GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lesson 7: Linear Regression

Hai Shu, PhD

10/31/2022

Topics

- Linear regression
- Ordinary Least Squares
- Linear regression in a Bayesian context
  - Using a semi-conjugate prior
  - What prior distribution to use
  - Prior distribution elicitation
- Model checking for a normal linear regression model
- Model selection
  - Bayesian variable selection
  - Bayesian model averaging
  - Criterion-based methods

Linear regression

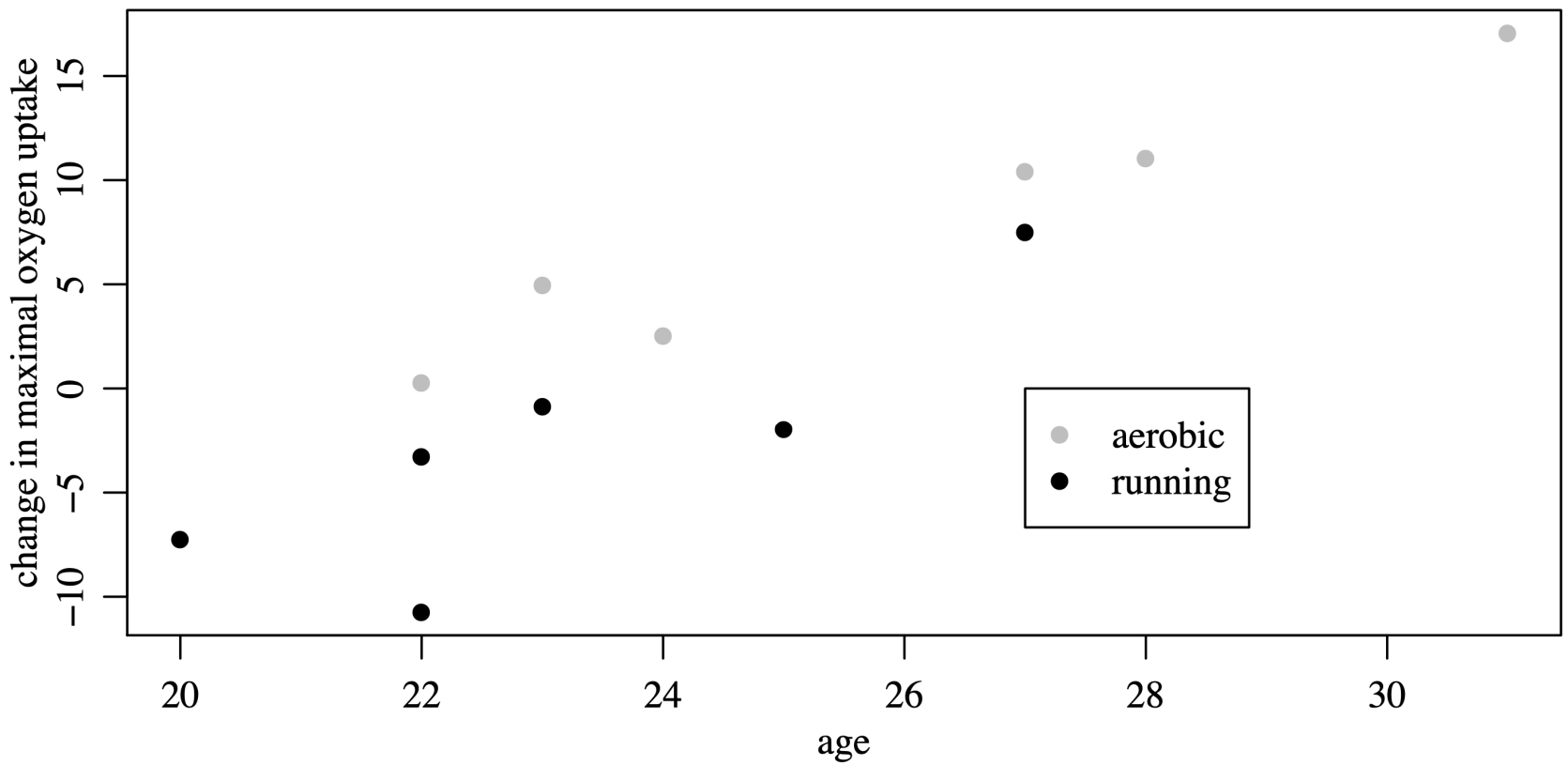
- The scenarios we have considered so far always assumed that $Y_1, Y_2, \ldots, Y_n | \theta \overset{iid}{\sim} p(y|\theta)$.
- However in some situations, $Y_1, Y_2, \ldots, Y_n$ might be independent but NOT identically distributed.
- We will now consider this situation by starting from the simplest of those cases.
- In many scientific studies we are interested in understanding the relationship between a variable Y and a second variable X or, in general, other variables $X_1, \ldots, X_p$.
- In particular we are interested in the conditional distribution of Y **given** $X = x$ or **given** $X_1 = x_1, \ldots, X_p = x_p$.
- In linear regression we provide a specific form for $p(y|x_1, \ldots, x_p) = p(y|\mathbf{x})$

An example

- Example: 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.
  Six of the twelve men were randomly assigned to a 12-week flat-terrain running program and the remaining six were assigned to a 12-week step aerobics program. The maximal oxygen uptake of each subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program.
- We are interested in how a subject's change in maximal oxygen uptake may depend on which program they were assigned to. Age may also affect the change in maximal oxygen uptake.

An example

- Change in maximal oxygen uptake as a function of age and exercise program.



Response variable and explanatory variables

- From the plot we can see that the change in maximal oxygen uptake increases linearly as a function of age in both groups.
- It is also possible that the rate of change as a function of age is different for the two groups.
- One way to model the relationship between Y, the response variable, in this case, the change in a subject's maximal oxygen uptake and $X_1, X_2, \ldots, X_p$, the explanatory variables or covariates, is to assume that the conditional expectation $E(Y|\mathbf{x})$ of Y given $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ is given by

$$E(Y|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \beta' \mathbf{x}$$

where $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_p \end{pmatrix}'$.

Normal linear regression model

- For the example we have just considered we will have the following variables:
  - Y is the change in maximal oxygen uptake in a subject
  - X_1 is equal to 1 for each subject
  - X_2 is an indicator for the program in which the subject is enrolled (0 for running and 1 for aerobic)
  - X_3 is the age of the subject
  - X_4 is the product of X_2 and X_3
- Then, in a linear regression framework, we assume that

$$E(Y|\mathbf{x}) = E(Y|x_1, \ldots, x_4) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

- This still does not specify the form of $p(y|\mathbf{x})$!
- In a normal linear regression model we assume that

$$p(y|\beta, \mathbf{x}, \sigma^2) = N(\beta'\mathbf{x}, \sigma^2)$$

or equivalently, we assume that for each $i = 1, \ldots, n$,

$$Y_i = \beta'\mathbf{x}_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Joint distribution $p(y_1, \ldots, y_n | \boldsymbol{\beta}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \sigma^2)$

- Note that assuming

$$Y_i = \beta' \mathbf{x}_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

  for each $i = 1, \ldots, n$ implies that $Y_i | \beta, \mathbf{x}_i, \sigma^2 \sim N(\beta' \mathbf{x}_i, \sigma^2)$.
- Since the mean for each Y_i is different, now $Y_1, \ldots, Y_n$ are independent but **NOT** identically distributed!
- If we have observations $y_1, \ldots, y_n$ for the response variable and observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for the explanatory variables, then the joint distribution $p(y_1, \ldots, y_n | \beta, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \sigma^2)$ is given by:

$$\begin{aligned} p(y_1, \ldots, y_n | \beta, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \sigma^2) &= \prod_{i=1}^n p(y_i | \beta, \mathbf{x}_i, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \cdot \frac{(y_i - \beta' \mathbf{x}_i)^2}{\sigma^2} \right] \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2} \sum_{i=1}^n \cdot \frac{(y_i - \beta' \mathbf{x}_i)^2}{\sigma^2} \right] \end{aligned}$$

Multivariate normal distribution

- Remember the multivariate normal distribution $N_p(\theta, \mathbf{\Sigma})$.
  If $\mathbf{Y}|\theta, \mathbf{\Sigma} \sim N_p(\theta, \mathbf{\Sigma})$ then its density is given by:

$$p(\mathbf{y}|\theta, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}(\mathbf{y}-\theta)'\mathbf{\Sigma}^{-1}(\mathbf{y}-\theta)\right]$$

  where the quadratic form $(\mathbf{y}-\theta)'\mathbf{\Sigma}^{-1}(\mathbf{y}-\theta)$ is given by:

$$\sum_{j,k=1}^{p} (y_j - \theta_j) \cdot (\mathbf{\Sigma}^{-1})_{jk} (y_k - \theta_k)$$

  where $(\mathbf{\Sigma}^{-1})_{jk}$ denotes the (j,k)-th element of the $p \times p$ matrix $\mathbf{\Sigma}^{-1}$

- Thus, we can write $p(\mathbf{y}|\theta, \mathbf{\Sigma})$ as

$$p(\mathbf{y}|\theta, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}\sum_{j,k=1}^{p} (y_j - \theta_j) \cdot (\mathbf{\Sigma}^{-1})_{jk} (y_k - \theta_k)\right]$$

Linear regression and multivariate normal distribution

- Compare $p(\mathbf{y}|\theta,\mathbf{\Sigma})$ when $\mathbf{Y}|\theta,\mathbf{\Sigma} \sim N_p(\theta,\mathbf{\Sigma})$:

$$p(\mathbf{y}|\theta,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}\sum_{j,k=1}^{p}(y_j-\theta_j)\cdot(\mathbf{\Sigma}^{-1})_{jk}(y_k-\theta_k)\right]$$

with

$$p(y_1,\ldots,y_n|\beta,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\cdot\frac{(y_i-\beta'\mathbf{x}_i)^2}{\sigma^2}\right]$$

We can write $p(y_1,\ldots,y_n|\beta,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\sigma^2)$ as a multivariate normal density for $\mathbf{Y}$ where $\mathbf{Y} = \begin{pmatrix} Y_1 & Y_2 & \ldots & Y_n \end{pmatrix}'$, $p = n$ and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 \cdot \mathbf{I}_n \qquad \theta = \begin{pmatrix} \beta'\mathbf{x}_1 \\ \beta'\mathbf{x}_2 \\ \vdots \\ \beta'\mathbf{x}_n \end{pmatrix} = \mathbf{X}\beta$$

Ordinary Least Squares

- Therefore, we can collect the n observations $y_1, \ldots, y_n$ in a $n \times 1$ vector $\mathbf{y}$ and consider it as a realization of a multivariate random vector $\mathbf{Y}|\beta, \mathbf{X}, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ where $\mathbf{I}_n$ is a diagonal $n \times n$ matrix with diagonal elements all equal to 1 and $\mathbf{X}$ is the $n \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

- In a classical framework, estimates of β and σ^2 are obtained via Ordinary Least Squares and are found by minimizing the sum of squared residuals SSR given by $SSR = (\mathbf{y} - \mathbf{X}\beta)' \cdot (\mathbf{y} - \mathbf{X}\beta)$.
- Taking the derivative of SSR with respect to β and setting it equal to 0 we obtain:

$$\beta_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Ordinary Least Squares and A Semi-conjugate Prior

- The Ordinary Least Squares estimator of σ^2 is given by

$$\sigma^2_{OLS} = \frac{SSR(\hat{\beta}_{OLS})}{n-p} = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})' \cdot (\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})}{n-p} = s^2$$

- The variance of β_{OLS} is given by $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. Since σ^2 is not known, standard errors for the components β are given by $(\mathbf{X}'\mathbf{X})^{-1} \cdot \hat{\sigma}^2_{OLS} = (\mathbf{X}'\mathbf{X})^{-1} \cdot s^2$.
- How do we proceed in a Bayesian framework instead?
- We will use our representation of our data $y_1, \ldots, y_n$ as one sample $\mathbf{y}$ from a multivariate normal distribution: $\mathbf{y}|\beta, \mathbf{X}, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$
- To carry out a Bayesian analysis we need to specify priors on β and σ^2.
- There are many choices that can be made as for the prior distribution on the regression parameters β and on the variance σ^2
- Let's consider the case where $p(\beta, \sigma^2) = p(\beta) \cdot p(\sigma^2) = N_p(\beta_0, \mathbf{\Sigma}_0) \cdot \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$.
- Then the joint posterior distribution cannot be derived in closed form (see the case of univariate normal data), so we approximate it via Gibbs sampling. Hence, we need to derive the full conditional distribution for β and σ^2.

Full conditional for β

- Let's derive the full conditional distribution $p(\beta|\mathbf{y},\mathbf{X},\sigma^2)$. We have:

$$\begin{aligned} p(\beta|\mathbf{y},\mathbf{X},\sigma^2) &= \frac{p(\beta,\mathbf{y},\mathbf{X},\sigma^2)}{p(\mathbf{y},\mathbf{X},\sigma^2)} \propto p(\mathbf{y}|\mathbf{X},\beta,\sigma^2)\cdot p(\beta) \\ &\propto \exp\left(-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\beta)\right) \\ &\cdot \exp\left(-\tfrac{1}{2}(\beta-\beta_0)'\mathbf{\Sigma}_0^{-1}(\beta-\beta_0)\right) \end{aligned}$$

- The full conditional distribution for β is $p(\beta|\mathbf{y},\mathbf{X},\sigma^2) = N_p\left(E\left[\beta|\mathbf{y},\mathbf{X},\sigma^2\right], \text{Var}\left[\beta|\mathbf{y},\mathbf{X},\sigma^2\right]\right)$ where

$$\begin{aligned} \text{Var}\left[\beta|\mathbf{y},\mathbf{X},\sigma^2\right] &= (\mathbf{\Sigma}_0^{-1} + \frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1} \\ E\left[\beta|\mathbf{y},\mathbf{X},\sigma^2\right] &= (\mathbf{\Sigma}_0^{-1} + \frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1}\cdot(\mathbf{\Sigma}_0^{-1}\cdot\beta_0 + \frac{\mathbf{X}'\mathbf{y}}{\sigma^2}) \end{aligned}$$

Full conditional for β

- As $p(\beta|\mathbf{y},\mathbf{X},\sigma^2) = N_p\left((\boldsymbol{\Sigma}_0^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1}\cdot(\boldsymbol{\Sigma}_0^{-1}\cdot\beta_0+\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}),(\boldsymbol{\Sigma}_0^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1}\right)$, and $\boldsymbol{\Sigma}_0^{-1}$ is the prior precision matrix, we can see that if the prior precision matrix has small elements (that is, we have a vague prior), then

$$\begin{aligned}(\boldsymbol{\Sigma}_0^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1} &\approx \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\\ (\boldsymbol{\Sigma}_0^{-1}+\frac{\mathbf{X}'\mathbf{X}}{\sigma^2})^{-1}\cdot(\boldsymbol{\Sigma}_0^{-1}\cdot\beta_0+\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}) &\approx \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\cdot(\frac{\mathbf{X}'\mathbf{y}}{\sigma^2})\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

- On the other hand, if the prior precision matrix has large entries (we have a very informative prior), then $E\left[\beta|\mathbf{y},\mathbf{X},\sigma^2\right]$ tends to β_0, the prior expectation.

Full conditional for σ^2

- Let's derive the full conditional distribution $p(\sigma^2|\mathbf{y},\mathbf{X},\beta)$ of σ^2. We have:

$$
\begin{aligned}
p(\sigma^2|\mathbf{y},\mathbf{X},\beta) &= \frac{p(\sigma^2,\mathbf{y},\mathbf{X},\beta)}{p(\mathbf{y},\mathbf{X},\beta)} \propto p(\mathbf{y}|\mathbf{X},\beta,\sigma^2)\cdot p(\sigma^2) \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\beta)\right) \\
&\cdot \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sigma^2)^{\frac{n+\nu_0}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\nu_0\sigma_0^2\right]\right)
\end{aligned}
$$

that is, the full conditional distribution of σ^2 is:

$$
p(\sigma^2|\mathbf{y},\mathbf{X},\beta) = \text{InverseGamma}\left(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2}\right)
$$

where $\nu_n = n+\nu_0$ and $\sigma_n^2 = \frac{1}{\nu_n}\cdot((\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\nu_0\sigma_0^2)$.

Gibbs sampling and prediction

- Now that we have derived the full conditionals we can devise the Gibbs sampling algorithm to approximate the joint posterior distribution $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$.
- The algorithm would work as follows:
  1. Choose a number S of iterations.
  2. Choose initial values $\sigma^{2(0)}$ and $\beta^{(0)}$ for σ^2 and β respectively.
  3. For $j = 1, \ldots, S$ repeat the following two steps:
     - sample $\sigma^{2(j)}$ from $p(\sigma^2 | \mathbf{y}, \mathbf{X}, \beta^{(j-1)})$
     - sample $\beta^{(j)}$ from $p(\beta | \mathbf{y}, \mathbf{X}, \sigma^{2(j)})$
- If we have m additional observations $\tilde{\mathbf{X}}$ of the p covariates, that is, we have $\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,m}, \tilde{x}_{2,1}, \ldots, \tilde{x}_{2,m}, \ldots, \tilde{x}_{p,1}, \ldots, \tilde{x}_{p,m}$ we can use the model and predict $\tilde{\mathbf{y}}$. We simply need to add this additional step to the Gibbs sampler:
  - sample $\tilde{\mathbf{y}}^{(j)}$ from $N_m(\tilde{\mathbf{X}}\beta^{(j)}, \sigma^{2(j)} \mathrm{I}_m)$.

Oxygen uptake example

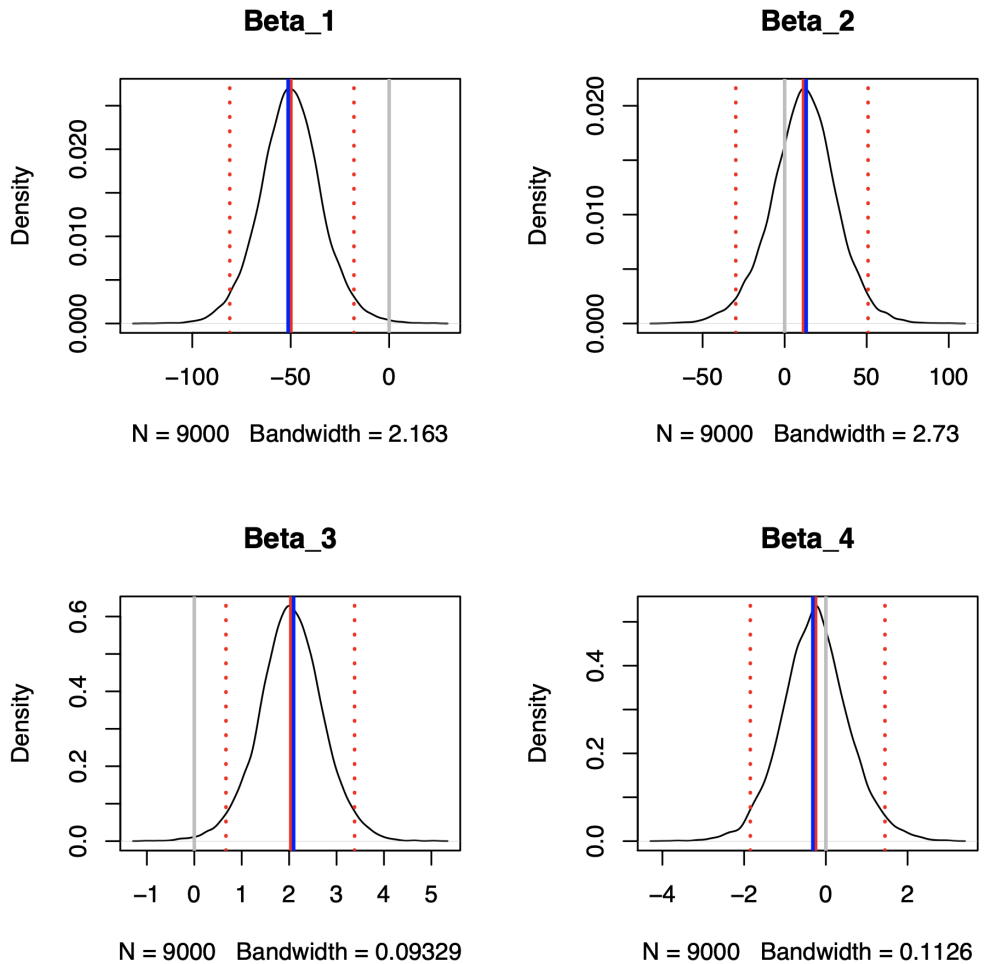
- Let's consider the oxygen uptake example: 12 subjects who did not exercise regularly were assigned to either a 12-week flat-terrain running program or to a 12-week step aerobics program. The change in their maximal oxygen uptake while running on an inclined treadmill between before and after participating in the 12-week program was recorded.
- We want to assess the effect of the program on the maximal oxygen uptake controlling for the subject's age.
- We consider a linear regression model that includes as covariates x_1, an intercept term, x_2, an indicator for the program in which the subject was enrolled, x_3, the subject's age, and x_4 the product of the program indicator times the subject's age.
- We make inference on the regression coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ and σ^2 by placing a semi-conjugate prior on (β, σ^2), thus we need to determine β_0, $\mathbf{\Sigma}_0$, ν_0 and σ_0^2.
- How do we choose them?

Oxygen uptake example

- Looking at literature of exercise physiology we learn that males in their 20s should have an oxygen uptake of around 150 liters per minute with a standard deviation of 15.
- This means that about 95% of males in their 20s have an oxygen uptake between $(150 - 2 \times 15 = 120, 150 + 2 \times 15 = 180)$.
- Hence, the changes in oxygen uptake should lie within $(-60, 60)$.
- If we consider the subjects in the running program, those for which $x_2 = 0$, this means that the line $\beta_1 + \beta_3 x_3$ should produce values between -60 and 60 for x_3 between 20 and 30.
- This leads to the values $\beta_{0,1} = 0$ and $\beta_{0,3} = 0$ and $\mathbf{\Sigma}_{0,(1,1)} = 150^2$ and $\mathbf{\Sigma}_{0,(3,3)} = 6^2$.
- We use again $\beta_{0,2} = 0$ and $\beta_{0,4} = 0$ with $\mathbf{\Sigma}_{0,(2,2)} = 150^2$ and $\mathbf{\Sigma}_{0,(4,4)} = 6^2$
- We assume no correlation a priori among these regression coefficients, that is, $\mathbf{\Sigma}_{0,(j,k)} = 0$ for $j \neq k$ and $j, k = 1, \ldots, 4$.

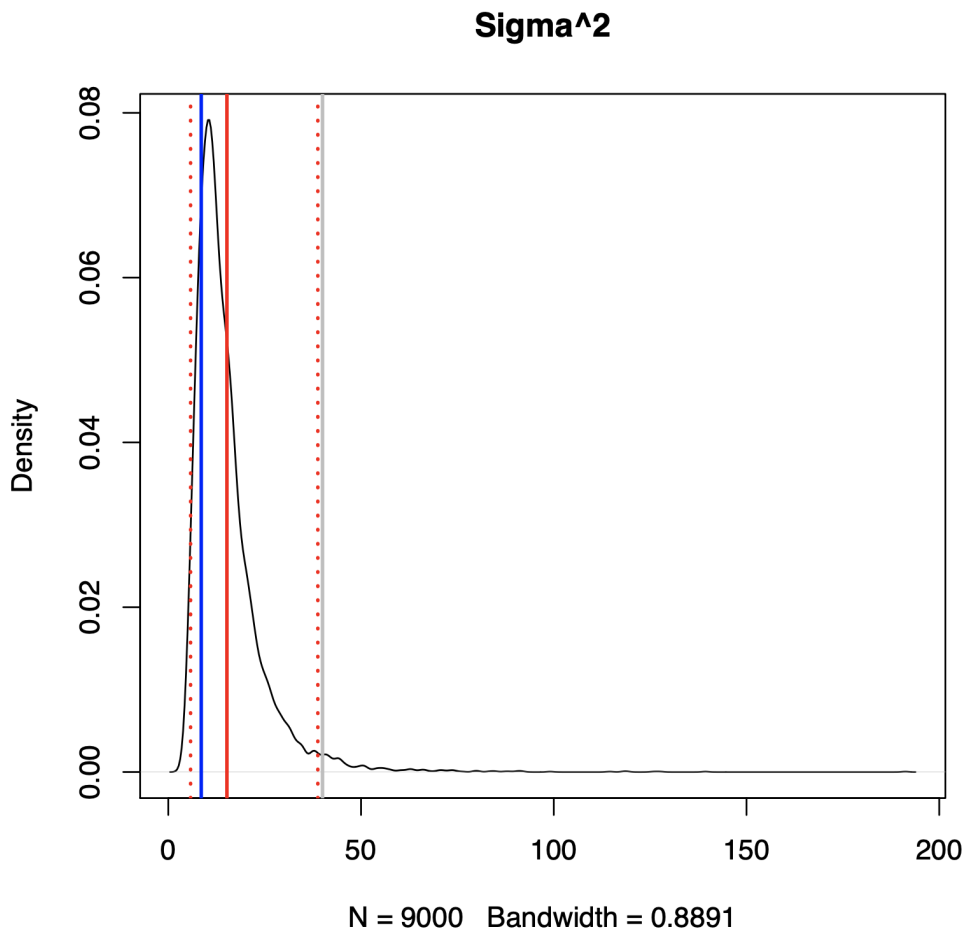
Oxygen uptake example

- We ran a Gibbs sampling algorithm for $B = 10,000$ iterations discarding the first 1,000 for burnin. We used the sample variance in the data and $(0\,0\,0\,0)'$ as initial values for σ^2 and β, respectively.
- Marginal posterior densities for $\beta_1, \beta_2, \beta_3$ and β_4: prior mean in grey, OLS estimate in blue and posterior mean in red.



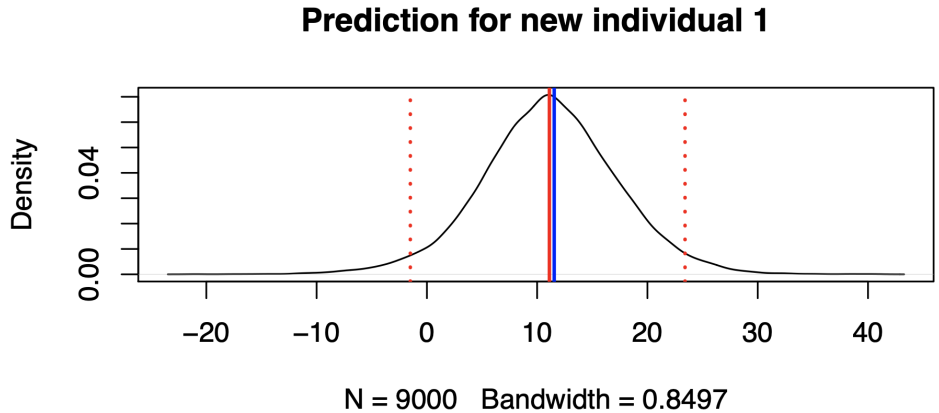
Oxygen uptake example

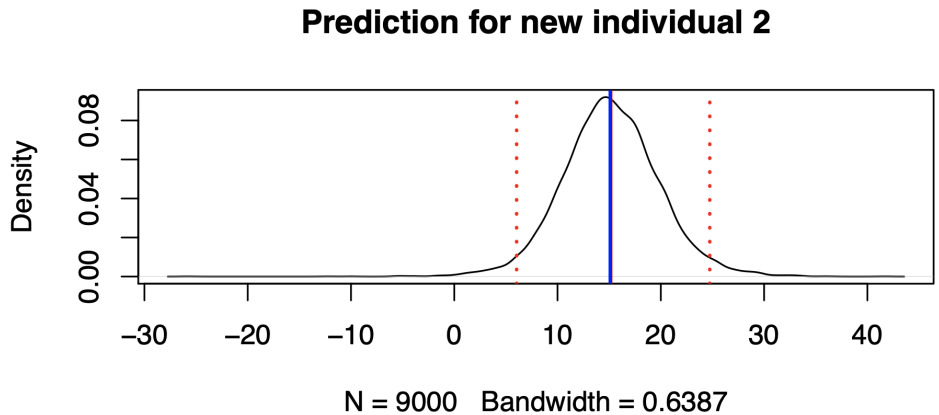
- Marginal posterior density for σ^2: prior mean in grey, OLS estimate in blue and posterior mean in red.



Oxygen uptake example

- I also predicted the change in maximal oxygen uptake for two individuals both of age 30, one who did the flat-terrain running program and one who did the step aerobics program.
- Marginal posterior predictive density for $\tilde{y}_1$ and $\tilde{y}_2$: OLS prediction in blue and posterior predictive mean in red.





Oxygen uptake example

Inferential approach	Parameters	Mean	Standard deviation
OLS	$(\sigma^2, \beta_1, \beta_2, \beta_3, \beta_4)$	$(8.5, -51.3, 13.1, 2.1, -0.32)$	$(?, 12.3, 15.8, 0.53, 0.65)$
Bayesian with semi-conjugate prior	$(\sigma^2, \beta_1, \beta_2, \beta_3, \beta_4)$	$(11.2, -50.0, 11.6, 2.0, -0.26)$	$(7.2, 13.8, 17.7, 0.6, 0.7)$
Bayesian with non-informative prior	$(\sigma^2, \beta_1, \beta_2, \beta_3, \beta_4)$	$(11.2, -51.1, 12.9, 2.1, -0.3)$	$(7.9, 14.3, 18.3, 0.62, 0.76)$

Oxygen uptake example

Inferential approach	Parameters	Mean	Standard deviation
OLS	$(\tilde{y}_1, \tilde{y}_2)$	$(11.5, 15.1)$	$(0.93, 0.98)$
Bayesian with semi-conjugate prior	$(\tilde{y}_1, \tilde{y}_2)$	$(11.2, 15.1)$	$(5.4, 4.1)$
Bayesian with non-informative prior	$(\tilde{y}_1, \tilde{y}_2)$	$(11.5, 15.1)$	$(3.40, 3.34)$

What prior to use

- We have seen how to make inference on the regression parameters β and σ^2 in a Bayesian framework using an informative prior.
- Some believe that using a noninformative prior is the best choice, because that would lead to a "more objective" result since the prior information does not carry any information.
- If one wants to go for a noninformative prior or for a prior with little information, choices could be:
  - the non-informative prior that we have seen
  - Jeffrey's prior: $p_J(\beta, \sigma^2) \propto |I(\beta, \sigma^2)|^{\frac{1}{2}}$
  - the unit information prior.
- The last one is not really a prior since it requires knowledge of the data in order to be constructed. However, one could could argue that it uses only a small amount of information since it use only the information contained in one observation.
- Your book also discusses Zellner's g prior for linear regression model. See pp. 157-159.

How to specify a semi-conjugate prior distribution

- Bayesian analysis is very advantageous when we don't have much data available (e.g. a Phase I clinical trial where very few observations are available or a situation where it is very costly to gather data).
  In this case, Ordinary Least Squares can produce very unstable solutions. Thus, using an informative prior can result useful in this case.
- How do we specify the parameters $\beta_0, \mathbf{\Sigma}_0, \nu_0, \sigma_0^2$ of such a prior distribution $p(\beta, \sigma^2) = N(\beta; \beta_0, \mathbf{\Sigma}_0) \cdot \text{InverseGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$?
- We can derive these parameters from the literature or previous similar studies. Or we can determine them if we have some previous experience with the data.

How to specify a semi-conjugate prior distribution

- If we have raw data $D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0)$ from a similar study we can determine $\beta_0, \mathbf{\Sigma}_0, \nu_0, \sigma_0^2$ as follows:
  - we can set β_0 equal to the OLS estimate $\hat{\beta}_{0,OLS}$ of β in a linear regression for $\mathbf{y}_0$ on $\mathbf{X}_0$.
  - we can set $\mathbf{\Sigma}_0$ equal to $a_0^{-1}(\mathbf{X}_0'\mathbf{X}_0)^{-1}$ for some predetermined constant a_0
  - we can set σ_0^2 equal to:

$$\sigma_0^2 = \frac{1}{c_0(n_0 - p)} \left(\mathbf{y}_0 - \mathbf{X}_0\hat{\beta}_{0,OLS}\right)' \left(\mathbf{y}_0 - \mathbf{X}_0\hat{\beta}_{0,OLS}\right)$$

    for some constant c_0
  - we can set ν_0 equal to some predetermined constant c_0.

Bayesian analysis with informative priors

- Note that in **any** Bayesian analysis, sensitivity analysis should be conducted by varying the choices of the parameters in the prior distributions (also called hyperparameters) and examining the impact of prior choices on the inference.
- This is particular important when one uses informative priors. it is **never** a good idea to carry out an informative Bayesian analysis based on only one set of chosen hyperparameters.
  One must examine sensitivity and robustness of the result to different choices of the prior parameters.
- It is always good to compare results derived under informative priors with results obtained under noninformative priors. The noninformative prior serves as benchmark for the comparison.
- Also, when possible, it is good to compare with results derived under a frequentist framework.

Model checking for the linear regression model

- In a classical framework, we check the adequacy of a linear regression model by looking at the residuals, examining the R^2 and so on.
- There are Bayesian counterparts to the frequentist model checking techniques.
- We have talked about some of these model checking techniques earlier when we discussed posterior predictive model checking.
- We are now going to add more to that discussion in the particular context of linear regression model

$$y_i = x_{1,i}\beta_1 + \ldots + x_{p,i}\beta_p + \varepsilon_i \quad \Longrightarrow \quad \mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$$

$$\Longrightarrow \mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- For each observation i, we can compute the residual:

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_1 x_{1,i} - \beta_2 x_{2,i} - \ldots - \hat{\beta}_p x_{p,i}$$

Model checking for the linear regression model

- We can then compute the posterior probability that the i-th observation is an outlier

$$p_i = P(|\varepsilon_i| > k\sigma|\mathbf{y}, \mathbf{X})$$

where k is a constant chosen so the prior probability the observations in the data are outlier is small.

- Since $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, for each $i = 1, \ldots, n$, $\varepsilon_i \sim N(0, \sigma^2)$, thus a priori the probability that the i-th observation is an outlier is given by:

$$P(|\varepsilon_i| > k\sigma) = 2 \cdot \Phi(-k)$$

where Φ is the cumulative distribution function of a $N(0,1)$ random variable.

- We can then compare the posterior probability that the i-th observation is an outlier with the prior probability that it is an outlier.

Model checking for the linear regression model

- If we have a significant amount of data that has a high posterior probability of being an outlier, then model is probably not adequate for the data.
- How do we compute the posterior probability p_i in practice?
- Other ways to determine if the i-th observation is an outlier consist in computing
  - the posterior predictive ordinate PPO_i, that is, the posterior predictive probability of observing **again** y_i after having observed $y_1, \ldots, y_n$
  - the conditional predictive ordinate CPO_i, that is, the posterior predictive probability of observing y_i after having observed $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$.
- We can also compute posterior p-values for the residual sum of squares, for example, or other test statistics.
- For more details on PPO_i, CPO_i and posterior p-values see Lecture 4.

Model checking for the linear regression model: example

- We now illustrate an example of model checking using the maximal oxygen uptake example.
- We have fit a linear regression model to these data using as covariates x_1, an intercept term, x_2, an indicator for the program in which a subject is enrolled, x_3, the subject's age and x_4 the interaction of a subject's age and the program in which he is enrolled.
- We are now checking the validity of the model by looking at the residuals and the posterior probabilty that they are larger than 3 times the posterior estimates of σ.
- We also look at the posterior predictive ordinate for each observation y_i. More precisely we look at the posterior probability of observing a normal random variable with mean equal to $\mathbf{X}\beta$ and variance σ^2 that is larger than y_i. This will give us an indication of how extremes the observation y_i is.

Model checking for the linear regression model: example

- Using the output from the Gibbs sampling algorithm, we obtain a posterior estimate of σ=3.2.

Observation	$P(\lvert\varepsilon_i\rvert > 3 * \sigma \lvert \mathbf{y}, \mathbf{X})$	PPO_i
1	1.1E-4	0.25
2	8.4E-3	0.94
3	1.1E-4	0.29
4	7.8E-4	0.80
5	8.0E-3	0.27
6	2.6E-3	0.30
7	2.4E-3	0.49
8	1.1E-3	0.25
9	0.0	0.43
10	1.1E-4	0.56
11	7.8E-4	0.57
12	3.3E-4	0.72

Variable selection

- Let's consider now this example: we have data on $N = 442$ individuals for which 10 variables (age, sex, body mass index, average blood pressure and six blood serum measurements, including ldl, hdl, total cholesterol, etc) were measured at baseline.
- These 442 people are diabetes patients for which a response variable, the change in some quantitative measure of the disease progression one year after baseline, was measured. The statisticians in the project were asked to develop a model that predicted the response y from the ten covariates $x_1, x_2, \ldots, x_{10}$.
- Two hopes were evident here: (1) that the model would produce accurate baseline predictions of response for future patients; (2) that the form of the model would suggest which covariates were important factors in disease progression.
- How do we decide which variables should we include in the model? Do we use all the 10 variables?
- It was also suggested that the relationship between the response variable and some of the covariates was **not** linear.

Mean squared predictive error

- To account for this and to allow for interaction among the variables, we can imagine a linear model that depends on all the 10 covariates $x_1, \ldots, x_{10}$, all the $45 = \binom{10}{2}$ interactions of the form $x_j x_k$ and 9 quadratic terms since one variable, $x_2 = \text{sex}$ is binary. This makes a total of $p = 64$ variables.
- We can envision a linear regression model for the response variable on the $p = 64$ covariates. We fit such a model where all the variables, the response as well as all the covariates, have been centered to have mean zero and variance 1.
- How we evaluate such a model? What are some of the problems we might incur with such a problem?
- To evaluate the performance of the model, we divide the $N = 442$ observations into two sets: a training set $(\mathbf{y}, \mathbf{X})$ made of a 342×1 vector $\mathbf{y}$ of observations for the response variable with the corresponding 342×64 matrix $\mathbf{X}$, and a test set $(\mathbf{y}_{test}, \mathbf{X}_{test})$ made of the remaining observations.

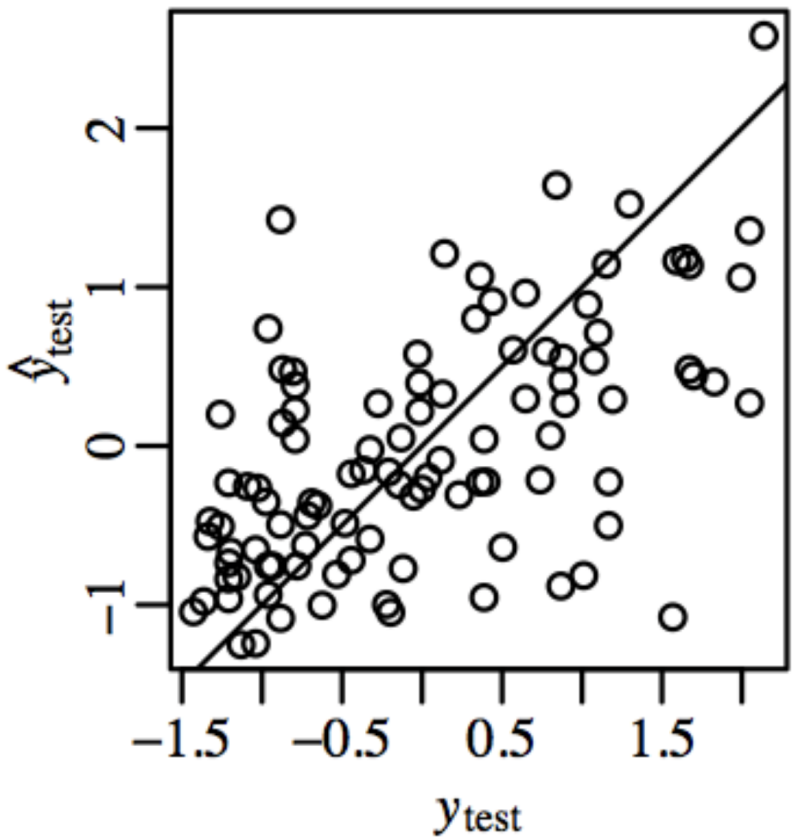
Assessing a model's predictive performance

- We assess the model's performance via cross-validation: we fit the linear regression model on the training data, estimate $\hat{\beta}_{OLS}$, we make predictions at the test set via $\hat{\mathbf{y}}_{test} = \mathbf{X}_{test} \cdot \hat{\beta}_{OLS}$ and we compare the mean squared predictive error:

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_{i,test} - \hat{y}_{i,test})^2$$

- We can also visually assess the fit by plotting the predicted values $\hat{y}_{test,i}$ versus the observed values $y_{test,i}$, $i = 1, 2, \ldots, 100$ in a scatterplot.

Assessing a model's predictive performance



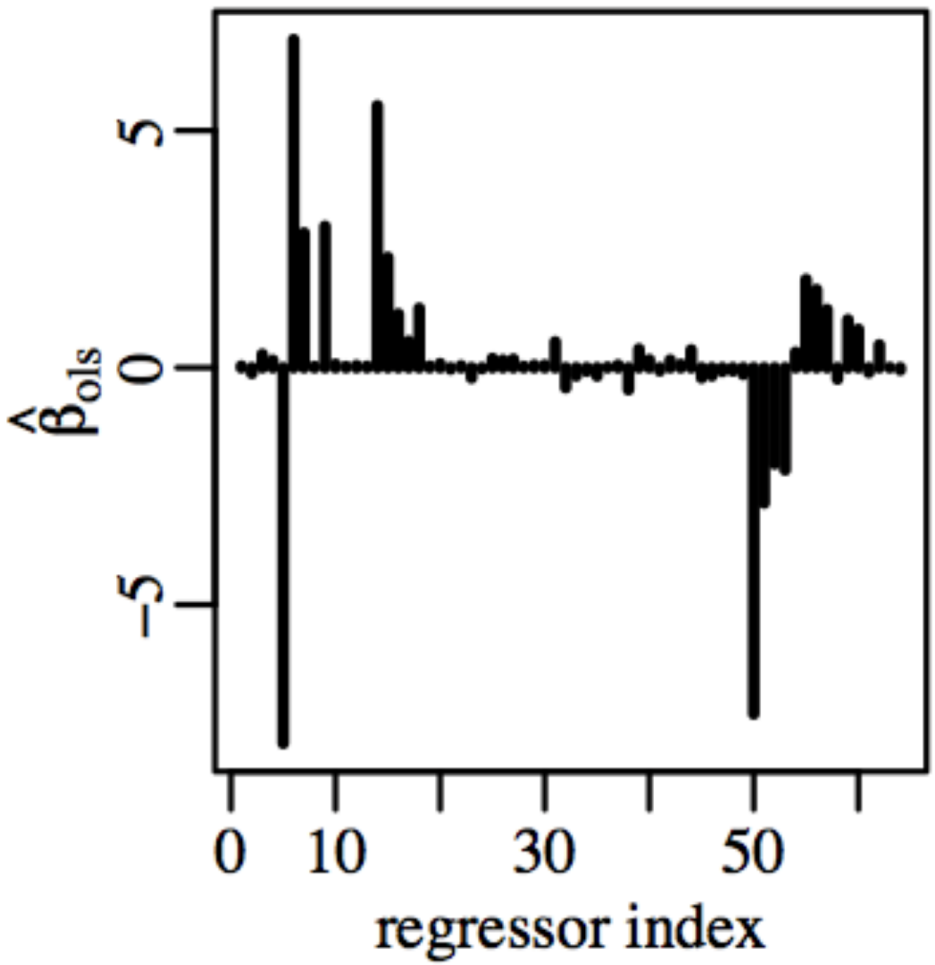
- The mean squared predictive error for the model with all 64 variables is 0.67, not much lower than what we would have obtained if we predicted $\hat{y}_{test,i} = 0$ for all $i = 1, \ldots, 100$. In this case, the mean squared predictive error would have been 0.97!

Backward variable selection

- It is clear that some of the variables in the model are not associated with the outcome variable once other variables are included in the model.
- We can improve the model performance by removing from the model covariates that are not necessary.
- We can perform backward selection.
- How does backward selection work? We perform the following procedure:
  1. We start with the model with all covariates. For each regression coefficient $\beta_1, \ldots, \beta_p$ we get the OLS estimate $\hat{\beta}_{OLS,j}$ and its standard error $\hat{\sigma^2}_{OLS} (\mathbf{X}'\mathbf{X})^{-1}_{(j,j)}$ and we evaluate the corresponding t statistic t_j.
  2. If there is any $|t_k|$ that is less than some cutoff t_{cutoff}, we find the covariate with the smallest $|t|$-value and we remove that variable from the model.
  3. We consider then the updated model. When the $|t|$ test statistic is above the cutoff for all the variables we stop.

Backward variable selection

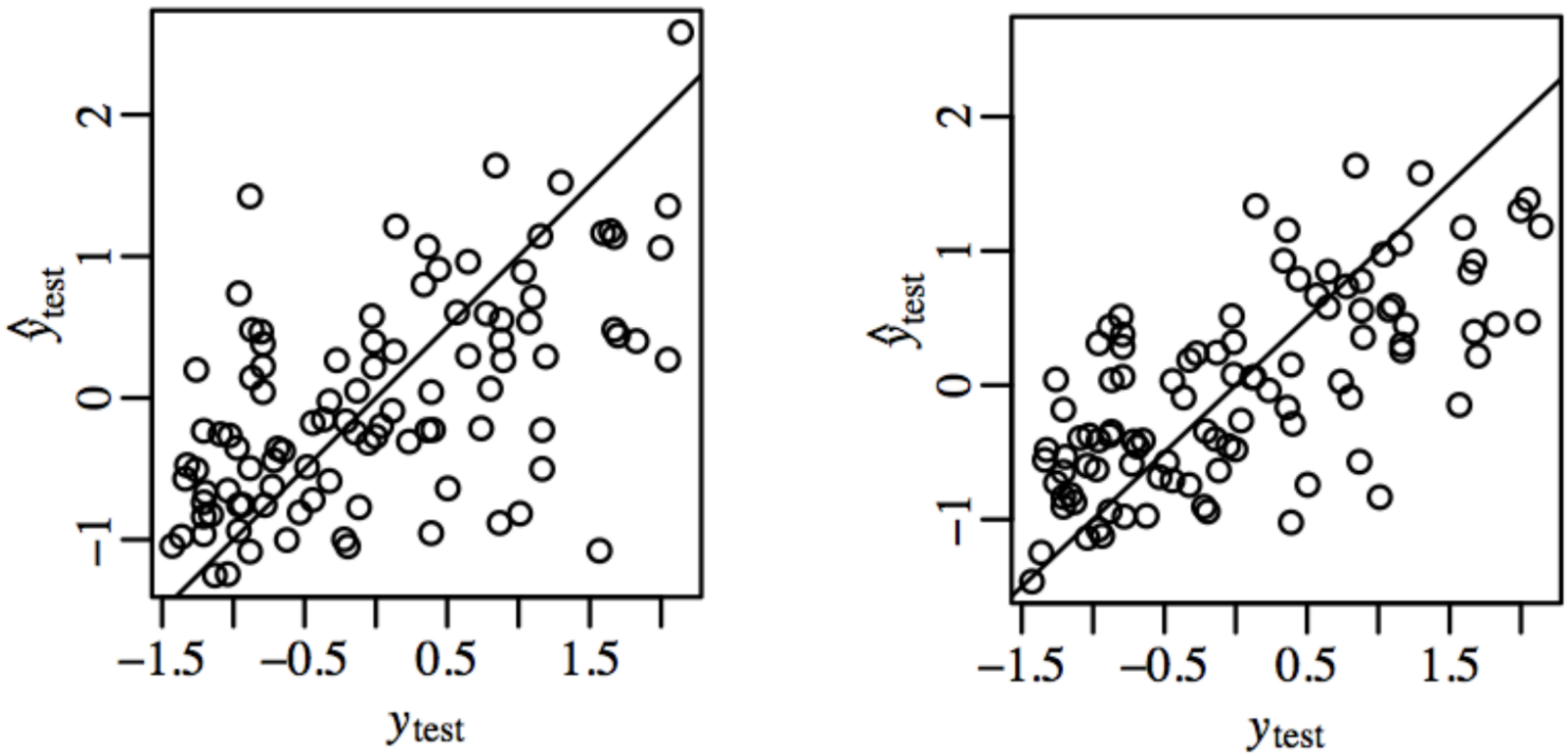
- Plot of the OLS estimates for each of the 64 covariates versus the covariate number.



Backward variable selection

- If we apply the backward selection procedure to the diabetes data using a cutoff $t_{cutoff} = 1.65$. We found that only 20 of the initial 64 variables are retained.
- We can assess the predictive performance of this new model by again comparing the predictions $\hat{\mathbf{y}}_{test}$ with the observed values $\mathbf{y}_{test}$.
- For the new model, we obtain a mean squared predictive error of 0.53 versus the initial 0.67!

Backward variable selection



- Pairwise plot of the predicted $\hat{\mathbf{y}}_{test,i}$ versus the observed $\mathbf{y}_{test,i}$ for (left panel) the regression model with all 64 covariates and (right panel) for the model with the 20 covariates selected via backward selection.

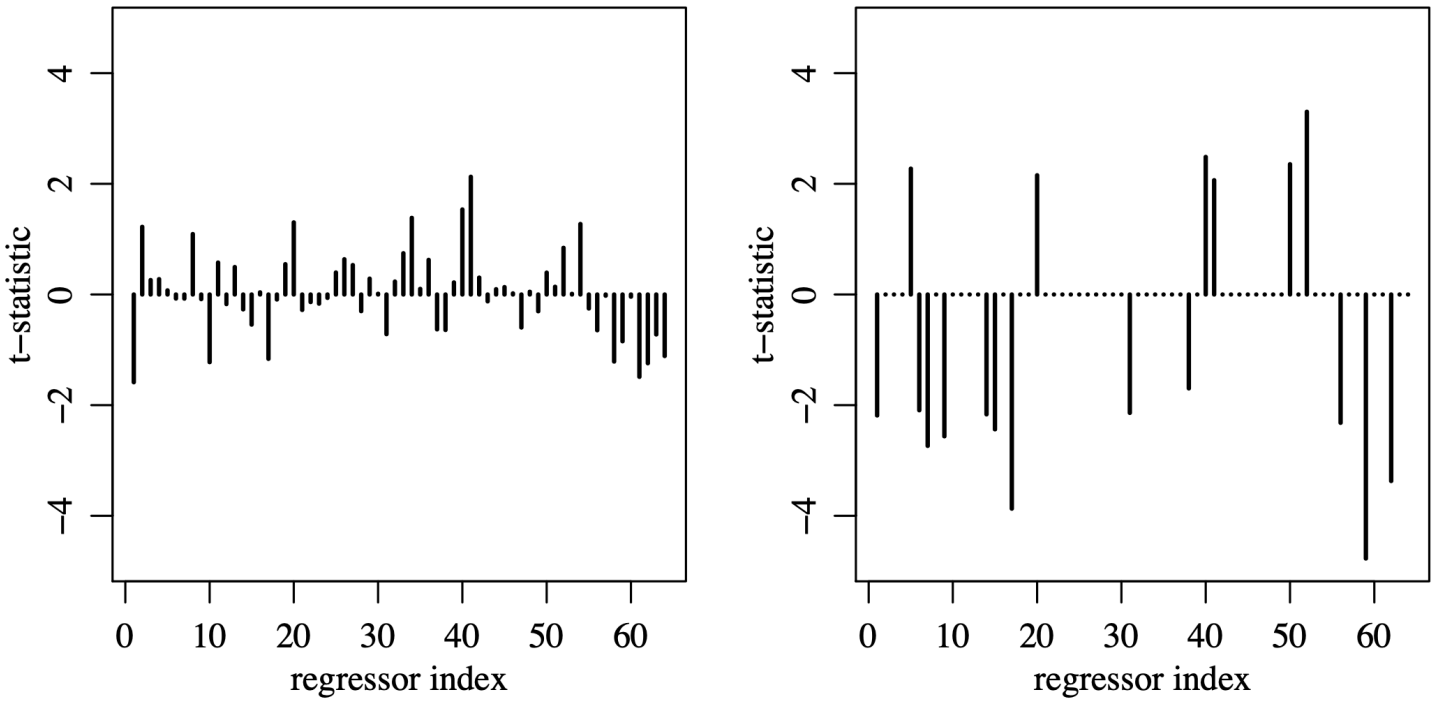
Backward variable selection

- Backward selection can also pick up some spurious association, that is some association between the response variable and the covariates that is only present in the training data but not in any other dataset.
- We can notice this by randomly permuting the responses associated to the covariates in the training set, e.g.

$$\begin{array}{cccccc} (x_{1,1}, & x_{2,1}, & \dots, & x_{64,1}) & \longrightarrow & y_{\pi(1)} \\ (x_{1,2}, & x_{2,2}, & \dots, & x_{64,2}) & \longrightarrow & y_{\pi(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (x_{1,342}, & x_{2,342}, & \dots, & x_{64,342}) & \longrightarrow & y_{\pi(342)} \end{array}$$

  where π is a permutation of the 342 indices.
- If we fit a linear regression model on the permutated response on the covariates, we are expecting most of the coefficients to be non significantly different from zero.
- Instead performing a backward selection on the permutated training set leads to a model with 18 regressors!

Backward variable selection



- Plot of the t statistic for the regression coefficients versus the covariate index before (left panel) and after (right panel) backward variable selection when we are using the permutated data $y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(342)}$ as response variable and the column of $\mathbf{X}$ as covariates.

Bayesian variable selection

- To perform variable selection in a Bayesian framework, we can proceed as follows.
- Let's consider the oxygen uptake example: there we have one response, y,the maximal oxygen uptake, and four variables, x_1, an intercept term, x_2, the program in which each subject is enrolled, x_3, a subject's age, and x_4 the interaction between a subject's age and the program in which he participated. In this case, $p = 4$ and $N = 12$
- We perform variable selection by introducing p auxiliary 0-1 variables z_j where $j = 1, \ldots, p$ such that $z_j = 1$ if the j-th variable x_j should be included in the linear regression model and 0 otherwise.
- Then, we can express the model as

$$y_i = z_1 b_1 x_{1,i} + z_2 b_2 x_{2,i} + \ldots + z_p b_p x_{p,i} + \varepsilon_i$$

where the b_j, $j = 1, \ldots, p$ are some real numbers.

Bayesian variable selection

- For example, in the maximal oxygen uptake:
  - $(z_1, z_2, z_3, z_4) = (1, 1, 1, 0)$ corresponds to the linear model

$$\begin{aligned} y_i &= b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + \varepsilon_i \\ \Longrightarrow y_i &= b_1 \cdot 1 + b_2 \cdot \text{program}_i + b_3 \cdot \text{age}_i + \varepsilon_i \end{aligned}$$

  - $(z_1, z_2, z_3, z_4) = (1, 0, 1, 0)$ corresponds to the linear model

$$\begin{aligned} y_i &= b_1 x_{1,i} + b_3 x_{3,i} + \varepsilon_i \\ \Longrightarrow y_i &= b_1 \cdot 1 + b_3 \cdot \text{age}_i + \varepsilon_i \end{aligned}$$

- Therefore, each configuration of $(z_1, \ldots, z_p)$ corresponds to a different model. Choosing which variable to include or not in a linear regression model then is equivalent to choose a model among the different models.
- How do we choose among these different models?

Bayesian variable selection

- We choose among these models by looking at the posterior probability of each model **given** the data.
- We start with a prior probability for each possible model $\mathbf{z} = (z_1, \ldots, z_p)$. If we don't have any prior belief on a specific model, we can assign all the possible models the same probability and have thus a uniform prior probability.
- In light of the data collected, we update our prior probability and derive the posterior probability for a model $\mathbf{z} = (z_1, \ldots, z_p)$ which is given by:

$$p(\mathbf{z}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}) \cdot p(\mathbf{z})}{\sum_{\tilde{z}} p(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{z}}) \cdot p(\tilde{\mathbf{z}})}$$

- The numerator is the product of the marginal likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{z})$ for the data under model $\mathbf{z}$ times the prior probability $p(\mathbf{z})$ of the model. The denominator is the sum over all the possible models $\tilde{\mathbf{z}}$ of the product of the marginal likelihood times the prior probability for the model.

Bayesian variable selection

- How do we compute the marginal likelihood $p(\mathbf{y}|\mathbf{X},\mathbf{z})$?
- By definition, this is given by the following integral

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

- The above integral is then the integral of the product of the likelihood for the data under model $\mathbf{z}$ times the prior distribution of (β,σ^2) under model $\mathbf{z}$.
- The integral can be evaluated in closed form for certain choices of prior distribution.
- Example: if $\mathbf{z}=(1,0,1,0)$, then $p=2$ and the model for the data is $\mathbf{y}|\mathbf{X},\beta,\sigma^2,\mathbf{z} \sim N_{n=12}(\mathbf{X}\beta_{\mathbf{z}},\sigma^2\mathbf{I}_{12})$ with $\beta_{\mathbf{z}} = \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix}$ and $\mathbf{X}$ a 12×2 matrix.

Bayesian variable selection

- If we use a semi-conjugate prior on (β, σ^2), then the marginal likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{z})$ is

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{z}) &= \int p(\mathbf{y}|\beta, \sigma^2, \mathbf{X}, \mathbf{z}) \cdot p(\beta, \sigma^2|\mathbf{X}, \mathbf{z}) d\beta d\sigma^2 \\ &= \int \Big[N_{12}(\mathbf{y}; \mathbf{X}\beta_{\mathbf{z}}, \sigma^2 \mathbf{I}_{12}) \cdot N_2(\beta; \beta_0, \mathbf{\Sigma}_0) \\ &\cdot \quad \text{InverseGamma}(\tfrac{\nu_0}{2}, \tfrac{\nu_0 \sigma_0^2}{2}) \Big] d\beta d\sigma^2 \end{aligned}$$

- Once the posterior probabilities for each model have been computed, we can choose the variables to include in our linear regression model by selecting the variables in the model $\mathbf{z}$ with the highest posterior probability.

Posterior odds

- If we want to compare only two models $\mathcal{M}_a = \mathbf{z}_a$ and $\mathcal{M}_b = \mathbf{z}_b$, we can compute the posterior odds of the two models:

$$\text{odds}(\mathbf{z}_a, \mathbf{z}_b|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{z}_a|\mathbf{y}, \mathbf{X})}{p(\mathbf{z}_b|\mathbf{y}, \mathbf{X})} = \frac{p(\mathbf{z}_a)}{p(\mathbf{z}_b)} \times \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_b)}$$

  where the first term, $\frac{p(\mathbf{z}_a)}{p(\mathbf{z}_b)}$, is the prior odds while the second term, $\frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_b)}$, is called Bayes factor.
- We compute each marginal likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{z}_a)$, $p(\mathbf{y}|\mathbf{X}, \mathbf{z}_b)$ as shown in the previous slide

An example

- Let's consider the maximal oxygen uptake example. We want to determine which variables should be included in the model.
- The book reports the marginal likelihood $p(\mathbf{y}|\mathbf{X},\mathbf{z})$ for each model $\mathbf{z}$ obtained assuming that β and σ^2 are independent and using a g-prior for β and a unit information prior for σ^2
- The marginal likelihood and the posterior probability for each model is reported in the table below:

$\mathbf{z}$	Model	$\log p(\mathbf{y}|\mathbf{X},\mathbf{z})$	$p(\mathbf{z}|\mathbf{y},\mathbf{X})$
(1,0,0,0)	$y_i = \beta_1$	-44.33	0.00
(1,1,0,0)	$y_i = \beta_1 + \beta_2 \cdot \text{program}_i$	-42.35	0.00
(1,0,1,0)	$y_i = \beta_1 + \beta_3 \cdot \text{age}_i$	-37.66	0.18
(1,1,1,0)	$y_i = \beta_1 + \beta_2 \cdot \text{program}_i + \beta_3 \cdot \text{age}_i$	-36.42	0.63
(1,1,1,1)	$y_i = \beta_1 + \beta_2 \cdot \text{program}_i + \beta_3 \cdot \text{age}_i + \beta_4 \cdot (\text{program} \times \text{age}_i)$	-37.60	0.19

An example

- Since the model with the highest posterior probability is the model that includes program and age as covariates, the variables program and age should be included in the model.
- Note that since the sum of the posterior probabilities of all models containing age is equal to 1, we conclude that there is clearly an effect of age on maximal uptake.
- There is less evidence of an effect of the aerobics program a subject is involved on maximal uptake since the sum of the posterior probabilities for all models that contain it is equal to $0.00 + 0.63 + 0.19 = 0.82$. Note that this is however larger than the prior probability for any model containing age as covariate since that was equal to $0.2 + 0.2 + 0.2 = 0.6$.

Bayesian model averaging

- In all the variable selection approaches we have just seen, the goal is to identify the variables that are associated to the response variable.
- We have seen that one way to do that is to compute the marginal posterior probability of each model, where we represent each model via its $p \times 1$ configuration vector $\mathbf{z} = (z_1, z_2, \ldots, z_p)$.
  Each z_j can take only two values: $z_j = 0$ or $z_j = 1$ for $j = 1, \ldots, p$. If $z_j = 0$, then the j-th covariate x_j is not included in the model.
- In Bayesian model averaging instead of a choosing and fitting each linear regression individually, we consider all the linear regressions simultaneously and we perform inference on all the regression parameters simultaneously.
- More precisely, each linear regression is identified by the $p \times 1$ vector $\mathbf{z}$. Our model has now as parameters, the entire vector of $p \times 1$ regression coefficients $\beta_1, \ldots, \beta_p$, the variance parameter σ^2 and the $p \times 1$ vector $\mathbf{z}$.

Bayesian model averaging

- We want to make inference on $\beta_1, \ldots, \beta_p, \sigma^2$ and $\mathbf{z} = (z_1, \ldots, z_p)$ **given** the data.
- In particular, inferring upon $\mathbf{z}$ will tell us which models are most supported by the data.
- Thus, we place a prior on $\beta_1, \ldots, \beta_p, \sigma^2$ and $\mathbf{z}$. We then run a Gibbs sampling algorithm, sampling sequentially from the full conditionals of all the parameters.
- To derive the full conditional distribution of $\mathbf{z} = (z_1, \ldots, z_p)$ we consider each z_j individually. Therefore, the full conditional of z_j should be

$$p(z_j | y_1, \ldots, y_n, \mathbf{X}, \mathbf{z}_{-j}, \beta_1, \ldots, \beta_p, \sigma^2)$$

where $\mathbf{z}_{-j}$ indicates the entire $p \times 1$ vector $\mathbf{z}$ except for the j-th component.

Bayesian model averaging

- In reality, in Bayesian model averaging, we don't sample z_j from $p(z_j|y_1,\ldots,y_n,\mathbf{X},\mathbf{z}_{-j},\beta_1,\ldots,\beta_p,\sigma^2)$ but we sample it from $p(z_j|y_1,\ldots,y_n,\mathbf{X},\mathbf{z}_{-j})$.
- Once we have sampled $\mathbf{z}$, we sample $\beta=(\beta_1,\ldots,\beta_p)$ and σ^2 respectively from $p(\beta|\mathbf{y},\mathbf{X},\sigma^2,\mathbf{z})$ (multivariate normal) and $p(\sigma^2|\mathbf{y},\mathbf{X},\beta,\mathbf{z})$ (Inverse Gamma).
- This sampling scheme still guarantees that we are sampling from the correct joint posterior distribution.
- Since z_j is a binary variable, we can express the probability that z_j is equal to 1 **given** $y_1,\ldots,y_n,\mathbf{X},\mathbf{z}_{-j}$ through the odds

$$o_j=\frac{P(z_j=1|y_1,\ldots,y_n,\mathbf{X},\mathbf{z}_{-j})}{P(z_j=0|y_1,\ldots,y_n,\mathbf{X},\mathbf{z}_{-j})}=\frac{P(z_j=1)\cdot P(y_1,\ldots,y_n|\mathbf{X},\mathbf{z}_{-j},z_j=1)}{P(z_j=0)\cdot P(y_1,\ldots,y_n|\mathbf{X},\mathbf{z}_{-j},z_j=0)}$$

Bayesian model averaging

- Therefore, the Gibbs sampling algorithm for Bayesian model averaging works as follows:
  - Choose a number S of iterations. For $k = 1, \ldots, S$ repeat the following steps
    1. For $j \in \{1, 2, \ldots, p\}$ in random order, replace $z_j^{(k-1)}$ with $z_j^{(k)}$ sampled from $p(z_j | \mathbf{y}, \mathbf{X}, \mathbf{z}_{-j}^{(k-1)})$
    2. Sample $\beta^{(k)}$ from $p(\beta | \mathbf{y}, \mathbf{X}, \sigma^{2(k-1)}, \mathbf{z}^{(k)})$
    3. Sample $\sigma^{2(k)}$ from $p(\sigma^2 | \mathbf{y}, \mathbf{X}, \beta^{(k)}, \mathbf{z}^{(k)})$
- The Gibbs sampler for $\mathbf{z}$ ensures that for large S the distribution of the samples $\mathbf{z}^{(B+1)}, \ldots, \mathbf{z}^{(S)}$ converge to the posterior distribution $p(\mathbf{z} | \mathbf{y}, \mathbf{X})$ where B is the burnin period.
- On the other hand, $(\beta^{(s)}, \sigma^{2(s)})$ is a sample from $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}, \mathbf{z}^{(s)})$. For large S the distribution of the samples $(\beta^{(s)}, \sigma^{2(s)})$ that correspond to a particular configuration of $\mathbf{z}$ converge to the posterior distribution $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$ corresponding to the particular linear regression model associated with $\mathbf{z}$.

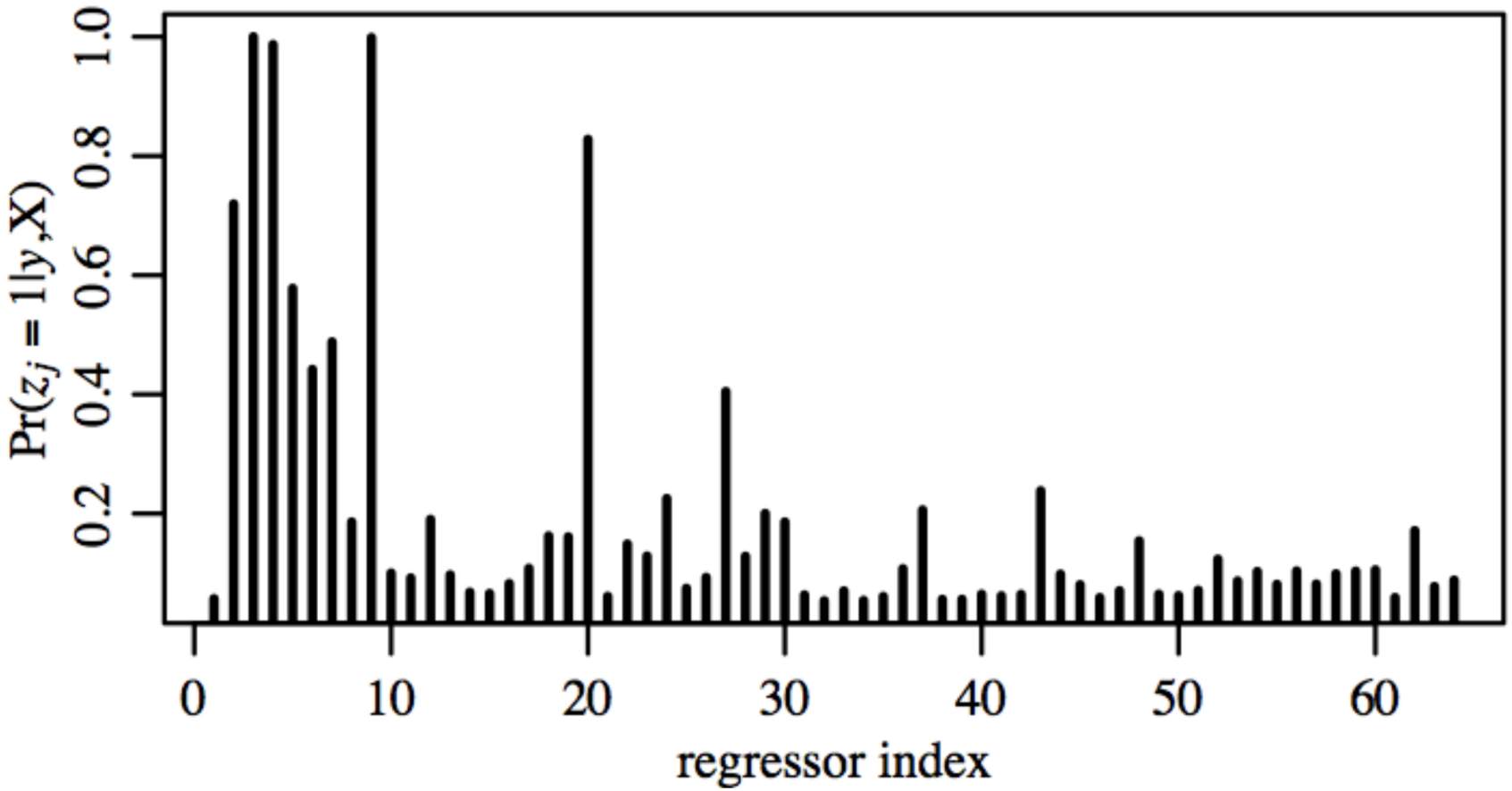
Bayesian model averaging: an example

- Let's consider again the diabetes example. We have data for 442 patients with diabetes and for those patients we have collected a response variable, which is some clinical measure of the progression of the disease during the course of a year. We also have data on 10 other covariates, six of which are blood serum measurements.
- Given that there might be some interaction between the covariates and the relationship between the response variable and the covariates might not be linear, from these 10 covariates we generated other 54 possible covariates, leading to $p = 64$ covariate in total.
- We want to now perform Bayesian model averaging. Thus, we take only the data for the training set and we place a uniform prior on $\mathbf{z}$. Given that $\mathbf{z}$ is made of 64 binary variables, there are 2^{64} possible configurations for $\mathbf{z}$.
- We ran the Gibbs sampling algorithm for $S = 10,000 < 2^{64}$ iterations. This means that there will be models that we never sample!

Bayesian model averaging: an example

- In particular, from the MCMC run, we can see that only 32 models were sampled more than once: 28 were sampled twice, 2 were sampled three times and 2 others were sampled five and six times.
- This means that if we have a large p, the Gibbs sampling scheme for Bayesian Model Averaging presented above provides a poor approximation to the posterior distribution of $\mathbf{z}$.

Bayesian model averaging: an example



- Posterior probability that $z_j = 1$ for $j = 1, \ldots, 64$ versus j, the regressor index.

Bayesian model averaging: an example

- We can also look at the posterior mean of each β_j.
- Note that if we take

$$\hat{\beta}_{j,BMA} = \frac{1}{(S-B)} \sum_{k=B+1}^{S} \beta_j^{(k)}$$

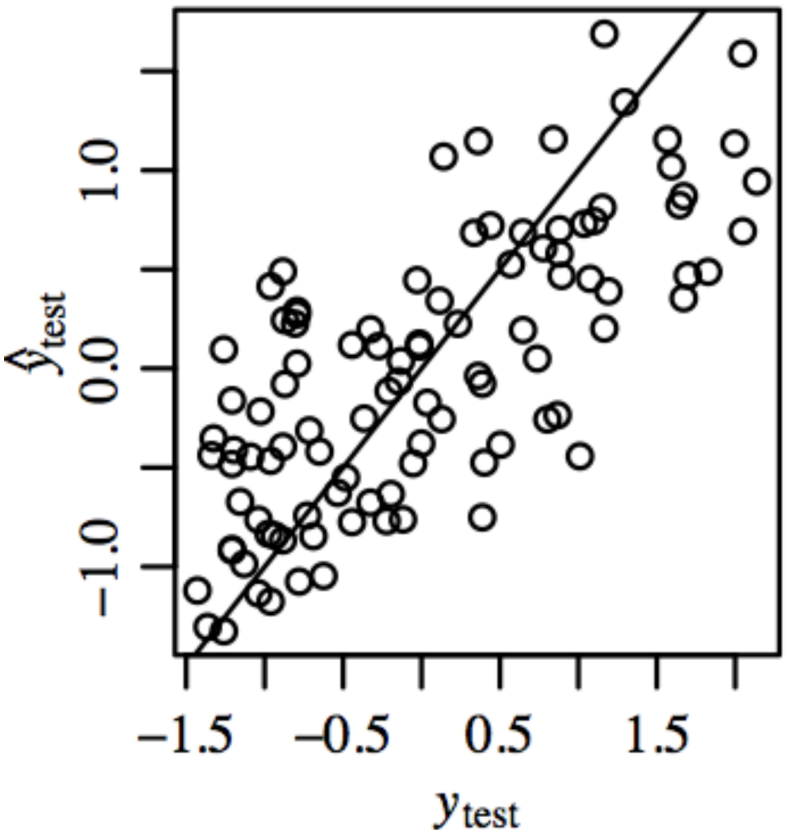
where B is the burnin period, the estimates that we get is called the (Bayesian) model average of β_j becuase it is averaged across different linear regression model.

- In terms of regression estimate, this estimate is certainly not comparable with the estimate we get if we fit a particular linear regression model. However, in terms of predictions, if we consider the predictions of the outcome variable for the test set given by

$$\hat{\mathbf{y}}_{test,BMA} = \mathbf{X}_{test}\hat{\beta}_{BMA}$$

these predictions often perform better than other predictions.

Bayesian model averaging: an example



- Pairwise scatterplot of $\hat{y}_{test,BMA,i}$ versus $y_{test,i}$. The mean square predictive error now is 0.45 compared to 0.53 that we obtained after backward variable selection.

Bayesian model averaging: an example

- We can also look at which regressor coefficients β_j has a posterior probability larger than 0.5 to be non-zero.
- This can be looked at it as a way to perform variable selection. If we apply this to the diabetes dataset, we obtained only six regression coefficients that satisfy this criteria. The six covariates associated with these six regression coefficients are a subset of the 20 variables selected using backward selection.

Criterion based methods

- Another way to compare and select between models is to use criterion based methods.
- For variable selection in linear models, the general form of many variable selection criteria for two nested models $m \subset m_0$ is given by

$$\Lambda(a) = \lambda - a(k_{m_0} - k_m)$$

where λ denotes the likelihood ratio statistics for testing m against m_0 and is given by

$$\lambda = -2\log\left[\frac{\max p(\mathbf{y}|\beta_m, \sigma^2)}{\max p(\mathbf{y}|\beta_{m_0}, \sigma^2)}\right]$$

where β_m and β_{m_0} refer to the regression coefficients for model m and model m_0, respectively, and k_m and k_{m_0} refer to the rank of matrices $\mathbf{X}_m$ and $\mathbf{X}_{m_0}$ with the covariates for models m and model m_0, respectively.

DIC

- The model based criterion is: $\Lambda(a) = \lambda - a(k_{m_0} - k_m)$
- The term a is a penalty term that quantifies the penalty for overfitting.
  If $a \geq 1$, then smaller models are favored compared to larger models.
- In Akaike Information Criteria (AIC), a is chosen to be equal to 2.
  In Bayesian Information Criteria (BIC), a is chosen to be equal to $log(n)$.
- For Bayesian model selection, the criterion often used is the Deviance Information Criteria (DIC). A model with a smaller DIC is better.
- DIC is given by the following expression:

$$DIC = p_D + \bar{D}$$

where p_D is called the effective number of parameters.

DIC

- If $\mathbf{y}$ represents the data, $p(\mathbf{y}|\theta)$ represents the sampling model for the data and θ is a parameter vector, then p_D is defined as the following difference

$$\begin{aligned} p_D &= \bar{D} - D(\bar{\theta}) \\ &= E_{\theta|\mathbf{y}}\left[-2\log p(\mathbf{y}|\theta)\right] - \left[-2\log\left(p(\mathbf{y}|E\left[\theta|\mathbf{y}\right])\right)\right] \end{aligned}$$

- Thus, $\bar{D}$ is the posterior mean of minus twice the log-likelihood, while $D(\bar{\theta})$ is minus twice the log-likelihood evaluated at the posterior mean of the parameters.
- DIC can be easily implemented within an MCMC algorithm: we just compute minus twice the log-likelihood for the given value of θ in the current MCMC iteration and then average it over the number of MCMC iterations!