Bayesian Modeling
Regression models

Yixin Wang

Preliminary Draft.
Please do not distribute.

9 Linear regression

9.1 Linear regression model

Regression problem is concerned with the relationship between a response variable Y and a collection of explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_p)$.

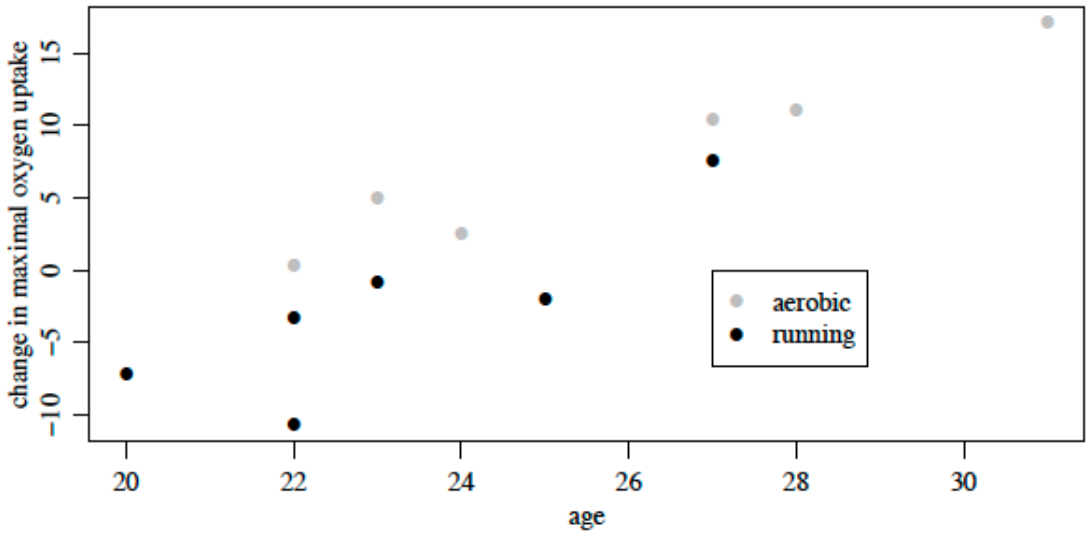


Figure 9.1: Change in maximal oxygen uptake as a function of age and exercise program.

Example 9.1. Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. The maximum oxygen uptake (liters per minute) of each subject was measured while running on an inclined treadmill, both before and after the program. See Fig. 9.1

A linear regression model assumes that $\mathbb{E}[Y|\boldsymbol{x}]$ takes a linear form:

def. of expectation

$$\mathbb{E}[Y|\boldsymbol{x}] = \int y p(y|\boldsymbol{x}) \mathrm{d}y = \beta_1 x_1 + \ldots \beta_p x_p = \boldsymbol{\beta}^\top \boldsymbol{x}.$$

In the above example, the explanatory variables (covariates) $\boldsymbol{x}$ may be taken to be

$$\begin{aligned} x_1 &= 1 \\ x_2 &= 0 \quad \text{if the subject is on the running program,} \quad 1 \quad \text{if on aerobic} \\ x_3 &= \text{age of subject} \\ x_4 &= x_2 \times x_3. \end{aligned}$$

We have not specified the distribution $p(y|\boldsymbol{x})$ beyond its conditional expectation. The *normal linear regression model* posits that in addition to $\mathbb{E}[Y|\boldsymbol{x}]$ being linear, the sampling variability around the mean is in fact i.i.d. from normal distribution:

$$\begin{aligned} \epsilon_1, \ldots, \epsilon_n &\sim \text{normal}(0, \sigma^2) \\ Y_i &= \boldsymbol{\beta}^\top \boldsymbol{x}_i + \epsilon_i. \end{aligned}$$

This gives the *conditional* likelihood, given the n-sample (noting that nothing is said about the marginal distribution of covariates $\boldsymbol{x}$:

$$\begin{aligned} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^n p(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^2 \Big\}. \end{aligned}$$

In customary matrix notations: $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is a $n \times 1$ column vector; $\boldsymbol{X}$ is the $n \times p$ design matrix whose ith row is $\boldsymbol{x}_i$. Then the above can be written as

$$\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I),$$

where I is the $n \times n$ identity matrix.

find β that predicts best

Parameter vector $\boldsymbol{\beta}$ may be estimated by minimizing the sum of squared residuals, SSR($\boldsymbol{\beta}$):

$$\begin{aligned} \text{SSR}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

To minimize the above expression, we take derivative with respect to $\boldsymbol{\beta}$ and set it to zero:

$$-2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = 0$$

resulting in

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

The value $\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is called the "ordinary least squares" (OLS) estimate of $\boldsymbol{\beta}$. This value is unique as long as the $p \times p$ matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is of full rank (and thus invertible). This happens when $n \geq p$ (and the columns of the design matrix $\boldsymbol{X}$ are linearly independent). The OLS estimate is a frequentist estimate, but it also plays a role in Bayesian estimation.

9.2 Semi-conjugate priors

The (conditional) likelihood function takes the form

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) &\propto \exp -\frac{1}{2\sigma^2}\text{SSR}(\boldsymbol{\beta}) \\ &= \exp -\frac{1}{2\sigma^2}(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

It is simple to see that a normal distribution can be used as a semi-conjugate prior for $\boldsymbol{\beta}$. Let $\boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{\beta}_0, \Sigma_0)$ a priori, then

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) &\propto p(\boldsymbol{\beta}) \times p(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta},\sigma^2) \\ &\propto \exp -\frac{1}{2}(-2\boldsymbol{\beta}^\top\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top\Sigma_0^{-1}\boldsymbol{\beta}) \times \exp -\frac{1}{2}(-2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y}/\sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}/\sigma^2) \\ &= \exp\{\boldsymbol{\beta}^\top(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top\boldsymbol{y}/\sigma^2) - \frac{1}{2}\boldsymbol{\beta}^\top(\Sigma_0^{-1} + \boldsymbol{X}^\top\boldsymbol{X}/\sigma^2)\boldsymbol{\beta}\}. \end{aligned}$$

This is a multivariate normal density with

$$\text{Var}[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2] = (\Sigma_0^{-1} + \boldsymbol{X}^\top\boldsymbol{X}/\sigma^2)^{-1}, \tag{50a}$$

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2] = (\Sigma_0^{-1} + \boldsymbol{X}^\top\boldsymbol{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top\boldsymbol{y}/\sigma^2). \tag{50b}$$

It is a simple exercise to see that the posterior expectation represents a combination of the prior expectation and the purely data driven estimate OLS.

It is also simple to see that the inverse-gamma distribution can be used as a semi-conjugate prior for σ^2. Let $\gamma = 1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$ a priori, then

$$\begin{aligned} p(\gamma|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) &\propto p(\gamma)p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \gamma) \\ &\propto \left[\gamma^{\nu_0/2-1}\exp(-\gamma\nu_0\sigma_0^2/2)\right] \times \left[\gamma^{n/2}\exp(-\gamma \times \text{SSR}(\boldsymbol{\beta})/2)\right] \\ &\propto \text{gamma}((\nu_0 + n)/2, (\nu_0\sigma_0^2/2 + \text{SSR}(\boldsymbol{\beta})/2)). \end{aligned}$$

A Gibbs sampler is simple to implement. Each Gibbs update consists of the following: given the current values $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$, for $s = 1, 2, \ldots$:

1. update $\boldsymbol{\beta}^{(s+1)} \sim \text{N}_p(\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}], \text{Var}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}])$.
2. update $\sigma^{2(s+1)} \sim \text{inverse-gamma}((\nu_0 + n)/2, (\nu_0\sigma_0^2/2 + \text{SSR}(\boldsymbol{\beta}^{(s+1)})/2))$.

9.3 Objective priors

In regression analysis it may be difficult to come up with a suitable prior distribution on $\boldsymbol{\beta}$ and σ^2.

Example 9.2. Continuing on the oxygen uptake example. Suppose we know from our prior knowledge (e.g., by consulting with experts on physiology) that males in their 20s have an oxygen uptake of around 150 liters per minute with a std of 15. We then take $150 \pm 2 \times 15 = (120, 180)$ as the prior expected range of oxygen uptake distribution, and so the changes in the oxygen uptake lies within $(-60, 60)$ with high probability.

Consider our subjects in the running group. This means the line $\beta_1 + \beta_3 x$ should produce values between -60 and 60 for all values of x between 20 and 30. A little algebra shows that we need a prior distribution on β_1 and β_3 so that $\beta_1 \in (-300, 300)$ and $\beta_3 \in (-12, 12)$ with high probability. From here we can find the suitable prior hyper-parameters $\boldsymbol{\beta}_0, \Sigma_0$. But this type of calculation becomes difficult when there are more explanatory variables. □

When we are in such a scenario, i.e., when it is difficult to come up with an informative prior specification, then one may consider prior specification that contains as little information as possible. This is the spirit of objective Bayes. [11] For linear regression, there are a number of objective priors that are commonly used in practice.

[11] We encountered this notion for the first time when we was discussing improper priors in Section 5. The ideas behind the derivation of both improper prior and unit information prior are basically the same, but the latter has the advantage of being proper.

Data driven prior

Unit information prior A unit information prior is one that contains the same amount of information as that would be contained in a single observation (Kass and Wasserman, 1995).

Recall $\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. Since $\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$, this implies that the variance (with $\boldsymbol{\beta}$ held fixed) of $\hat{\boldsymbol{\beta}}_{ols}$ is $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

The precision of $\hat{\boldsymbol{\beta}}_{ols}$ is its inverse variance: $(\boldsymbol{X}^\top \boldsymbol{X})/\sigma^2$. Viewing this as the amount of information contained in n observations, the amount of information in one observation should be $1/n$ as much. Thus, we set

$$\Sigma_0^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})/(n\sigma^2).$$

To complete the prior specification $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \Sigma_0)$, we set $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{ols}$.

In a similar way, the prior distribution of σ^2 is given by $\sigma^2 \sim$ inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$, where $\nu_0 = 1$ and $\sigma_0^2 := \hat{\sigma}^2_{\text{ols}}$, which is obtained as an unbiased estimate of σ^2:

$$\hat{\sigma}^2_{\text{ols}} = \text{SSR}(\hat{\boldsymbol{\beta}}_{ols})/(n-p).$$

Some remarks

- the unit information prior is not purely Bayesian, since the prior is derived from the data. It provides some sort of protection against misleading prior specification.
- however, it uses only a very small amount of the information gleaned from the data due to suitable scale $1/n$ of information. Thus, its influence on the posterior inference is expected to be weak.

g-prior g-prior is another popular choice proposed by Arnold Zellner. It is motivated from another principle of objective Bayesian statistics: the relevant distributions of interest should remain invariant to changes in parameterization of the model.[12]

Example 9.3. Continue on the regression model for oxygen uptake. Suppose that someone were to analyze the data using explanatory variable $\tilde{x}_3$ = age in months, instead of x_3 = age in years. The role of this variable in the model for the response Y is in the linear term $\tilde{\beta}_3\tilde{x}_3$, as opposed to $\beta_3 x_3$. Since now $\tilde{x}_3 = 12 \times x_3$, it makes sense that the posterior distribution for $12 \times \tilde{\beta}_3$ in the model with $\tilde{x}_3$ should be the same as the posterior distribution for β_3 based on the model with x_3.

For many modelers, due to the lack of domain knowledge, the same form of prior specification may be given to $\tilde{\beta}_3$ as would be the case for β_3. Thus, it is important the impart the kind of prior so that the posterior inference is robust against such rescaling in the explanatory variables. □

posterior($\tilde{\beta}_3 \tilde{x}_3$) = posterior($\beta_3 x_3$)

one example of parameterization of model.

[12] Jeffreys' prior is another example.

Let us proceed to a formulation of the g-prior that arises in the normal linear regression model.

- Suppose $\boldsymbol{X}$ is the given $n \times p$ design matrix. Under this design,

$$\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I).$$

- Alternatively, due to a change of explanatory variables, $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{H}$ is a modified design matrix, for some $p \times p$ matrix $\boldsymbol{H}$. Under this design,

$$\boldsymbol{y}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\beta}}, \sigma^2 \sim \mathrm{N}_n(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 I) = \mathrm{N}_n(\boldsymbol{X}\boldsymbol{H}\tilde{\boldsymbol{\beta}}, \sigma^2 I).$$

- We need the *same* conditional prior on $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ (*conditionally given* $\boldsymbol{X}$ *or* $\tilde{\boldsymbol{X}}$) such that under such prior specification, the posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{H}\tilde{\boldsymbol{\beta}}$ are equal for all $\boldsymbol{H}$:

$$[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2] \overset{d}{=} [\boldsymbol{H}\tilde{\boldsymbol{\beta}}|\boldsymbol{y}, \tilde{\boldsymbol{X}}, \sigma^2]. \tag{51}$$

Suppose the prior is of the form $\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{\beta}_0, \Sigma_0)$. Recall from Eq. (50) the posterior distribution $\boldsymbol{\beta}$ is a multivariate normal with

$$\mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2)^{-1}, \tag{52a}$$

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{y}/\sigma^2). \tag{52b}$$

The answer is that if we put $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = g\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, where $g > 0$ is an arbitrary constant, then the invariance property expressed in Eq. (51) is satisfied (It is a straightforward exercise to verify this.)

- to be clear, we set the prior for $\boldsymbol{\beta}$ as $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$. The prior for $\tilde{\boldsymbol{\beta}}$ would be of the form $\tilde{\boldsymbol{\beta}} \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1})$.
- in fact,

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2] &= (\boldsymbol{X}^\top \boldsymbol{X}/(g\sigma^2) + \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2)^{-1}, \\
&= \frac{g}{g+1}\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&=: \boldsymbol{V}; \\
\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2] &= (\boldsymbol{X}^\top \boldsymbol{X}/(g\sigma^2) + \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2)^{-1}(\boldsymbol{X}^\top \boldsymbol{y}/\sigma^2) \\
&= \frac{g}{g+1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\
&= \frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}} \\
&=: \boldsymbol{m}.
\end{aligned}$$

In short,

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2 \sim \mathrm{N}_p(\boldsymbol{m}, \boldsymbol{V}). \tag{53}$$

- For σ^2, suppose that an inverse-gamma prior is given: $\sigma^2 \sim$ inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$. It is a very nice feature of g-prior that the induced posterior distribution of σ^2 is again an inverse-gamma distribution (Exercise: verify this):

$$[\sigma^2|\boldsymbol{y}, \boldsymbol{X}] \sim \text{inverse-gamma}((\nu_0 + n)/2, (\nu_0\sigma_0^2 + \text{SSR}_g)/2), \tag{54}$$

  where the term

$$\text{SSR}_g := \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{m}^\top\sigma^2\boldsymbol{V}^{-1}\boldsymbol{m} = \boldsymbol{y}^\top(\mathbf{I} - \frac{g}{g+1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}. \tag{55}$$

  when $g \to \infty$, this term tends to the SSR corresponding to the OLS estimate $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

- We observe a form of shrinkage for both parameters $\boldsymbol{\beta}$ and σ^2.
- MCMC is not needed, as we can obtain Monte Carlo samples for $(\sigma^2, \boldsymbol{\beta})$ from the above computation.

Example 9.4. Back to our example of regression analysis of the oxygen uptake data.

Set the g-prior with $g = n = 12$, $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{ols}} = 8.54$. The posterior mean for $\boldsymbol{\beta}$ does not depend on σ^2 and can be computed directly. The posterior standard deviations of these parameters are obtained. Some observations:

- the posterior distributions seem to suggest only weak evidence of a difference between the two groups, as the 95% quantile-based posterior intervals for β_2 and β_4 both contain zero.
- however, there seems to be a relatively strong evidence on the effect of age. According to our model, the average difference in y between two people of the same age x but in different training programs is $\beta_2 + \beta_4 x$. The box plots of the posterior distribution of this quantity is given for each x. It suggests a strong evidence of a difference at young ages, but less so at the older ones.

$$E(Y \mid x, \beta, \sigma^2) = \beta^T x$$

$x_1 = 1$

$x_2 = \{0, 1\}$

$x_3 = \text{age} = x$

$x_4 = x_2 x_3$

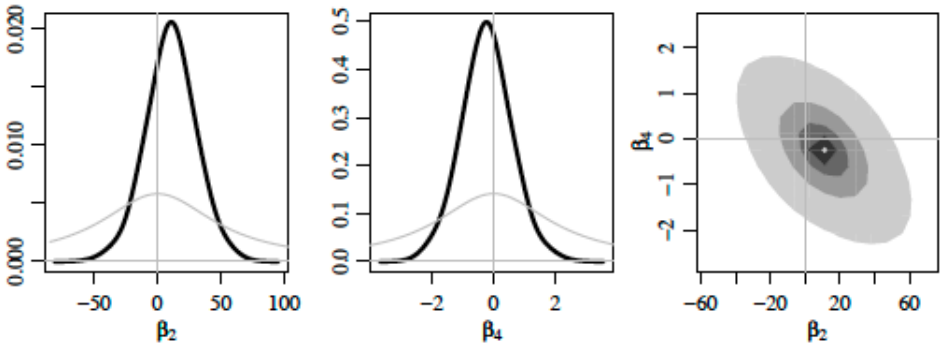


Figure 9.2: Posterior distributions of β_2 and β_4, with the marginal prior distributions in gray.

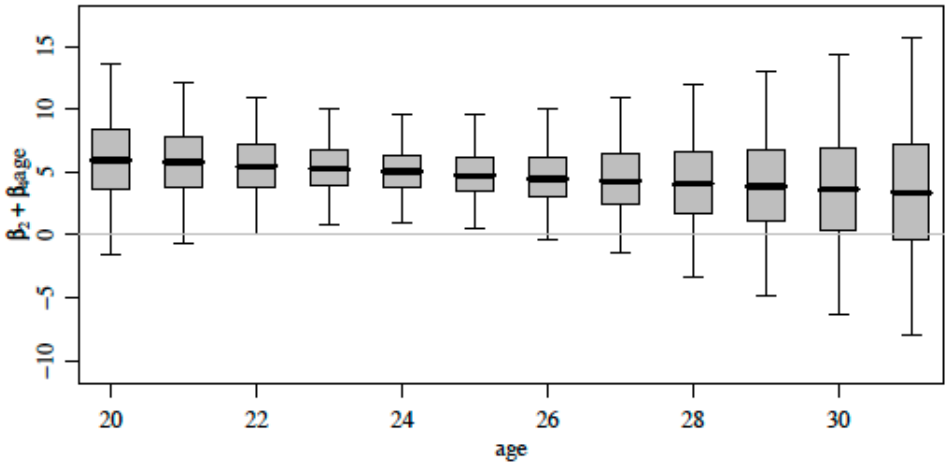


Figure 9.3: Ninety-five percent confidence intervals for the difference in expected change scores between aerobic subjects and running subjects.

For more details, see Hoff (2009).

9.4 Model selection

too little data, too complex model

In regression problems we may encounter a large number of possible explanatory variables/ regressors $x_1, \ldots, x_p$, many of which may be irrelevant to the response variable y. Although we may fit a regression model with all such potential regressors, such a technique will likely produce a poor result in terms of both prediction and parameter estimation, due to overfitting. Thus, selecting only the most relevant subset of variables x_i's for predictive and interpretative purposes is an extremely important task. The broad term for this task is called "model selection".

Example 9.5. (Diabetes data) There are ten variables $x_1, \ldots, x_{10}$ on a group of $= 442$ diabetes patients, and a variable y representing the disease progression taken one year after the baseline measurements x_i's.

It is suspected that the relationship between x_i's and y may be nonlinear, so a common practice is utilize a linear regression model using regressors $x_1, \ldots, x_{10}$ (a.k.a. *main effects*), as well as nonlinear terms that represent the *interactions* between the main effects, namely $x_j x_k$, and the quadratic terms x_j^2 for $j, k = 1, \ldots, 10$. One of regressors, $x_2 =$ sex, is binary so x_2^2 is unnecessary.

This gives a total of $p = 10 + \binom{10}{2} + 9 = 64$ potential regressors among

$$\{x_j, x_j^2, x_j x_k\}$$

Naive OLS approach Randomly split the 442 diabetes subjects into 342 *training samples* and 100 *test samples*, resulting in training data set $(\boldsymbol{y}, \boldsymbol{X})$ and test set $(\boldsymbol{y}_{\text{test}}, \boldsymbol{X}_{\text{test}})$.

Apply the OLS approach to the training data with all 64 regressors to obtain $\hat{\boldsymbol{\beta}}_{\text{ols}}$ (cf. Section 9.1), and then generate the predictive responses $\hat{\boldsymbol{y}}_{\text{test}} = \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{ols}}$.

The average sequared preditive error is $\frac{1}{100}\|\boldsymbol{y}_{\text{test}} - \hat{\boldsymbol{y}}_{\text{test}}\|^2 = 0.67$. This is not good, since if we simply put the predicted responses to be zero, our predictive error would already be $\frac{1}{100}\|\boldsymbol{y}_{\text{test}}\|^2 = 0.97$.

always important baseline to consider

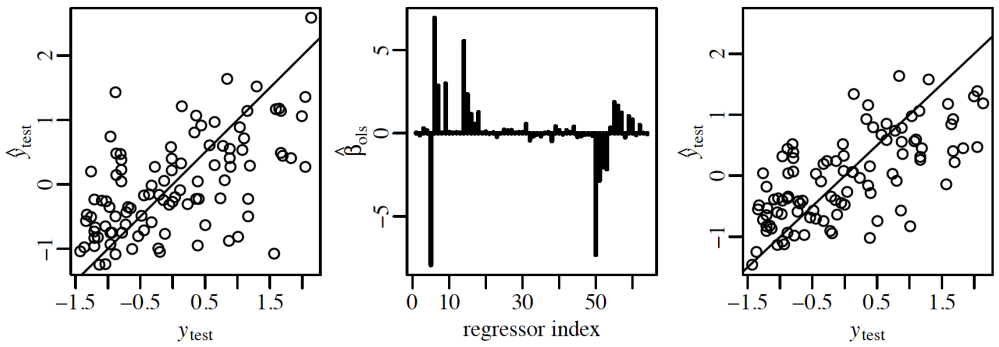


Figure 9.4: Left and middle panels: Predicted values and regression coefficients for the diabetes data via OLS. Right panel: Results based on a backwards elimination procedure.

The second panel shows that most of the estimated regression coefficients are quite small — this suggests we should remove them. A simple way is a greedy procedure known as *backwards elimination*.

Backwards elimination procedure This is a sequential procedure for assessing the relevance of the regression coefficients based on the current model's fit, and eliminating one variable at a time.

A standard way of assessing the evidence that the true value of coefficient β_j is non-zero is via a t-statitic, which is obtained by dividing the OLS estimate $\hat{\beta}_j$ by its standard error. Since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$, and $\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta} \sim \mathrm{N}_n(0, \sigma^2 \boldsymbol{I})$, we put

$$t_j = \frac{\hat{\beta}_j}{(\hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj})^{1/2}}.$$

(Note: $\hat{\sigma}^2$ is the corresponding OLS estimate of the residual variance σ^2. Also, the response vector $\boldsymbol{y}$ and all columns of $\boldsymbol{X}$ have been centered to have mean zero.)

Now, if $|t_j|$ is below a certain cutoff threshold, $|t_j| < t_{\text{cutoff}}$, then the evidence for $\beta_j \neq 0$ is weak; variable x_j is removed from the model.

A version of the overall backwards elimination procedure is as follows

1. Obtain OLS estimate $\hat{\boldsymbol{\beta}}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| \leq t_{\text{cutoff}}$,
   a) find the regressor j that has the smallest value of t_j and remove column j from $\boldsymbol{X}$.
   b) return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all j, then stop.

Example 9.6. Apply this procedure to diabetes data, using $t_{\text{cutoff}} = 1.65$ (corresponding roughly to a p-value of $2 \times 0.05 = 0.10$ according to a t distribution with a very large number of degrees of freedom, or the standard normal distribution). We obtain that 44 of the 64 variables are eliminated, leaving 20 variables in the regression model. The third plot of Fig. 9.4 shows $\hat{\boldsymbol{y}}_{\text{test}}$ according to the reduced-model regression coefficients. The prediction error for the model is 0.53, which is an improvement from the standard OLS error of 0.67.

The backwards elimination procedure described above is a fast heuristic, but it may pick up many spurious associations between selected x_j's and y.

Example 9.7. Let's consider the following experiment: we create a new data vector $\tilde{\boldsymbol{y}}$ by randomly permuting the values of $\boldsymbol{y}$. Thus, the value of $\boldsymbol{x}_i$ has no effect on $\tilde{y}_i$. There is no true association between $\tilde{\boldsymbol{y}}$ and the columns of $\boldsymbol{X}$. The left figure of Fig. 9.5 shows the t-statistics for one randomly generated $\tilde{\boldsymbol{y}}$ of $\boldsymbol{y}$. Initially, only one regressor has a t-statistic greater than 1.65, but as we sequentially remove the columns of $\boldsymbol{X}$, the estimated variance of the remaining regressors decreases and their t-statistics increase in value. With $t_{\text{cutoff}} = 1.65$, the procedure arrives at a regression model with 18 regressors. See the illustration in the right panel. All such regressors are spurious, of course.

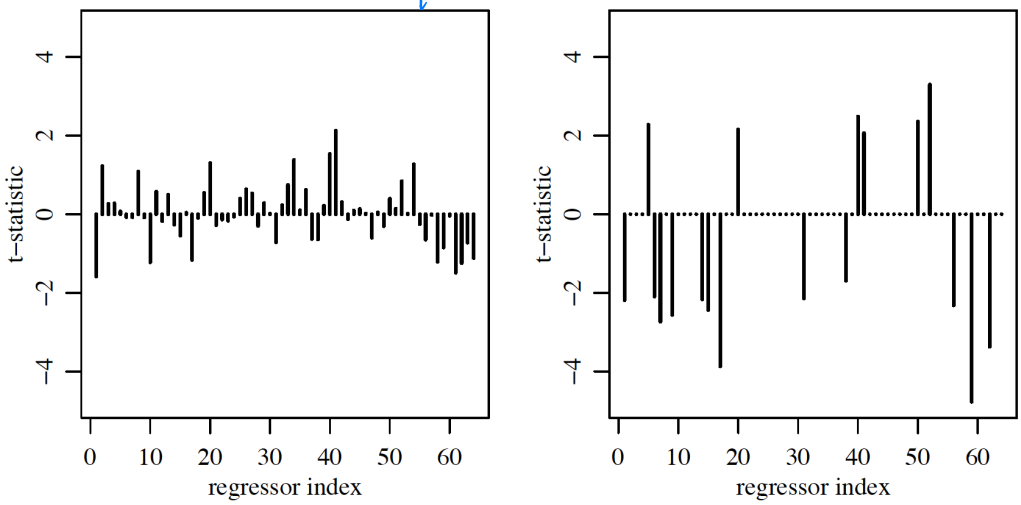


Figure 9.5: t-statistics for the regression of $\tilde{y}$ on $\boldsymbol{X}$, before and after backwards elimination.

9.4.1 Bayesian model comparison

The Bayesian approach is conceptually straightforward: we do not know which variables are spurious or not; such information will be represented by random variables (parameters) which are then endowed with some prior distributions. The model selection problem is essentially no different from the inference of an unknown parameter(s).

Let $z_j = 0$ if the explanatory variable x_j is spurious and $z_j = 1$ otherwise (that is, if x_j is *active*). We may express the regression coefficients as $z_j\beta_j$, so the regression equation becomes

$$y = z_1\beta_1 x_1 + \ldots z_p\beta_p x_p + \epsilon.$$

As before, the conditional distribution of the response is given by $Y|\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}(\sum_{j=1}^{p} z_j\beta_j x_j, \sigma^2)$. We need a prior specification for $\{\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2\}$.

The prior distribution over $\boldsymbol{z}$ can be viewed as a prior over the space of models, while the conditional prior distribution of $\boldsymbol{\beta}, \sigma^2$ given a model represented by $\boldsymbol{z}$ can be specified as in the previous subsections, e.g., via semi-conjugate priors or objective priors, etc.

Then, by Bayes' rule, we can compute a posterior probability for each regression model:

$$p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z})}{\sum_{\tilde{\boldsymbol{z}}} p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{Z}, \tilde{\boldsymbol{z}})}. \tag{56}$$

The posterior computation may be a challenging issue: the normalizing constant involves the integration over the space of potential models. Moreover, the computation of the *marginal likelihood* term $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z})$ may be far from being straightforward, due to the need of integration over the remaining parameters $\boldsymbol{\beta}$ and σ^2. The specific modeling choices will play crucial role in mitigating such computational challenges.

Model comparison via the posterior odds is relatively simpler computationally, because the difficult normalizing constants are cancelled out:

$$\begin{aligned} \text{odds}(\boldsymbol{z}_a, \boldsymbol{z}_b|\boldsymbol{y}, \boldsymbol{X}) &= \frac{p(\boldsymbol{z}_a|\boldsymbol{y}, \boldsymbol{X})}{p(\boldsymbol{z}_b|\boldsymbol{y}, \boldsymbol{X})} \\ &= \frac{p(\boldsymbol{z}_a)}{p(\boldsymbol{z}_b)} \times \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_a)}{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_b)} \\ \text{posterior odds} &= \text{prior odds} \times \text{Bayes factor}. \end{aligned}$$

$$p(y|X, z) = \int p(y|\beta, \sigma^2, X, z)\, p(\beta, \sigma^2|X, z)\, d\beta\, d\sigma^2$$

Computing the marginal likelihood We have

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

Some notations: For a given $\boldsymbol{z}$ with $p_{\boldsymbol{z}}$ non-zero entries, let $\boldsymbol{X}_{\boldsymbol{z}}$ be the $n \times p_{\boldsymbol{z}}$ design matrix corresponding the active explanatory variable x_j's, and $\boldsymbol{\beta}_{\boldsymbol{z}}$ the $p_{\boldsymbol{z}} \times 1$ vector consisting of the entries of $\boldsymbol{\beta}$ for the active variables.

Let's consider a (conditional) g-prior for $\boldsymbol{\beta}$ given $\boldsymbol{z}$:

$$\boldsymbol{\beta}_{\boldsymbol{z}}|\boldsymbol{X},\boldsymbol{z},\sigma^2 \sim \mathrm{N}_{p_{\boldsymbol{z}}}(\boldsymbol{0}, g\sigma^2[\boldsymbol{X}_{\boldsymbol{z}}^\top\boldsymbol{X}_{\boldsymbol{z}}]^{-1}).$$

In addition, give $\gamma := 1/\sigma^2$ a gamma prior: gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$. Then we have

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{z}) &= \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{z},\sigma^2)p(\sigma^2)\mathrm{d}\sigma^2 \\ &= \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{z},\gamma)p(\gamma)\mathrm{d}\gamma \\ &= \int (2\pi)^{-n/2}(1+g)^{-p_{\boldsymbol{z}}/2} \times \left[\gamma^{n/2}e^{-\gamma \mathrm{SSR}_g^{\boldsymbol{z}}/2}\right] \times \\ &\quad (\nu_0\sigma_0^2/2)^{\nu_0/2}\Gamma(\nu_0/2)^{-1}\left[\gamma^{\nu_0/2-1}e^{-\gamma\nu_0\sigma_0^2/2}\right]\mathrm{d}\gamma, \end{aligned}$$

where $\mathrm{SSR}_g^{\boldsymbol{z}}$ is the same as in Eq. (55), with $\boldsymbol{X}$ being replaced by $\boldsymbol{X}_{\boldsymbol{z}}$ (exercise: verify this!):

$$\mathrm{SSR}_g^{\boldsymbol{z}} = \boldsymbol{y}^\top(\mathbf{I} - \frac{g}{g+1}\boldsymbol{X}_{\boldsymbol{z}}(\boldsymbol{X}_{\boldsymbol{z}}^\top\boldsymbol{X}_{\boldsymbol{z}})^{-1}\boldsymbol{X}_{\boldsymbol{z}}^\top)\boldsymbol{y}.$$

Now, using the normalizing constant identity for Gamma density leads to

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{z}) = \pi^{-n/2}\frac{\Gamma((\nu_0+n)/2)}{\Gamma(\nu_0/2)}(1+g)^{-p_{\boldsymbol{z}}/2}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + \mathrm{SSR}_g^{\boldsymbol{z}})^{(\nu_0+n)/2}}. \quad (*)$$

With particular choices of priors on $\boldsymbol{z}$, we get closed form for marginal likelihood.

With the marginal likelihood calculation completed, we can proceed to model comparison by computing the posterior odds defined earlier. Suppose that we set $g = n$, $\nu_0 = 1$ for all $\boldsymbol{z}$, while σ_0^2 is the estimated residual variance under the least squares estimate for a given model $\boldsymbol{z}$. That is, given $\boldsymbol{z}$, $\nu_0\sigma_0^2 := s_{\boldsymbol{z}}^2$.

To compare the two models represented by $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$, the Bayes factor is given by

$$\frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_a)}{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_b)} = (1+n)^{(p_{\boldsymbol{z}_b} - p_{\boldsymbol{z}_a})/2} \left(\frac{s_{\boldsymbol{z}_a}^2}{s_{\boldsymbol{z}_b}^2}\right)^{1/2} \times \left(\frac{s_{\boldsymbol{z}_b}^2 + \text{SSR}_g^{\boldsymbol{z}_b}}{s_{\boldsymbol{z}_a}^2 + \text{SSR}_g^{\boldsymbol{z}_a}}\right)^{(n+1)/2}. \tag{57}$$

The ratio of marginal probabilities associated with the two models reflect the balance between model complexity and goodness of fit. In particular, the ratio improves for $\boldsymbol{z}_a$ (i.e., increases) if

- $\text{SSR}_g^{\boldsymbol{z}_a}$ becomes small relatively to $\text{SSR}_g^{\boldsymbol{z}_b}$, i.e., the goodness of fit improves for $\boldsymbol{z}_a$. This happens when the model becomes more complex, i.e., $p_{\boldsymbol{z}_a}$ increases relatively to $p_{\boldsymbol{z}_b}$.
- on the other hand, the term $(1+n)^{(p_{\boldsymbol{z}_b} - p_{\boldsymbol{z}_a})/2}$ penalizes large $p_{\boldsymbol{z}_a}$.

It is important to note that this observation on the balancing act present in the marginal likelihood (and their ratios) is a very general characteristic: by the virtue of integrating over the unknown parameters, the marginal likelihood captures the tension between both model complexity and goodness of fit in its expression.

specific to Bayesian statistics

without a priori specifying any prior preference to simpler models

Example 9.8. Consider the oxygen uptake example. Recall our regression model

$$\begin{aligned}\mathbb{E}[Y|\boldsymbol{\beta}, \boldsymbol{x}] &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ &= \beta_1 + \beta_2 \times \text{group} + \beta_3 \times \text{age} + \beta_4 \times \text{group} \times \text{age}.\end{aligned}$$

The model selection question is whether or not β_2 and β_4 are non-zero (i.e., are there effects of grouping according training programs on oxygen uptake change?). Recall from our earlier analyses that the answer was somewhat ambiguous: the posterior coverage of both β_2 and β_4 contain zero in their 95% confidence intervals. However, also according to the posterior joint distribution, the two parameters are negatively correlated, so whether or not$\beta_2 = 0$ affects our inference about β_4.

We consider 5 candidate models, giving them equal prior probabilities $1/5$. The remaining prior specification is as described above. Then we may obtain the relavant marginal likelihood and posterior odds as following:

$\boldsymbol{z}$	model	$\log p(\boldsymbol{y}\|\mathbf{X}, \boldsymbol{z})$	$p(\boldsymbol{z}\|\boldsymbol{y}, \mathbf{X})$
(1,0,0,0)	β_1	-44.33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{group}_i$	-42.35	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times \text{age}_i$	-37.66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i$	-36.42	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{group}_i \times \text{age}_i$	-37.60	0.19

According to the posterior computation, the best model is $(1, 1, 1, 0)$. There is a strong evidence for age effect, as the posterior probabilities for the three models that include age is essentially 1. The group effect is relatively weaker, as the posterior probabilities of the three models that include group information is $0.00 + 0.63 + 0.19 = 0.82$. This is still substantially higher than the prior probability of 0.60 for the three models combined.

9.4.2 Model averaging via MCMC

Given p explanatory variables, each of which may be either zero or non-zero, there are 2^p model candidates to consider. If p is large, it is challenging to compute the marginal likelihood for each model.

The posterior distribution of interest is then $\Pr(\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X})$. We can derive a Markov chain that enables us to approximate this distribution. However, $\boldsymbol{z}$ is high-dimensional, finding an approximation of the joint posterior distribution for $\boldsymbol{z}$ may be impractical.

Instead, we want to do the following:

1. finding the high probability density region for any variable z_j of interest
2. finding a good estimate for parameters $\boldsymbol{\beta}$ and σ^2 (presumably residing near a low-dimensional subspace) by integrating over $\boldsymbol{z} \in \{0, 1\}^p$

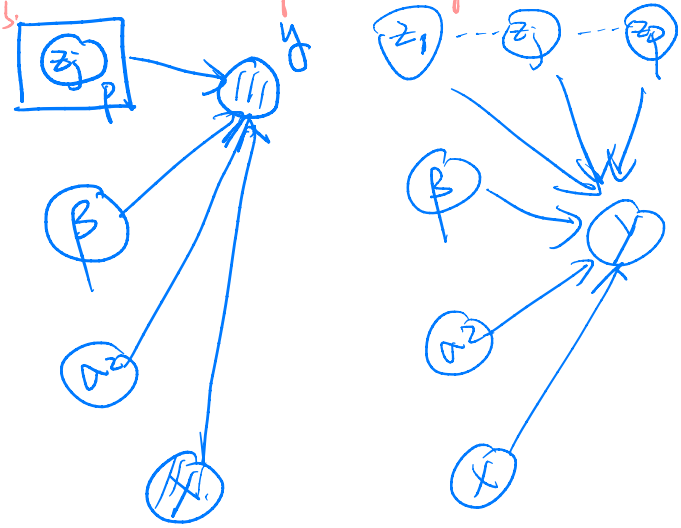
Deriving a Gibbs sampler for the posterior distribution of this model is simple. The full conditional distribution for each z_j is

$$\Pr(z_j = 1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_{-j}) = \frac{o_j}{1 + o_j} \tag{58}$$

where the odds o_j is given by

$$\begin{aligned} o_j &= \frac{\Pr(z_j = 1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_{-j})}{\Pr(z_j = 0 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_{-j})} \\ &= \frac{\Pr(z_j = 1)}{\Pr(z_j = 0)} \times \frac{p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 1)}{p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 0)} := A \times B. \end{aligned}$$

Note that B was already computed via Eq. (57) for a g-prior specification. If we put an (independent) uniform prior probability on each variable x_j, so that $\Pr(z_j = 1) = \Pr(z_j = 0) = 1/2$, then $A = 1$. The posterior samples for $\boldsymbol{\beta}$ and σ^2 were given in 9.3 for the g-prior.



To summarize the Gibbs sampling procedure using the g-prior for $\boldsymbol{\beta}$ and σ^2, and the uniform prior for $\boldsymbol{z}$: Given the sample $(\boldsymbol{z}^{(s)}, \boldsymbol{\beta}^{(s)}, \sigma^{2(s)})$, the sample at step $s+1$ is generated as follows

1. Set $\boldsymbol{z} = \boldsymbol{z}^{(s)}$;
2. For $j \in \{1, \ldots, p\}$ in random order, replace z_j with a sample from $p(z_j | \boldsymbol{z}_{-j}, \boldsymbol{y}, \boldsymbol{X})$ given by Eq. (58);
3. Set $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}$;
4. Sample $\sigma^{2(s+1)} \sim p(\sigma^2 | \boldsymbol{z}^{(s+1)}, \boldsymbol{y}, \boldsymbol{X})$ given by Eq. (54);
5. Sample $\boldsymbol{\beta}^{(s+1)} \sim p(\boldsymbol{\beta} | \boldsymbol{z}^{(s+1)}, \sigma^{2(s+1)}, \boldsymbol{y}, \boldsymbol{X})$ given by Eq. (53);

Additional step in Gibbs for model averaging

Sample a choice of model

This is the R codes for the Gibbs sampling procedure above (only the portion for sampling z is included)

```
##### a function to compute the marginal probability
lpy.X<-function(y,X,g=length(y),
   nu0=1,s20=try(summary(lm(y~-1+X))$sigma^2,silent=TRUE))
{
  n<-dim(X)[1] ; p<-dim(X)[2]
  if(p==0) { Hg<-0 ; s20<-mean(y^2) }
  if(p>0)  { Hg<-(g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X) }
  SSRg<- t(y)%*%( diag(1,nrow=n) - Hg )%*%y

  -.5*( n*log(pi)+p*log(1+g)+(nu0+n)*log(nu0*s20+SSRg)-
        nu0*log(nu0*s20) ) +
  lgamma( (nu0+n)/2 ) - lgamma(nu0/2)
}
#####

##### starting values and MCMC setup
z<-rep(1,dim(X)[2] )
lpy.c<-lpy.X(y,X[,z==1,drop=FALSE])
S<-10000
Z<-matrix(NA,S,dim(X)[2])
#####

##### Gibbs sampler
for(s in 1:S)
{
  for(j in sample(1:dim(X)[2]))
    {
      zp<-z ; zp[j]<-1-zp[j]
      lpy.p<-lpy.X(y,X[,zp==1,drop=FALSE])
      r<- (lpy.p - lpy.c)*(-1)^(zp[j]==0)
      z[j]<-rbinom(1,1,1/(1+exp(-r)))
      if(z[j]==zp[j]) {lpy.c<-lpy.p}
    }
  Z[s,]<-z
}
#####
```

log p(y|X,z) marginalize out β.

(*) equation on p310

$r = \log \frac{p(y|X,z_j)}{p(y|X,1-z_j)}$

draw samples of β, σ^2 given z. samples.

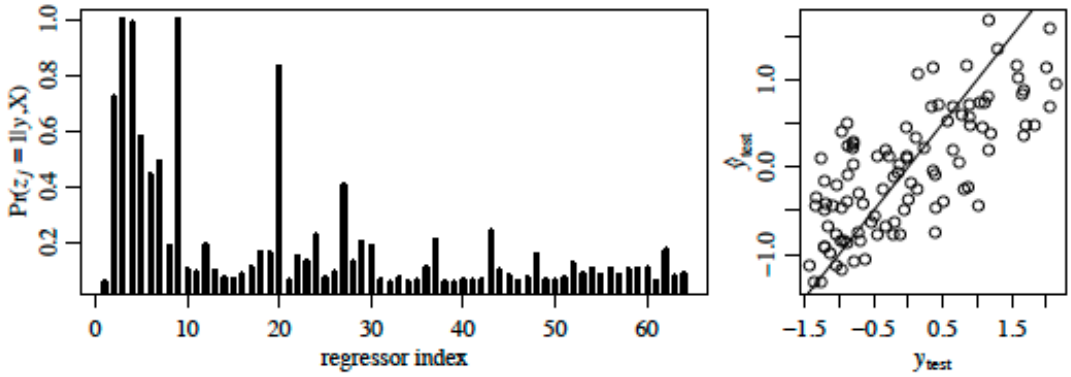
distribution of "best models"

Example 9.9. We return to the diabetes data example.

- Recall that we have $p = 64$ potential regressors, resulting in $2^{64} \approx 10^{19}$ total number of models.
- It is impossible to explore this space: if we generate 10,000 Gibbs samples, these samples account for only $1/10^{15}$ total number of models.
- Our intuition is that if there are only a small number of relevant regressors, and so they will be present in many of the most likely models among the 2^{64} candidates. Averaging over the most likely candidates will still give us a good estimate of the marginal posterior probabilities of each of the regressor z_j's as well as the corresponding $\boldsymbol{\beta}$. (Recent theoretical developments on Bayesian asymptotics confirmed this intuition).

circumvent the need to explore all 2^{64} models.

- The estimate for $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}_{\text{bma}} = \sum_{s=1}^{S} \boldsymbol{\beta}^{(s)}/S$, where S is the MCMC sample size.
  - This is called the Bayesian model averaged estimate of $\boldsymbol{\beta}$, because it does not correspond to any particular value of $\boldsymbol{z}$, but an average of regression parameters from different values of $\boldsymbol{z}$. By averaging the regression coefficients from multiple high-probability models, the resulting estimate often performs better than a point estimate that corresponds to only a single model.
  - The test error for the model averaging technique is 0.452, which is better than both OLS and backwards elimination.



- More on Bayesian robustness: recall that the backwards elimination procedure also produced 18 spurious associations in a randomization experiment (cf. Example 9.7). Using the Bayesian model averaging technique, it was found that the (approximated) posterior probabilities $\Pr(z_j = 1|\boldsymbol{y}, \boldsymbol{X})$ are less than 1/2 for all $j = 1, \ldots, 64$, and all but two of which are less than 1/4. The model averaging technique did not erroneously identify any regressors as having an effect on the distribution of $\tilde{\boldsymbol{y}}$.

original Bayesian regression
non-hierarchical
(X_i, y_i) iid data
$y_i \sim N(X_i\beta, \sigma^2)$

hierarchical Bayesian regression
grouped data
group j: (X_{ji}, y_{ji})
i^{th} sample of j^{th} group
$y_{ji} \sim N(X_{ji}\beta_j, \sigma_j^2)$
$\beta_j \overset{iid}{\sim} N(\beta, \Sigma)$
$\beta \sim$ prior
$\Sigma \sim$ prior (eg. Inverse Wishart)
$\sigma_j^2 \overset{iid}{\sim} IG(\cdots)$
priors of IG parameters
Each group has its own regression parameters, that share same distri-

aka random effects models
mixed effect

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.