Linear and Logistic Regression

Michael Anderson, PhD

Department of Biostatistics and Epidemiology
The University of Oklahoma Health Sciences Center

Sept 16, 2016

Outline

Bayesian Linear Regression

When a response variable is continuous and predictors are continuous or categorical.

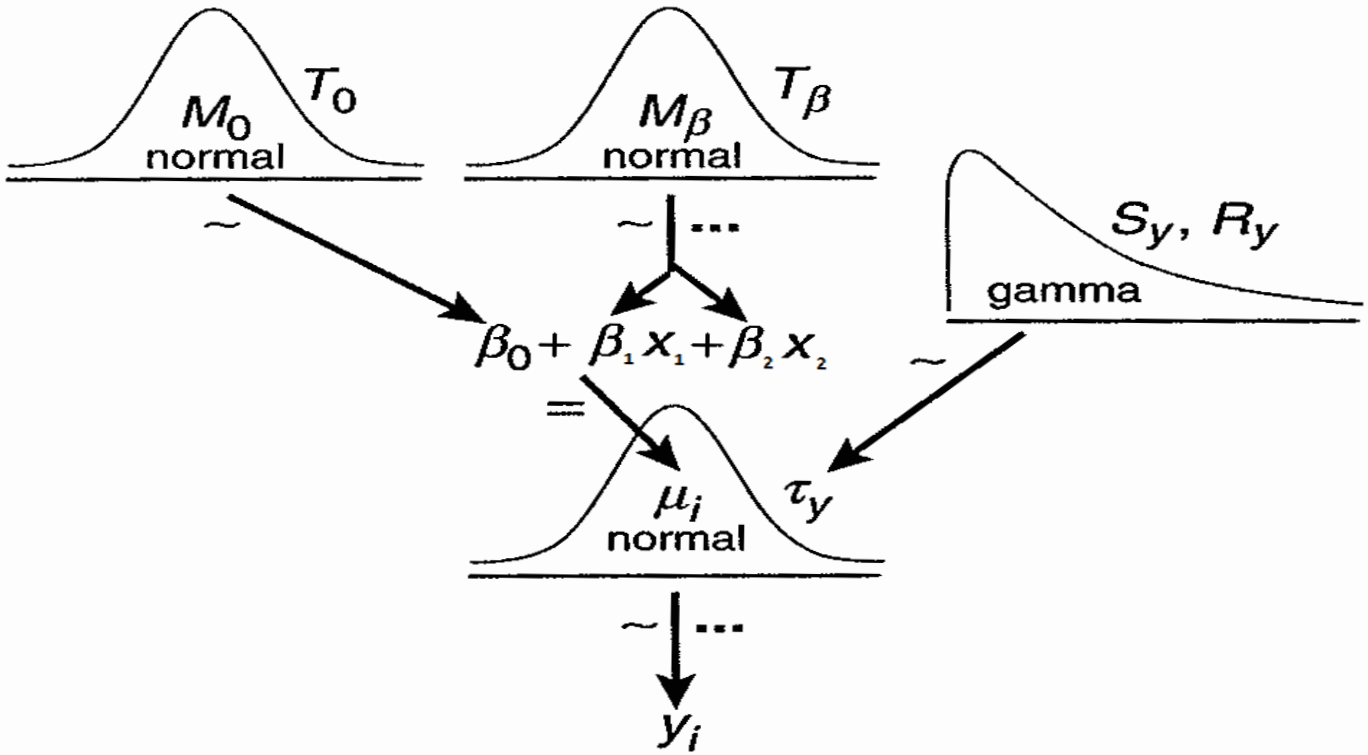
- $y_i \sim N(\mu, \tau)$ and $\mu = \beta_0 + \beta_1 x_1 + \ldots$.
  - where $\tau = 1/\sigma^2$.
  - σ^2 is the "residual" variance.
- Typically put Normal priors on βs.
- Using standardized data can make MCMC more efficient.
  - Correlation between intercept and slope can cause MCMC to stall.
  - Standardizing the data $z_y = \frac{y-\bar{y}}{SD_y}$ and $z_x = \frac{x-\bar{x}}{SD_x}$ can help.

Bayesian Linear Regression

Why consider using a Bayesian approach here?

- You may want to control for known confounders.
- Posterior will describe the uncertainty of parameter rather than just give point estimates.
- Complex relationship between predictor and response may exist.
- Missing data can be easily handled in the Bayesian framework.
- We can model the variance $\sigma^2 = \frac{1}{\tau}$.
  - Estimating variances is non trival but is often trivialized with frequentist approaches.
  - Bayesian framework provides a sophisticated treatment of variances.
  - Priors with appropriate (positive) support can be used as prior distributions for variances.
  - Common prior distributions for variances include the Gamma and Uniform.

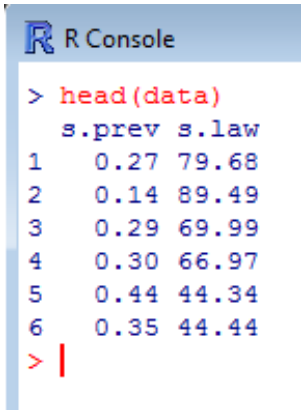
Bayesian Linear Regression



Bayesian Linear Regression

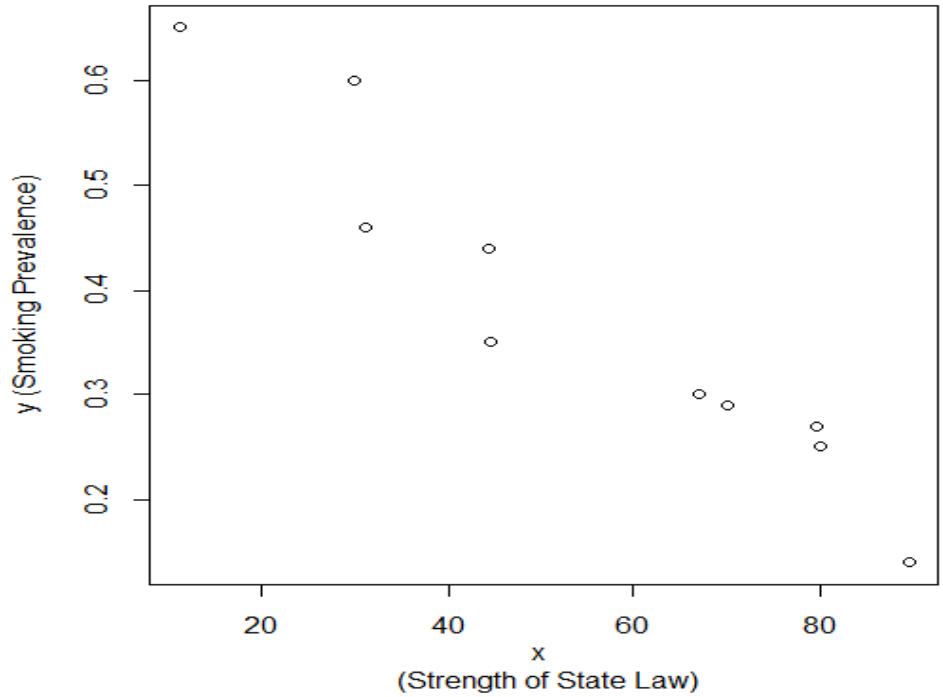
Example

- Locate the files "smoking slr.R" and "smoking.bugs.txt"
- 10 US states with smoking prevalence (y) and strength of state law (x).



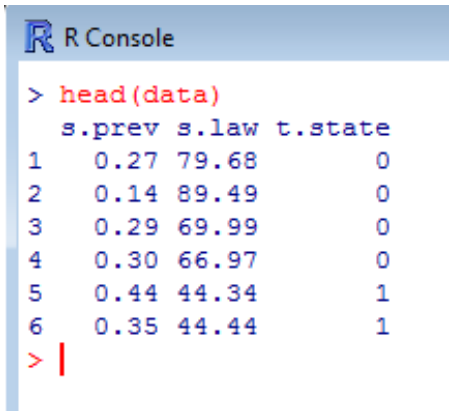
Bayesian Linear Regression

Example:
10 US states with smoking prevalence (y) and strength of state law (x).



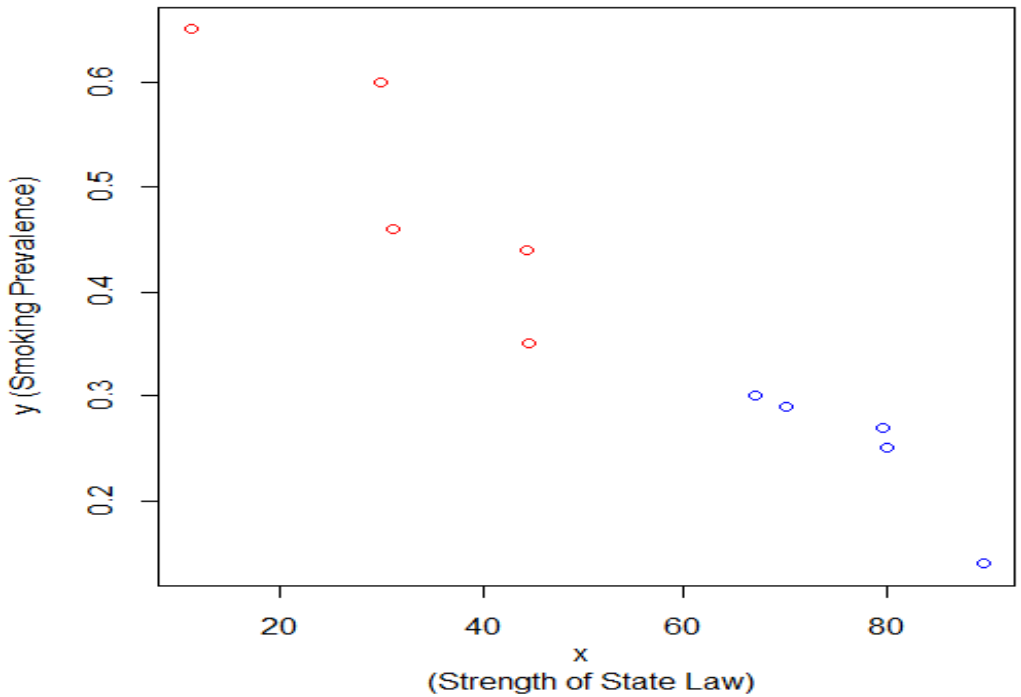
Bayesian Linear Regression

Suppose we also had information on whether the state produces tobacco. Smoking prevalence (y), Strength of state law (x1), Tobacco Production (x2).



Bayesian Linear Regression

Example:
Smoking prevalence (y), Strength of state law (x) and Tobacco production (red=Yes).



See the files "smoking mlr.R", "smoking2.bugs.txt", and

Model Diagnostics: Deviance Information Criterion (DIC)

Q: When comparing 2 or more models, is the improvement in fit large engough to justify the added difficulty in fitting?
A: Examine the DIC

- Deviance $= D(y, \theta_l) = -2logp(y|\theta_l)$
  - Is a numerical measure of model fit.
  - It is $\propto$ MSE when the likelihood is Normal w/constant variance
- DIC=$2\widehat{D}_{avg}(y) - D_{\widehat{\theta}}(y)$
  - where $\widehat{D}_{avg}(y) = \frac{1}{L}\sum D(y, \theta_l)$ which is the deviance averaged over all posterior draws.
  - and $D_{\widehat{\theta}}(y) = D(y, \widehat{\theta})$ which is the deviance at the posterior mean.
- Alternatively, DIC=$p_D - \widehat{D}_{avg}(y)$ where p_D is the effective number of parameters.
- DIC can be thought of as the expected prediction error. Smaller is better.

Bayesian Linear Regression

In the Smoking Prevalence example we fit two models.

- Model 1: Simple linear regression (x1=strength of state law) DIC=-21.9
- Model 2: Multivariable linear regression (x1 as above and x2=tabacco producing state) DIC=-16.1

We see that adding x2 to the model does NOT improve the model fit enough to justify the added compexitiy of including x2 in the model. Here the simple linear regresssion model would be the better fitting model.

Random Component

Regression involves relating observed responses $(y_1, \ldots, y_n)$ to predictors $\alpha + \beta_1 x_1$.

- The random component specifies a probability distribution for $(y_1, \ldots, y_n)$.
  - $y_i \sim N(\mu, \tau)$.
    - This is general linear regression.
  - $y_i \sim Bin(n, \pi)$.
  - $y_i \sim Poi(\lambda)$.

Systematic Component

Regression involves relating observed responses $(y_1, \ldots, y_n)$ to predictors $\alpha + \beta_1 x_1$.

- The systematic component specifies the predictors $\alpha + \beta_1 x_1$.
- This is known as a linear predictor.

Link Component

Regression involves relating observed responses $(y_1, \ldots, y_n)$ to predictors $\alpha + \beta_1 x_1$.

- The link connects the random and systematic components.
- It specifies how $\mu = E(Y)$ relates to $\alpha + \beta_1 x_1$.
- $g(\mu) = \alpha + \beta_1 x_1$ where $g(.)$ is the link function.
  - $g(\mu) = \mu$ is the *identity link*.
  - $g(\mu) = log(\mu)$ is the *log link*.
    - used when $0 < \mu$ as with count data.
    - produces "log-linear" models.
  - $g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$ is the *logit link*.
    - used when $0 < \mu < 1$ as when μ is a probability.
    - produces "logit" models.
    - referred to as logistic regression.

Logistic Regression Model

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1$$

- $0 < \pi < 1$.
- $-\infty < log\left(\frac{\pi}{1-\pi}\right) < \infty$.
- So this relates predictors to π in a common sense way.

Note:

$$log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1$$

implies

$$\frac{\pi}{1-\pi} = e^{\alpha+\beta_1 x_1} = e^{\alpha} e^{\beta_1 x_1}$$

and

$$\pi = \frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}$$

Relationship Between Snoring and Heart Disease

- PG Norton and EV Dunn (1985) collected data from 2484 subjects to determine the association between snoring (Never=1, Occasionally=2, Nearly Every Night=3, Every night=4) and heart disease (Yes=1/No=0).
  - Never: 24 HD 1355 No HD (1.7%)
  - Occasionally: 35 HD 603 No HD (5.5%)
  - Nearly Every Night: 21 HD 192 No HD (9.9%)
  - Every Night: 30 HD 224 No HD (11.8%)
- Let's use JAGS to determine whether there is an association between snoring and heart disease.

In Class Practice Problems

See files "snoring.R" "snoring.data.txt" "snoring.bugs.txt"