Linear Regression Models II

Statistics 220

Spring 2005

Comparing Regression Models

Model 1:

$$E[\text{CityFuel}] = \beta_1\text{Weight} + \beta_2\text{EngSize} + \beta_3\text{Domestic}$$
$$+\beta_4 I(\text{Type} = \text{Compact}) + \beta_5 I(\text{Type} = \text{Large}) + \beta_6 I(\text{Type} = \text{Midsize})$$
$$+\beta_7 I(\text{Type} = \text{Small}) + \beta_8 I(\text{Type} = \text{Sporty}) + \beta_9 I(\text{Type} = \text{Van})$$

Model 2:

$$E[\text{CityFuel}] = \beta_1\text{Weight} + \beta_2\text{EngSize} + \beta_3\text{Domestic} + \beta_4$$

Do we get significantly better fit when we include the car type in the model.

There are a couple of ways of examining this:

- Examine the distributions of $\beta_i - \beta_j | y;\ i, j = 4, \ldots, 9$ in Model 1
- Compare DICs for the two models.

Implementation:

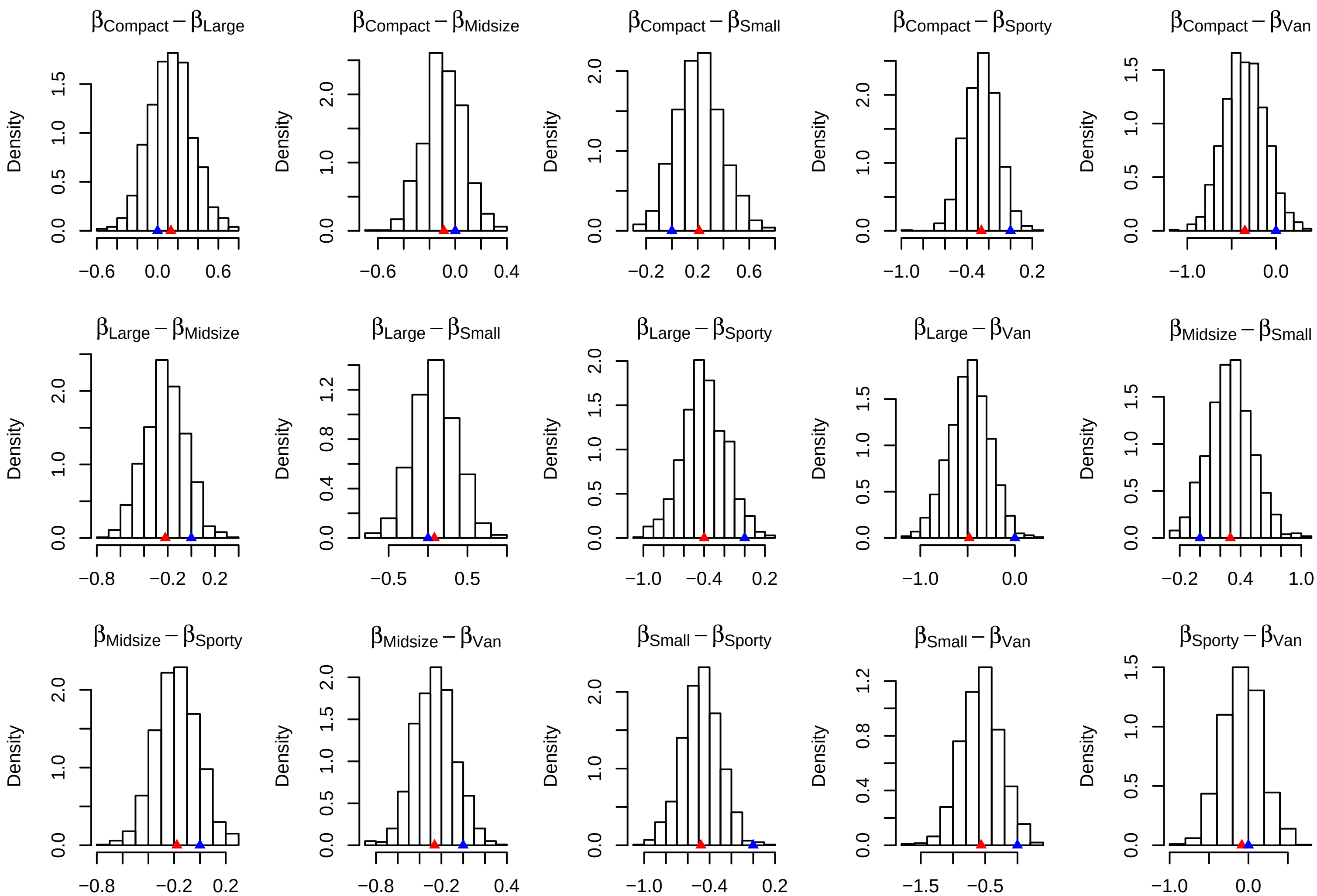
Both models where examined with WinBUGS with the non-informative prior

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

approximated by

$$\begin{aligned} \beta_i &\sim N(0, 10^6) \\ \sigma^2 &\sim \text{Inv}-\text{Gamma}(0.001, 0.001) \end{aligned}$$

Posterior distributions of $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j|\mathbf{y}$



```
5 chains, each with 2000 iterations (first 1000 discarded),
n.thin = 5, n.sims = 1000 iterations saved
Time difference of 9 secs
         mean  sd 2.5%  25%  50%   75% 97.5% Rhat n.eff
beta[1]   1.1 0.2  0.6  0.9  1.1   1.2   1.5    1  1000
beta[2]   0.2 0.1  0.0  0.1  0.2   0.2   0.4    1  1000
beta[3]   0.1 0.1 -0.1  0.0  0.1   0.1   0.3    1   710
beta[4]   0.9 0.6 -0.1  0.6  0.9   1.3   2.1    1  1000
beta[5]   0.8 0.7 -0.5  0.4  0.8   1.3   2.1    1  1000
beta[6]   1.0 0.6 -0.2  0.6  1.0   1.5   2.2    1  1000
beta[7]   0.7 0.5 -0.1  0.4  0.7   1.0   1.7    1  1000
beta[8]   1.2 0.6  0.2  0.8  1.2   1.6   2.3    1  1000
beta[9]   1.3 0.7 -0.1  0.8  1.3   1.8   2.7    1  1000
sigma     0.4 0.0  0.4  0.4  0.4   0.4   0.5    1   370
deviance 99.5 4.9 92.1 95.9 98.9 102.3 110.8    1   650
 pD = 11.8 and DIC = 111.3 (using the rule, pD = var(deviance)/2)

5 chains, each with 2000 iterations (first 1000 discarded),
n.thin = 5, n.sims = 1000 iterations saved
```

```
Time difference of 5 secs
          mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
beta[1]    1.4 0.1   1.1   1.3   1.4   1.5   1.7    1  1000
beta[2]    0.1 0.1  -0.1   0.0   0.1   0.1   0.2    1  1000
beta[3]    0.1 0.1  -0.1   0.0   0.1   0.2   0.3    1  1000
beta[4]    0.3 0.3  -0.2   0.1   0.3   0.5   0.8    1  1000
sigma      0.4 0.0   0.4   0.4   0.4   0.5   0.5    1   750
deviance 107.6 3.1 103.2 105.3 107.2 109.4 114.7    1  1000
 pD = 4.7 and DIC = 112.4 (using the rule, pD = var(deviance)/2)
```

Based on the distributions of $\beta_i - \beta_j | y$, it appears that some types of cars do get different gas mileage, such as Compacts and Vans or Small and Sporty.

However, from a prediction point of view, it doesn't seem to be a big difference as the increase in DIC for Model 2 is very small, suggesting that the we are not getting a great improvement in fit with the extra 5 parameters.

Including Prior Information

It is possible (of course) to include informative priors in regression models. While any proper prior could be used, a common approach is to us an analogue to the semi-conjugate normal model discussed in Chapter 3.

This prior is of the form

$$\begin{aligned} \beta &\sim N(\beta_0, \Sigma_\beta) \\ \sigma^2 &\sim \mathrm{Inv}-\chi^2(n_0, \sigma_0^2) \end{aligned}$$

While Σ_β can be any valid variance-covariance matrix, often it will be diagonal (e.g. $\Sigma_\beta = \mathrm{diag}(\sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_k})$), implying all parameters are independent apriori.

When putting a proper prior on β you often will want to use different variances for the different parameters for a number of reasons

- The values of the individual β_is will depend on the scale of the predictor variables, x_i. For example if you change the scale of an x_i from pounds to kilograms, you need to adjust the variance by a factor of 4.852.

- Different prior beliefs on the different βs

The analysis of this model needs be done by Monte Carlo techniques such as the Gibbs Sampler, as the marginal posteriors aren't nice.

However the conditional posteriors are as

- $\beta|\sigma^2, y \sim N(\mu, \Lambda)$ with

$$\begin{aligned}\Lambda &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1} \\ \mu &= \Lambda \left(\Sigma_\beta^{-1}\beta_0 + \frac{1}{\sigma^2} X^T y\right)\end{aligned}$$

- $\sigma^2|\beta, y$

$$\sigma^2|\beta, y \sim \text{Inv}-\chi^2 \left(n_0 + n, \frac{n_0\sigma_0^2 + ns^2}{n_0 + n}\right)$$

where

$$s^2 = \frac{1}{n}(y - X\beta)^T (y - X\beta)$$

Different Measurement Variance Structures

As mentioned earlier, the error structure of the observations does not have to to be independent with equal variance. In general

$$y|\beta, \Sigma_y \sim N(X\beta, \Sigma_y)$$

where Σ_y is a symmetric, positive definite matrix.

This matrix can come from many different approaches

- Variance matrix known up to a scalar factor

  $$\Sigma_y = Q_y \sigma^2$$

  where Q_y is a known fixed matrix and σ^2 is unknown.

  Inference in this case reduces to what we have seen before. Let $Q_y^{1/2}$ be a matrix square root of Q_y (e.g. $(Q_y^{1/2})^T Q_y^{1/2} = Q_y$). Then

$$Q_y^{-1/2} y | \beta, \sigma^2 \sim N(Q_y^{-1/2} X\beta, \sigma^2 I)$$

For example, if the $p(\beta, \sigma^2) \propto \sigma^{-2}$ noninformative prior is used, the earlier approach with

$$\begin{aligned}
\hat{\beta} &= (X^T Q_y^{-1} X)^{-1} X^T Q_y^{-1} y \\
V_\beta &= (X^T Q_y^{-1} X)^{-1} \\
s^2 &= \frac{1}{n-k} (y - X\hat{\beta})^T Q_y^{-1} (y - X\hat{\beta})
\end{aligned}$$

Note that the matrix inversions do not usually need to be calculated directly as $Q_y^{1/2}$ is usually determined by the Cholesky decomposition or the Singular Value decomposition and the inverse can be based on these.

One example where this approach is reasonable is Weighted regression where

$$Q_y = \text{diag}\left(\frac{1}{w_1}, \ldots, \frac{1}{w_n}\right)$$

where w_i are known as weights. This can occur if y_i is the average of w_i observations.

- Parametric models

  Instead of Q_y being a fixed matrix, it can be a function of a parameter ϕ. Examples of this include

  – Equal correlation

$$Q_y = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

– AR(1)

$$Q_y = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

If the $p(\beta, \sigma^2) \propto \sigma^{-2}$ noninformative prior is used for β and σ^2, the previous results can be used to get $p(\beta, \sigma^2|\phi, y)$. Then it can be shown that in this case

$$\begin{aligned} p(\phi|y) &= \frac{p(\beta, \sigma^2, \phi|y)}{p(\beta, \sigma^2|\phi, y)} \\ &\propto \frac{p(\phi)N(y|X\beta, \sigma^2 Q_y)}{\text{Inv}-\chi^2(\sigma^2|n-k, s^2)N(\beta|\hat{\beta}, V_\beta \sigma^2)} \\ &\propto p(\phi)|V_\beta|^{1/2}(s^2)^{-(n-k)/2} \end{aligned}$$

Note that $\hat{\beta}, V_\beta$, and s^2 are functions of ϕ so the posterior density is non-standard.

If an informative prior is put on β and/or σ^2, sampling will need to be done by an MCMC routine. Gibbs is often useful here, particularly if the $\mathrm{N{-}Inv{-}}\chi^2$ prior is placed on β, σ^2. In this case the conditional posteriors are

– $\beta|\sigma^2, \phi, y \sim N(\mu, \Lambda)$ with

$$\begin{aligned} \Lambda &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} X^T Q_y^{-1} X\right)^{-1} \\ \mu &= \Lambda \left(\Sigma_\beta^{-1}\beta_0 + \frac{1}{\sigma^2} X^T Q_y^{-1} y\right) \end{aligned}$$

– $\sigma^2|\beta, \phi, y$

$$\sigma^2|\beta, \phi, y \sim \mathrm{Inv{-}}\chi^2 \left(n_0 + n, \frac{n_0\sigma_0^2 + ns^2}{n_0 + n}\right)$$

where

$$s^2 = \frac{1}{n}(y - X\beta)^T Q_y^{-1}(y - X\beta)$$

Again $\hat{\beta}, V_\beta$, and s^2 are functions of ϕ in these two conditional posteriors.

– $\phi|\beta, \sigma^2, y$

This depends on the situation be will probably will have to be handled by something like acceptance - rejection sampling as a conjugate structure will be difficult in many situations

- Arbitrary matrices

It is possible for Σ_y to be an arbitrary, symmetric, positive definite matrix. Depending on the form of the prior on β and Σ_y, the posterior $p(\beta, \Sigma_y|y)$ can be difficult to handle, leading to MCMC approaches. However there are some cases where the posterior can be handled somewhat more easily.

– $p(\beta|\Sigma_y) \propto 1$

$$\beta|\Sigma_y, y \sim N((X^T\Sigma_y^{-1}X)^{-1}X^Ty, (X^T\Sigma_y^{-1}X)^{-1})$$

$$p(\Sigma_y|y) \propto p(\Sigma_y)|(X^T\Sigma_y^{-1}X)|^{-1/2}\exp\left(-\frac{1}{2}(y-X\hat{\beta})^T\Sigma_y^{-1}(y-X\hat{\beta})\right)$$

Usually this is difficult to handle, but is feasible if

$$\Sigma_y \sim \mathrm{Inv-Wishart}_\nu(S^{-1})$$

as this is a conjugate distribution in this case.

– $\beta|\Sigma_y \sim N(\beta_0, \Sigma_\beta)$

  This has a similar structure to before as $\beta|\Sigma_y, y \sim N(\mu, \Lambda)$ with

$$\begin{aligned}\Lambda &= \left(\Sigma_\beta^{-1} + X^T\Sigma_y^{-1}X\right)^{-1} \\ \mu &= \Lambda\left(\Sigma_\beta^{-1}\beta_0 + X^T\Sigma_y^{-1}y\right)\end{aligned}$$

(Let $\Sigma_y \to \infty \times I$ in above and the formula reduce to the uniform prior case.)

And again $p(\Sigma_y|y)$ will probably be tough to handle, except when $\Sigma_y \sim \mathrm{Inv-Wishart}_\nu(S^{-1})$

Posterior Predictive Distribution

As noted in the text, the posterior predictive distribution is more difficult as you need to consider the correlation between y and $\tilde{y}$.

However, the approach is the same regardless of the structure of Σ_y.

Assume that

$$\left(\begin{array}{c} y \\ \tilde{y} \end{array} \middle| X, \tilde{X}, \theta \right) \sim N\left(\left(\begin{array}{c} X\beta \\ \tilde{X}\beta \end{array} \right), \left(\begin{array}{cc} \Sigma_y & \Sigma_{y,\tilde{y}} \\ \Sigma_{\tilde{y},y} & \Sigma_{\tilde{y}} \end{array} \right) \right)$$

Then $\tilde{y}|\beta, \Sigma_y, y \sim N(\mu, \Lambda)$ with

$$\begin{aligned} \mu &= \tilde{X}\beta + \Sigma_{\tilde{y},y}\Sigma_y^{-1}(y - X\beta) \\ \Lambda &= \Sigma_{\tilde{y}} - \Sigma_{\tilde{y},y}\Sigma_y^{-1}\Sigma_{y,\tilde{y}} \end{aligned}$$

Thus simulation is not difficult, assuming that sampling from $p(\beta, \Sigma_y)$ is possible.

Also note that if y_i are independent, then the formulas reduce to the simpler cases we've seen before, except possibly for an adjustment if the y_i don't have equal variance.