Example of Bayesian Model Selection

- Example in `R` with Oxygen Data Set
- We can consider all possible subsets of set of predictor variables:
- We can consider only certain subsets (here, we only consider including the interaction term when both first-order terms appear):

The Posterior Predictive Distribution of the Data

- Suppose we have built our Bayesian regression model using response data $\mathbf{y}$ and explanatory data matrix $\mathbf{X}$.
- Suppose we consider future observations whose explanatory variable values are in the matrix $\mathbf{X}^*$.
- What is the marginal distribution of the corresponding future response values $\mathbf{y}^*$?
- This is the **posterior predictive distribution**

$$\pi(\mathbf{y}^* | \mathbf{y}, \mathbf{X}^*, \mathbf{X}).$$

- We will use this later as a tool for checking the fit of our regression model.

The Posterior Predictive Distribution of the Data

- In our analysis with the noninformative priors, note that

$$\pi(\mathbf{y}^*, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}^*, \mathbf{X}) = \pi(\mathbf{y}^* | \boldsymbol{\beta}, \sigma^2, \mathbf{X}^*) \pi(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y})$$

- Then integrating out $\boldsymbol{\beta}$ and σ^2, it can be shown that the posterior predictive distribution of $\mathbf{y}^*$ is multivariate-t with $(n-k)$ degrees of freedom so that

$$E(\mathbf{y}^*) = \mathbf{X}^* \hat{\mathbf{b}} \text{ and}$$

$$\text{covariance matrix } = \frac{(n-k)\hat{\sigma}^2}{n-k-2}[\mathbf{I} + \mathbf{X}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{*'}]$$

- **Intuition**: Our original data are multivariate normal, given the model.
- Our future predictions are multivariate-t (reflects added uncertainty about the model).

CHAPTER 5 SLIDES START HERE

The Bayesian Prior

- A prior distribution **must** be specified in a Bayesian analysis.
- The choice of prior can substantially affect posterior conclusions, especially when the sample size is not large.
- We now examine several broad methods of determining prior distributions.

Conjugate Priors

- We know that **conjugacy** is a property of a prior **along with a likelihood** that implies the posterior distribution will have the same *distributional form* as the prior (just with different parameter(s)).
- We have seen some examples of conjugate priors:
  Data/Likelihood **Prior**
  1. Bernoulli $\longrightarrow$ Beta for p
  2. Poisson $\longrightarrow$ Gamma for λ
  3. Normal $\longrightarrow$ Normal for μ
  4. Normal $\longrightarrow$ Inverse gamma for σ^2

Conjugate Priors

Other examples:

1. Multinomial $\rightarrow$ Dirichlet for $p_1, p_2, \ldots, p_k$
2. Negative Binomial $\rightarrow$ Beta for p
3. Uniform(0, θ) $\rightarrow$ Pareto for upper limit
4. Exponential $\rightarrow$ Gamma for β
5. Gamma (β unknown) $\rightarrow$ Gamma for β
6. Pareto (α unknown) $\rightarrow$ Gamma for α
7. Pareto (β unknown) $\rightarrow$ Pareto for β

Conjugate Priors: Exponential Family

- Consider the family of distributions known as the **one-parameter exponential family**.
- This family consists of any distribution whose p.d.f. (or p.m.f.) can be written as:

$$f(x|\theta) = e^{[t(x)u(\theta)]}r(x)s(\theta)$$

  where $t(x)$ and $r(x)$ do not depend on the parameter θ and $u(\theta)$ and $s(\theta)$ do not depend on x.
- Note that any such density can be written as

$$f(x|\theta) = e^{\{t(x)u(\theta)+\ln[r(x)]+\ln[s(\theta)]\}}$$

Conjugate Priors: Exponential Family

- If we observe an iid sample $X_1, \ldots, X_n$, the joint density of the data is thus

$$f(\mathbf{x}|\theta) = e^{\{u(\theta)\sum_{i=1}^{n} t(x_i) + \sum_{i=1}^{n} \ln[r(x_i)] + n \ln[s(\theta)]\}}$$

- Consider a prior for θ (with the prior parameters k and γ) having the form:

$$p(\theta) = c(k, \gamma) e^{\{k u(\theta) \gamma + k \ln[s(\theta)]\}}$$

Conjugate Priors: Exponential Family

Then the posterior is

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)p(\theta) \\
&\propto \exp\left\{ u(\theta)\sum t(x_i) + n\ln[s(\theta)] + ku(\theta)\gamma + k\ln[s(\theta)] \right\} \\
&= \exp\left\{ u(\theta)\left[\sum t(x_i) + k\gamma\right] + (n+k)\ln[s(\theta)] \right\} \\
&= \exp\left\{ (n+k)u(\theta)\left[\frac{\sum t(x_i) + k\gamma}{n+k}\right] + (n+k)\ln[s(\theta)] \right\}
\end{aligned}$$

which is of the same form as the prior, except with "k"$= n + k$ and "γ"$= \dfrac{\sum t(x_i) + k\gamma}{n+k}$.

$\Rightarrow$ If our data are iid from a one-parameter exponential family, then a conjugate prior will exist.

Conjugate Priors

- Conjugate priors are mathematically convenient.
- Sometimes they are quite flexible, depending on the specific hyperparameters we use.
- But they reflect very specific prior knowledge, so we should be wary of using them unless we truly possess that prior knowledge.

Uninformative Priors

- These priors intentionally provide very little specific information about the parameter(s).
- A classic uninformative prior is the *uniform* prior.
- A *proper* uniform prior integrates to a finite quantity.
- **Example 1**: For Bernoulli(θ) data, a uniform prior on θ is

$$p(\theta) = 1, \;\; 0 \le \theta \le 1.$$

- This makes sense when θ has **bounded support**.

Uninformative Priors

- **Example 2**: Consider $N(0, \sigma^2)$ data. If it is "reasonable" to assume, that, say $\sigma^2 < 100$, we could use the uniform prior

$$p(\sigma^2) = \frac{1}{100}, \quad 0 \le \sigma^2 \le 100$$

(even though σ^2 is not intrinsically bounded).
- An **improper** uniform prior integrates to ∞:
- **Example 3**: $N(\mu, 1)$ data with

$$p(\mu) = 1, \quad -\infty < \mu < \infty.$$

- This is fine as long as the resulting **posterior** is proper.
- But be careful: Sometimes an improper prior will yield an improper posterior.

Invariance Property

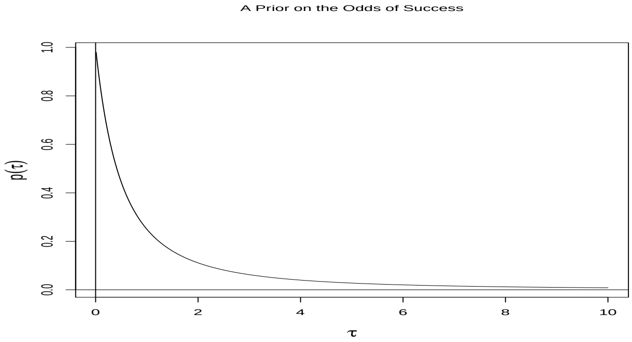
- A problem with the uniform prior is that its "lack of information" is **not invariant** under transformation.
- **Example 1 again**: Consider the **odds** of success $\tau = \dfrac{\theta}{1-\theta}$.
- Then if $p(\theta) = 1$, with the Jacobian

$$J = \left| \frac{\mathrm{d}}{\mathrm{d}\tau} \left(\frac{\tau}{1+\tau} \right) \right| = \frac{1}{(1+\tau)^2},$$

$$\text{then } p(\tau) = \frac{1}{(1+\tau)^2}, \;\; 0 < \tau < \infty :$$

Invariance Property

- Picture:



- This same prior is now an "informative" prior for the odds.
- (However, note that $P(0 < \tau < 1) = P(\tau > 1) = 0.5$.)

Jeffreys Prior

- Jeffreys (1961) developed a class of priors that were invariant under transformation.
- For a single parameter θ and data having joint density $f(\mathbf{x}|\theta)$, the Jeffreys prior

$$p_J(\theta) \propto \left[-E\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln f(\mathbf{x}|\theta)\right)\right]^{1/2} = [I(\theta)]^{1/2}$$

  (square root of Fisher information)
- For a parameter vector $\boldsymbol{\theta}$:

$$p_J(\boldsymbol{\theta}) \propto \left[E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}} \ln f(\mathbf{x}|\boldsymbol{\theta})\right]' \left[\frac{\partial}{\partial\boldsymbol{\theta}} \ln f(\mathbf{x}|\boldsymbol{\theta})\right]\right\}\right]^{1/2}$$

Jeffreys Prior

- **Example 1 yet again**: For $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$,

$$f(\mathbf{x}|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad 0 \leq \theta \leq 1,$$

where $y = \sum_{i=1}^{n} x_i$.

$$\Rightarrow \ln f(\mathbf{x}|\theta) = \ln \binom{n}{y} + y \ln(\theta) + (n-y) \ln(1-\theta)$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f(\mathbf{x}|\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln f(\mathbf{x}|\theta) = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

Jeffreys Prior

$$\Rightarrow -E\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f(\mathbf{x}|\theta)\right] = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta}$$

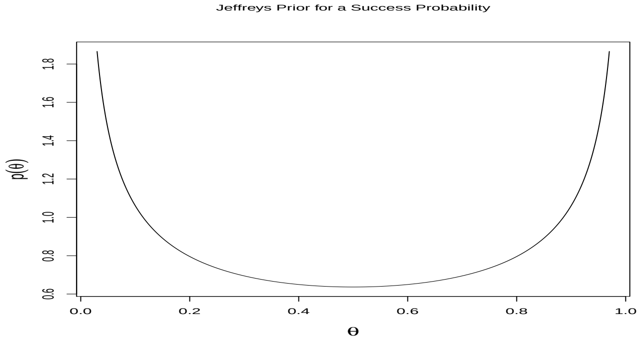
$$= \frac{n(1-\theta) + n\theta}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)}$$

$$\Rightarrow p_J(\theta) \propto \left[\frac{n}{\theta(1-\theta)}\right]^{1/2}$$

$$\Rightarrow p_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} = \theta^{1/2-1}(1-\theta)^{1/2-1}$$

Jeffreys Prior

$\Rightarrow$ Jeffreys prior for θ is a Beta($1/2, 1/2$):
Picture:



Jeffreys Prior

- **Invariance**: If $p_J(\theta)$ is the Jeffreys prior for θ, for any transformation $\phi = g(\theta)$,

$$p_J(\theta) = p_J(\phi)\left|\frac{\mathrm{d}\phi}{\mathrm{d}\theta}\right|.$$