Part 7

Model selection

ST740

North Carolina State University

Model selection

- We now have many potential models in our arsenal
- For a given dataset, how do determine whether a simple model is sufficient or if we need to bring out the “big guns”?
- Is there a “right” model? Probably not
- A **statistical model** is a mathematical representation of the system that includes errors and biases in the observation process
- All models are simplifications of reality
- Why fit models at all?
- We want a model that is as simple as possible yet seems to fit the data reasonably well

Outline

Model selection

- **Bayes factors**
- Model averaging
- Selection criteria
- Cross validation

Model evaluation

- Measures of fit
- Posterior predictive checks

Bayes factors (BF)

- In some sense BFs are the gold standard
- Say we are comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$
- For example, $Y \sim \text{Binomial}(n, \theta)$ and the two models are

$$\mathcal{M}_1 : \theta = 0.5 \quad \text{and} \quad \mathcal{M}_2 : \theta \neq 0.5$$

- Another example, $Y_1, Y_2, ..., Y_n$ is a time series and

$$\mathcal{M}_1 : \text{Cor}(Y_{t+1}, Y_t) = 0 \quad \text{and} \quad \mathcal{M}_2 : \text{Cor}(Y_{t+1}, Y_t) > 0$$

- Another example,

$$\mathcal{M}_1 : \text{E}(Y) = \beta_0 + \beta_1 X \quad \text{and} \quad \mathcal{M}_2 : \text{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

Bayes factors (BF)

- This is really the same as hypothesis testing, and in fact Bayes factors are the gold standard for hypothesis testing
- As before we proceed by computing the posterior probability of the two models
- This require priors probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$
- This is not prior on a parameter, it is a prior on the model!
- This approach permits statements such “Given the data we have observed, the quadratic model is 5 times more likely than a linear model”

Bayes factors (BF)

- The Bayes factor for model 2 compared to model 1 is

$$BF = \frac{\text{Posterior odds}}{\text{Prior odds}} = \frac{p(\mathcal{M}_2|\mathbf{Y})/p(\mathcal{M}_1|\mathbf{Y})}{p(\mathcal{M}_2)/p(\mathcal{M}_1)} = \frac{p(\mathbf{Y}|\mathcal{M}_2)}{p(\mathbf{Y}|\mathcal{M}_1)}$$

- Rule of thumb: $BF > 10$ is strong evidence for $\mathcal{M}_2$

- Rule of thumb: $BF > 100$ is decisive evidence for $\mathcal{M}_2$

- In linear regression, *BIC* approximates the BF comparing a model to the null model

Example

- $Y \sim \text{Binomial}(n, \theta)$ with

$$\mathcal{M}_1 : \theta = 0.5 \text{ and } \mathcal{M}_2 : \theta \neq 0.5$$

- $p(Y|\mathcal{M}_1)$ is just the binomial density with $\theta = 0.5$

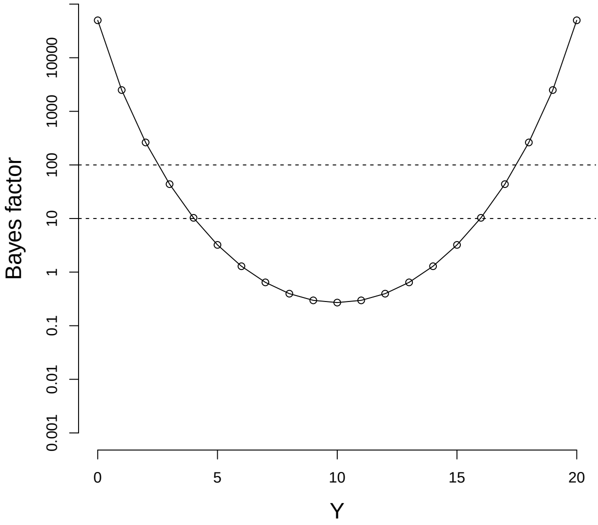
- $\mathcal{M}_2$ involves an unknown parameter θ

- This requires a prior, say $\theta \sim \text{Beta}(a, b)$, and integration

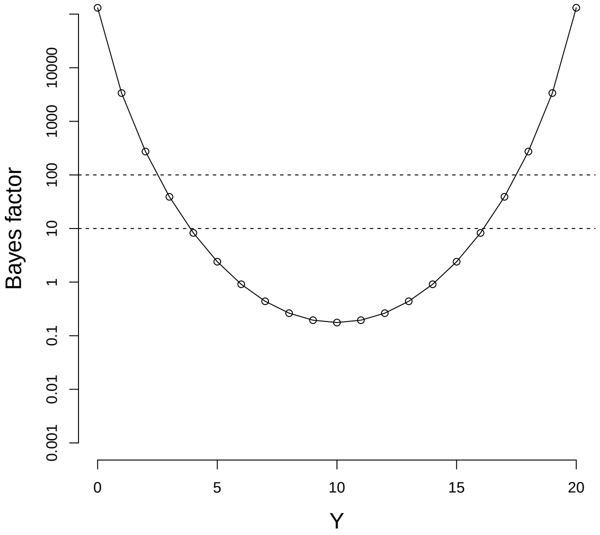
$$p(Y|\mathcal{M}_2) = \int p(Y, \theta) d\theta = \int p(Y|\theta) p(\theta) d\theta$$

- See “BF Beta-binomial” in the online derivations

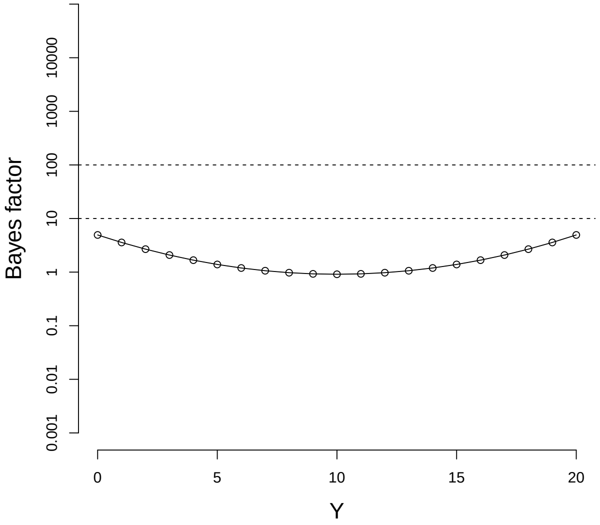
BF by Y with $n = 20$ and prior $a = 1$ and $b = 1$



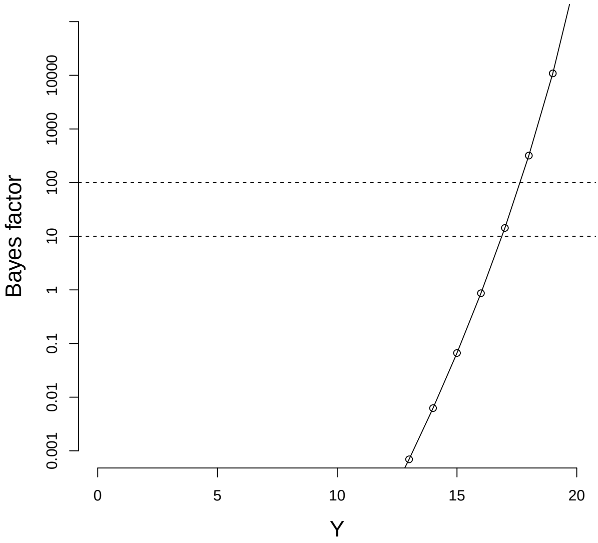
BF by Y with $n = 20$ and prior $a = 0.5$ and $b = 0.5$



BF by Y with $n = 20$ and prior $a = 50$ and $b = 50$



BF by Y with $n = 20$ and prior $a = 50$ and $b = 1$



Problems with Bayes factors

- Often hard to compute the required integrals which is only feasible for simple models

- Requires proper priors

- Can be very sensitive to priors (Lindley's paradox)

- In most cases, I prefer computing posterior intervals from the full model and testing by comparing these to the null

Lindley's paradox

- Lindley's paradox is when Bayesian and frequentist hypothesis tests give vastly different results
- For example, $Y \sim \text{Normal}(\mu, \sigma^2)$ with $H_o : \mu = 0$ versus $H_a : \mu \neq 0$
- The Bayesian approach requires a prior under H_a, say $\mu \sim \text{Normal}(0, \tau^2)$
- Paradox: For any Y, $\text{Prob}(H_0|Y) \rightarrow 0$ as $\tau \rightarrow \infty$

Lindley’s paradox

Derivation

Computing Bayes factors using MCMC

- If models can be written as nested, then MCMC can be used to approximate model probabilities

- For example, say

  $\mathcal{M}_1 : \text{E}(Y) = \beta_0 + \beta_1 X$ and $\mathcal{M}_2 : \text{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$

- Both model can be written as

$$\text{E}(Y) = \beta_0 + \beta_1 X + \gamma\beta_2 X^2$$

  where $\gamma \in \{0, 1\}$ indicates the model

- The prior on models becomes $\gamma \sim$ Bernoulli(0.5)

- Then Prob($\gamma = 1|\mathbf{Y}$) = Prob($\mathcal{M}_2|\mathbf{Y}$) can be approximated using MCMC

Stochastic search variable selection (SSVS)

- This is the Bayesian analog of forward/backward/stepwise variable selection
- We place a prior on all 2^p models using p variable inclusion indicators γ_j
- MCMC returns the approximate posterior probability of each model
- With large p all models will have low probability and so this requires long MCMC runs
- As with Bayesian factors, SSVS can be sensitive to priors

Multiple testing

- Bayesian model selection can be extended to multiple tests

- Example: A study measures expression at p genetic markers

- Let θ_j the difference in mean expression for cancer and control subjects for marker j

- For $j \in \{1, ..., p\}$ the hypotheses are

$$\mathcal{M}_{j1} : \theta_j = 0 \quad \text{versus} \quad \mathcal{M}_{j2} : \theta_j \neq 0$$

- In addition to multiple testing, correlation based on the markers' position on the chromosome is a challenge

Frequentist multiple testing criteria

- Instead of controlling error rates of individual tests, we could consider error rates across all p tests

- Global Type I error is the probability of rejecting any of the p tests given $\theta_j = 0$ for all $j \in \{1, ..., p\}$

- Global Type I error can be controlled via the threshold for the individual tests

- Or we would conduct one global test

$$
\begin{aligned}
\mathcal{M}_1 &: \theta_j = 0 \text{ for all } \ j \in \{1, ..., p\} \\
\mathcal{M}_2 &: \theta_j \neq 0 \text{ for at least one } \ j \in \{1, ..., p\}
\end{aligned}
$$

- A Bayesian global test computes the BF for $\mathcal{M}_2$ versus $\mathcal{M}_1$

Frequentist multiple testing criteria

- In most multiple testing cases, the global null is unrealistic
- False discovery rate (FDR) control is more common
- For a given dataset and testing decision

$$\mathit{FDP}(\mathbf{Y}) = \frac{\text{Number of rejections where the null is true}}{\text{Number of rejections}}$$

- For a given testing procedure, its frequentist FDR is

$$\mathit{FDR} = \mathrm{E}\{P(\mathbf{Y})\}$$

  where the expectation is wrt **Y**

- The testing procedure is tuned (e.g., p-value thresholds are set) so that $\mathit{FDR} \approx \alpha$

Bayesian false discovery rate (BFDR)

- Let $\delta_j = 1$ if the alternative $\mathcal{M}_{j2}$ is true and $\delta_j = 0$ otherwise

- In the Bayesian setting, $\delta = (\delta_1, ..., \delta_p)$ is a random variable and MCMC gives its joint posterior distribution

- We use Bayesian decision theory to summarize δ

- Say our decision is $r_j = 1$ if we reject $\mathcal{M}_{j1}$ in favor of $\mathcal{M}_{j2}$

- The false discovery proportion is

$$FDP(r, \delta) = \frac{\sum_{j=1}^{p} r_j(1 - \delta_j)}{\sum_{j=1}^{p} r_j}$$

Bayesian false discovery rate (BFDR

- To make the problem tractable, say our decision rule is

$$r_j(t) = \mathcal{I}(\text{reject } \mathcal{M}_{j1} \text{ in favor of } \mathcal{M}_{j2} \text{ if } \pi_j > t),$$

where $\pi_j = \text{Prob}(\delta_j = 1|\text{data}) = \text{Prob}(\mathcal{M}_{j2} = 1|\text{data})$

- Given this rule, the FDP is a function only of the threshold t

$$FDP(t, \delta) = \frac{\sum_{j=1}^{p} r_j(t)(1 - \delta_j)}{\sum_{j=1}^{p} r_j(t)}$$

Bayesian false discovery rate (BFDR)

- From a Bayesian perspective, the random variable in $FDP(t, \delta)$ is δ

- BFDR is the posterior mean $FDP(t, \delta)$

$$BFDR(t) = E_{\delta|Y}\{FDP(t, \delta)\}$$

- Since this expectation is wrt the joint posterior of δ it accounts for dependence between tests

- We can select t so that $BFDR(t) \approx \alpha$

- This controls posterior FDR, not frequentist FDR, although connections can be made [1]

[1] Storey, 2003, The positive false discovery rate: A Bayesian interpretation and the q-value

Outline

Model selection

- Bayes factors
- **Model averaging**
- Selection criteria
- Cross validation

Model evaluation

- Measures of fit
- Posterior predictive checks

Model averaging

- Let's go back to the linear regression example

  $\mathcal{M}_1 : \mathrm{E}(Y) = \beta_0 + \beta_1 X$ and $\mathcal{M}_2 : \mathrm{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$

- Say we have fit both models and found that both are about equally likely, but that $\mathcal{M}_1$ is slightly preferred

- For prediction, $\hat{Y}$, we could simply take the prediction that comes from fitting $\mathcal{M}_1$

- But the prediction from $\mathcal{M}_2$ is likely different and nearly as accurate

- Also, taking the prediction from $\mathcal{M}_1$ suppresses our uncertainty about the form of the model

Model averaging

- Let $\hat{Y}_k$ be the prediction from model $\mathcal{M}_k$ for $k = 1, 2$
- The model averaged predictor is

$$\hat{Y} = w\hat{Y}_1 + (1 - w)\hat{Y}_2$$

- It can be shown that the optimal weight w is the posterior probability of $\mathcal{M}_1$
- Madigan and Raftery[2] show that BMA gives better prediction than any individual model
- In regression with p predictors, there are 2^p models and all model probabilities will likely be small

[2] https://www.jstor.org/stable/pdf/2291017.pdf

Outline

Model selection

- Bayes factors
- Model averaging
- **Selection criteria**
- Cross validation

Model evaluation

- Measures of fit
- Posterior predictive checks

Information criteria

- Several information criteria have been proposed that do not require fitting the model several times
- Many are functions of the **deviance**, i.e., twice the negative log likelihood

$$D(\mathbf{Y}|\theta) = -2\log[f(\mathbf{Y}|\theta)]$$

- Ideally, models will have small deviance
- However, if a model is too complex it will have small deviance between be unstable (over-fitting)
- The Akaike information criteria has a complexity penalty

$$AIC = D(\mathbf{Y}|\hat{\theta}) + 2p$$

  where $\hat{\theta}$ is the MLE
- Model with smaller *AIC* are preferred

Bayesian information criteria (BIC)

- The Bayesian information criteria is similar

$$BIC = D(\mathbf{Y}|\hat{\theta}) + \log(n)p$$

- This is motivated as an approximation to the log Bayes factor of the model compared to the null model
- However, this is only an asymptotic (large *n*) approximation
- With large *n* the prior is irrelevant, and so this is not satisfying to a subjective Bayesian

Deviance information criteria (DIC)

- *DIC* is a popular Bayesian analog of *AIC* or *BIC*
- Unlike CV, *DIC* requires only one model fit
- Unlike BF, it can be applied to complex models
- However, proceed with caution
- *DIC* really only applies when the posterior is approximately normal, and will give misleading results when the posterior far from normality, e.g., bimodal
- *DIC* is also criticized for selecting overly-complex models

Deviance information criteria (DIC)

- Let $\bar{D} = \text{E}[D(Y|\theta)|\mathbf{Y}]$ be the posterior mean of the deviance
- Denote $\hat{\boldsymbol{\theta}}$ as the posterior mean of $\boldsymbol{\theta}$
- The effective number of parameters is

$$p_D = \bar{D} - D(\mathbf{Y}|\hat{\boldsymbol{\theta}})$$

- DIC can be written like *AIC*,

$$DIC = \bar{D} + p_D = D(\mathbf{Y}|\hat{\boldsymbol{\theta}}) + 2p_D$$

- Models with small $\bar{D}$ fit the data well
- Models with small p_D are simple
- We prefer models that are simple and fit well, so we select the model with smallest *DIC*

DIC

- The effective number of parameters is a useful measure of model complexity
- Intuitively, if there are p parameters and we have uninformative priors then $p_D \approx p$
- However, $p_D << p$ if there are strong priors
- For example, how many free degrees of freedom do we have with $\theta \sim \text{Beta}(1, 1)$ versus $\theta \sim \text{Beta}(1000, 1000)$?
- In some cases p_D has a nice closed form
- A few examples are worked out in “DIC” on the online derivations

DIC

- As with *AIC* or *BIC*, we compute *DIC* for all models under consideration and select the one with smallest *DIC*

- Rule of thumb: a difference of *DIC* of less than 5 is not definitive and a difference greater than 10 is substantial

- As with *AIC* or *BIC*, the actual value is meaningless, only differences are relevant

- *DIC* can only be used to compare models with the same likelihood

Watanabe-Akaike information criteria (WAIC)

- *WAIC* is an alternative to *DIC*
- It is motivated as an approximation to leave-one-out CV
- In the end *WAIC* has model-fit and model-complexity components
- It is used the same as *DIC* with smaller *WAIC* begin preferred
- In practice the two often give similar results, but *WAIC* is arguably more theoretically justified

Watanabe-Akaike information criteria (WAIC)

- *WAIC* is written in terms of the posterior of the likelihood rather than parameters

- Let m_i and v_i be the posterior mean and variance of

$$\log[f(Y_i|\boldsymbol{\theta})]$$

- The effective model size is $p_W = \sum_{i=1}^{n} v_i$

- The criteria is

$$WAIC = -2\sum_{i=1}^{n} m_i + 2p_W$$

Outline

Model selection

- Bayes factors
- Model averaging
- Selection criteria
- **Cross validation**

Model evaluation

- Measures of fit
- Posterior predictive checks

Cross validation

- Another very common approach is cross-validation
- This is exactly the same procedure used in classical statistics
- This operates under the assumption that the “true” model likely produces better out-of-sample predictions than competing models
- Advantages: Simple, intuitive, and broadly applicable
- Disadvantages: Slow because it requires several model fits and it is hard to say a difference is statistically significant

K-fold Cross validation

0 Split the data into K equally-sized groups

1 Set aside group k as test set and fit the model to the remaining $K - 1$ groups

2 Make predictions for the test set k based on the model fit to the training data

3 Repeat steps 1 and 2 for $k = 1, ..., K$ giving a predicted value $\hat{Y}_i$ for all n observations

4 Measure prediction accuracy, e.g.,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Variants

- Usually K is either 5 or 10

- $K = n$ is called “leave-one-out” cross-validation, which is great but slow

- The predicted value $\hat{Y}_i$ can be either the posterior predictive mean or median

- Mean squared error (MSE) can be replaced with Mean absolute deviation

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- Also common to compute Cor(Y_i, Y_i), the average posterior variance and coverage of prediction intervals

Measures of fit for point predictions

- $\text{Corr}(Y_i, \hat{Y}_i)$
- $\text{Corr}(Y_i, \hat{Y}_i)^2$
- Bias: $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)/n$
- MSE: $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2/n$
- MAD: $\sum_{i=1}^{n}|\tilde{Y}_i - Y_i|/n$

where $\hat{Y}_i$ and $\tilde{Y}_i$ are the posterior predictive mean and median

Measures of fit for binary data

- Classification accuracy, $\sum_{i=1}^{n} I(Y_i - \hat{Y}_i)/n$

- True and false positive rates comparing Y_i and $\hat{Y}_i$

- Area under the receiver-operator and precision-recall curves comparing Y_i and p_i

- Brier score, $BS = \sum_{i=1}^{n} (Y_i - p_i)^2/n$

where p_i is the posterior predictive probability that $Y_i = 1$ and $\hat{Y}_i = I(p_i > 0.5)$

Measures of fit for quantiles and intervals

- Coverage of 95% prediction interval, $\sum_{i=1}^{n} I\{l_i < Y_i < u_i\}$
- Interval width $\sum_{i=1}^{n}(u_i - l_i)/n$
- Interval score

$$\frac{1}{n}\sum_{i=1}^{n}(u_i - l_i) + \frac{2}{\alpha}\{(l_i - Y_i)I(Y_i < l_i) + (Y_i - u_i)I(Y_i > u_i)\}$$

- Quantile score $\sum_{i=1}^{n} \rho_\tau\{Y_i - q_i(\tau)\}/n$ for check function

$$\rho_\tau(e) = \begin{cases} \tau|e| & e \geq 0 \\ (1-\tau)|e| & e < 0 \end{cases}$$

where $q_i(\tau)$ is the τ quantile of the PPD,$l_i = q_i(\alpha/2)$ and $u_i = q_i(1-\alpha/2)$

Measures of overall fit of the PPD

- The log score, $\sum_{i=1}^{n} \log(\hat{f}_i(Y_i)\}/n$
- Probability Integral Transform (PIT) Histogram, i.e., a histogram of

  $$PIT_i = \hat{F}_i(Y_i)$$

  which should be roughly uniform
- Continuous rank probability score (CRPS)

  $$\int \{\hat{F}(y) - I(Y_i < y)\}^2 dy$$

where $\hat{f}$ and $\hat{F}$ are the posterior predictive PDF and CDF

Outline

Model selection

- Bayes factors
- Model averaging
- Selection criteria
- Cross validation

Model evaluation

- **Measures of fit**
- Posterior predictive checks

Measures of fit

- In least squares, fit is measured using (adjusted) R^2
- A Bayesian version is proposed in Gelman (2019)
- Let $E(Y_i|\boldsymbol{\theta}) = \mu_i(\boldsymbol{\theta})$ and $\text{Var}(Y_i|\boldsymbol{\theta}) = \sigma^2(\boldsymbol{\theta})$
- Then

  $$R^2 = \frac{\text{V}\{\mu_1(\boldsymbol{\theta}), ..., \mu_n(\boldsymbol{\theta})\}}{\text{V}\{\mu_1(\boldsymbol{\theta}), ..., \mu_n(\boldsymbol{\theta})\} + \text{M}\{\sigma_1^2(\boldsymbol{\theta}), ..., \sigma_n^2(\boldsymbol{\theta})\}},$$

  where M and V are the sample mean and variance operators, respectively
- Mixing over $\boldsymbol{\theta}$ gives a posterior distribution of R_2

Outline

Model selection

- Bayes factors
- Model averaging
- Selection criteria
- Cross validation

Model evaluation

- Measures of fit
- **Posterior predictive checks**

Posterior predictive checks

- After comparing a few models, we settle on the one that seems to fit the best

- Given this model, we then verify it is adequate

- The usual residual checks are appropriate here: qq-plots; added variable plots; etc.

- A uniquely Bayesian diagnostic is the posterior predictive check

- This leads to the Bayesian p-value

Posterior predictive distributions

- Before discussing posterior predictive checks, let's review Bayesian prediction in general

- The plug-in approach would fix the parameters θ at the posterior mean $\hat{\theta}$ and then predict $Y_{new} \sim f(y|\hat{\theta})$

- This suppresses uncertainty in θ

- We would like to propagate this uncertainty through to the predictions

Posterior predictive distributions (PPD)

- We really want the PPD

$$f(Y_{new}|\mathbf{Y}) = \int f(Y_{new}, \theta|\mathbf{y})d\theta = \int f(Y_{new}|\theta)f(\theta|\mathbf{y})d\theta$$

- MCMC easily produces draws from this distribution

- To make S draws from the PPD, for each of the S MCMC draws of θ we draw a Y_{new}

- This gives draws from the PPD and clearly accounts for uncertainty in θ.

Posterior predictive checks

- Posterior predictive checks sample many datasets from the PPD with the identical design (same n, same $\mathbf{X}$) as the original data set

- We then define a statistic describing the dataset, e.g.,

$$d(\mathbf{Y}) = \max\{Y_1, ..., Y_n\}$$

- Denote the statistic for the original data set as d_0 and the statistic from simulated data set number s as d_s

- If the model is correct, then d_0 should fall in the middle of the $d_1, ..., d_S$

Posterior predictive checks

- A measure of how extreme the observed data is relative to this sampling distribution is the Bayesian p-value

$$p = \frac{1}{S}\sum_{s=1}^{S} I(d_s > d_0)$$

- If *p* is near zero or one the model doesn't fit

- This is repeated for several *d* to give a comprehensive evaluation of model fit