellite}}$
- Model informs of ambient concentration at the new location: $\widetilde{\mathrm{PM}}_{2.5}$
- Example: frequentists use MLE (OLS) estimates $(\widehat{\beta}_0, \widehat{\beta}_1)$ and predict

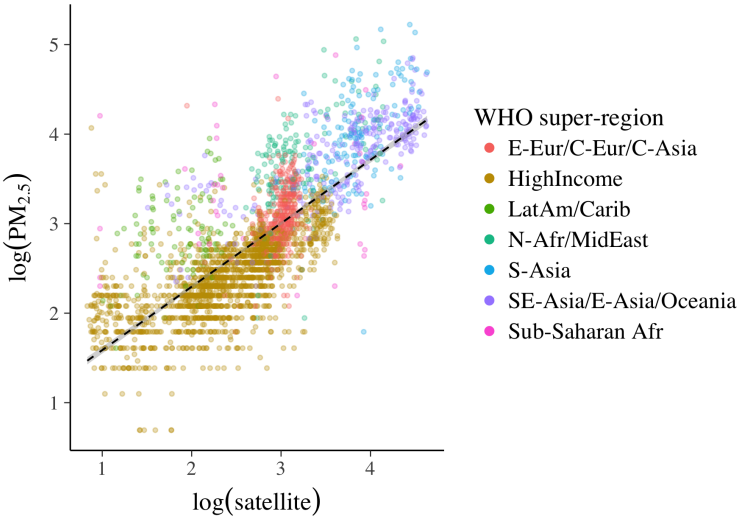
$$\log(\widetilde{\mathrm{PM}}_{2.5}) = \widehat{\beta}_0 + \widehat{\beta}_1 \log(\widetilde{\text{Satellite}})$$

- Bayesians can use posterior predictive distributions
- Use posterior checking to evaluate quality of predicted concentrations

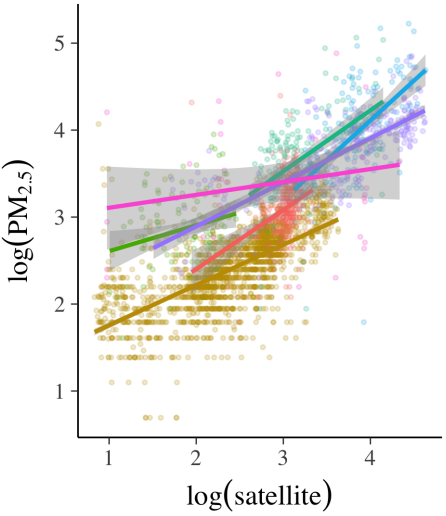
Example: Exposure to air pollution



Example: Exposure to air pollution



Example: Exposure to air pollution

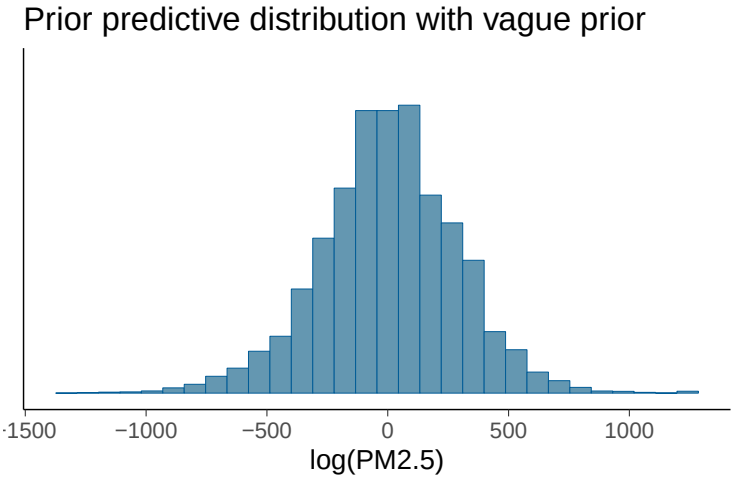


Example: Exposure to air pollution

- Remember, the objective is to build a model for worldwide concentrations of PM 2.5
- Let y_{ij} be the satellite-measured concentration for PM 2.5 in area i that belongs to region j
- Our covariate is the ground-measured concentrations at the available stations averaged over areas
- Three modeling approaches:
  1. Pooled model using all data: $y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$
  2. Hierarchical model based on true regions: $y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ with μ_j and σ_j^2 assigned a prior as before. Regions j are defined using the true regions
  3. Hierarchical model based on clustered regions: same as before, except now regions j were themselves found using a clustering algorithm

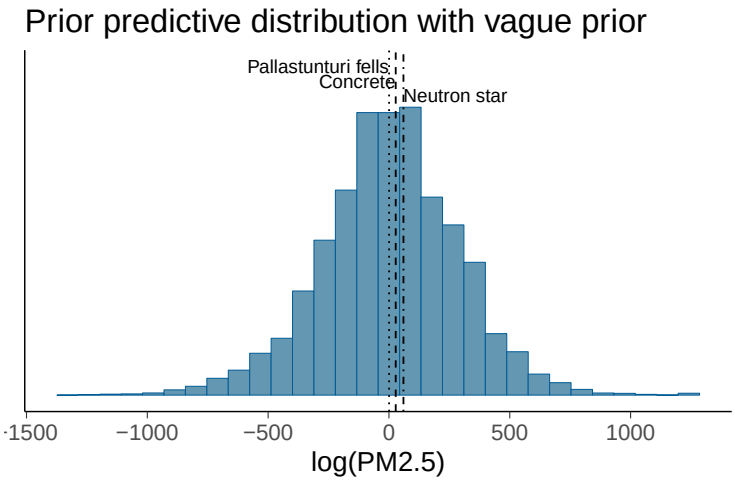
Example: Exposure to air pollution

Prior predictive checking



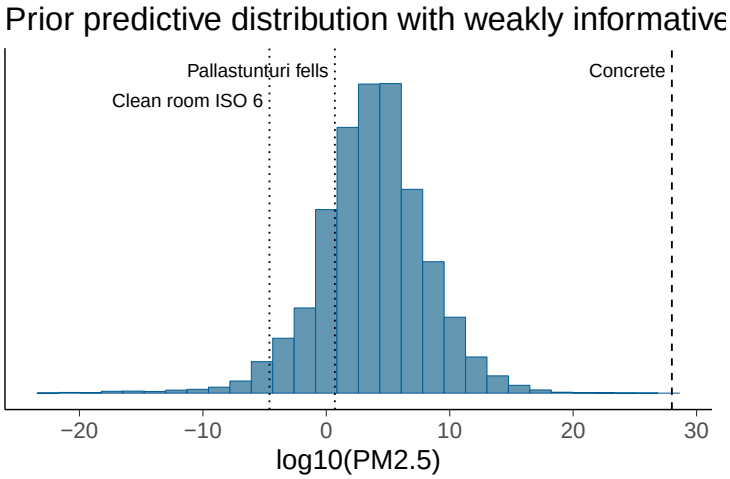
Example: Exposure to air pollution

Prior predictive checking



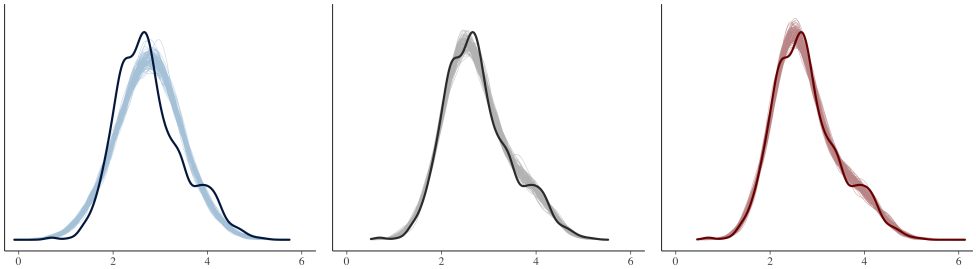
Example: Exposure to air pollution

Prior predictive checking



Example: Exposure to air pollution

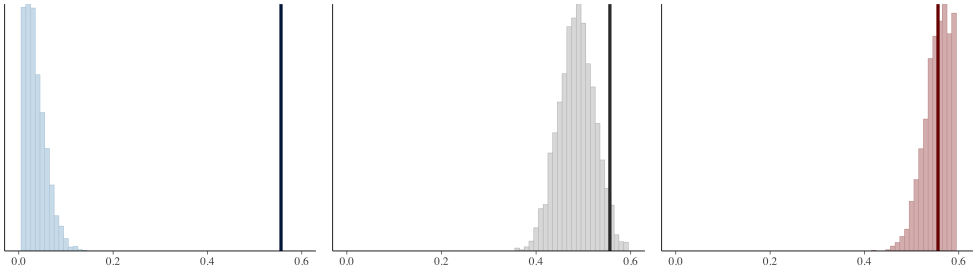
Posterior predictive checking – marginal predictive distributions



(a) Model 1 (b) Model 2 (c) Model 3

Example: Exposure to air pollution

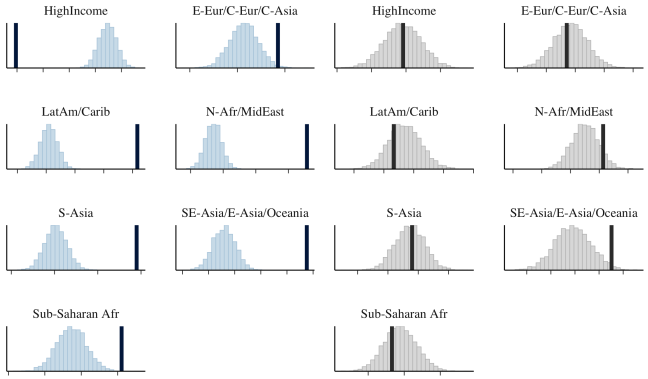
Posterior predictive checking – test statistic (skewness)



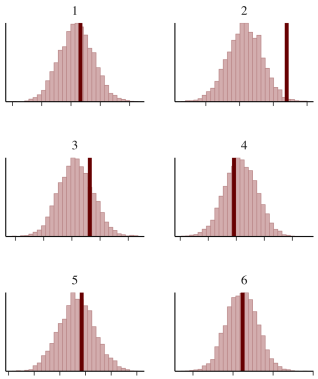
(a) Model 1 (b) Model 2 (c) Model 3

Example: Exposure to air pollution

Posterior predictive checking – test statistic (median for groups)



(a) Model 1 (b) Model 2



(c) Model 3