Linear Models

- A linear model is defined by the expression

$$x = F\beta + \epsilon.$$

- where $x = (x_1, x_2, \ldots, x_n)'$ is vector of size n usually known as the *response vector*.
- $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the transpose of a vector of dimension p also known as parameter vector.
- F is a matrix of known elements and of dimension $n \times p$ with rows denoted by f_i' also known as design matrix.
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ is a vector of size n that contain the models errors.
- These errors are usually assumed iid with $\epsilon_i \sim N(0, \sigma^2)$

- *Least squares:* Find the value of β so that the sum of squares $S = (x - F\beta)^{'}(x - F\beta)$ reaches its minimum.
- *MLE:* Find the value of β that produces the maximum likelihood.
- Under normality and assuming σ^2 known, the log-likelihood for β is given by

$$l(\beta) = c - (n/2)log(\sigma^2) - (1/2\sigma^2)(x - F\beta)^{'}(x - F\beta)$$

- Also under normality, the MLE and LSE of β is

$$b = (F^{'}F)^{-1}F^{'}x(= \hat{\beta})$$

- Other concepts: The residual sum of squares is $R = (x - Fb)^{'}(x - Fb)$

- This sum of squares is associated with $n-p$ degrees of freedom since $R/\sigma^2 \sim \chi^2_{n-p}$.
- An unbiased estimate of the variance is $s^2 = R/(n-p)$ which is not the same to the MLE of σ^2 ($\hat{\sigma^2} = R/n$).
- We also have the *sum of squares factorization* $x^{'}x = R + b^{'}F^{'}Fb$.
- "Total sum of squares is equal to the residual sum of squares plus the regression sum of squares".

Bayesian Statistics

- The main goal of Bayesian analysis is to incorporate prior information into statistical modeling.
- This leads into the treatment of “observations” and “parameters” as random variables.
- A Bayesian model is established in a *hierarchical* way.
- First we define a probability distribution for the observations (likelihood) given a specific value of the parameter
$$f(x_1, x_2, \ldots, x_n|\theta)$$
- We also specify a probability distribution for θ known, the prior distribution $p(\theta)$, which reflects the current state of knowledge for θ.

- After a *likelihood* and a *prior* are specified, Bayesians compute the posterior distribution given by *Bayes Theorem*

$$p(\theta|x_1, x_2, \ldots, x_n) \propto f(x_1, x_2, \ldots, x_n|\theta)p(\theta)$$

- The proportionality constant is given by the *marginal* distribution of the data

$$p(x_1, x_2, \ldots, x_n) = \int f(x_1, x_2, \ldots, x_n|\theta)p(\theta)d\theta$$

- All the inferences are based on the posterior distribution.
- If we wish to estimate θ, we could use the *posterior expectation*

$$E(\theta|x) = \int \theta \; p(\theta|x)d\theta$$

where x represents the data vector $x = (x_1, x_2, \ldots, x_n)$.

- If we want to predict a future value x_f we use the *predictive distribution* of x_f given the data x,

$$p(x_f|x) = \int p(x_f|\theta, x)p(\theta|x)d\theta$$

- Usually the computations related to Bayesian Statistics require numerical evaluation of complicated integrals except in specific cases known as *conjugate models*.
- Example of conjugate model: Binomial data-Beta prior.
- Outside conjugate models it is usually hard to determine a prior distribution (requires scientific and probability knowledge).
- To deal with this problem, some statisticians appeal to

non-informative (objective) prior distributions.

- Non-informative priors have the purpose of reflecting lack of prior knowledge. A starting point to run Bayes machinery.

- Non-informative priors are also known as *reference priors*.

- An intuitive choice for a non-informative prior is the *Uniform*

$$p(\theta) \propto 1$$

  also known as *flat* prior.

- Both Bayes and Laplace proposed this prior as a default non-informative prior.

- However, this prior distribution is not invariant for one-to-one transformations.

- A famous probabilist, Jeffreys, derived the invariant non-informative prior.
- *Jeffreys' rule* $p(\theta) \propto |I(\theta)|^{1/2}$ where $I(\theta)$ denotes *Expected information.* $I(\theta) = E_{X|\theta}\left(-\frac{d^2 log\ f(x|\theta)}{d^2\theta}\right)$
- In the Binomial-Beta example, Jeffreys' prior is:

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

- If we have a probability model with a location parameter μ and a scale parameter σ^2, Jeffreys' prior becomes:

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

- For more information about Bayesian Statistics you may want to check Tim Hanson's course page. http://www.stat.unm.edu/~hanson/sta579/sta579.html

Summary of Bayes results for the Linear Model

- For the linear model, β and σ^2 are essentially location/scale parameters.
- The default non-informative prior for β and σ^2 is:

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

- With Bayes theorem the posterior distribution is given by

$$p(\beta, \sigma^2 | x, F) \propto f(x|\beta, \sigma^2)(1/\sigma^2)$$

- Under this prior, the posterior distribution for (β, σ^2) is a *Normal-Gamma* distribution.
- Conditional on σ^2, the posterior for β is a p-dimensional Normal with mean b and a covariance matrix $\sigma^2(F^{'}F)^{-1}$

or $\beta \sim N(b, \sigma^2(F'F)^{-1})$.

- The marginal posterior distribution for σ^2 is an Inverse Gamma with shape parameter $n/2$ and scale parameter $R/2$ or $\sigma^2 \sim IG(n/2, R/2)$
- The product of this p-dimensional Normal and the Inverse Gamma defines the Normal/Gamma posterior.
- For the marginal posterior distribution of β we need

$$p(\beta|x, F) = \int p(\beta, \sigma^2|x, F)d\sigma^2$$

- After some algebraic manipulation, it can be shown that

$$p(\beta|x, F) = c(n, p)|F'F|^{1/2}/(1+(\beta-b)'F'F(\beta-b)/ps^2)^{n/2}$$

- Roughly, for n large $p(\beta|x, F) \approx N(b, s^2(F'F)^{-1})$.

- The marginal density of x given F is,

$$p(x|F) = \int p(x|\beta, \sigma^2) p(\beta, \sigma^2) d\beta d\sigma^2 = c|F^{'}F|^{-1/2}/R^{(n-p)/2}$$

- Due to the sum of squares factorization, we can establish that

$$p(x|F) \propto |F^{'}F|^{-1/2}(1 - b^{'}F^{'}Fb/(x^{'}x))^{(p-n)/2}$$

- If we think of F as a "parameter", $p(x|F)$ is a likelihood that could be used to produce inferences on F or on quantities that determine F (*marginal likelihood*).
- Under orthogonality of the F matrix, the evaluation of $p(x|F)$ becomes really easy.
- F orthogonal means that $F^{'}F = kI$