STAT 431 — Applied Bayesian Analysis — Course Notes

Topics in Model Comparison and Assessment

Spring 2019

Bayes Factors

Recall: Bayesians don't use frequentist p-values.

What can we use instead?

Earlier, we used posterior probabilities of one-sided hypotheses (some of which happened to equal p-values in some noninformative cases).

But we would like a more general approach ...

Simple-vs-Simple Case

Consider data y which may follow one of two different models,

$$M_0 \quad \text{and} \quad M_1$$

Assume each of these models fully specifies a distribution for y, and the distributions have densities

$$p(y \mid M_0) \quad \text{and} \quad p(y \mid M_1)$$

Then

$$H_0 : \ M_0 \text{ is true} \qquad\qquad H_1 : \ M_1 \text{ is true}$$

are two **simple** hypotheses.

Let the models have prior probabilities

$$\mathrm{P}(M_0) \quad (= \mathrm{P}(H_0)) \qquad\qquad \mathrm{P}(M_1) \quad (= \mathrm{P}(H_1))$$

We assume

$$\mathrm{P}(M_0) \; > \; 0 \qquad \text{and} \qquad \mathrm{P}(M_1) \; > \; 0$$

but it is not necessary that they sum to 1.

The **prior odds in favor of** M_1 are

$$\frac{\mathrm{P}(M_1)}{\mathrm{P}(M_0)}$$

By Bayes' rule,

$$\begin{aligned} \mathrm{P}(M_0 \mid y) &\propto \mathrm{P}(M_0)\, p(y \mid M_0) \\ \mathrm{P}(M_1 \mid y) &\propto \mathrm{P}(M_1)\, p(y \mid M_1) \end{aligned}$$

with the same normalizing constant ($p(y)$) in both cases.

The **posterior odds in favor of** M_1 are

$$\begin{aligned} \frac{\mathrm{P}(M_1 \mid y)}{\mathrm{P}(M_0 \mid y)} &= \frac{\mathrm{P}(M_1)\, p(y \mid M_1)}{\mathrm{P}(M_0)\, p(y \mid M_0)} \\ &= \text{prior odds} \times \frac{p(y \mid M_1)}{p(y \mid M_0)} \end{aligned}$$

The **Bayes factor in favor of M_1 versus M_0** is

$$BF_{1,0} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p(y \mid M_1)}{p(y \mid M_0)}$$

Interpretation: $BF_{1,0}$ is the factor by which the "odds" of M_1 (relative to M_0) change due to the data.

The **Bayes factor in favor of M_1 versus M_0** is

$$BF_{1,0} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p(y \mid M_1)}{p(y \mid M_0)}$$

Interpretation: $BF_{1,0}$ is the factor by which the "odds" of M_1 (relative to M_0) change due to the data.

So, for example,

- $BF_{1,0} \approx 1$ means that the data do not distinguish between the models very well
- $BF_{1,0} \gg 1$ means that the data strongly support M_1 over M_0

Notice: In this simple-vs-simple case, the Bayes factor $BF_{1,0}$

- equals the likelihood ratio

$$\frac{L(M_1\,;\,y)}{L(M_0\,;\,y)}$$

- does not depend on the prior — it is the same for any valid values of $\mathrm{P}(M_0)$ and $\mathrm{P}(M_1)$

Example: Does Waldo ride the bus?

$$M_1 = \text{does ride} \qquad M_0 = \text{doesn't ride}$$

$$y = \begin{cases} 1 & \text{if lives in an apartment} \\ 0 & \text{if not} \end{cases}$$

Based on the class survey (our best guess),

$$p(y \mid M_1) = \begin{cases} 0.91, & y = 1 \\ 0.09, & y = 0 \end{cases}$$

$$p(y \mid M_0) = \begin{cases} 0.81, & y = 1 \\ 0.19, & y = 0 \end{cases}$$

If Waldo lives in an apartment $(y = 1)$,

$$BF_{1,0} = \frac{0.91}{0.81} \approx 1.12$$

and, if not $(y = 0)$,

$$BF_{1,0} = \frac{0.09}{0.19} \approx 0.47$$

So living in an apartment increases the odds of riding the bus, while not living in an apartment decreases them.

An Interpretation Scale

$BF_{1,0}$	data evidence for M_1 (H_1) vs. M_0 (H_0)
1 to 3.2	Barely worth mentioning
3.2 to 10	Substantial
10 to 100	Strong
> 100	Decisive

More General Case

Consider modeling data $\boldsymbol{y}$.

Suppose models M_0 and M_1 have (unknown) parameters:

$$\theta_0 \quad \text{for} \quad M_0 \qquad\qquad \theta_1 \quad \text{for} \quad M_1$$

We will assume the models are "disjoint": They don't share any distributions for $\boldsymbol{y}$.

Suppose the models have (conditional) priors

$$p(\theta_0 \mid M_0) \qquad\qquad p(\theta_1 \mid M_1)$$

Then

$$\begin{aligned} p(\boldsymbol{y} \mid M_0) &= \int p(\boldsymbol{y}, \theta_0 \mid M_0)\, d\theta_0 \\ &= \int \underbrace{p(\theta_0 \mid M_0)}_{\text{prior}}\ \underbrace{p(\boldsymbol{y} \mid \theta_0, M_0)}_{M_0 \text{ model}}\, d\theta_0 \end{aligned}$$

Suppose the models have (conditional) priors

$$p(\theta_0 \mid M_0) \qquad\qquad p(\theta_1 \mid M_1)$$

Then

$$\begin{aligned} p(\boldsymbol{y} \mid M_0) &= \int p(\boldsymbol{y}, \theta_0 \mid M_0)\, d\theta_0 \\ &= \int \underbrace{p(\theta_0 \mid M_0)}_{\text{prior}}\ \underbrace{p(\boldsymbol{y} \mid \theta_0, M_0)}_{M_0 \text{ model}}\, d\theta_0 \end{aligned}$$

and similarly

$$p(\boldsymbol{y} \mid M_1) = \int p(\theta_1 \mid M_1)\ p(\boldsymbol{y} \mid \theta_1, M_1)\, d\theta_1$$

These are the **marginal likelihoods** of M_0 and M_1 (under their respective priors).

The **Bayes factor in favor of M_1 versus M_0** is

$$BF_{1,0} = \frac{p(\boldsymbol{y} \mid M_1)}{p(\boldsymbol{y} \mid M_0)}$$

Notes:

- Unlike in the simple-vs-simple case, this Bayes factor does depend on the priors — it is not purely a measure of the evidence in the data.
- Both priors must be proper — otherwise, the Bayes factor would depend on an arbitrary scaling.

Unfortunately, Bayes factors are generally difficult to compute, requiring specialized methods.

But they can be easily computed for certain types of hypothesis tests ...

For Hypothesis Testing

Consider a model with data densities $p(\boldsymbol{y} \mid \theta)$ and prior $p(\theta)$.

Consider testing

$$H_0 : \theta \in \Theta_0 \qquad\qquad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$, and both have positive prior probability.

Regard this as a test of two data model/prior combinations:

$$\begin{aligned} M_0 : \quad & p(\boldsymbol{y} \mid \theta), \quad \theta \in \Theta_0 \\ & \text{with prior } p(\theta) \textit{ restricted} \text{ to } \Theta_0 \\ M_1 : \quad & p(\boldsymbol{y} \mid \theta), \quad \theta \in \Theta_1 \\ & \text{with prior } p(\theta) \textit{ restricted} \text{ to } \Theta_1 \end{aligned}$$

Proposition

In this case, the Bayes factor in favor of M_1 versus M_0 is

$$BF_{1,0} = \frac{\mathrm{P}(H_1 \mid \boldsymbol{y}) \,/\, \mathrm{P}(H_0 \mid \boldsymbol{y})}{\mathrm{P}(H_1) \,/\, \mathrm{P}(H_0)}$$

$$\left(= \frac{\textit{posterior odds}}{\textit{prior odds}} \right)$$

We call this the **Bayes factor in favor of H_1 (versus H_0)**.

Proof.

$$p(\boldsymbol{y} \mid M_1) = \int_{\Theta_1} \underbrace{\frac{p(\theta)}{\mathrm{P}(H_1)}}_{\substack{p(\theta) \text{ restricted} \\ \text{to } \Theta_1}} \underbrace{p(\boldsymbol{y} \mid \theta)}_{\substack{M_1 \text{ on} \\ \Theta_1}} \, d\theta$$

Proof.

$$p(\boldsymbol{y} \mid M_1) = \int_{\Theta_1} \underbrace{\frac{p(\theta)}{\mathrm{P}(H_1)}}_{\substack{p(\theta) \text{ restricted} \\ \text{to } \Theta_1}} \underbrace{p(\boldsymbol{y} \mid \theta)}_{\substack{M_1 \text{ on} \\ \Theta_1}} \, d\theta$$

$$= \frac{1}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta)\, p(\boldsymbol{y} \mid \theta)\, d\theta$$

Proof.

$$
\begin{aligned}
p(\boldsymbol{y} \mid M_1) &= \int_{\Theta_1} \underbrace{\frac{p(\theta)}{\mathrm{P}(H_1)}}_{\substack{p(\theta) \text{ restricted to } \Theta_1}} \underbrace{p(\boldsymbol{y} \mid \theta)}_{M_1 \text{ on } \Theta_1} \, d\theta \\
&= \frac{1}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta)\, p(\boldsymbol{y} \mid \theta) \, d\theta \\
&= \frac{p(\boldsymbol{y})}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta \mid \boldsymbol{y}) \, d\theta
\end{aligned}
$$

Proof.

$$
\begin{aligned}
p(\boldsymbol{y} \mid M_1) &= \int_{\Theta_1} \underbrace{\frac{p(\theta)}{\mathrm{P}(H_1)}}_{\substack{p(\theta) \text{ restricted to } \Theta_1}} \underbrace{p(\boldsymbol{y} \mid \theta)}_{M_1 \text{ on } \Theta_1} \, d\theta \\
&= \frac{1}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta)\, p(\boldsymbol{y} \mid \theta)\, d\theta \\
&= \frac{p(\boldsymbol{y})}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta \mid \boldsymbol{y})\, d\theta \quad = \quad p(\boldsymbol{y})\, \frac{\mathrm{P}(H_1 \mid \boldsymbol{y})}{\mathrm{P}(H_1)}
\end{aligned}
$$

Proof.

$$p(\boldsymbol{y} \mid M_1) = \int_{\Theta_1} \underbrace{\frac{p(\theta)}{\mathrm{P}(H_1)}}_{\substack{p(\theta) \text{ restricted} \\ \text{to } \Theta_1}} \underbrace{p(\boldsymbol{y} \mid \theta)}_{\substack{M_1 \text{ on} \\ \Theta_1}} \, d\theta$$

$$= \frac{1}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta)\, p(\boldsymbol{y} \mid \theta)\, d\theta$$

$$= \frac{p(\boldsymbol{y})}{\mathrm{P}(H_1)} \int_{\Theta_1} p(\theta \mid \boldsymbol{y})\, d\theta = p(\boldsymbol{y}) \frac{\mathrm{P}(H_1 \mid \boldsymbol{y})}{\mathrm{P}(H_1)}$$

and similarly

$$p(\boldsymbol{y} \mid M_0) = p(\boldsymbol{y}) \frac{\mathrm{P}(H_0 \mid \boldsymbol{y})}{\mathrm{P}(H_0)}$$

so the result follows by taking the ratio.

Example: Are shark attacks becoming more frequent?

Recall model for yearly number of attacks:

$$Y \mid \lambda \sim \text{Poisson}(\lambda)$$

$$\ln(\lambda) = \beta_0 + \beta_1 (\text{year} - \overline{\text{year}})$$

so we consider

$$H_0 : \beta_1 \leq 0 \quad \text{(no)} \qquad\qquad H_1 : \beta_1 > 0 \quad \text{(yes)}$$

Recall prior on β_1:

$$\beta_1 \sim \text{N}(0, 100^2)$$

So

$$\mathrm{P}(H_0) \;=\; 0.5 \qquad\qquad \mathrm{P}(H_1) \;=\; 0.5$$

From JAGS (Example 10.1):

$$\begin{aligned} \mathrm{P}(H_1 \mid \boldsymbol{y}) &\approx 0.99999 \\ \mathrm{P}(H_0 \mid \boldsymbol{y}) &\approx 1 - 0.99999 \;=\; 0.00001 \end{aligned}$$

So

$$BF_{1,0} \;\approx\; \frac{0.99999/0.00001}{0.5/0.5} \;=\; 99999$$

representing decisive evidence that shark attacks are becoming more frequent.

What about *point-null hypotheses*, like

$$H_0 : \beta_1 = 0$$

(perhaps corresponding to a two-sided test)?

Historically, Bayesians generally have considered this to be problematic: If β_1 is a continuous parameter (with a continuous prior), it should have posterior probability 0 of being any specified value.

See discussion in Cowles, Sec. 11.2.2.

The Deviance Information Criterion

You may have heard of model selection criteria like AIC or BIC ...

Several models for the *same* data $\boldsymbol{y}$ are under consideration — possibly nested, possibly intersecting, possibly unrelated. They can differ in type and number of parameters.

Model selection criteria aim to answer the question

Which is the "best" model for the data?

Bayesians want a criterion that evaluates the *prior*, too ...

Suppose you have several candidate models for the *same* data $\boldsymbol{y}$:

$$M_1, \quad M_2, \quad \ldots \quad M_m$$

Suppose each model has its own prior.

Goal: Choose the "best" model/prior combination for predicting new data (of the same kind).

Let M be a particular model, parameterized by θ (continuous), under which the data have a density

$$p(\boldsymbol{y} \mid \theta)$$

and let $p(\theta)$ be a prior density.

Let $p(\theta \mid \boldsymbol{y})$ be the resulting posterior density.

Define

$$D(\boldsymbol{y}, \theta) \;\; = \;\; -2 \, \ln p(\boldsymbol{y} \mid \theta)$$

(This is analytically computable, provided the density can be evaluated.)

Define

$$\begin{aligned} \hat{D}_{\mathsf{avg}}(\boldsymbol{y}) &= \int D(\boldsymbol{y}, \theta)\, p(\theta \mid \boldsymbol{y})\, d\theta \\ &= D \text{ averaged over the posterior} \end{aligned}$$

and

$$D_{\hat{\theta}}(\boldsymbol{y}) = D(\boldsymbol{y}, \hat{\theta})$$

where $\hat{\theta}$ is the posterior mean:

$$\hat{\theta} = \mathrm{E}(\theta \mid \boldsymbol{y})$$

Note: Both $\hat{D}_{\mathsf{avg}}(\boldsymbol{y})$ and $D_{\hat{\theta}}(\boldsymbol{y})$ can be approximated from simulation output.

Define

$$p_D \;=\; \hat{D}_{\mathsf{avg}}(\boldsymbol{y}) \;-\; D_{\hat{\theta}}(\boldsymbol{y})$$

as the **effective number of parameters**.

Note: This is *not* what a frequentist would call the "effective number of parameters." It is usually not an integer, and could possibly be negative!

The **deviance information criterion (DIC)** value (for model M with prior $p(\theta)$) is

$$\begin{aligned} DIC &= \hat{D}_{\text{avg}}(\boldsymbol{y}) + p_D \\ &= 2\,\hat{D}_{\text{avg}}(\boldsymbol{y}) - D_{\hat{\theta}}(\boldsymbol{y}) \\ &= D_{\hat{\theta}}(\boldsymbol{y}) + 2p_D \end{aligned}$$

Usage: Choose the model/prior with minimal DIC.

Remark: Motivation for DIC is similar to the frequentist-motivated AIC —

$$AIC \;=\; D(\boldsymbol{y}, \hat{\theta}_{\mathsf{MLE}}) \;+\; 2p$$

where p is the "true" (frequentist) number of free parameters (in θ), and $\hat{\theta}_{\mathsf{MLE}}$ is the maximum likelihood estimate.

- the first term penalizes lack of fit
- the second term penalizes complexity

Warning: Models compared using DIC (or AIC or BIC) must be for *exactly* the same data $\boldsymbol{y}$.

For example, if the data is transformed, exactly the same transformation must be used for all models being compared.

Remark: Software usually assumes that θ contains *all* parameters and hyperparameters (at all levels) of a hierarchical model.

Example: Baby Rat Weights

Model 1: Bivariate Formulation (Separate Lines)

$$\begin{aligned} Y_{ij} &= \text{weight (mass) of rat } i \text{ at time } x_j \\ &\sim \text{indep. } \mathrm{N}\big(\alpha_{0i} + \alpha_{1i}(x_j - \bar{x}),\, \sigma_y^2\big) \end{aligned}$$

$$\boldsymbol{\alpha}_i = \begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \;\Big|\; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\alpha \quad \sim \quad \text{indep. } \mathrm{N}_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\alpha)$$

(with appropriate priors on σ_y^2, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_\alpha$)

Model 2: Univariate Formulation, Separate Lines

$$Y_{ij} \sim \text{indep. } \mathrm{N}\big(\alpha_{0i} + \alpha_{1i}(x_j - \bar{x}),\, \sigma_y^2\big)$$

$$\left.\begin{aligned} \alpha_{0i} \mid \beta_0, \sigma^2_{\alpha_0} &\sim \mathrm{N}(\beta_0, \sigma^2_{\alpha_0}) \\ \alpha_{1i} \mid \beta_1, \sigma^2_{\alpha_1} &\sim \mathrm{N}(\beta_1, \sigma^2_{\alpha_1}) \end{aligned}\right\} \text{ all conditionally independent}$$

(with vague independent priors on σ_y^2, β_0, β_1, $\sigma^2_{\alpha_0}$, $\sigma^2_{\alpha_1}$)

Model 3: Separate Intercepts, Common Slope

$$Y_{ij} \sim \text{indep. } \mathrm{N}\big(\alpha_{0i} + \beta_1(x_j - \bar{x}),\, \sigma_y^2\big)$$

$$\alpha_{0i} \mid \beta_0, \sigma^2_{\alpha_0} \sim \text{indep. } \mathrm{N}(\beta_0, \sigma^2_{\alpha_0})$$

Model 4: Same Line (SLR)

$$Y_{ij} \sim \text{indep. } \mathrm{N}\big(\beta_0 + \beta_1(x_j - \bar{x}),\, \sigma_y^2\big)$$

R/JAGS Example 11.1:

DIC for Hierarchical Normal Regressions

Remark: JAGS uses a different version of p_D suggested by Plummer (in the discussion of Spiegelhalter et al., 2002).

Posterior Predictive Assessment

So far, we considered only ways to compare different model/prior combinations to each other.

How can we check whether a model/prior combination is a good fit to the data?

Frequentist approach: *lack-of-fit test* (such as a *chi-square test* — later)

Usually produces a p-value — smaller indicates more evidence against the model.

A Bayesian wants to assess the prior and model together.

For notation:

$$\boldsymbol{y} \;=\; \text{the data (vector)}$$

$$\boldsymbol{\theta} \;=\; \text{the parameter (vector) in model } M$$

We also need to choose a function called a **discrepancy**:

$$T(\boldsymbol{y}; \boldsymbol{\theta})$$

It is intended to measure how far observed data $\boldsymbol{y}$ depart from what would be expected under model M with parameter value $\boldsymbol{\theta}$. Larger values should indicate greater departures from the model.

An example of a discrepancy:

$$T(\boldsymbol{y}; \boldsymbol{\theta}) \;\; = \;\; \sum_{i=1}^{n} \frac{\big(y_i - \mathrm{E}(Y_i \mid \boldsymbol{\theta})\big)^2}{\mathrm{Var}(Y_i \mid \boldsymbol{\theta})}$$

where the observed data $\boldsymbol{y} = (y_1, \ldots y_n)$ is a numerical vector, and $(Y_1, \ldots Y_n)$ has its distribution under the model.

This is larger when the y_is are generally farther from their means than their variances would suggest, under the model with parameter value $\boldsymbol{\theta}$.

(Note: Similar in form to the classical *chi-square statistic*.)

A frequentist can't use $T(\boldsymbol{y};\boldsymbol{\theta})$ directly, since it depends on the unknown $\boldsymbol{\theta}$.

Replacing $\boldsymbol{\theta}$ with an estimate $\hat{\boldsymbol{\theta}}$ might yield a reasonable test statistic

$$\hat{T}(\boldsymbol{y}) \;=\; T(\boldsymbol{y};\hat{\boldsymbol{\theta}})$$

but its distribution under model M might still depend on the unknown $\boldsymbol{\theta}$.

If the distribution is (approximately) known, a frequentist can compute the (approximate) p-value

$$p \;=\; \mathrm{P}\big(\hat{T}(\boldsymbol{Y}^{\mathsf{rep}}) \geq \hat{T}(\boldsymbol{y})\big)$$

where $\boldsymbol{Y}^{\mathsf{rep}}$ is a *replication* of the data under the model.

In the earlier example:

$$\hat{T}(\boldsymbol{y}) \;=\; \sum_{i=1}^{n} \frac{\left(y_i - \mathrm{E}(Y_i \mid \hat{\boldsymbol{\theta}})\right)^2}{\mathrm{Var}(Y_i \mid \hat{\boldsymbol{\theta}})}$$

is the classical (Pearson) chi-square statistic (where $\hat{\boldsymbol{\theta}}$ is usually the MLE).

When the model is correct, asymptotic theory often suggests

$$\hat{T}(\boldsymbol{Y}) \mid \boldsymbol{\theta} \quad \dot{\sim} \quad \chi^2_{n-k}$$

if $\boldsymbol{\theta}$ effectively has k elements.

When this approximation is valid, an approximate p-value is the χ^2_{n-k} density tail area to the right of $\hat{T}(\boldsymbol{y})$.

In contrast, a Bayesian wants to

- assess the prior, not just the data model
- average over the distribution of $\boldsymbol{\theta}$ (to avoid substituting an estimate $\hat{\boldsymbol{\theta}}$)
- avoid any asymptotic approximations

Suppose

$$\boldsymbol{Y}^{\text{rep}} \mid \boldsymbol{\theta} \;\sim\; M(\boldsymbol{\theta})$$

and is conditionally independent of the data.

Note: Averaging its distribution over the posterior gives the *posterior predictive distribution* of the data.

Note: We can generate $\boldsymbol{Y}^{\text{rep}}$ for a given $\boldsymbol{\theta}$ by simulating from the data model, which is usually easy.

A simulated value $\boldsymbol{y}^{\text{rep}}$ would be called a **replicate** data set.

Then, instead of a frequentist p-value, a Bayesian could use a **posterior predictive p-value**

$$p_b \;=\; \mathrm{P}\big(T(\boldsymbol{Y}^{\mathsf{rep}};\boldsymbol{\theta}) \geq T(\boldsymbol{y};\boldsymbol{\theta}) \;\big|\; \boldsymbol{y}\big)$$

where the probability is over the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{Y}^{\mathsf{rep}}$.

Sufficiently small p_b indicates a problem with the model and/or prior.

Usually p_b can't be directly computed. Instead, it can be approximated by posterior simulation (e.g., MCMC):

1. For each posterior-generated value $\boldsymbol{\theta}$, generate a replicate data set $\boldsymbol{y}^{\text{rep}}$ (conditionally) and compute

$$T(\boldsymbol{y}^{\text{rep}}; \boldsymbol{\theta}) \quad \text{and} \quad T(\boldsymbol{y}; \boldsymbol{\theta})$$

2. Approximate p_b as the fraction of generated pairs $(\boldsymbol{\theta}, \boldsymbol{y}^{\text{rep}})$ for which

$$T(\boldsymbol{y}^{\text{rep}}; \boldsymbol{\theta}) \geq T(\boldsymbol{y}; \boldsymbol{\theta})$$

Example: Shark Attacks

Recall:

$$Y_i = \text{number of shark attacks (worldwide)}$$

$$x_i = \text{year (2005–2017)}$$

Our chosen model and prior:

$$Y_i \mid \lambda_i \sim \text{indep. Poisson}(\lambda_i)$$

$$\ln(\lambda_i) = \beta_0 + \beta_1(x_i - \bar{x})$$

$$\beta_0, \beta_1 \sim \text{indep. N}(0, 100^2)$$

We will investigate fit using the chi-square discrepancy

$$T(\boldsymbol{y}; \beta_0, \beta_1) \;\; = \;\; \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i}$$

(Why is this the chi-square discrepancy?)

If there are problems with the Poisson regression or the (vague) normal priors, we expect $T(\boldsymbol{y}; \beta_0, \beta_1)$ to be large relative to its replicate version, leading to small p_b.

The JAGS code:

```
data {
  xmean <- mean(x)
  for(i in 1:length(x)) {
    xcent[i] <- x[i] - xmean
  }
}
model {
  for(i in 1:length(y)) {
    y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- beta0 + beta1 * xcent[i]

    yrep[i] ~ dpois(lambda[i])
  }

  chisq <- sum((y - lambda)^2 / lambda)
  chisqrep <- sum((yrep - lambda)^2 / lambda)
  pb.ind <- chisqrep >= chisq

  beta0 ~ dnorm(0, 0.0001)
  beta1 ~ dnorm(0, 0.0001)
}
```

Note: Arithmetic operations are automatically vectorized in JAGS.

Note: The data set used here has only years 2005 to 2017 (and not the missing 2018 observation).

We will also try an alternative version of the model that has a badly mis-specified prior ...

R/JAGS Example 11.2:

Assessment for Poisson Regression

Remark: For some choices of discrepancy T, obtaining an especially *large* value of p_b would also indicate a problem with the model and/or prior. See Cowles, Sec. 11.4, for an example.