Univariate and Multivariate Linear Regression
Univariate Linear Regression
Regression

$$P(\beta, \sigma^2 | y, X) = \left(2\pi \sigma^2 \right)^{-\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right)$$

\begin{align*}
\hat{\beta} &= (X'X)^{-1}X'y \\
\hat{\sigma}^2 &= \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-k}
\end{align*}

The LS/ML estimators

$$\text{Var} \left( \hat{\beta} \right) = \hat{\sigma}^2 (X'X)^{-1}$$

X is a n×k matrix of rank k
y is a n×1 vector
β is a k×1 vector
Regression Likelihood

\[(y - X\beta)'(y - X\beta) = (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta) + (y - X\hat{\beta})'(y - X\hat{\beta}) + 2(y - X\hat{\beta})'(X\hat{\beta} - X\beta)\]

\[= \nu s^2 + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)\]

where \(\nu s^2 = SSE = (y - X\hat{\beta})'(y - X\hat{\beta})\)

\(\nu = n - k\)

Substitution yields the likelihood:

\[p(y \mid X, \beta, \sigma^2) \propto (\sigma^2)^{-k/2} \exp\left[-\frac{1}{2\sigma^2} (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\]

\[\times (\sigma^2)^{-\nu/2} \exp\left[-\frac{\nu s^2}{2\sigma^2}\right]\]
Regression Model

\[ y \mid X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I) \]

For non-experimental data, don’t we need a model for the joint distribution of \(y\) and \(x\)?

\[ p(x, y) = p(x \mid \psi) p(y \mid x, \beta, \sigma^2) \]

This implies: \(p(x \mid \beta) = p(x)\) if \(\Psi\) is a priori independent of \((\beta, \sigma)\),

\[ p(\psi, \beta, \sigma^2 \mid y, X) \propto \left[ p(\psi) \prod_i p(x_i \mid \psi) \right] p(\beta, \sigma^2) \prod_i p(y_i \mid x_i, \beta, \sigma^2) \]

The likelihood factors: two separate analyses

[\(E\)] Are there cases where the likelihood does not factor?
Flat Priors

Classic “non-informative” prior (improper):

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

Several Issues:

• The prior is not strictly non-informative, because it says that $\beta$ is large with high prior “probability”

• It may be problematic computationally if there is singularity in $X'X$. There is no shrinkage.

$$p(\beta \mid y, X) \propto \left(1 + \frac{1}{(n-k)\hat{\sigma}^2} (\beta - \hat{\beta})' X' X (\beta - \hat{\beta})\right)^{-\frac{n}{2}} \sim T_{n-k} \left(\hat{\beta}, \hat{\sigma}^2 (X'X)^{-1}\right)$$

$$p(\sigma^2 \mid y, X) \propto (\sigma^2)^{-\frac{1}{2}(n-k-1)} \exp \left[\frac{1}{2\sigma^2} \hat{\sigma}^2 (n-k)\right] \sim IG \left(\frac{n-k-1}{2}, \frac{\hat{\sigma}^2 (n-k)}{2}\right)$$

[E:] Is there another way to write these posteriors

[hint: condition $p(\beta|y)$ on $\sigma$]?
Regression with Conjugate Priors

\[ P(\beta | \sigma^2) = N(b_0, \sigma^2 V_0) \]
\[ P(\sigma^2) = IG\left( \frac{r_0}{2}, \frac{s_0}{2} \right) \]
\[ P(\beta | y, \sigma^2) = N(b_p, \sigma^2 V_p) \]
\[ b_p = V_p \left( X'y + V_0^{-1}b_0 \right) \]
\[ V_p = \left( X'X + V_0^{-1} \right)^{-1} \]
\[ P(\sigma^2 | y) = IG\left( \frac{r_0 + n}{2}, \frac{s_0 + y'y - b'_p V_p b_p + b'_0 V_0^{-1}b_0}{2} \right) \]

Note that the setup is similar to that of the Normal distribution with conjugate prior.

Usually the prior parameters are set \( b_0=0, V_0=cl, r_0=6; s_0=2 \), (but var(y) may be better).

Caveat: this specification is less robust and sensitive to outliers.
Shrinkage

The Bayes Estimator is the posterior mean of $\beta$.

This is a “shrinkage” estimator (compare ridge regression).

$$b_p = \left( X' X + V_0^{-1} \right)^{-1} \left( X' y + V_0^{-1} b_0 \right)$$

shrinks $\hat{\beta} \to b_0$

as $n \to \infty$, $b_p \to \hat{\beta}$, because $X' X$ is of order $n$

$$\text{Var}(b_p \mid \sigma^2) = \sigma^2 \left( X' X + V_0^{-1} \right)^{-1} \leq \sigma^2 \left( X' X \right)^{-1} \text{ or } \sigma^2 V_0$$

[E:] How can the posterior variance be smaller than that of the LS/ML estimator, as the latter is efficient (James -Stein)?
Zellner’s g-Prior

\[ P(\beta \mid \sigma^2) = N(b_0, n_0 \sigma^2 (X'X)^{-1}) \quad \text{Conjugate prior} \]

\[ P(\sigma^2) \propto \sigma^{-2} \quad \text{Flat prior} \]

\[ P(\beta \mid y, \sigma^2) = N(b_p, \sigma^2 V_p) \]

\[ b_p = \frac{n_0}{n_0 + 1} \left( \hat{b} + \frac{b_0}{n_0} \right) \]

\[ V_p = \frac{n_0}{n_0 + 1} (X'X)^{-1} \]

\[ \hat{b} = (X'X)^{-1} X' y \]

\[ P(\sigma^2 \mid y) = IG \left( \frac{n-1}{2}, \left[ s^2 + \frac{1}{n_0+1} (b_0 - \hat{b}) X' X (b_0 - \hat{b}) \right] \right) \]

\( 1/n_0 \) ("g" in Zellner’s notation) is the amount of prior information.

\( n_0 = n \): prior has same weight as 1 observation

\( n_0 = 1 \): prior has same weight as \( n \) observations
Simulations

Scheme: \[ y | X, \beta, \sigma^2 \] \[ \beta | \sigma^2 \] \[ \sigma^2 \]

1) Draw \[ \sigma^2 | y, X \]
2) Draw \[ \beta | \sigma^2, y, X \]

[E:] Write an R-function that implements this for the uninformative and conjugate cases.
Predictive Distribution

The posterior predictive distribution of a new value $y^*$ with predictor values $X^*$ is a multivariate t-distribution:

$$p(y^* | X^*, X, y) = T_{n-k} \left(X^* b_p, \hat{\sigma}^2 \left(I_m - X^* X \right) \right)^{-1}$$

$$b_p = V_p \left( X'y + V_0^{-1} b_0 \right)$$

$$V_p = \left( X'X + V_0^{-1} \right)^{-1}$$

$$M = X'X + X^* X$$

[E]: Is there a way to do this computationally [hint what is $p(y^*|\sigma^2)$ ]?
Regression with Non-Conjugate Priors

\[ P(\beta) = N(b_0, V_0) \]
\[ P(\sigma^2) = IG(r_0/2, s_0/2) \]
\[ P(\beta | y, \sigma^2) = N(b_p, V_p) \]

\[ b_p = V_p \left( \frac{1}{\sigma^2} X' y + V_0^{-1} b_0 \right) \]
\[ V_p = \left( \frac{1}{\sigma^2} X' X + V_0^{-1} \right)^{-1} \]
\[ P(\sigma^2 | y, \beta) = IG \left( \frac{r_0 + n}{2}, \frac{s_0 + (y - X\beta)'(y - X\beta)}{2} \right) \]

Usually the prior parameters are set \( b_0 = 0, V_0 = c \cdot I, r_0 = 6; s_0 = 2. \)

The predictive distribution does not have a known form.

Note that the setup is similar to that of the Normal distribution with non-conjugate prior.
Simulations

Scheme: \[y|X, \beta, \sigma^2\] \[\beta|\sigma^2\] \[\sigma^2|\beta\]

1) Draw \[\sigma^2 | \beta, y\]

2) Draw \[\beta | \sigma^2, y\]

[E:] Write an R-function that implements this
Theory-Based Priors

Theory may predict, for example, price effects are negative. We can formulate a truncated prior, which leads to a truncated posterior, with c and d are vectors of lower and upper truncation bounds

\[
\begin{align*}
P(\beta) &= N(b_0, V_0) \times I(c < \beta < d) \\
P(\beta \mid y, \sigma^2) &= N(b_p, V_p) \times I(c < \beta < d)
\end{align*}
\]

\[
\begin{align*}
b_p &= \left( \frac{1}{\sigma^2} X' X + V_0^{-1} \right)^{-1} \left( \frac{1}{\sigma^2} X' y + V_0^{-1} b_0 \right) \\
V_p &= \left( \frac{1}{\sigma^2} X' X + V_0^{-1} \right)^{-1}
\end{align*}
\]

We can sample from the posteriors by

- discarding draws for which the constraint does not hold (inefficient)
- draw from the truncated MVN using the inverse cdf method (efficient)

[E:] Write an R-function that implements this
Imputing Missing Data $y^m$

In Bayesian statistics, missing data are treated as random variables. This allows a posterior distribution to be calculated which admits sampling, leading to multiple imputations.

We have $y = (y^o, y^m)'$

Scheme: $[y^m | X, \beta, \sigma^2] [\beta | \sigma^2] [\sigma^2 | \beta]$

1) Draw $[\sigma^2 | \beta, y]$

2) Draw $[\beta | \sigma^2, y]$

3) Draw $[y^m | X, \beta, \sigma^2]$

[E:] What prior is implied for $y^m$?

[E:] Write an R-function that implements this
Multivariate Regression
Suppose we have m regression equations:

\[ y_1 = X\beta_1 + \epsilon_1 \]

\[ \vdots \]

\[ y_c = X\beta_c + \epsilon_c \]

\[ \vdots \]

\[ y_m = X\beta_m + \epsilon_m \]

which can be written in matrix notation as:

\[ Y = XB + E, \]

\[ Y = \begin{bmatrix} y_1, \ldots, y_c, \ldots, y_m \end{bmatrix} \]

\[ B = \begin{bmatrix} \beta_1, \ldots, \beta_c, \ldots, \beta_m \end{bmatrix} \]

\[ E = \begin{bmatrix} \epsilon_1, \ldots, \epsilon_c, \ldots, \epsilon_m \end{bmatrix} \]

\[ \epsilon_{row} \sim \text{iid } N(0, \Sigma) \]

\[ \hat{B} = (X'X)^{-1}X'Y \]

\[ \hat{S} = \frac{1}{n-k} \left( Y - X\hat{B} \right)' \left( Y - X\hat{B} \right) \]

The LS/ML estimators
Multivariate Regression
Conjugate Prior and Posterior

The form of the likelihood suggests that a conjugate prior for \( \Sigma \) is an **Inverted Wishart**, and that for \( B \) is a **MV-Normal**.

**Prior:**

\[
\tilde{\beta} = \text{vec}(B'), \quad \tilde{\beta}_0 = \text{vec}(B'_0)
\]

\[
P(\tilde{\beta} | \Sigma) = N(\tilde{\beta}_0, V_0 \otimes \Sigma)
\]

\[
P(\Sigma) = \text{IW}(r_0, S_0)
\]

**Posterior:**

\[
P(\tilde{\beta} | Y, \Sigma) = N(\text{vec}(B_p), V_p \otimes \Sigma)
\]

\[
B_p = V_p \left( X'Y + V_0^{-1}B_0 \right)
\]

\[
V_p = \left( X'X + V_0^{-1} \right)^{-1}
\]

\[
P(\Sigma | Y, B) = \text{IW}(r_0 + n, \left[ S_0 + (Y'Y - B'_0V_0^{-1}B_0 - B'_pV_p^{-1}B_p)^{-1} \right])
\]

The prior setup is similar to that of the univariate regression.
Multivariate Regression
non-Conjugate Prior and Posterior

Prior: \[ \tilde{\beta} = \text{vec}(B'), \quad \tilde{y} = \text{vec}(Y') \]
\[ P(\tilde{\beta}) = N(b_0, V_0) \]
\[ P(\Sigma) = IW(r_0, S_0) \]

Posterior: \[ P(\tilde{\beta} | \tilde{y}, \Sigma) = N(b_p, V_p) \]
\[ b_p = V_p \left( (X' \otimes \Sigma^{-1}) \tilde{y} + V_0^{-1} b_0 \right) \]
\[ V_p = \left( X' X \otimes \Sigma^{-1} + V_0^{-1} \right)^{-1} \]
\[ P(\Sigma | Y, B) = IW\left( r_0 + n, \left[ S_0 + (Y - XB)'(Y - XB) \right] \right) \]

The prior setup is similar to that of the univariate regression.
Gibbs Samplers

The samplers for the conjugate and non-conjugate multivariate regression models are similar to those in the univariate case:

1. Conjugate draw: \([\mathbf{B} | \Sigma], [\Sigma]\)
2. Non-Conjugate draw: \([\mathbf{B} | \Sigma], [\Sigma | \mathbf{B}]\)

[E:] Write R-functions that implement this
Summary

• Univariate Regression
  – Conjugate and non-Conjugate priors

• Multivariate Regression
  – Conjugate and non-Conjugate priors
Some review questions

- How do you set the prior parameters, can you choose the data mean (for the Normal)?
- How can you impute missing values?
- How do you choose starting values?
- When do you use (non)conjugate priors?