Lecture 9 - Key

Up until now, we've primarily concerned ourselves with estimation type problems. However, many perform hypothesis tests.

Say, $x \sim N(0,1)$ there are three types of tests you might consider for testing μ:

1. $\mu < \mu_0$ (one-sided test)
2. $\mu = \mu_0$ (point test)
3. $\mu \in [\mu_1, \mu_2]$ (interval test)

In a Bayesian framework, we will use point mass priors for Bayesian hypothesis testing.

Example: Consider testing the hypothesis $H_0 : \theta = \theta_0$ vs $H_1 : \theta_0 \neq \theta_1$. Say you observe data, $\tilde{x} = (x_1, \ldots, x_n)$, where $x_i \sim N(\theta, \sigma^2)$ with σ^2 known.

- Q: How would this question be addressed in a classical framework?

- If we want to be Bayesian, we need a prior. Suppose we choose a flat prior, $p(\theta) \propto 1$. Then $p(\theta = \theta_0|\tilde{x}) \sim N(\bar{x}, \sigma^2/n)$. With this distribution, we compute $Pr(H_0|\tilde{x}) = 0$. This is a result of a continuous prior on a continuous parameter.

- Consider a different prior that places mass on $H_0 : \theta = \theta_0$ which is non-zero. Specifically let $Pr(\theta = \theta_0) = p_{\theta_0}(\theta) = \pi_{\theta_0} > 0$. Say $\pi_{\theta_0} = 0.5$.

- We also need a prior for the alternative space. Let's choose a conjugate prior. Let $\theta_1 \sim N(\mu, \tau^2)$.

- Combining these the prior is:

$$p(\theta) = \pi_0 \delta(\theta = \theta_0) + (1 - \pi_0) p_1(\theta),$$

  where $\delta(\theta = \theta_0)$ is an indicator function for $\theta = \theta_0$ and $p_1(\theta) = N(\mu_1, \tau^2)$. This prior is known as a point mass prior. It is also a special case of another prior known as a spike-and-slab prior.

- Recall $x_i \sim N(\theta, \sigma^2)$. We want to know $Pr(H_0|\tilde{x})$ and $Pr(H_1|\tilde{x})$.

$$\begin{aligned} Pr(H_0|\tilde{x}) &= \frac{p(\tilde{x}|H_0)p(H_0)}{p(\tilde{x})} \\ &\propto p(\tilde{x}|H_0)\pi_0 \\ &\propto \int_{\theta \in H_0} p(\tilde{x}|\theta_0)p(\theta_0)d\theta\pi_0 \end{aligned}$$

  and similarly,

$$Pr(H_1|\tilde{x}) \propto \int_{\theta \in H_1} p(\tilde{x}|\theta)p_1(\theta)d\theta(1 - \pi_0)$$

- So how do we pick $p_1(\theta)$? In this course we will use $p_1(\theta) \sim N(\mu_1, \tau^2)$. So how do we pick the parameters of this distribution μ and τ^2?

- This point mass mixture prior can be written as:

$$p(\theta) = \pi_0 \delta(\theta = \theta_0) + (1 - \pi_0) p_1(\theta) \tag{1}$$

- Consider the ratio:

$$\frac{p(\tilde{x}|H_0)}{p(\tilde{x}|H_1)} = \frac{\int_{\theta \in H_0} p(\tilde{x}|\theta_0) p_0(\theta) d\theta}{\int_{\theta \in H_1} p(\tilde{x}|\theta) p_1(\theta) d\theta} = \left(\frac{p(H_0|\tilde{x})}{p(H_1|\tilde{x})}\right) / \left(\frac{p(H_0)}{p(H_1)}\right) \tag{2}$$

  This is known as a Bayes Factor.

- Recall the maximum-likelihood has a related form:

$$\frac{max : \theta \in H_0 \quad \mathcal{L}(\theta|\tilde{x})}{max : \theta \in H_1 \quad \mathcal{L}(\theta|\tilde{x})} \tag{3}$$

  In a likelihood ratio test we compare the difference for specific values of θ that maximize the ratio, whereas the Bayes factor (BF) integrates out the parameter values - in effect averages across the parameter space.

- In this example, let's choose $\mu_1 = \theta_1$ and set $\tau^2 = \psi^2$. Note $\bar{x}$ is a sufficient statistic, so we consider $p(\bar{x}|\theta)$.Then:

$$\begin{aligned}
BF &= \frac{\int_{\theta \in H_0} p(\bar{x}|\theta) p_{\theta_0}(\theta) d\theta}{\int_{\theta \in H_1} p(\bar{x}|\theta) p_1(\theta) d\theta} = \frac{\sqrt{n}/\sigma \exp\left(-\frac{(\bar{x}-\theta_0)^2}{2\sigma^2 n}\right)}{1/\sqrt{\sigma^2/n + \psi^2} \exp\left(-\frac{(\bar{x}-\theta_0)^2}{2(\sigma^2/n + \psi^2)}\right)} \\
&= \left(\frac{\sigma^2/n}{\sigma^2/n + \psi^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{(\bar{x}-\theta_0)^2}{\sigma^2/n} - \frac{(\bar{x}-\theta_0)^2}{\sigma^2/n + \psi^2}\right]\right) \\
&= \left(1 + \frac{\psi^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left[\frac{(\bar{x}-\theta_0)^2}{\sigma^2/n}\left(1 + \frac{\sigma^2}{n\psi^2}\right)^{-1}\right]\right) \\
&= \left(1 + \frac{\psi^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left[z^2\left(1 + \frac{\sigma^2}{n\psi^2}\right)^{-1}\right]\right).
\end{aligned}$$

  Note that $Pr(H_0|Data) = \left(1 + \frac{1-\pi_0}{\pi_0} BF^{-1}\right)$.

- Example. Let $\pi_0 = 1/2$, $\sigma^2 = \psi^2$, $N = 15$, $Z = 1.96$, plugging this all in we get BF = 0.66. This implies:

$$Pr(H_0|\bar{x}) = (1 + .66^{-1})^{-1} = 0.4.$$

- **Q:** Reject or not? **Q:** What is the corresponding p-value here?

- Consider the following scenarios with $z = 1.96$.

N	5	10	50	100	1000
$Pr(H_0\|\bar{x})$	.331	.367	.521	.600	.823

In each case the p-value is 0.05. Note that for a given effect size (ψ^2) the Bayes Factor is effect size calibrated. For a given effect size, a p-value goes to zero. Hence the disagreement between "practical significance" and "statistical significance".

- So in this case the relevant question is how to choose ψ^2. ψ is the distance we find meaningful for rejecting H_0. That is, if inferences about θ tend (with high probability) to be larger than ψ from θ_0 then reject.

- **Q:** What happens as $\psi^2 \to \infty$? Recall from our example:

$$BF = \left(1 + \frac{N\psi^2}{\sigma^2}\right)^{1/2} \exp\left(-1/2z^2\left[1 + \frac{\sigma^2}{n\psi^2}\right]^{-1}\right)$$

so the BF $\to \infty$. This implies that we need to put proper priors on the parameters when using BF. \item Consider two models: $M_1 \& M_2$, each with parameters sets $\Theta^{(M_1)}$ and $\Theta^{(M_2)}$. The Bayes Factor is:

$$BF = \frac{\int \mathcal{L}(\Theta^{(M_1)}|\tilde{x}) p_{M_1}(\theta^{(M_1)}) d\Theta^{(M_1)}}{\int \mathcal{L}(\Theta^{(M_2)}|\tilde{x}) p_{M_2}(\theta^{(M_2)}) d\Theta^{(M_2)}} \tag{4}$$

When $p_{M_1}(\Theta^{(M_1)}$ and $p_{M_2}(\theta^{(M_2)}$ are proper, the BF is well defined.

- **Q:** Can you ever specify an improper prior on any of the parameters in $\Theta^{(M_1)}$ and $\Theta^{(M_2)}$? Yes, so long as the parameter appears in both models (and that parameter has the same meaning in both models). Otherwise, the BF is meaningless if the parameter does not appear in both models.

Bayesian Regression

Linear modeling is an important element in a statistician's toolbox. We are going to discuss the impact of different priors versus the classical regression setting.

A common challenge in regression framework is variable selection (or model selection).

Q: How do you currently handle variable selection?

The Bayesian paradigm, through placing priors on the model space, provide a natural way to carryout model selection as well as model averaging.

Notation

In a regression framework the goal is to model the relationship between a variable of interest, y, and a set of covariates $\mathcal{X}$. Specifically, we are modeling the conditional expectation for y given a set of parameters $\mathcal{X}$, which can be formulated as:

$$E[y|\mathcal{X}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \tilde{\beta}^T \tilde{x}.$$

While this model is linear in the parameters, transformation of the covariates and basis functions give a great deal of flexibility and make linear regression a powerful tool.

Typically, the model is stated as:

$$y_i = \tilde{\beta}^T \tilde{x}_i + \epsilon_i$$

where the $\epsilon_i's$ are i.i.d. from a $N(0, \sigma^2)$ distribution. Recall, we could also think about regression with error terms from the $t-$ distribution as well.

Using the normal distributional assumptions then the joint distribution of the observed data, given the data $x_1, \ldots, x_n$ along with β and σ^2 can be written as:

$$\begin{aligned} p(y_1, \ldots, y_n | \tilde{x_1}, \ldots, \tilde{x_n}, \tilde{\beta}, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \tilde{x}_i, \tilde{\beta}, \sigma^2) && (5) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}^T \tilde{x}_i \right)^2 \right]. && (6) \end{aligned}$$

Note this is the same as the sampling, or generative model, that we have seen earlier in class.

Given our newfound excellence in linear algebra, the model is often formulated using matrix expressions and a multivariate normal distribution. Let

$$\tilde{y} | X, \tilde{\beta}, \sigma^2 \sim MVN(X\tilde{\beta}, \sigma^2 I), \quad (7)$$

where $\tilde{y}$ is an $n \times 1$ vector of the responses, X is an $n \times p$ matrix of the covariates where the i^{th} row is $\tilde{x}_i$, and I is a $p \times p$ identity matrix.

In a classical setting, typically least squares methods are used tom compute the values of the covariates in a regression setting. Note in a normal setting these correspond to maximum likelihood estimates. Specifically, we seek to minimize the sum of squared residuals (SSR), where $SSR(\tilde{\beta}) = \left(\tilde{y} - X\tilde{\beta}\right)^T \left(\tilde{y} - X\tilde{\beta}\right)$.

Thus we will take the derivative of this function with respect to β to minimize this expression.

$$\begin{aligned}
\frac{d}{d\tilde{\beta}} SSR(\tilde{\beta}) &= \frac{d}{d\tilde{\beta}} \left(\tilde{y} - X\tilde{\beta}\right)^T \left(\tilde{y} - X\tilde{\beta}\right) && (8)\\
&= \frac{d}{d\tilde{\beta}} \left(\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T X^T X\tilde{\beta}\right) && (9)\\
&= -2X^T\tilde{y} + 2X^T X\tilde{\beta} && (10)\\
\text{then set} = 0 \quad \text{which implies} \quad & X^T X\beta = X^t\tilde{y} && (11)\\
\text{and} \quad & \tilde{\beta} = (X^T X)^{-1} X^t\tilde{y}. && (12)
\end{aligned}$$

This value is the OLS estimate of $\tilde{\beta}_{OLS} = (X^T X)^{-1} X^T\tilde{y}$. Under the flat prior $p(\tilde{\beta}) \propto 1$, $\tilde{\beta}_{OLS}$ is the mean of the posterior distribution.

Bayesian Regression

As we have seen, the sampling distribution is:

$$\begin{aligned}
p(\tilde{y}|X, \tilde{\beta}, \sigma^2) &\propto \exp\left[-\frac{1}{2\sigma^2}(\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})\right] && (13)\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T X^T X\tilde{\beta}\right)\right] && (14)
\end{aligned}$$

Given this looks like the kernel of a normal distribution for $\tilde{\beta}$, we will consider a prior for $\tilde{\beta}$ from the normal family.

Let $\tilde{\beta} \sim MVN(\tilde{\beta}_0, \Sigma_0)$, then

$$\begin{aligned}
p(\tilde{\beta}|\tilde{y}, X, \sigma^2) &\propto p(\tilde{y}|X, \tilde{\beta}, \sigma^2) \times p(\tilde{\beta}) && (15)\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\tilde{\beta}^T X^T X\tilde{\beta} - 2\tilde{\beta}^T X^T\tilde{y}\right) - \frac{1}{2}\left(\tilde{\beta}^T\Sigma_0^{-1}\tilde{\beta} - 2\tilde{\beta}\Sigma_0^{-1}\tilde{\beta}_0\right)\right] && (16)\\
&\propto \exp\left[-\frac{1}{2}\left(\tilde{\beta}^T\left(\Sigma_0^{-1} + X^T X/\sigma^2\right)\tilde{\beta} - 2\tilde{\beta}\left(\Sigma_0^{-1}\tilde{\beta}_0 + X^t\tilde{y}/\sigma^2\right)\right)\right] && (17)
\end{aligned}$$

Thus, using the properties of the multivariate normal distribution from a previous chapter, we can identify the mean and variance of the posterior distribution. - $Var(\tilde{\beta}|\tilde{y}, X, \sigma^2) = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}$. - $E[\tilde{\beta}|\tilde{y}, X, \sigma^2] = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1} \times \left(\Sigma_0^{-1}\tilde{\beta}_0 + X^T\tilde{y}/\sigma^2\right)$

For a sanity check, let's look at the posterior distribution under a flat prior, $p(\tilde{\beta}) \propto 1$. Then $p(\tilde{\beta}|-) \sim N\left((X^TX)^{-1}X^t\tilde{y}, (X^TX)^{-1}\sigma^2\right)$. Note these are the OLS estimates.

We still need to consider a prior on σ^2. As we have seen in other scenarios the semi-conjugate prior is from the Inverse Gamma distribution. Let $\sigma^2 \sim IG(\nu_0/2, \nu_0\sigma_0^2/2)$ then

$$
\begin{aligned}
p(\sigma^2|-) &\propto p(\tilde{y}|X, \tilde{\beta}, \sigma^2)p(\sigma^2) && (18)\\
&\propto \left[(\sigma^2)^{-n/2}\exp(-SSR(\tilde{\beta})/2\sigma^2)\right] \times \left[(\sigma^2)^{-\nu_0/2-1}\exp(-\nu_0\sigma_0^2/2\sigma^2)\right] && (19)\\
&\propto \left(\sigma^2\right)^{-\frac{\nu_0+n}{2}-1}\exp\left(-\left[SSR(\tilde{\beta}) + \nu_0\sigma_0^2\right]/2\sigma^2\right) && (20)
\end{aligned}
$$

We recognize this distribution as an

$$IG\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + SSR(\tilde{\beta})}{2}\right)$$

.

Given the full conditional distributions, we need to sketch out a Gibbs sampler to take draws from the full conditional distributions.

1. Update $\tilde{\beta}$:
   (a) Compute $V = Var(\tilde{\beta}|\tilde{y}, X, \sigma^2)$ and $E = E[\tilde{\beta}|\tilde{y}, X, \sigma^{2(s)}]$.
   (b) Sample $\tilde{\beta}^{(s+1)} \sim MVN(E, V)$
2. Update σ^2:
   (a) Compute $SSR(\tilde{\beta}^{(s+1)})$
   (b) Sample $\sigma^{2(s+1)} \sim IG\left([\nu_0+n]/2, [\nu_0\sigma_0^2 + SSR(\tilde{\beta}^{(s+1)})]/2\right)$.

Exercise

Write code to take posterior samples in a regression setting. Note we will talk more about priors next, but now go ahead and use the values in the script below

```
# Simulate Data
set.seed(10262018)
p <- 2
n <- 100
beta <- c(10, 3)
X <- cbind(rep(1,n),rnorm(n))
sigmasq <- 1
y <- X %*% beta + rnorm(n, mean = 0, sd = sqrt(sigmasq))
reg.df <- data.frame(y= y, x = X[,2])
ggplot(data = reg.df, aes(y=y, x=x)) + geom_point() + geom_smooth(method = 'lm')

# Initialization and Prior
num.mcmc <- 1000
beta.0 <- rep(0,2)
Sigma.0 <- diag(2) * 100
nu.0 <- .01
sigmasq.0 <- .01
beta.samples <- matrix(0, nrow = num.mcmc, ncol = p)
sigmasq.samples <- rep(1, num.mcmc)
```

```
for (iter in 2:num.mcmc){
  # sample beta
  beta.samples[iter,] <- beta

  # sample sigmasq
  sigmasq.samples[iter] <- sigmasq
}

# Look at trace plots
```

Priors for Bayesian Regression

For this model we need to think about priors for σ^2 and $\tilde{\beta}$. From similar situations, we have a good handle on how to think about the prior on σ^2. In particular, the Inverse-Gamma distribution gives us a semi-conjugate prior and the parameterization using ν_0 and σ_0^2 provides and intuitive way to think about the parameters in this distribution. Recall, $p\tilde{\beta} = N(\tilde{\beta}_0, \Sigma_0)$. The challenge is how to come up for values for $\tilde{\beta}_0$ and Σ_0.

In an applied setting, often some information is available about the potential magnitude of the covariates. This allows reasonable values for $\tilde{\beta}_0$ and the variance components in Σ_0, but still requires some thought for the covariance terms in Σ_0. As p, the number of covariates, increases this becomes more and more difficult.

The textbook discusses the *unit information prior*, that injects information proportional to a single observation - similar to how we have used ν_0 and η_0 previously in the course. One popular prior from this principle is known as Zellner's g-prior, where $\Sigma_0 = g\sigma^2(X^TX)^{-1}$. Using Zellner's g-prior the marginal distribution $p(\sigma^2|\tilde{y}, X)$ can be derived directly. This allows the conditional draws from $p(\tilde{\beta}|\sigma^2, \tilde{y}, X)$ in a similar fashion to our first normal settings when $p(\beta) = N(\mu_0, \sigma^2/\kappa_0)$.

Other common strategies are to use a weakly informative prior on Σ_0, say $\Sigma_0 = \tau_0^2 \times I_p$.

Bayesian Modeling and Regularization

Ordinary Least Squares (OLS) regression can be written as:

$$\hat{\tilde{\beta}}_{OLS} = \arg\min_{\hat{\tilde{\beta}}} \; ||\tilde{y} - X\hat{\tilde{\beta}}||_2^2 \rightarrow \hat{\tilde{\beta}} = (X^TX)^{-1}X^T\tilde{y},$$

where $||\tilde{x}||_p = (|x_1|^p + \cdots + |x_m|^p)^{1/p}$ is an LP norm. So the L2 norm is $||\tilde{x}||_2 = \sqrt{x_1^2 + \ldots x_m^2}$.

Recall ridge regression is a form of penalized regression such that:

$$\hat{\tilde{\beta}}_R = \arg\min_{\hat{\tilde{\beta}}_R} ||\tilde{y} - X\hat{\tilde{\beta}}||_2^2 + \lambda||\hat{\tilde{\beta}}_R||_2^2 \rightarrow \hat{\tilde{\beta}}_R = \left(X^T X + \lambda I\right)^{-1} X^T \tilde{y}, \tag{21}$$

where λ is a tuning parameter that controls the amount of shrinkage.

- As λ gets large all of the values are shrunk toward 0.

- As λ goes to 0, the ridge regression estimator results in the OLS estimator.

- It can be shown that ridge regression results better predictive ability than OLS by reducing variance of the predicted values at the expense of bias. Note that typically the X values are assumed to be standardized, so that the intercept is not necessary.

- **Q:** How do we choose λ?

An alternative form of penalized regression is known as Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO uses an L1 penalty such that:

$$\hat{\tilde{\beta}}_L = \arg\min_{\hat{\tilde{\beta}}} ||\tilde{y} - X\hat{\tilde{\beta}}||_2^2 + \lambda||\hat{\tilde{\beta}}_L||_1, \tag{22}$$

the L1 penalty results in $||\tilde{x}||_1 = |x_1| + \cdots + |x_m|$, which minimizes the absolute differences.

The nice feature of LASSO, relative to ridge regression, is that coefficients are shrunk to 0 providing a way to do *variable selection.*

One challenge with LASSO is coming up with proper distributional assumptions for inference about variables.

Consider the following prior $p(\tilde{\beta}) = N(0, I_p\tau^2)$. How does this relate to ridge regression? First compute the posterior distribution for $\tilde{\beta}$.

$$\begin{aligned} p(\tilde{\beta}|-) &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\tilde{\beta}^T X^T X\tilde{\beta} - \frac{1}{\sigma^2}\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T\frac{I_p}{\tau^2}\tilde{\beta}\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\tilde{\beta}^T\left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)\tilde{\beta} - \frac{1}{\sigma^2}\tilde{\beta}^T X^T\tilde{y}\right)\right] \end{aligned}$$

Thus $Var(\tilde{\beta}|-) = \left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)^{-1}\sigma^2$ and $E(\tilde{\beta}|-) = \left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)^{-1} X^t\tilde{y}$.

Q: does this look familiar?

Define: $\lambda = \frac{\sigma^2}{\tau^2}$. How about now? This is essentially the ridge regression point estimate.

Note that in a similar fashion we can use a specific prior on $\tilde{\beta}$ to achieve LASSO properties. It is also important to clarify the differences between classical ridge regression (and Lasso) with the Bayesian analogs. In the Bayesian case we can still easily compute credible intervals to account for the uncertainty in our estimation. Interval calculations for inference are difficult in the these settings, particularly for Lasso.

Bayesian Model Selection

Recall, we discussed common model selection techniques:

- All subsets (use aic, bic) works for moderate p
- Backward selection
- Forward selection
- Backward - Forward Selection
- Cross Validation

However, in a Bayesian framework we have a coherent way to talk about model selection. Specifically, given a prior on the model space we can compute posterior probability for a given model.

In model selection for linear regression the goal is to decide which covariates to include in the model.
- To do this, we introduce a parameter $\tilde{z}$, where $z_i = 1$ if covariate i is in the model, otherwise $z_i = 0$.

- Then define $\beta_i = z_i \times b_i$. Note the b_i' are the real-values regression coefficients. For now we will ignore the intercept (standardizing the covariates).

- The regression equation now becomes:

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i. \tag{23}$$

- Again thinking about the regression model as a conditional expectation, then for $p = 3$:

$$\begin{aligned} E[y|\tilde{x}, \tilde{b}, \tilde{z} = (1,0,1)] &= b_1 x_1 + b_3 x_3 \\ E[y|\tilde{x}, \tilde{b}, \tilde{z} = (0,1,0)] &= b_2 x_2. \end{aligned}$$

Note that the vector $\tilde{z}_a$ defines a model and is interchangeable with the notation M_a.

- Now the goal is a probabilistic statement about $\tilde{z}_a$, specifically:

$$Pr(\tilde{z}_a|\tilde{y}, X) = \frac{p(\tilde{y}|X, \tilde{z}_a)p(\tilde{z}_a)}{\int_{z^*} p(\tilde{y}|X, \tilde{z}^*)p(\tilde{z}^*)d\tilde{z}^*}, \tag{24}$$

where X is a matrix of observed covariates. Of course, this requires a prior on z_a, which we will see momentarily.

- For model comparison between model a and model b, an alternative way to express this is through the following (familiar) ratio:

$$BF(a,b) = \frac{p(\tilde{y}|X, \tilde{z}_a)}{p(\tilde{y}|X, \tilde{z}_b)} = \left(\frac{Pr(\tilde{z}_a|\tilde{y}, X)}{Pr(\tilde{z}_b|\tilde{y}, X)}\right) / \left(\frac{p(\tilde{z}_a)}{p(\tilde{z}_b)}\right). \tag{25}$$

Of course this is a Bayes Factor.

Now the question is, how do we think about the priors for $\tilde{z}$ or equivalently for M_a?

The textbook does not explicitly discuss this, but a couple common parameters are the discrete uniform prior, where each model has the same prior probability. In terms of the $z_i's$ this would be equivalent to prior inclusion probability of 0.5. In other situations, a prior can be placed on the total number of parameters in the model.

Bayesian Model Comparison

The posterior probability for model a is a function of the prior $p(z_a)$ and $p(\tilde{y}|X, \tilde{z}_a)$ which is known as the marginal probability.

In a regression setting the marginal probability is computed by integrating out the parameters as:

$$\begin{aligned} p(\tilde{y}|X, \tilde{z}_a) &= \int\int p(\tilde{y}, \tilde{\beta}_a, \sigma^2|X, \tilde{z}_a) d\tilde{\beta}_a d\sigma^2 \\ &= \int\int p(\tilde{y}|\tilde{\beta}_a, X) p(\tilde{\beta}_a|\tilde{z}_a, \sigma^2) p(\sigma^2) d\tilde{\beta}_a d\sigma^2, \end{aligned}$$

where $\tilde{\beta}_a$ is a $p_{z_a} \times 1$ vector containing the p_{z_a} elements in M_a.

In general this integration is very difficult, particularly when p, the dimension of $\tilde{\beta}$, is large. However, recall Zellner's g-prior had a form that facilitates efficient integration. Under this prior $p(\tilde{\beta}_a|\sigma^2, \tilde{z}_a) = MVN_{p_{z_a}}\left(\tilde{\beta}_0, g\sigma^2(X^TX)^{-1}\right)$.

It can be shown that integrating out $\tilde{\beta}_a$, $\int p(\tilde{y}|X, \tilde{z}_a, \sigma^2) = p(\tilde{y}|\tilde{\beta}_a, X) p(\tilde{\beta}_a|\tilde{z}_a, \sigma^2) d\tilde{\beta}_a$ is fairly straightforward.

This leaves:

$$p(\tilde{y}|X, \tilde{z}_a) = \int p(\tilde{y}|X, \tilde{z}_a, \sigma^2) p(\sigma^2) d\sigma^2.$$

Due to the form of the priors, this can also be easily integrated such that the marginal probability is:

$$p(\tilde{y}|X, \tilde{z}_a) = \pi^{-n/2} \frac{\Gamma([\nu_0 + n]/2}{\Gamma(\nu_0/2} (1+g)^{-p_{z_a}/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}}, \tag{26}$$

where $SSR_g^z = \tilde{y}^T(I_{p_z} - \frac{g}{g+1} X(X^TX)^{-1}X^T)\tilde{y}$.

Given the marginal likelihoods, we can compute the posterior probability of a given model, M_a as:\

$$Pr(M_a = \tilde{y}, X) = Pr(\tilde{z}_a|\tilde{y}, X) = \frac{p(\tilde{y}|X, \tilde{z}_a) p(\tilde{z}_a)}{\int_{z^*} p(\tilde{y}|X, \tilde{z}^*) p(\tilde{z}^*) d\tilde{z}^*}. \tag{27}$$

Using this formulation we can choose the most probable model or a set of the most probable models.

Bayesian Model Averaging

A powerful tool in Bayesian modeling, particularly for predictive settings, is Bayesian model averaging. Rather than choosing a single model, or set of models, we will average across models according to their posterior probability.

Assume we are interested in some quantity of interest Δ, which can be computed as a function of the posterior distribution. Then:

$$p(\Delta|X,\tilde{y}) = \sum_{i=1}^{k} p(\Delta|M_k, X, \tilde{y}) Pr(M_k|X,\tilde{y}) \tag{28}$$

For example let Δ represent the posterior predictive distribution for $\tilde{y}^*$, given X^*. Then the model averaged posterior predictive distribution can be written as,

$$p(\tilde{y}^*|, X^* X, \tilde{y}) = \sum_{i=1}^{k} p(\tilde{y}^*|X^*, M_k, X, \tilde{y}) Pr(M_k|X,\tilde{y}). \tag{29}$$

This model averaged prediction is a special type of an ensemble method, which have nice predictive properties *and* in this case account for uncertainty in the model selection process.

General Model Selection and Averaging

In many cases, the g-prior framework is too restrictive or the models will not allow closed form solutions for the marginal likelihood. For instance consider the following model:

$$\tilde{y}|- \sim N(XB, \sigma^2 H(\phi) + \tau^2 I), \tag{30}$$

where $H(\phi)$ is a correlation matrix. Finding the marginal (integrated) likelihood analytically would be very difficult in this case. It turns out this model is often used in Bayesian spatial modeling. By introducing an infinite-dimensional Gaussian distribution known as a Gaussian Process (GP), a posterior predictive distribution can be computed for any point in space. This is a Bayesian analog to Kriging.

In situations like this, where model selection is often conducted using MCMC. The basic idea is that each iteration you:

- Update each z_i
- Sample σ^2
- Sample $\tilde{\beta}|\tilde{z}$,

where for each z_i the $Pr(z_i = 1|\tilde{y}, X, \tilde{z}_{-i})$ can be computed and $\tilde{z}_{-i}$ is all of the elements excluding i.