Bayesian Theory and Computation

Lecture 7: Decision Theory and Model Selection

Cheng Zhang

School of Mathematical Sciences, Peking University

March 19, 2025

- In the Bayesian paradigm, estimation, hypothesis testing, and model selection are special cases of decision problems.
- Decision theory provides a mathematical framework for making decision under uncertainty; that is, when the outcome of an event is not known.
- However, we assume that we know our loss (or gain) when one of the possible outcomes occur.

- We use $\mathcal{V}$ to denote the set of all possible values, v, for unknown variables. We refer to $\mathcal{V}$ as the *outcome space*.
- $\mathcal{V}$ could be the value of future observations. For example, $\mathcal{V} = \{\text{Head}, \text{Tail}\}$ when you are tossing a coin.
- Or it could be the value of a parameter in a model. For example $\mathcal{V} = \mathbb{R}$, when we want to estimate μ, the mean of a normal distribution.
- We present the set of all possible actions, a, as $\mathcal{A}$. We refer to $\mathcal{A}$ as the *action space*.
- If we are predicting the outcome of the next coin toss, $\mathcal{A} = \{\text{Head}, \text{Tail}\}$; if we want to estimate μ (i.e., point estimation), our action space would be $\mathcal{A} = \mathbb{R}$.
- For hypothesis testing, we can define our action $\mathcal{A} = \{0, 1\}$, where 0 means do not reject the null hypothesis $H_0 : \mu \leq 0$ and 1 means rejecting it.

- We define *utility* as a function $u = U(v, a)$ that maps the product of outcome space and action space to a real number $u \in \mathbb{R}$ representing how much we gain if we choose action a and the outcome v occurs.
- It is more common to choose a loss function instead of utility (e.g., negative of utility) representing our loss when we choose action a and the outcome v occurs.
- In the coin tossing experiment, the loss function, $L(v, a)$ can be defined as follows

$$L(\text{Head}, \text{Head}) = L(\text{Tail}, \text{Tail}) = 0$$

$$L(\text{Head}, \text{Tail}) = L(\text{Tail}, \text{Head}) = 1$$

- This is known as 0-1 loss function.

- Now, assume that we have observed data y, for example,

$$y = HHTHTHHT$$

  which is the outcome from a sequence of coin tossing. Using this data, we want to make a decision about what the outcome of the next toss would be (or what is θ, the probability of head for this coin).
- The tool for making decision is called *decision rule*, and it's denoted as $\delta(y)$. Note that δ is a function of data only.
- For example, given y, we might define our decision rule for guessing what would be the outcome of the next toss as follows

$$\delta(y) = \begin{cases} \text{Head} & \text{if the frequency of Heads is} \geq 0.5 \\ \text{Tail} & \text{if the frequency of Heads is} < 0.5. \end{cases}$$

- Posterior risk for a decision rule δ is

$$r(\delta|y) = \int_{\mathcal{V}} L(v, \delta(y))p(v|y)dv$$

- Note that we replaced the action a with the decision rule $\delta(y)$ since our action now depends on our decision rule which itself depends on the observed data.
- Also, note that $p(v|y)$ is the posterior predictive probability if v is future observation (i.e., what is the outcome of the next toss), or it is posterior probability if v is the parameter of a model (i.e., μ, the mean of a normal distribution).

Bayes Action

- The expected loss principle: In deciding between different rules, choose the one with the smallest posterior risk.
- Bayes action $\delta^*(y)$ is the action that minimizes the posterior risk: $r(\delta^*|y) \leq r(\delta|y)$, $\forall y$ and δ.
- In theory, this is all we need to know for all sorts of decision problems (e.g., prediction, point estimation, and hypothesis setting).
- For example, as we will see later, if we have a simple 0-1 loss function and a discrete action space such as the coin tossing example, the Bayes action is choosing the mode of the posterior distribution $p(v|y)$.

- Many decision problems in statistics deal with estimating the parameter of a probability model (e.g., the mean of a normal model, or the coefficients in a linear regression model), i.e. we have $\mathcal{V} = \theta$.
- A possible loss function is the *squared error loss* function: $L(\theta, a) = \|\theta - a\|^2$.
- In general, the Bayes action for this specific loss function is to choose the mean of the posterior distribution

$$\mathbb{E}_{\theta|y}(L(\theta, a)) = \mathbb{E}_{\theta|y}(\theta^2 - 2a\theta + a^2) = \mathbb{E}_{\theta|y}(\theta^2) - 2a\mathbb{E}_{\theta|y}(\theta) + a^2$$

We take the derivative with respect to a and set it to zero:

$$-2\mathbb{E}(\theta) + 2a = 0 \Rightarrow a = \mathbb{E}_{\theta|y}(\theta)$$

- That's the reason we usually use posterior mean for point estimate.

Absolute Error Loss

- Now suppose we want to use the *absolute error loss* function: $L(\theta, a) = |\theta - a|$.
- Therefore, we need to minimize $\mathbb{E}_{\theta|y}(|\theta - a|)$.
- Using Leibniz's rule

$$\frac{\partial}{\partial t}\int_{a(t)}^{b(t)} f(x,t)dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)dx - f(a(t),t)a'(t) + f(b(t),t)b'(t)$$

we have

$$\begin{aligned}\frac{\partial}{\partial a}\mathbb{E}_{\theta|y}(|\theta - a|) &= \frac{\partial}{\partial a}\int_{-\infty}^{a}(a-\theta)f(\theta|y)d\theta + \frac{\partial}{\partial a}\int_{a}^{\infty}(\theta - a)f(\theta|y)d\theta \\ &= \int_{-\infty}^{a} f(\theta|y)d\theta - \int_{a}^{\infty} f(\theta|y)d\theta\end{aligned}$$

- Bayes estimator in this case is the posterior median.

- Given a prior distribution π, it is also possible to define the *integrated risk*, which is the frequentist risk averaged over the values of θ according to their prior distribution

$$r(\pi, \delta) = \mathbb{E}_\pi R(\theta, \delta) = \int_\Theta \int_\mathcal{Y} L(\theta, \delta(y)) p(y|\theta) dy \, \pi(\theta) d\theta$$

- This introduces a total ordering on the set of estimators, allowing for the direct comparison of estimators.
- Bayes Rule. A Bayes rule is a function δ_π, that minimizes the integrated risk.

- Another way of deriving the Bayes estimator

$$\begin{aligned} r(\pi, \delta) &= \int_{\Theta} \int_{\mathcal{Y}} L(\theta, \delta(y)) p(y|\theta) dy \, \pi(\theta) d\theta \\ &= \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) p(\theta|y) d\theta \, p(y) dy \\ &= \int_{\mathcal{Y}} r(\delta|y) p(y) dy \end{aligned}$$

- Note that the last equation is the posterior risk averaged over the marginal distribution of y. This implies that the Bayes rule can be obtained by taking the Bayes estimator for each particular y!
- The value $r(\pi) = r(\pi, \delta_\pi)$ is called the *Bayes risk*.

- An estimator δ_0 is inadmissible if there exist δ_1 which dominates δ_0, that is $R(\theta, \delta_0) \geq R(\theta, \delta_1)$ and, for at least one value θ_0 of the parameter, $R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$. Otherwise, δ_0 is said to be admissible.
- If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ_π is admissible.
- Sketch of proof. Suppose δ_π is inadmissible and consider δ' which uniformly dominates δ_π. Then $R(\theta, \delta') \leq R(\theta, \delta_\pi)$ and, in an open set C of Θ, $R(\theta, \delta') < R(\theta, \delta_\pi)$. Hence

$$r(\pi, \delta') = \int_\Theta R(\theta, \delta')\pi(\theta)d\theta < \int_\Theta R(\theta, \delta_\pi)\pi(\theta)d\theta = r(\pi, \delta_\pi),$$

which is impossible.

- Another type of decision problem, as we mentioned above, is hypothesis testing.
- If we want to choose between two hypothesis $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \notin \Theta_0$, all we need to do again is to choose the hypothesis whose posterior risk is smaller.
- Let's assume a simple 0-1 loss function, that is, the penalty associated with an estimate δ is 0 if the answer is correct and 1 otherwise.

$$L(\theta, \delta) = \begin{cases} 1 - \delta & \text{if } \theta \in \Theta_0 \\ \delta & \text{otherwise} \end{cases}$$

- The Bayes estimator is

$$\delta(y) = \begin{cases} 1 & \text{if } p(\theta \in \Theta_0 | y) > p(\theta \notin \Theta_0 | y) \\ 0 & \text{otherwise.} \end{cases}$$

- In general, the loss due to *type I error* (i.e., rejecting H_0 when it is true) is different from that of *type II error* (i.e., accepting H_0 when it is not true).
- In this case, although we might not choose the one with a higher posterior probability, the principle of choosing the one with a smaller posterior risk remains as before.
- Let's assume the loss due to type I error is 19 and the loss due to type II error is 1. We accept H_1 if its posterior risk is smaller than the posterior risk of H_0,

$$0 \times p(H_1|y) + 19 \times p(H_0|y) < 0 \times p(H_0|y) + 1 \times p(H_1|y)$$
$$p(H_0|y) < 1/20 = 0.05$$

- That is, for this specific loss function, we reject the null hypothesis if its posterior probability is less than 0.05

Example: p-value as a Bayes Estimator

- Consider $x \sim \mathcal{N}(\theta, 1)$ and the null hypothesis $H_0 : \theta \leq 0$ is tested against the alternative hypothesis $H_1 : \theta > 0$. This testing problem is an estimation problem if we consider the estimation of the indicator function $1_{H_0}(\theta)$.
- Under the quadratic loss $(1_{H_0}(\theta) - \delta(x))^2$, we can propose the following estimator

$$P_{\text{value}}(x) = p(X > x) = 1 - \Phi(x)$$

- This is also a generalized Bayesian estimator under Lebesgue measure and quadratic loss

$$\begin{aligned}\delta_\pi(x) &= \mathbb{E}_{\theta|x}(1_{H_0}(\theta)) = p(\theta < 0|x) \\ &= p(\theta - x < -x|x) = 1 - \Phi(x)\end{aligned}$$

Therefore, p-value in this case is admissible.

- Now let's consider a simple hypothesis testing problem formalized as a decision problem between two possible models: $p(y|\theta_0)$ and $p(y|\theta_1)$. That is, we think the model parameter θ could take one of the two possible values.
- A *priori*, we believe the probabilities of $\theta = \theta_0$ and $\theta = \theta_1$ are $p(\theta_0)$ and $p(\theta_1)$ respectively.
- With a simple 0-1 loss function, we choose the model with a higher posterior probability. We could compare posterior probabilities by presenting them in the form of a posterior odds as follows

$$\frac{p(\theta_0|y)}{p(\theta_1|y)} = \frac{p(\theta_0)p(y|\theta_0)/p(y)}{p(\theta_1)p(y|\theta_1)/p(y)} = \frac{p(\theta_0)p(y|\theta_0)}{p(\theta_1)p(y|\theta_1)}$$

That is, the posterior odds is the product of the prior odds and the likelihood ratio.

- Traditionally, statisticians avoid expressing a prior odds in favor of either alternatives (especially if we are not making a decision, rather, we are reporting our findings): $p(\theta_0)/p(\theta_1) = 1$, and rely only on

$$\frac{p(y|\theta_0)}{p(y|\theta_1)}$$

  which is known as **Bayes factor**.
- This is analogous (not the same in general settings though) to the likelihood ratio test that is commonly used in the frequentist framework.
- When H_0 and H_1 are not single point hypothesis, the Bayes factor is defined in general as

$$\text{BF}(H_0; H_1) = \frac{p(y|H_0)}{p(y|H_1)} = \frac{\int p(y|\theta_0, H_0)p(\theta_0|H_0)d\theta_0}{\int p(y|\theta_1, H_1)p(\theta_1|H_1)d\theta_1}$$

- We can also use BF to choose between two alternative models

$$\text{BF}_{12} = \frac{p(y|M_1)}{p(y|M_2)}$$

- In general, when the models are specified in terms of unknown parameters, θ, we have

$$\text{BF}_{12} = \frac{\int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(y|\theta_2, M_2)p(\theta_2|M_2)d\theta_2}$$

- In this case, BF is the ratio of *prior predictive distributions*.

- Generally speaking, the Bayes factor is a summary of the evidence provided by the data, in favor of one scientific theory, represented by a statistical model, as opposed to another.
- Jeffreys (1961) provided interpretive ranges for the BF analogous to what frequentist use for p-values:
  - $1 < \text{BF} < 3$: slight evidence
  - $3 < \text{BF} < 10$: positive evidence
  - $\text{BF} > 10$: strong evidence
- Using the BF has some difficulties. For example, in general we cannot use improper prior distributions.
- Other alternatives such as fractional Bayes Factor (O'Hagan 1995) are more appropriate (this is beyond the scope of this course, but you can refer to O'Hagan's paper for more details).

- Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance. Consider a woman who has an affected brother and unaffected father.
- Let θ be an indicator that the woman is a carrier of the gene. Based on the information thus far, we may assume $p(\theta = 1) = p(\theta = 0) = \frac{1}{2}$.
- Suppose she has two sons, neither of whom is affected, i.e., $y_1 = y_2 = 0$. Now consider the two competing models $H_1 : \theta = 1$, and $H_2 : \theta = 0$.
- The prior odds are $p(H_2)/p(H_1) = 1$, and the Bayes factor of the data is

$$\frac{p(y|H_2)}{p(y|H_1)} = \frac{1.0}{0.5 \times 0.5} = 4$$

- The posterior odds are thus $p(H_2|y)/p(H_1|y) = 4$.

- For many data analyses, explicit benefit or cost information is not available, and the predictive performance of a model is assessed by generic scoring functions and rules.
- In *point prediction*, scoring functions are often used. One typical example is mean square error:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}(y_i|\theta))^2$$

(or a weighted version according to the variance). Easy to compute and interpret, but could be less appropriate for non-Gaussian models.

- In *probabilistic prediction*, people often use score rules to account for the uncertainty. One commonly used rule is the log predictive density $\log p(y|\theta)$, which is proportional to the mean square error for Gaussian models with constant variance.

- Let $\tilde{y}_i$ be a new data point, the out-of-sample predictive fit is

$$\log p(\tilde{y}_i|y) = \log \mathbb{E}_{\theta|y}(p(\tilde{y}_i|\theta)) = \log \int p(\tilde{y}_i|\theta)p(\theta|y)d\theta.$$

  Here $p(\tilde{y}_i|y)$ is the predictive density for $\tilde{y}_i$ induced by the posterior distribution $p(\theta|y)$.

- Let $f(y)$ be the true data distribution. We can define the expected out-of-sample log predictive density as

$$\text{elpd} = \mathbb{E}_f(\log p(\tilde{y}_i|y)) = \int \log p(\tilde{y}_i|y) f(\tilde{y}_i) d\tilde{y}_i$$

- One can also define a measure of predictive accuracy for n data points

$$\text{elppd} = \sum_{i=1}^{n} \mathbb{E}_f(\log p(\tilde{y}_i|y)) \quad (1)$$

- Since $f(y)$ is generally unknown, we need to find an estimate for it. A typical choice is to use the observed data

$$\text{lppd} = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|y)d\theta$$

- In practice, we can evaluate the expectation using draws from $p(\theta|y)$, $\theta^1, \ldots, \theta^S$

$$\text{computed lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right)$$

- the lppd of observed data y is an overestimate of the elppd for future data. We need to apply some sort of bias correction to get a reasonable estimate of (1).

- Deviance (the log predictive density of the data given a point estimate of the fitted model) is defined as

$$D(y, \theta) = -2 \log p(y|\theta)$$

  which is a measure of discrepancy (i.e., lack of fit, lower is better).
- The deviance measure as described above, depends on both y and θ. If we want to use a measure that depends only on y, we can integrate the deviance over the posterior

$$D_{\text{post-avg}} = \mathbb{E}_{\theta|y}(D(y, \theta))$$

- We can estimate this by using simulated samples from the posterior distribution

$$\hat{D}_{\text{post-avg}}(y) \approx \frac{1}{L} \sum_{\ell=1}^{L} D(y, \theta^{\ell})$$

- Deviance is especially useful when we compare nested models; that is, when we are deciding whether to include the predictor x in the model or not, i.e.:

$$M_0 : y = \beta_0 + \epsilon$$
$$M_1 : y = \beta_0 + \beta_1 x + \epsilon$$

- However, we could decrease deviance by arbitrarily increasing the complexity of model, for example, by adding more predictors into the model.
- In general, it is recommended to use more complex models only when they result in substantial (i.e., statistically significant) improvement in performace (i.e., substantial decrease in deviance).
- The above principle is widely known as Occam's razor: "everything equal, we should use the simplest solution".

Deviance Information Criterion (DIC)

- When we are relying on deviance, we need a measurement that accounts for the trade-off between complexity and goodness-of-fit.
- In a decision model, this could be done by using a loss function that penalizes larger models.
- A simple measure, which does this automatically, is called *deviance information criterion* (DIC) defined as follows (Spiegelhalter et al., 2002)

$$\text{DIC} = \hat{D}_{\text{post-avg}}(y) + p_{\text{DIC}}$$

- p_{DIC} is called *effective number of parameters* and is a measure of complexity

$$p_{\text{DIC}} = \hat{D}_{\text{post-avg}}(y) - D_{\hat{\theta}_{\text{Bayes}}}(y)$$

Deviance Information Criterion (DIC)

- Here, $D_{\hat{\theta}}(y)$ is the deviance when we first average posterior parameters and then calculate deviance (as opposed to integrating deviance over posterior parameters).
- Therefore, we can obtain DIC as follows

$$\begin{aligned}\text{DIC} &= 2\hat{D}_{\text{post-avg}}(y) - D_{\hat{\theta}_{\text{Bayes}}}(y) \\ &= D_{\hat{\theta}_{\text{Bayes}}}(y) + 2p_{\text{DIC}}\end{aligned}$$

- Caution! Although it is easy to use DIC for model evaluation, remember that the best approach is still to use problem specific loss function, and based on the posterior risk, to find the optimal decision rule. Use DIC only when you don't have a better loss function or you simply want to report your findings.

- Akaike Information Criterion (AIC):

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

  where k is the number of parameters in the model.
- Bayesian Information Criterion (BIC)

$$\text{BIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + k \log n$$

- BIC penalizes more on model complexity when n is large.
- See Gelman et al. (2003) for more criteria.

Example: Forecasting Presidential Elections

- We use a simple linear regression for example where vote share y is predicted from economic performance x solely as follows

$$y_i \sim \mathcal{N}(a + bx_i, \sigma^2), \quad i = 1, \ldots, n$$

with a noninformative prior $p(a, b, \sigma^2) \propto \sigma^{-2}$.

- The conditional posterior of $\beta = (a, b)$ is

$$\beta|\sigma^2, y \sim \mathcal{N}(\hat{\beta}, V_\beta \sigma^2)$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$, $V_\beta = (X^T X)^{-1}$.

- The marginal posterior distribution of σ^2 is

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-2, s^2), \quad s^2 = \frac{1}{n-2}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

Example: Forecasting Presidential Elections

- For the data at hand, $n = 15, s = 3.6, \hat{\beta} = (45.9, 3.2)$, and $V_\beta = \begin{pmatrix} 0.21 & -0.07 \\ -0.07 & 0.04 \end{pmatrix}$.
- AIC. The MLE of $(\hat{a}, \hat{b}, \hat{\sigma})$ is $(45.9, 3.2, 3.6)$.

$$\text{AIC} = -2 \sum_{i=1}^{15} \log \mathcal{N}(y_i | x_i^T \hat{\beta}, s^2) + 2 \times 3 = 86.6$$

- BIC. Similarly,

$$\text{BIC} = -2 \sum_{i=1}^{15} \log \mathcal{N}(y_i | x_i^T \hat{\beta}, s^2) + \log(15) \times 3 = 88.7$$

- DIC. Using MCMC samples, the estimated DIC is 87.0. The estimate of p_{DIC} is 3, which is exactly the number of parameters in the model.

Example: Titanic Survival

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

The first 5 data in Titanic dataset.

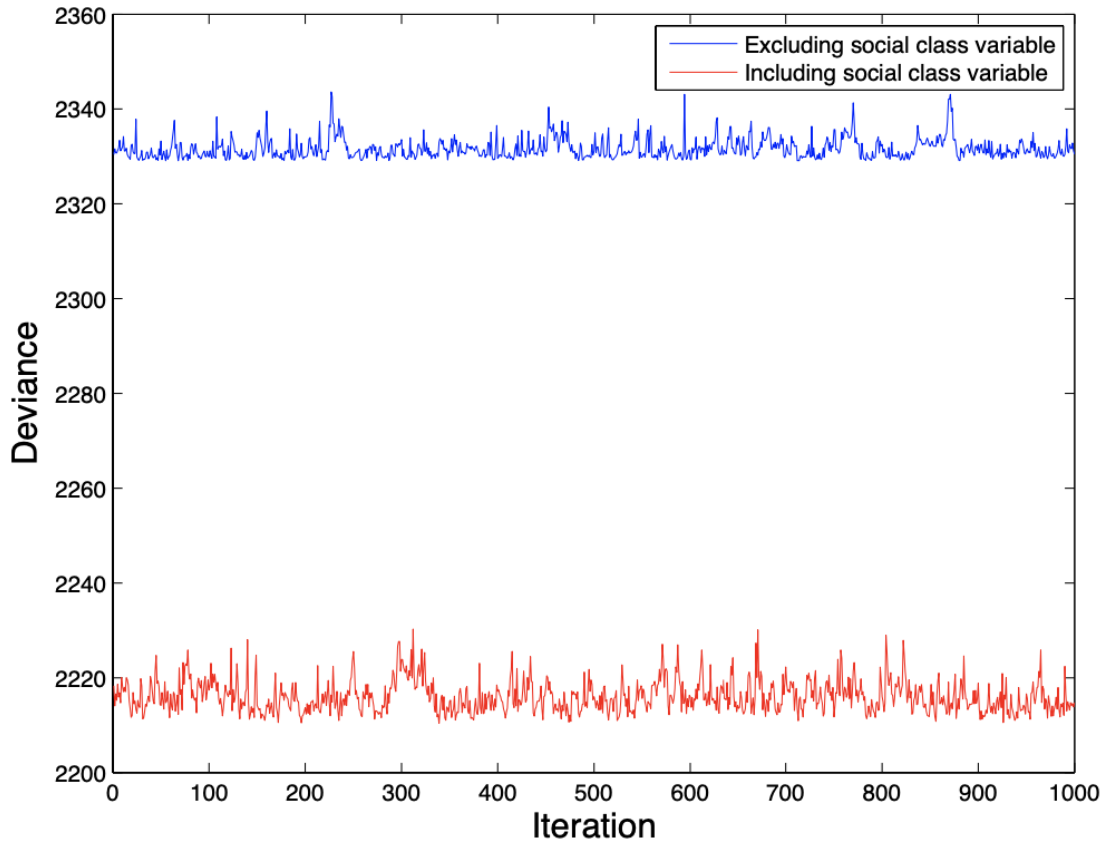
- One of the commonly used exemplar dataset for logistic regression is the Titanic dataset.
- We consider two nested logistic regression models: Model M_0, which does not include the social class predictor (i.e., only the intercept, age and gender are included), and Model M_1, which includes the social class as well as other variables.

- We fit these two models separately and present the results in the following table

Model	$\hat{D}_{avg}$	$D_{\hat{\theta}}$	p_D	DIC
M_0	2331.6	2329.1	2.5	2334.1
M_1	2216.2	2210.1	6.1	2222.4

- As we can see, M_1 has a smaller DIC, and therefore, provides a better fit. This could be interpreted as statistical significance of social class.

- We can also compare the posterior distribution of deviance for different models. The following graph shows the trace plot of deviance for models M_0 and M_1.



- When the parameter space is finite, $\Theta = \{\theta_0, \theta_1, \ldots, \theta_k\}$ with prior $p_j = p(\theta = \theta_j) > 0$, then given n observed samples from $p(y|\theta_0)$, we can show that

$$\lim_{n\to\infty} p(\theta = \theta_0|y) = 1 \qquad \lim_{n\to\infty} p(\theta = \theta_j|y) = 0 \; \forall j \neq 0$$

- To see this, consider the log-posterior odds with respect to θ_0 (i.e., true value of model parameter)

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

- For $\theta \neq \theta_0$, the expectation of each summand in the second term is negative.

$$\mathbb{E}_{y_i|\theta_0} \log \left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)} \right) = C < 0, \quad \forall i$$

- So the right hand side $\to -\infty$ as $n \to \infty$. Therefore,

$$\lim_{n\to\infty} p(\theta = \theta_j|y) = 0 \quad \forall j \neq 0$$

- Because the sum of probabilities is 1,

$$\lim_{n\to\infty} p(\theta = \theta_0|y) = 1$$

- When the parameter space is continuous, we can show that the posterior probability of θ becomes more and more concentrate about θ_0 as $n \to \infty$.
- **Theorem**. If θ is defined on a compact set Θ and A is a neighborhood of θ_0 with nonzero prior probability, then $\Pr(\theta \in A|y) \to 1$ as $n \to \infty$.
- Since Θ is compact, we can find a finte subcovering of it, with A being the only neighborhood that includes θ_0. The proof then simply follows the discrete case.
- Moreover, we can show that (under some regularity conditions) for large n, the posterior is asymptotically normal

$$\theta|y \overset{\text{asy}}{\sim} \mathcal{N}(\hat{\theta}_n, I(\hat{\theta}_n)^{-1})$$

where $\hat{\theta}_n$ is the MLE and $I(\hat{\theta}_n)$ is the observed Fisher information.

References

- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). Bayesian Data Analysis. Chapman and Hall
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795. doi:10.1080/01621459.1995.10476572
- O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). J. Roy. Statist. Soc. Ser. B 57:99–138.
- Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelita van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit." Journal of the Royal Statistical Society, Series B 64:1-34.