Chapter 7: Evaluating, comparing, and expanding models
Discussion of homework due beginning of Class 6b

- Theory problem
- Computing problem
- Applied problem
Discussion of homework due beginning of Class 6b

- Theory problem
- Computing problem
- Applied problem
Discussion of homework due beginning of Class 6b

- Theory problem
- Computing problem
- Applied problem
Discussion of homework due beginning of Class 6b

- Theory problem
- Computing problem
- Applied problem
Stopping rules and the distribution of $y^{rep}$ for predictive checks
Stoping rules and the distribution of $y_{rep}$ for predictive checks
Computing problem

- Fitting simple linear models in Stan
- Checking the coverage of posterior intervals
Fitting simple linear models in Stan
Checking the coverage of posterior intervals
Computing problem

- Fitting simple linear models in Stan
- Checking the coverage of posterior intervals
Applied problem

 Estimate probability of knowing someone gay, given age, sex, and race
Estimate probability of knowing someone gay, given age, sex, and race
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
- Continuous model expansion
- Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
- Continuous model expansion
- Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
  - Model comparison
  - Bayes factors
  - Continuous model expansion
  - Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
- Continuous model expansion
- Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
  - Continuous model expansion
- Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
- Continuous model expansion
- Model checking, robustness, and transformations
7. Evaluating, comparing, and expanding models

- Summarizing predictive accuracy using expected log probability of data
- Deviance, information criteria, and effective number of parameters
- Model comparison
- Bayes factors
- Continuous model expansion
- Model checking, robustness, and transformations
### Example of within-sample and out-of-sample prediction

<table>
<thead>
<tr>
<th>Election</th>
<th>Incumbent party's share of the popular vote</th>
<th>Income growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johnson vs. Goldwater (1964)</td>
<td></td>
<td>more than 4%</td>
</tr>
<tr>
<td>Reagan vs. Mondale (1984)</td>
<td></td>
<td>3% to 4%</td>
</tr>
<tr>
<td>Nixon vs. McGovern (1972)</td>
<td></td>
<td>3% to 4%</td>
</tr>
<tr>
<td>Humphrey vs. Nixon (1968)</td>
<td></td>
<td>3% to 4%</td>
</tr>
<tr>
<td>Eisenhower vs. Stevenson (1956)</td>
<td></td>
<td>2% to 3%</td>
</tr>
<tr>
<td>Stevenson vs. Eisenhower (1952)</td>
<td></td>
<td>2% to 3%</td>
</tr>
<tr>
<td>Gore vs. Bush, Jr. (2000)</td>
<td></td>
<td>1% to 2%</td>
</tr>
<tr>
<td>Bush, Sr. vs. Dukakis (1988)</td>
<td></td>
<td>1% to 2%</td>
</tr>
<tr>
<td>Bush, Jr. vs. Kerry (2004)</td>
<td></td>
<td>1% to 2%</td>
</tr>
<tr>
<td>Ford vs. Carter (1976)</td>
<td></td>
<td>1% to 2%</td>
</tr>
<tr>
<td>Clinton vs. Dole (1996)</td>
<td></td>
<td>1% to 2%</td>
</tr>
<tr>
<td>Nixon vs. Kennedy (1960)</td>
<td></td>
<td>0% to 1%</td>
</tr>
<tr>
<td>Bush, Sr. vs. Clinton (1992)</td>
<td></td>
<td>0% to 1%</td>
</tr>
<tr>
<td>McCain vs. Obama (2008)</td>
<td></td>
<td>0% to 1%</td>
</tr>
<tr>
<td>Carter vs. Reagan (1980)</td>
<td></td>
<td>negative</td>
</tr>
</tbody>
</table>
7.1. Measures of predictive accuracy

- Log predictive density as a measure of fit
- Out-of-sample predictive accuracy as a gold standard
- Why log data density rather than log posterior density?
- Why log density rather than mean squared error?
7.1. Measures of predictive accuracy

- Log predictive density as a measure of fit
  - Out-of-sample predictive accuracy as a gold standard
  - Why log data density rather than log posterior density?
  - Why log density rather than mean squared error?
7.1. Measures of predictive accuracy

- Log predictive density as a measure of fit
- Out-of-sample predictive accuracy as a gold standard
  - Why log data density rather than log posterior density?
  - Why log density rather than mean squared error?
7.1. Measures of predictive accuracy

- Log predictive density as a measure of fit
- Out-of-sample predictive accuracy as a gold standard
- Why log data density rather than log posterior density?
- Why log density rather than mean squared error?
7.1. Measures of predictive accuracy

- Log predictive density as a measure of fit
- Out-of-sample predictive accuracy as a gold standard
- Why log data density rather than log posterior density?
- Why log density rather than mean squared error?
Accounting for posterior uncertainty

Log predictive density, $p(y | \theta)$
7.2. Information criteria and cross-validation

- Estimates of out-of-sample predictive accuracy:
  - Within-sample predictive accuracy
  - Subtracting an adjustment
  - Cross-validation
7.2. Information criteria and cross-validation

- Estimates of out-of-sample predictive accuracy:
  - Within-sample predictive accuracy
  - Subtracting an adjustment
  - Cross-validation
7.2. Information criteria and cross-validation

- Estimates of out-of-sample predictive accuracy:
  - Within-sample predictive accuracy
  - Subtracting an adjustment
  - Cross-validation
7.2. Information criteria and cross-validation

- Estimates of out-of-sample predictive accuracy:
  - Within-sample predictive accuracy
  - Subtracting an adjustment
  - Cross-validation
7.2. Information criteria and cross-validation

- Estimates of out-of-sample predictive accuracy:
  - Within-sample predictive accuracy
  - Subtracting an adjustment
  - Cross-validation
Akaike information criterion (AIC)

- \( \text{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)
- elpd = expected log predictive density
- Based on fit to observed data given mle
- Goal: elpd = \( E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}})) \)
Akaike information criterion (AIC)

\[ \hat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \]

- \( \text{elpd} = \) expected log predictive density
- Based on fit to observed data given mle
- Goal: \( \text{elpd} = E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}})) \)
Akaike information criterion (AIC)

- \( \text{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)

- \( \text{elpd} = \) expected log predictive density

- Based on fit to observed data given mle

- Goal: \( \text{elpd} = E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}})) \)

  - Expectation averages over the predictive distribution of \( \tilde{y} \)
Akaike information criterion (AIC)

- \( \hat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)
- \( \text{elpd} = \text{expected log predictive density} \)
- Based on fit to observed data given mle
- Goal: \( \text{elpd} = E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}})) \)
  - Expectation averages over the predictive distribution of \( \tilde{y} \)
Akaike information criterion (AIC)

\[ \text{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \]

- \text{elpd} = \text{expected log predictive density}
- Based on fit to observed data given mle
- Goal: \text{elpd} = E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}}))
  - Expectation averages over the predictive distribution of \( \tilde{y} \)
Akaike information criterion (AIC)

- \( \text{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)
- elpd = expected log predictive density
- Based on fit to observed data given mle
- Goal: \( \text{elpd} = E(\log p(\tilde{y}|\hat{\theta}_{\text{mle}})) \)
  - Expectation averages over the predictive distribution of \( \tilde{y} \)
Effective number of parameters

- Fitting a function with 30 parameters given 30 data points:
  - $y_t \sim \text{Poisson}(N_t \theta_t)$, for $t = 35, \ldots, 64$
  - Uniform prior: $p(\theta) \propto 1$
  - Constraint of increasing convexity: $p(\theta) \propto 1$ under constraint
- How many parameters are being estimated?
Fitting a function with 30 parameters given 30 data points:

- \( y_t \sim \text{Poisson}(N_t \theta_t) \), for \( t = 35, \ldots, 64 \)
- Uniform prior: \( p(\theta) \propto 1 \)
- Constraint of increasing convexity: \( p(\theta) \propto 1 \) under constraint

How many parameters are being estimated?
Fitting a function with 30 parameters given 30 data points:

- \( y_t \sim \text{Poisson}(N_t \theta_t) \), for \( t = 35, \ldots, 64 \)
- Uniform prior: \( p(\theta) \propto 1 \)
- Constraint of increasing convexity: \( p(\theta) \propto 1 \) under constraint

How many parameters are being estimated?
Fitting a function with 30 parameters given 30 data points:

- $y_t \sim \text{Poisson}(N_t \theta_t)$, for $t = 35, \ldots, 64$
- Uniform prior: $p(\theta) \propto 1$
- Constraint of increasing convexity: $p(\theta) \propto 1$ under constraint

How many parameters are being estimated?
Fitting a function with 30 parameters given 30 data points:

- $y_t \sim \text{Poisson}(N_t \theta_t)$, for $t = 35, \ldots, 64$
- Uniform prior: $p(\theta) \propto 1$
- Constraint of increasing convexity: $p(\theta) \propto 1$ under constraint

How many parameters are being estimated?
Effective number of parameters

- Fitting a function with 30 parameters given 30 data points:
  - \( y_t \sim \text{Poisson}(N_t \theta_t) \), for \( t = 35, \ldots, 64 \)
  - Uniform prior: \( p(\theta) \propto 1 \)
  - Constraint of increasing convexity: \( p(\theta) \propto 1 \) under constraint

- How many parameters are being estimated?
Data and posterior mode estimate
Posterior mode and posterior simulations
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:

- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
  - Uniform prior on second differences
  - Distribution of 30 order statistics from a uniform distribution
  - Things get even worse as "30" gets larger:

- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And ... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And ... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
    - And ... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And ... less data at each age point!
- Seemingly weak prior is extremely informative!
Problem with model of increasing convexity

- Prior is strongly concentrated around quadratic curves
- Uniform prior on second differences
- Distribution of 30 order statistics from a uniform distribution
- Things get even worse as “30” gets larger:
  - Consider data by month instead of by year
  - 30 becomes 360
  - And... less data at each age point!
- Seemingly weak prior is extremely informative!
Deviance information criterion (DIC)

- $\text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}$
- Based on fit to observed data given posterior mean
- Effective number of parameters $p_{\text{DIC}}$ computed based on normal approx ($\chi^2_k$ approximation to $-2\log$ likelihood):
  - Either is asymptotically ok in expectation
  - Advantages and disadvantages of each $p_{\text{DIC}}$ formula
Deviance information criterion (DIC)

\[ \text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}} \]

- Based on fit to observed data given posterior mean
- Effective number of parameters \( p_{\text{DIC}} \) computed based on normal approx (\( \chi^2_k \) approximation to \(-2\log \text{likelihood})
  \[ p_{\text{DIC}} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{pred density given } \theta) \right] \]
  Or, \( p_{\text{DIC}} = \text{var}_\text{post}(\log p(y|\theta)) \)
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each \( p_{\text{DIC}} \) formula
Deviance information criterion (DIC)

\[ \text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}} \]

-Based on fit to observed data given posterior mean
- Effective number of parameters \( p_{\text{DIC}} \) computed based on normal approx (\( \chi_k^2 \) approximation to \(-2\log \text{likelihood}):
  \[ p_{\text{DIC}} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{pred density given } \theta) \right] \]
  Or, \( p_{\text{DIC,alt}} = \text{var}_{\text{post}}(\log p(y|\theta)) \)
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each \( p_{\text{DIC}} \) formula
Deviance information criterion (DIC)

- \( \text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}} \)
- Based on fit to observed data given posterior mean
- Effective number of parameters \( p_{\text{DIC}} \) computed based on normal approx (\( \chi^2_k \) approximation to \(-2\log \text{likelihood})
  - \( p_{\text{DIC}} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{pred density given } \theta) \right] \)
  - Or, \( p_{\text{DIC,alt}} = \text{var}_{\text{post}}(\log p(y|\theta)) \)
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each \( p_{\text{DIC}} \) formula

Andrew Gelman  Bayesian Data Analysis, class 6b
Deviance information criterion (DIC)

- \( \text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}} \)
- Based on fit to observed data given posterior mean
- Effective number of parameters \( p_{\text{DIC}} \) computed based on normal approx (\( \chi_k^2 \) approximation to \(-2\log\) likelihood):
  - \( p_{\text{DIC}} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{pred density given } \theta) \right] \)
  - Or, \( p_{\text{DIC,alt}} = \text{var}_{\text{post}}(\log p(y|\theta)) \)
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each \( p_{\text{DIC}} \) formula
Deviance information criterion (DIC)

- $\hat{\text{elpd}}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}$
- Based on fit to observed data given posterior mean
- Effective number of parameters $p_{\text{DIC}}$ computed based on normal approx ($\chi^2_k$ approximation to $-2\log$ likelihood):
  - $p_{\text{DIC}} = 2 \left[ (\log \text{ pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{ pred density given } \theta) \right]$
  - Or, $p_{\text{DIC,alt}} = \text{var}_{\text{post}}(\log p(y|\theta))$
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each $p_{\text{DIC}}$ formula
Deviance information criterion (DIC)

- $\text{elpd}_{\text{DIC}} = \log p(y|\hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}$
- Based on fit to observed data given posterior mean
- Effective number of parameters $p_{\text{DIC}}$ computed based on normal approx ($\chi^2_k$ approximation to $-2\log$ likelihood):
  - $p_{\text{DIC}} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{\text{Bayes}}) - \text{avg}(\log \text{pred density given } \theta) \right]$
  - Or, $p_{\text{DIC},\text{alt}} = \text{var}_{\text{post}}(\log p(y|\theta))$
- Either is asymptotically ok in expectation
- Advantages and disadvantages of each $p_{\text{DIC}}$ formula
Deviance information criterion (DIC)

\[ \text{elpd}_{DIC} = \log p(y|\hat{\theta}_{Bayes}) - p_{DIC} \]

- Based on fit to observed data given posterior mean
- Effective number of parameters \( p_{DIC} \) computed based on normal approx (\( \chi^2_k \) approximation to \(-2\log\text{likelihood})
  
  \[ p_{DIC} = 2 \left[ (\log \text{pred density given } \hat{\theta}_{Bayes}) - \text{avg} (\log \text{pred density given } \theta) \right] \]
  
  Or, \( p_{DIC,alt} = \text{var}_{post} (\log p(y|\theta)) \)

- Either is asymptotically ok in expectation
- Advantages and disadvantages of each \( p_{DIC} \) formula
Watanabe-Akaike information criterion (WAIC)

\[ \hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}} \]

- \( \hat{\text{elppd}}_{\text{WAIC}} \) = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[ \text{elppd}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}} \]

- elppd = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[
\hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}}
\]

- elppd = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[ \hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}} \]

- elppd = expected log posterior predictive density
- Based on posterior predictive fit to observed data

\[ p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \]

- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[ \hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}} \]

- elppd = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[ \hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}} \]

- \( \hat{\text{elppd}}_{\text{WAIC}} \) = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[
elppd_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}}
\]

- elppd = expected log posterior predictive density
- Based on posterior predictive fit to observed data
- \( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)
- Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations
- Requires data partition
- Connection to leave-one-out cross-validation
Watanabe-Akaike information criterion (WAIC)

\[
\hat{\text{elppd}}_{\text{WAIC}} = \left( \sum_{i=1}^{n} \log p_{\text{post}}(y_i) \right) - p_{\text{WAIC}}
\]

\( \hat{\text{elppd}} = \) expected log posterior predictive density

Based on posterior predictive fit to observed data

\( p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)) \)

Compute \( p_{\text{post}} \) and \( \text{var}_{\text{post}} \) using simulations

Requires data partition

Connection to leave-one-out cross-validation
Recall AIC: $\text{elpd}_{AIC} = \log p(y | \hat{\theta}_{mle}) - k$

- BIC subtracts $\frac{k}{2} \log n$ instead of $k$
- Not an estimate of out-of-sample predictive accuracy
- Favors smaller models
“Bayesian” information criterion (BIC)

Recall AIC: \( \text{elpd}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)

BIC subtracts \( \frac{k}{2} \log n \) instead of \( k \)

Not an estimate of out-of-sample predictive accuracy

Favors smaller models
“Bayesian” information criterion (BIC)

- Recall AIC: $\text{elpd}_{AIC} = \log p(y|\hat{\theta}_{mle}) - k$
- BIC subtracts $\frac{k}{2} \log n$ instead of $k$
- Not an estimate of out-of-sample predictive accuracy
- Favors smaller models
“Bayesian” information criterion (BIC)

- Recall AIC: \( \hat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \)
- BIC subtracts \( \frac{k}{2} \log n \) instead of \( k \)
- Not an estimate of out-of-sample predictive accuracy
- Favors smaller models
“Bayesian” information criterion (BIC)

- Recall AIC: $\hat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k$
- BIC subtracts $k/2 \log n$ instead of $k$
- *Not* an estimate of out-of-sample predictive accuracy
- Favors smaller models
Cross-validation

For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:

- Fit model to training set, get posterior sims
- Compute log posterior predictive density of $y_{\text{holdout}}$
- Average over simulations to get $\hat{\text{elpd}}_{\text{xval}}$

AIC, DIC, WAIC, and LOO-CV for election forecasting example
For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:

- Fit model to training set, get posterior sims
- Compute log posterior predictive density of $y_{\text{holdout}}$
- Average over simulations to get $\hat{\text{elpdp}}_{\text{xval}}$
- AIC, DIC, WAIC, and LOO-CV for election forecasting example
Cross-validation

- For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:
  - Fit model to training set, get posterior sims
  - Compute log posterior predictive density of $y_{\text{holdout}}$
  - Average over simulations to get $\hat{\text{elpd}}_{\text{xval}}$
- AIC, DIC, WAIC, and LOO-CV for election forecasting example
Cross-validation

- For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:
  - Fit model to training set, get posterior sims
  - Compute log posterior predictive density of $y_{\text{holdout}}$
  - Average over simulations to get $\hat{\text{elpd}}_{\text{xval}}$
  - AIC, DIC, WAIC, and LOO-CV for election forecasting example
Cross-validation

For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:

- Fit model to training set, get posterior sims
- Compute log posterior predictive density of $y_{\text{holdout}}$
- Average over simulations to get $\hat{\text{elpd}}_{x\text{val}}$
- AIC, DIC, WAIC, and LOO-CV for election forecasting example
Cross-validation

- For many partitions of the data into $y_{\text{train}}$ and $y_{\text{holdout}}$:
  - Fit model to training set, get posterior sims
  - Compute log posterior predictive density of $y_{\text{holdout}}$
- Average over simulations to get $\hat{\text{elpd}}_{\text{xval}}$
- AIC, DIC, WAIC, and LOO-CV for election forecasting example
Using log predictive probability and information criteria to compare three models for the 8 schools:

- No pooling ($\tau = \infty$)
- Complete pooling ($\tau = 0$)
- Hierarchical (average over $\tau$)
7.3. Model comparison based on predictive performance

- Using log predictive probability and information criteria to compare three models for the 8 schools:
  - No pooling ($\tau = \infty$)
  - Complete pooling ($\tau = 0$)
  - Hierarchical (average over $\tau$)
7.3. Model comparison based on predictive performance

- Using log predictive probability and information criteria to compare three models for the 8 schools:
  - No pooling ($\tau = \infty$)
  - Complete pooling ($\tau = 0$)
  - Hierarchical (average over $\tau$)
7.3. Model comparison based on predictive performance

- Using log predictive probability and information criteria to compare three models for the 8 schools:
  - No pooling ($\tau = \infty$)
  - Complete pooling ($\tau = 0$)
  - Hierarchical (average over $\tau$)
7.3. Model comparison based on predictive performance

- Using log predictive probability and information criteria to compare three models for the 8 schools:
  - No pooling ($\tau = \infty$)
  - Complete pooling ($\tau = 0$)
  - Hierarchical (average over $\tau$)
Comparing three models for the 8 schools

Discuss:

<table>
<thead>
<tr>
<th></th>
<th>No pooling ($\tau = \infty$)</th>
<th>Complete pooling ($\tau = 0$)</th>
<th>Hier. model ($\tau$ est’d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AIC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-2 \text{lpd} = -2 \log p(y</td>
<td>\hat{\theta}_{mle})$</td>
<td>54.6</td>
<td>59.4</td>
</tr>
<tr>
<td>$k$</td>
<td>8.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>AIC = $-2 \text{elpd}_{AIC}$</td>
<td>70.6</td>
<td>61.4</td>
<td></td>
</tr>
<tr>
<td><strong>DIC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-2 \text{lpd} = -2 \log p(y</td>
<td>\hat{\theta}_{Bayes})$</td>
<td>54.6</td>
<td>59.4</td>
</tr>
<tr>
<td>$p_{DIC}$</td>
<td>8.0</td>
<td>1.0</td>
<td>2.8</td>
</tr>
<tr>
<td>DIC = $-2 \text{elpd}_{DIC}$</td>
<td>70.6</td>
<td>61.4</td>
<td>63.0</td>
</tr>
<tr>
<td><strong>WAIC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-2 \text{lppd} = -2 \sum_i \log p_{post}(y_i)$</td>
<td>60.2</td>
<td>59.8</td>
<td>59.2</td>
</tr>
<tr>
<td>$p_{WAIC 1}$</td>
<td>2.5</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>$p_{WAIC 2}$</td>
<td>4.0</td>
<td>0.7</td>
<td>1.3</td>
</tr>
<tr>
<td>WAIC = $-2 \text{elppd}_{WAIC 2}$</td>
<td>68.2</td>
<td>61.2</td>
<td>61.8</td>
</tr>
<tr>
<td><strong>LOO-CV</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-2 \text{lppd}$</td>
<td></td>
<td>59.8</td>
<td>59.2</td>
</tr>
<tr>
<td>$p_{\text{loo-cv}}$</td>
<td></td>
<td>0.5</td>
<td>1.8</td>
</tr>
<tr>
<td>$-2 \text{lppd}_{\text{loo-cv}}$</td>
<td>60.8</td>
<td>62.8</td>
<td></td>
</tr>
</tbody>
</table>
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:
  \[ \frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)} \]

  - \( \frac{p(H_2)}{p(H_1)} \) is "prior odds"
  - \( \frac{p(y|H_2)}{p(y|H_1)} \) is "Bayes factor"

- Problem with \( p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta \)
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)}
\]

- \( \frac{p(H_2)}{p(H_1)} \) is “prior odds”
- \( \frac{p(y|H_2)}{p(y|H_1)} \) is “Bayes factor”
- Problem with \( p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta \)
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)}
\]

- \(\frac{p(H_2)}{p(H_1)}\) is “prior odds”
- \(\frac{p(y|H_2)}{p(y|H_1)}\) is “Bayes factor”

- Problem with \(p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta\)
  - Integral depends on irrelevant tail properties of the prior density.

Consider simple example: \(y \sim N(\theta, \sigma^2/n)\)
\(p(\theta) \propto U(-A, A)\), for some large \(A\).
Marginal \(p(y)\) is proportional to \(\frac{1}{A}\).
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2) \cdot p(y|H_2)}{p(H_1) \cdot p(y|H_1)}
\]

- \( \frac{p(H_2)}{p(H_1)} \) is “prior odds”
- \( \frac{p(y|H_2)}{p(y|H_1)} \) is “Bayes factor”

- Problem with \( p(y|H) = \int p(\theta|H)p(y|\theta,H)d\theta \)
  - Integral depends on irrelevant tail properties of the prior density
  - Consider simple example: \( y \sim \text{N}(\theta, \sigma^2/n) \)
  - \( p(\theta) \propto \text{U}(-A, A) \), for some large \( A \)
  - Marginal \( p(y) \) is proportional to \( \frac{1}{A} \)
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2) \ p(y|H_2)}{p(H_1) \ p(y|H_1)}
\]

- \(\frac{p(H_2)}{p(H_1)}\) is “prior odds”
- \(\frac{p(y|H_2)}{p(y|H_1)}\) is “Bayes factor”

- Problem with \(p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta\)
  - Integral depends on irrelevant tail properties of the prior density
  - Consider simple example: \(\bar{y} \sim N(\theta, \sigma^2/n)\)
  - \(p(\theta) \propto U(-A, A)\), for some large \(A\)
  - Marginal \(p(y)\) is proportional to \(\frac{1}{A}\)
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)}
\]

- \( \frac{p(H_2)}{p(H_1)} \) is “prior odds”
- \( \frac{p(y|H_2)}{p(y|H_1)} \) is “Bayes factor”

- Problem with \( p(y|H) = \int p(\theta|H)p(y|\theta, H)\,d\theta \)
  - Integral depends on irrelevant tail properties of the prior density
  - Consider simple example: \( \bar{y} \sim N(\theta, \sigma^2/n) \)
  - \( p(\theta) \propto U(-A, A) \), for some large \( A \)
  - Marginal \( p(y) \) is proportional to \( \frac{1}{A} \)
7.4. Model comparison using Bayes factors

- Comparing (or averaging over) two or more models:
  \[
  \frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)}
  \]

  - \(\frac{p(H_2)}{p(H_1)}\) is “prior odds”
  - \(\frac{p(y|H_2)}{p(y|H_1)}\) is “Bayes factor”

- Problem with \(p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta\)
  - Integral depends on irrelevant tail properties of the prior density
  - Consider simple example: \(\bar{y} \sim N(\theta, \sigma^2/n)\)
    - \(p(\theta) \propto U(-A, A)\), for some large \(A\)
    - Marginal \(p(y)\) is proportional to \(\frac{1}{A}\)
7.4. Model comparison using Bayes factors

Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2) p(y|H_2)}{p(H_1) p(y|H_1)}
\]

- \(\frac{p(H_2)}{p(H_1)}\) is “prior odds"
- \(\frac{p(y|H_2)}{p(y|H_1)}\) is “Bayes factor”

Problem with \(p(y|H) = \int p(\theta|H)p(y|\theta, H)d\theta\):

- Integral depends on irrelevant tail properties of the prior density
- Consider simple example: \(\bar{y} \sim N(\theta, \sigma^2/n)\)
- \(p(\theta) \propto U(-A, A)\), for some large \(A\)
- Marginal \(p(y)\) is proportional to \(\frac{1}{A}\)
7.4. Model comparison using Bayes factors

Comparing (or averaging over) two or more models:

\[
\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \frac{p(y|H_2)}{p(y|H_1)}
\]

▶ \( \frac{p(H_2)}{p(H_1)} \) is “prior odds”
▶ \( \frac{p(y|H_2)}{p(y|H_1)} \) is “Bayes factor”

Problem with \( p(y|H) = \int p(\theta|H) p(y|\theta, H) d\theta \)
▶ Integral depends on irrelevant tail properties of the prior density
▶ Consider simple example: \( \bar{y} \sim \text{N}(\theta, \sigma^2/n) \)
▶ \( p(\theta) \propto \text{U}(-A, A) \), for some large \( A \)
▶ Marginal \( p(y) \) is proportional to \( \frac{1}{A} \)
An example where the Bayes factor is good

Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene

What makes the Bayes factor work here?
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene
  - What makes the Bayes factor work here?
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene
- What makes the Bayes factor work here?
  - Truly discrete parameter space
  - Model of probabilities; no unbounded parameters
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene

- What makes the Bayes factor work here?
  - Truly discrete parameter space
  - Model of probabilities; no unbounded parameters
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene

- What makes the Bayes factor work here?
  - Truly discrete parameter space
  - Model of probabilities; no unbounded parameters
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene

- What makes the Bayes factor work here?
  - Truly discrete parameter space
  - Model of probabilities; no unbounded parameters
An example where the Bayes factor is good

- Genetics example
  - $H_1$: a woman carries a certain gene
  - $H_2$: she does not have the gene

- What makes the Bayes factor work here?
  - Truly discrete parameter space
  - Model of probabilities; no unbounded parameters
An example where the Bayes factor is bad

- 8 schools example: $y_j \sim N(\theta_j, \sigma_j^2)$, for $j = 1, \ldots, 8$
  - $H_1$: no pooling, $p(\theta_1, \ldots, \theta_8) \propto 1$
  - $H_2$: complete pooling, $\theta_1 = \cdots = \theta_j = \theta$, $p(\theta) \propto 1$
- Bayes factor is 0/0
- Instead, express flat priors as $N(0, A^2)$ and let $A$ get large
8 schools example: \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)

- \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
- \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_8 = \theta, p(\theta) \propto 1 \)

Bayes factor is 0/0

Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
An example where the Bayes factor is bad

- **8 schools example:** \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)
  - \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
  - \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_8 = \theta, p(\theta) \propto 1 \)
- Bayes factor is 0/0
- Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
An example where the Bayes factor is bad

- 8 schools example: $y_j \sim N(\theta_j, \sigma_j^2)$, for $j = 1, \ldots, 8$
  - $H_1$: no pooling, $p(\theta_1, \ldots, \theta_8) \propto 1$
  - $H_2$: complete pooling, $\theta_1 = \cdots = \theta_J = \theta$, $p(\theta) \propto 1$

- Bayes factor is 0/0
- Instead, express flat priors as $N(0, A^2)$ and let $A$ get large
  - Now Bayes factor strongly depends on $A$
    - As $A \to \infty$, complete pooling model gets 100% of the probability
      - for any data!
    - Also a horrible dependence on $J$
An example where the Bayes factor is bad

- **8 schools example:** $y_j \sim N(\theta_j, \sigma_j^2)$, for $j = 1, \ldots, 8$
  - $H_1$: no pooling, $p(\theta_1, \ldots, \theta_8) \propto 1$
  - $H_2$: complete pooling, $\theta_1 = \cdots = \theta_J = \theta$, $p(\theta) \propto 1$
- Bayes factor is 0/0
- Instead, express flat priors as $N(0, A^2)$ and let $A$ get large
  - Now Bayes factor strongly depends on $A$
  - As $A \to \infty$, complete pooling model gets 100% of the probability
  - ... for any data!
  - Also a horrible dependence on $J$
An example where the Bayes factor is bad

- 8 schools example: \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)
  - \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
  - \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_J = \theta \), \( p(\theta) \propto 1 \)
- Bayes factor is 0/0
- Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
  - Now Bayes factor strongly depends on \( A \)
  - As \( A \to \infty \), complete pooling model gets 100% of the probability . . .
  - . . . for any data!
  - Also a horrible dependence on \( J \)
An example where the Bayes factor is bad

- 8 schools example: \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)
  - \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
  - \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_J = \theta \), \( p(\theta) \propto 1 \)
- Bayes factor is 0/0
- Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
  - Now Bayes factor strongly depends on \( A \)
  - As \( A \to \infty \), complete pooling model gets 100% of the probability . . .
  - . . . for any data!
  - Also a horrible dependence on \( J \)
An example where the Bayes factor is bad

- **8 schools example**: $y_j \sim N(\theta_j, \sigma_j^2)$, for $j = 1, \ldots, 8$
  - $H_1$: no pooling, $p(\theta_1, \ldots, \theta_8) \propto 1$
  - $H_2$: complete pooling, $\theta_1 = \cdots = \theta_J = \theta$, $p(\theta) \propto 1$

- Bayes factor is 0/0

- Instead, express flat priors as $N(0, A^2)$ and let $A$ get large
  - Now Bayes factor strongly depends on $A$
  - As $A \to \infty$, complete pooling model gets 100% of the probability . . .
  - . . . for any data!
  - Also a horrible dependence on $J$
An example where the Bayes factor is bad

- 8 schools example: \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)
  - \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
  - \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_J = \theta \), \( p(\theta) \propto 1 \)

- Bayes factor is 0/0
- Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
  - Now Bayes factor strongly depends on \( A \)
  - As \( A \to \infty \), complete pooling model gets 100% of the probability . . .
  - . . . for any data!
  - Also a horrible dependence on \( J \)
An example where the Bayes factor is bad

- 8 schools example: \( y_j \sim N(\theta_j, \sigma_j^2) \), for \( j = 1, \ldots, 8 \)
  - \( H_1 \): no pooling, \( p(\theta_1, \ldots, \theta_8) \propto 1 \)
  - \( H_2 \): complete pooling, \( \theta_1 = \cdots = \theta_J = \theta \), \( p(\theta) \propto 1 \)
- Bayes factor is 0/0
- Instead, express flat priors as \( N(0, A^2) \) and let \( A \) get large
  - Now Bayes factor strongly depends on \( A \)
  - As \( A \to \infty \), complete pooling model gets 100\% of the probability . . .
  - . . . for any data!
  - Also a horrible dependence on \( J \)
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples

- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or t
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a *continuous* model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or t
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or $t$
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or $t$
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a *continuous* model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
    - Normal or $t$
    - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or $t$
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a *continuous* model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or $t$
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
7.5. Continuous model expansion

- 8-schools example: build a bridge between $H_1$ and $H_2$
- Embed original choices in a continuous model
- More examples
  - A regression including one set of predictors or another
  - Constant variance or varying variance
  - Normal or $t$
  - Spline or Gaussian process
- Can be hard to construct a joint prior distribution over the larger space
### 7.6. Implicit assumptions and model expansion

<table>
<thead>
<tr>
<th></th>
<th>Population $(N = 804)$</th>
<th>Sample 1 $(n = 100)$</th>
<th>Sample 2 $(n = 100)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>13,776,663</td>
<td>1,966,745</td>
<td>3,850,502</td>
</tr>
<tr>
<td>mean</td>
<td>17,135</td>
<td>19,667</td>
<td>38,505</td>
</tr>
<tr>
<td>sd</td>
<td>139,147</td>
<td>142,218</td>
<td>228,625</td>
</tr>
<tr>
<td>lowest</td>
<td>19</td>
<td>164</td>
<td>162</td>
</tr>
<tr>
<td>5%</td>
<td>336</td>
<td>308</td>
<td>315</td>
</tr>
<tr>
<td>25%</td>
<td>800</td>
<td>891</td>
<td>863</td>
</tr>
<tr>
<td>median</td>
<td>1,668</td>
<td>2,081</td>
<td>1,740</td>
</tr>
<tr>
<td>75%</td>
<td>5,050</td>
<td>6,049</td>
<td>5,239</td>
</tr>
<tr>
<td>95%</td>
<td>30,295</td>
<td>25,130</td>
<td>41,718</td>
</tr>
<tr>
<td>highest</td>
<td>2,627,319</td>
<td>1,424,815</td>
<td>1,809,578</td>
</tr>
</tbody>
</table>
Analysis of sample 1

- $y_{total} = N\bar{y} = n\bar{y}_{obs} + (N-n)\bar{y}_{mis}$
  - $\bar{y}_{obs}$ is known
  - Need inference for $\bar{y}_{mis}$
- Classical 95% interval: $[-5.4 \times 10^6, 37.0 \times 10^6]$
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
Analysis of sample 1

- $y_{total} = N\bar{y} = n\bar{y}_{obs} + (N - n)\bar{y}_{mis}$
  - $\bar{y}_{obs}$ is known
  - Need inference for $y_{mis}$

- Classical 95% interval: $[-5.4 \times 10^6, 37.0 \times 10^6]$

- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
Analysis of sample 1

- $y_{total} = N \bar{y} = n \bar{y}_{obs} + (N - n) \bar{y}_{mis}$
  - $\bar{y}_{obs}$ is known
  - Need inference for $y_{mis}$

- Classical 95% interval: $[-5.4 \times 10^6, 37.0 \times 10^6]$

- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
Analysis of sample 1

\[ y_{total} = N\bar{y} = n\bar{y}_{obs} + (N - n)\bar{y}_{mis} \]

- \( \bar{y}_{obs} \) is known
- Need inference for \( y_{mis} \)

- Classical 95\% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\]

- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\]
Analysis of sample 1

- \( y_{\text{total}} = N\bar{y} = n\bar{y}_{\text{obs}} + (N - n)\bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)
- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
- But the model doesn't fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
Analysis of sample 1

- \( y_{total} = N \bar{y} = n \bar{y}_{obs} + (N - n) \bar{y}_{mis} \)
  - \( \bar{y}_{obs} \) is known
  - Need inference for \( y_{mis} \)

- Classical 95\% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data

- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)

- Posterior predictive check using sample total:
  \( T(y_{obs}) = \sum_{i=1}^{N} y_{obs,i} \)
  Observed \( T(y_{obs}) = 1,966,745 \)
  100 replications: \( T(y_{rep}^{(i)}) \) are all lower than this value
  Model doesn’t fit!
Analysis of sample 1

- \( y_{\text{total}} = N\bar{y} = n\bar{y}_{\text{obs}} + (N - n)\bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)
- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Posterior predictive check using sample total, \( T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs},i} \)
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{rep},i}) \) are all lower than this value
  - Model doesn’t fit!
Analysis of sample 1

- \( y_{\text{total}} = N \bar{y} = n \bar{y}_{\text{obs}} + (N - n) \bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)
- Classical 95% interval: \( [−5.4 \times 10^6, 37.0 \times 10^6] \)
  - But the model doesn’t fit the data
- Lognormal, Bayes inference: \( [5.4 \times 10^6, 9.9 \times 10^6] \)
  - Posterior predictive check using sample total,
    \( T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs}i} \)
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{obs}}^{\text{rep}}) \) are all lower than this value
  - Model doesn’t fit!
Analysis of sample 1

\[ y_{\text{total}} = N\bar{y} = n\bar{y}_{\text{obs}} + (N - n)\bar{y}_{\text{mis}} \]

- \( \bar{y}_{\text{obs}} \) is known
- Need inference for \( y_{\text{mis}} \)

- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Posterior predictive check using sample total,
    \[ T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs}i} \]
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{obs}}^{\text{rep}}) \) are \textit{all} lower than this value
  - Model doesn’t fit!
Analysis of sample 1

- \( y_{\text{total}} = N\bar{y} = n\bar{y}_{\text{obs}} + (N - n)\bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)

- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data

- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Posterior predictive check using sample total,
    \( T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs}i} \)
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{obs}}^{\text{rep}}) \) are all lower than this value
  - Model doesn’t fit!

Andrew Gelman
Bayesian Data Analysis, class 6b
Analysis of sample 1

- \( y_{\text{total}} = N\bar{y} = n\bar{y}_{\text{obs}} + (N - n)\bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)
- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Posterior predictive check using sample total,
    \( T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs}i} \)
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{obs}}^{\text{rep}}) \) are all lower than this value
  - Model doesn’t fit!
Analysis of sample 1

- \( y_{\text{total}} = N \bar{y} = n \bar{y}_{\text{obs}} + (N - n) \bar{y}_{\text{mis}} \)
  - \( \bar{y}_{\text{obs}} \) is known
  - Need inference for \( y_{\text{mis}} \)
- Classical 95% interval: \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - But the model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Posterior predictive check using sample total, \( T(y_{\text{obs}}) = \sum_{i=1}^{n} y_{\text{obs} i} \)
  - Observed \( T(y_{\text{obs}}) = 1,966,745 \)
  - 100 replications, \( T(y_{\text{obs}}^{\text{rep}}) \) are all lower than this value
  - Model doesn’t fit!
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)
  - 95% posterior interval: $[5.18 \times 10^6, 31.8 \times 10^6]$
  - Posterior predictive check: $T(y_{rep}) > T(y_{obs})$ in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): \([-5.4 \times 10^6, 37.0 \times 10^6]\]
  - Model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\]
  - Model doesn’t fit
- Power transformation (normal model for \(y^\phi\), estimate \(\phi\) from data)
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
  - Model doesn’t fit
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)
Power-transformation model

- Classical interval (normal model): \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - Model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Model doesn’t fit
- Power transformation (normal model for \(y^\phi\), estimate \(\phi\) from data)
  - Best fit, \(\phi = -1/4\)
  - 95% posterior interval: \([5.8 \times 10^6, 31.8 \times 10^6]\)
- Posterior predictive check: \(T(y_{rep}) > T(y_{obs})\) in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
  - Model doesn’t fit
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)
  - Best fit, $\phi = -1/4$
  - 95% posterior interval: $[5.8 \times 10^6, 31.8 \times 10^6]$
  - Posterior predictive check: $T(y_{rep}) > T(y_{obs})$ in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
  - Model doesn’t fit
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)
  - Best fit, $\phi = -1/4$
  - 95% posterior interval: $[5.8 \times 10^6, 31.8 \times 10^6]$
  - Posterior predictive check: $T(y_{\text{rep}}^{\text{obs}}) > T(y_{\text{obs}})$ in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$
  - Model doesn’t fit
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)
  - Best fit, $\phi = -1/4$
  - 95% posterior interval: $[5.8 \times 10^6, 31.8 \times 10^6]$
  - Posterior predictive check: $T(y_{rep}^{obs}) > T(y_{obs})$ in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - Model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Model doesn’t fit
- Power transformation (normal model for \(y^\phi\), estimate \(\phi\) from data)
  - Best fit, \(\phi = -1/4\)
  - 95% posterior interval: \([5.8 \times 10^6, 31.8 \times 10^6]\)
  - Posterior predictive check: \(T(y_{\text{rep}}^{\text{obs}}) > T(y_{\text{obs}})\) in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): $[-5.4 \times 10^6, 37.0 \times 10^6]$  
  - Model doesn’t fit the data
- Lognormal, Bayes inference: $[5.4 \times 10^6, 9.9 \times 10^6]$  
  - Model doesn’t fit
- Power transformation (normal model for $y^\phi$, estimate $\phi$ from data)  
  - Best fit, $\phi = -1/4$
  - 95% posterior interval: $[5.8 \times 10^6, 31.8 \times 10^6]$
  - Posterior predictive check: $T(y_{\text{obs}}^{\text{rep}}) > T(y_{\text{obs}})$ in 15 of 100 replications
  - Success!
Power-transformation model

- Classical interval (normal model): \([-5.4 \times 10^6, 37.0 \times 10^6]\)
  - Model doesn’t fit the data
- Lognormal, Bayes inference: \([5.4 \times 10^6, 9.9 \times 10^6]\)
  - Model doesn’t fit
- Power transformation (normal model for \(y^\phi\), estimate \(\phi\) from data)
  - Best fit, \(\phi = -1/4\)
  - 95% posterior interval: \([5.8 \times 10^6, 31.8 \times 10^6]\)
  - Posterior predictive check: \(T(y_{\text{rep}}^\text{obs}) > T(y_{\text{obs}})\) in 15 of 100 replications
  - Success!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$ 
  - Model doesn’t fit the data

- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$ 
  - Model doesn’t fit the data

- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$

- But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - Model seems to fit
- But there's a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - Model seems to fit
- But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - Model seems to fit
- But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - Model seems to fit
- But there’s a big problem!
Analysis of sample 2

- Classical interval (normal model): $[-3.4 \times 10^6, 65.3 \times 10^6]$
  - Model doesn’t fit the data
- Lognormal fit to data, Bayes inference: $[8.2 \times 10^6, 19.6 \times 10^6]$
  - Model doesn’t fit the data
- Power transformation (best fit, $\phi = -1/4$): $[10^7, 10^{15}]$
  - Model seems to fit
- But there’s a big problem!
Why did the classical interval work sort of ok?
Why did the lognormal model not work?
Why did the power-transformed model blow up—without getting caught by the model check?
How could we do better?

Statistics is the science of defaults
Andrew Gelman
Why did the classical interval work sort of ok?

Why did the lognormal model not work?

Why did the power-transformed model blow up—without getting caught by the model check?

How could we do better?
Why did the classical interval work sort of ok?

Why did the lognormal model not work?

Why did the power-transformed model blow up—without getting caught by the model check?

How could we do better?

Statistics is the science of defaults.
Why did the classical interval work sort of ok?
Why did the lognormal model not work?
Why did the power-transformed model blow up—without getting caught by the model check?
How could we do better?

Statistics is the science of defaults

Andrew Gelman
Bayesian Data Analysis, class 6b
Why did the classical interval work sort of ok?

Why did the lognormal model not work?

Why did the power-transformed model blow up—without getting caught by the model check?

How could we do better?

Statistics is the science of defaults
Why did the classical interval work sort of ok?
Why did the lognormal model not work?
Why did the power-transformed model blow up—without getting caught by the model check?
How could we do better?
  Statistics is the science of defaults
Summary of Chapter 7

- Predictive accuracy, information criteria, and effective number of parameters
- Bayes factors and discrete model averaging
- Continuous model expansion
- Implicit assumptions in statistical procedures
Summary of Chapter 7

- Predictive accuracy, information criteria, and effective number of parameters
  - Bayes factors and discrete model averaging
  - Continuous model expansion
  - Implicit assumptions in statistical procedures
Summary of Chapter 7

- Predictive accuracy, information criteria, and effective number of parameters
- Bayes factors and discrete model averaging
  - Continuous model expansion
  - Implicit assumptions in statistical procedures
Summary of Chapter 7

- Predictive accuracy, information criteria, and effective number of parameters
- Bayes factors and discrete model averaging
- Continuous model expansion
  - Implicit assumptions in statistical procedures
Summary of Chapter 7

- Predictive accuracy, information criteria, and effective number of parameters
- Bayes factors and discrete model averaging
- Continuous model expansion
- Implicit assumptions in statistical procedures
Homework due beginning of class 7b

- All assignments are at http://www.stat.columbia.edu/~gelman/bda.course/homeworks.pdf
  - Theory problem: Bayes factors
  - Computing problem: Stan
  - Applied problem: clustered data
Homework due beginning of class 7b

- All assignments are at http://www.stat.columbia.edu/~gelman/bda.course/homeworks.pdf
  - Theory problem: Bayes factors
  - Computing problem: Stan
  - Applied problem: clustered data
All assignments are at http://www.stat.columbia.edu/~gelman/bda.course/homeworks.pdf

- Theory problem: Bayes factors
- Computing problem: Stan
- Applied problem: clustered data
Homework due beginning of class 7b

- All assignments are at http://www.stat.columbia.edu/~gelman/bda.course/homeworks.pdf
  - Theory problem: Bayes factors
  - Computing problem: Stan
  - Applied problem: clustered data
Homework due beginning of class 7b

- All assignments are at http://www.stat.columbia.edu/~gelman/bda.course/homeworks.pdf
  - Theory problem: Bayes factors
  - Computing problem: Stan
  - Applied problem: clustered data