Regression Models

- Bayesian analysis of the "ordinary linear regression" (Chapter 8)
- Example: analysis of radon measurements
- Hierarchical Linear Regression Models (Chapter 13)
- Simple random effect model
- Mixed effect model
- Bayesian Variable Selection

Ordinary Linear Regression

- Question: how does one quantity, y, vary as a function of another quantity x?
- $y = (y_1, \ldots, y_n)$ is the continuous outcome
- $x_i = (x_{i1}, \ldots, x_{ik})$ is the vector of explanatory variables, discrete or continuous
- X is the $n \times k$ matrix of the explanatory variables

$$\begin{aligned} E[y_i \mid \beta, X] &= \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ Var[y_i \mid \beta, X] &= \sigma^2, \text{ for } i = 1, \ldots, n \\ \theta &= (\beta_1, \ldots, \beta_k, \sigma^2) \end{aligned}$$

1. define x and y (possibly using transformations) so that $E[y_i \mid \beta, X]$ is approximately linear, with normal errors

2. setting up a prior for θ that accurately reflects substantive knowledge

Ordinary Linear Regression, Basic Model

$$\begin{aligned} y \mid \beta, \sigma^2, X &\sim N(X\beta, \sigma^2 I) \\ p(\beta, \sigma^2 \mid X) &\sim \sigma^{-2} \end{aligned}$$

conditional distribution of β given σ

$$\begin{aligned} \beta \mid \sigma^2, y &\sim N(\hat{\beta}, V_\beta \sigma^2) \\ \hat{\beta} &= (X^t X)^{-1} X^t y \\ V_\beta &= (X^t X)^{-1} \end{aligned}$$

conditional distribution of σ

$$\begin{aligned} \sigma^2 \mid y &\sim \chi^{-2}(n-k, s^2) \\ s^2 &= \tfrac{1}{n-k}(y - X\hat{\beta})^t (y - X\hat{\beta}) \end{aligned}$$

$\hat{\beta}$ and s^2 are the MLE.

When the posterior is proper?

For any analysis based on a improper prior distribution, it is important to check that the posterior distribution is proper.

$p(\beta, \sigma^2 \mid y)$ is a proper as long as

1. $n > k$
2. rank$(X) = k$ (e.g. columns of X must be linearly independent)

Sampling from the posterior distribution

- draw σ^2 from $\chi^{-2}(n-k, s^2)$
- draw β conditionally to σ^2 from $N(\hat{\beta}, V_\beta \sigma^2)$

Posterior Predictive distribution

- $\tilde{X}$ new data
- $\tilde{y}$ future outcomes
- posterior predictive simulations

To draw a random sample $\tilde{y}$ from $p(\tilde{y} \mid y)$, we

1. draw $\sigma^{2(j)}, \beta^{(j)}$ from $p(\sigma^2, \beta \mid y)$
2. draw $\tilde{y} \sim N(\tilde{X}\beta^{(j)}, \sigma^{2(j)} I)$ for $j = 1, \ldots, N$

Analytic form of the posterior predictive distribution

$$\tilde{y} \mid y \sim t_{n-k}(\tilde{X}\hat{\beta}, s^2(I + \tilde{X}V_\beta\tilde{X}^t))$$

$$\begin{aligned} E[\tilde{y} \mid y] &= E[E(\tilde{y} \mid \beta, \sigma^2, y) \mid \sigma^2, y] \\ &= E[\tilde{X}\beta \mid \sigma^2, y] \\ &= \tilde{X}\hat{\beta} \end{aligned}$$

$$\begin{aligned} Var[\tilde{y} \mid y, \sigma^2] &= E[Var(\tilde{y} \mid \beta, \sigma^2, y) \mid \sigma^2, y] \\ &+ Var[E(\tilde{y} \mid \beta, \sigma^2, y) \mid \sigma^2, y] = \\ E[\sigma^2 I \mid \sigma^2, y] &+ Var[\tilde{X}\beta \mid \sigma^2, y] = \\ \sigma^2 I &+ \sigma^2 \tilde{X} V_\beta \tilde{X}^t \end{aligned}$$

$p(\tilde{y} \mid y)$, has two components of uncertainty

1. variability of the model not accounted by $X\beta$ (σ^2)
2. posterior uncertainty in β and σ^2 due to the finite sample size of y

Table 1: *Short-term measurements of radon concentration in a sample of houses in three counties in Minnesota. All measurements were recorded on the basement level of the houses, except for those indicated with *, which were recorded on the first floor*

County	Radon Measurements
Blue Earth	5.0, 13.0, 7.2,6.8,12.8,5.8*,9.5,6.0,3.8,14.3*,1.8,6.9,4.7,9.5
Clay	0.9*,12.9,2.6,3.5*,26.6,1.5,13.0,8.8,19.5,2.5*,9.0,13.1,3.6,6.9*
Goodhue	14.3,6.9*,7.6,9.8*,2.6,43.5,4.9,3.5,4.8,5.6,3.5,3.9,6.7

Analysis of Radon measurements

- Exercise 8.1: fit a linear regression to the logarithms of the radon measurements, with indicator variables for the three counties and for whether a measurement was recorded on the first floor.
- Basic model

$$log(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

where

- x_{i1} is the basement indicator

- $x_{i2},\ x_{i3}$ and x_{i4} are counties indicators
- $\exp(\beta_2)$ denotes the geometric mean of the radon measurements (no in the basement) in the county Blue Earth
- $\exp(\beta_2+\beta_1)$ denotes the geometric mean of the radon measurements (in the basement) in the county Blue Earth
- Prior

$$\beta_1, \beta_2, \beta_3, \beta_4, \log(\sigma) \sim \text{constant}$$


```
#bayesian approach to ordinary linear regression, radon data-set pag.189 Gelman boox,
#ex 8.1(a) exercise 8.1(b) homework
#explanatory variables:basement indicator and three counties indicators
make.indicators_function(x){
  ux_unique(x)
  mat1_matrix(x,nrow=length(x),ncol=length(ux))
  mat2_matrix(ux,nrow=length(x),ncol=length(ux),byrow=T)
  (mat1==mat2)*1}

bayesian.regression_function(nsim=1000){
  y.1_c(5.0, 13.0, 7.2,6.8,12.8,5.8,9.5,6.0,3.8,14.3,1.8,6.9,4.7,9.5)
  y.2_c(0.9,12.9,2.6,3.5,26.6,1.5,13.0,8.8,19.5,2.5,9.0,13.1,3.6,6.9)
  y.3_c(14.3,6.9,7.6,9.8,2.6,43.5,4.9,3.5,4.8,5.6,3.5,3.9,6.7)
  basement.1_c(1,1,1,1,1,0,1,1,1,0,1,1,1,1)
  basement.2_c(0,1,1,0,1,1,1,1,1,0,1,1,1,0)
  basement.3_c(1,0,1,0,1,1,1,1,1,1,1,1,1,1)
  counties_rep(1:3,c(length(y.1),length(y.2),length(y.3)))
  y_c(y.1,y.2,y.3)
  x_cbind(c(basement.1,basement.2,basement.3),make.indicators(counties))
  ls.out_lsfit(x,log(y),intercept=F)
  lsd_ls.diag(ls.out)
  n_nrow(x)
  k_ncol(x)
  sigmasqr_rep(NA,nsim)
  beta_matrix(NA,nsim,k)
  for ( i in 1:nsim){
    sigmasqr[i]_lsd$std.dev*(n-k)/rchisq(1,n-k)
    PP_ t(x)%*%x
    VV_solve(PP)
    VV_.5*(VV+t(VV))
    beta[i,]_simulate.multnorm(as.numeric(ls.out$coef),sigmasqr[i]*VV)
}
output_exp(cbind(beta[,2],beta[,1]+beta[,2],beta[,3],beta[,1]+beta[,3],
                 beta[,4],beta[,1]+beta[,4],sigmasqr))
    for(i in 1:ncol(output)) print(round(quantile(output[,i],c(.25,.5,.75)),1))
return(beta)
}
```



```
                                                 Posterior quantiles
                                                25%  50%   75%
 geometric mean for Blue earth (no basement) exp(beta2)       3.9   5    6.6
 geometric mean for Blue earth (basement) exp(beta1 + beta2)  6     7    8.2
 geometric mean for Clay (no basement) exp(beta3)             3.7   4.8  6.1
geometric mean for Clay (basement) exp(beta1 + beta3)         5.5   6.5  7.9
 geometric mean for  Goodhue (no basement) exp(beta4)         3.8   4.9  6.2
 geometric mean for Goodhue (basement) exp(beta1 + beta4)     5.7   6.8  7.8
geometric std of predictions, exp(sigma)                      2     2.3  2.6
```


Hierarchical Regression Models (Chapter 13)

Hierarchical Model	+
Regression Model	=
Hierarchical Regression Model	

Hierarchical Regression Models are useful tools when:

- we have covariate information at different levels of variation
- data are obtained by stratified or cluster sampling

Posterior Distributions of the coefficients

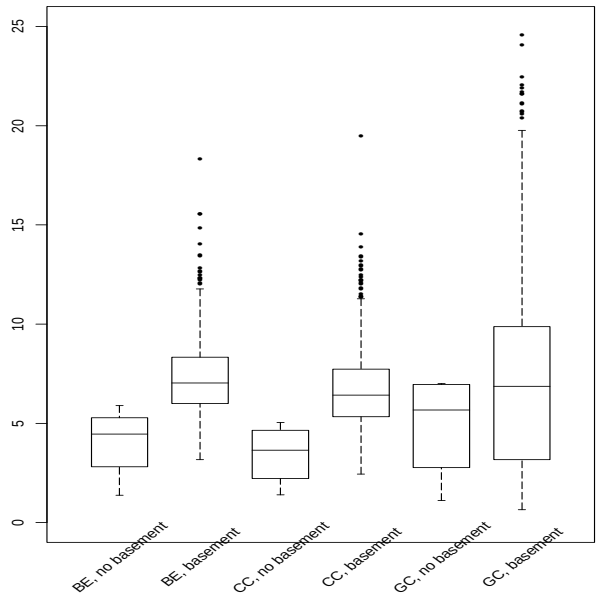

Figure 1: Boxplots of sample of the posterior distributions of the coefficients of "basement" and "no basement" for three counties

For example, in studying scholastic achievement we may have information about

1. individual students (family background)
2. class-level informations (characteristics of the teacher)
3. information about the school (educational policy, type of neighborhood)

rat tumor data set in Gelfand et al. (1990) is an example of hierarchical regression model

$$
\begin{aligned}
Y_{ij} &\sim N(\alpha_i + \beta_i x_{ij}, \sigma^2),\ i = 1, \ldots, k = 30 \\
\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &\sim N\left(\begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}, \Sigma \right),\ i = 1, \ldots, k \\
\sigma^2 &\sim IG(a, b) \\
\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} &\sim N\left(\begin{bmatrix} \eta_0 \\ \eta_1 \end{bmatrix}, C \right) \\
\Sigma^{-1} &\sim W((\rho R)^{-1}, \rho),\ E(\Sigma^{-1}) = R^{-1},\ var(\Sigma) \propto \rho^{-1}
\end{aligned}
$$

Hierarchical Regression Model

We assume a normal linear regression model for the likelihood

$$
\begin{aligned}
y \mid \beta, X, \Sigma_y &\sim N(X\beta, \Sigma_y) \\
dim(X) &= n \times J \text{ explanatory variables} \\
\beta &= (\beta_1, \ldots, \beta_J) \text{ regression coefficients} \\
E(y_i \mid \beta, X) &= \beta_1 x_{i1} + \ldots + \beta_J x_{iJ} \\
Var(y_i \mid \beta, X) &= [\Sigma_y]_{ii} \\
Cov(y_i, y_j \mid \beta, X) &= [\Sigma_y]_{ij}
\end{aligned}
$$

Simple random effect model

$$y \mid \beta, X, \Sigma_y \sim N_n(X\beta, \Sigma_y)$$
$$\beta \mid \alpha, \sigma_\beta^2 \sim N_J(\mathbf{1}\alpha, \sigma_\beta^2 I)$$

- $\mathbf{1}$ is a vector of ones
- $\sigma_\beta^2 = 0 \rightarrow$ all $\beta_j's$ equal
- $\sigma_\beta^2 = \infty \rightarrow$ unrelated $\beta_j's$
- $p(\log \sigma_\beta) =$ constant $\rightarrow$ improper posterior (sec.5.4) with all its mass on $\sigma_\beta = 0$
- we can use $\sigma_\beta^2 \sim \chi^2$ scaled with $df \leq 2$
- *always check that the posterior inferences are not sensitive to the choice of the prior distribution for* σ^2

Intraclass correlation

- data $y_1, \ldots, y_n$ fall into J batches

$$y \mid \alpha, \Sigma_y \sim N(\alpha, \Sigma_y)$$
$$Var(y_i) = \sigma^2, \quad \text{for all } i$$
$$Cov(y_{i1}, y_{i2}) = \begin{cases} \rho\sigma^2 & i_1, i_2 \text{ are in the same batch} \\ 0 & \text{otherwise} \end{cases}$$

- if $\rho \geq 0$, this is equivalent to the model

$$y \sim N_n(X\beta, \sigma^2 I)$$
$$\beta \sim N(1\alpha, \sigma_\beta^2 I)$$

where X is a $n \times J$ matrix of indicator variables such that $X_{ij} = 1$ if unit i is in batch j and 0 otherwise.

- the equivalence of the two models occurs when $\rho = \sigma_\beta^2/(\sigma^2 + \sigma_\beta^2)$.

Positive intraclass correlation in a linear regression can be subsumed into a random effect model by augmenting the regression with J indicator variables whose coefficients have the population distribution $\beta \sim N(1\alpha, \sigma_\beta^2 I)$

Mixed effect model

$$\beta = \begin{bmatrix} \beta_{J_1} \\ \beta_{J_2} \end{bmatrix}$$
$$\beta_{J_1} \sim \text{Unif}(-\infty, \infty)$$
$$\beta_{J_2} \sim N_{J_2}(1\alpha, \sigma_\beta^2 I), \; J_2 = J - J_1$$

- β_{J_1} are exchangeable with infinite prior variance and they are labeled *fixed effects*
- the random effect model with the school means normally distributed and a uniform prior density assumed for their mean α is equivalent to a mixed-effect model with a single constant *fixed-effect* and a set of *random effects* of mean 0.

Several sets of random effects

$$\beta = \begin{bmatrix} \beta_{J_1} \\ \vdots \\ \beta_{J_K} \end{bmatrix}$$

$$\beta_{J_k} \sim N_{J_k}(1\alpha_k, \sigma^2_{\beta_k} I)$$

- a mixed effect is obtained by setting the variance to ∞ for one of the clusters of random effects

General notation and computation for hierarchical linear models

$$\begin{aligned} y \mid X, \beta, \Sigma &\sim N_n(X\beta, \Sigma_y) && \textit{likelihood} \\ \beta \mid X_\beta, \alpha, \Sigma_\beta &\sim N_J(X_\beta \alpha, \Sigma_\beta) && \textit{population dist.} \\ \alpha \mid \alpha_0, \Sigma_\alpha &\sim N_K(\alpha_0, \Sigma_\alpha) && \textit{hyperprior dist.} \end{aligned}$$

- n data points y_i
- J parameters β_j
- K parameters α_k

Interpretation as a single regression

We consider the hierarchical model as a single normal regression model using a "larger" data set that includes as observations the information added by the population and hyperprior distributions

$$y_\star \mid X_\star, \gamma, \Sigma_\star \sim N(X_\star \gamma, \Sigma_\star)$$

where

- $\gamma = (\beta, \alpha),\ dim(\gamma) = J + K$
- $y_\star = (y, 0, \alpha_0)^t$
- $X_\star = \begin{pmatrix} X & 0 \\ I_j & -X_\beta \\ 0 & I_K \end{pmatrix},\ \Sigma_\star^{-1} = \begin{pmatrix} \Sigma_y^{-1} & 0 & 0 \\ 0 & \Sigma_y^{-1} & 0 \\ 0 & 0 & \Sigma_\alpha^{-1} \end{pmatrix}$

If any of the components of β or α have non informative prior distributions, the corresponding rows in $y_\star$ and $X_\star$, as well as the corresponding rows and columns in $\Sigma_\star^{-1}$, can be eliminated, because they correspond to observations with infinite variance. The resulting regression then has $n + J_\star + K_\star$ "observations" where $J_\star$ and $K_\star$ are the number of components of β and α with informative prior distributions.

Bayesian Variable Selection
(George and McCulloch JASA 1993)

- The problem is to find and fit the "best" model of the form

$$Y = X_1^\star \beta_1 + X_2^\star \beta_2 + \ldots + X_q^\star \beta_q + \epsilon$$

- where $(X_1^\star, X_2^\star, \ldots, X_q^\star)$ is a "selected" subset of $X_1, \ldots, X_p$.
- If we consider the canonical regression setup

$$y \sim N_n(X\beta, \sigma^2 I)$$

where $\beta = (\beta_1, \ldots, \beta_p)$.

- Selecting a subset of predictor is equivalent to setting to 0 those β_is corresponding to non selected predictors.

- Bayesian variable selection introduce a binary latent variable γ_i and assumes

$$\begin{aligned} \beta_i \gamma_i \quad &\sim (1-\gamma_i) N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2) \\ p(\gamma_i = 1) \; &= \; 1 - p(\gamma_i = 0) = p_i \end{aligned}$$

- p_i is the prior probability that β_i will require a non zero estimate, or equivalently that X_i will be included in the model.
- proming subset of predictors can be identified as those with higher posterior probability
- In practice, a common sense Bayesian perspective indicates that the key is to use substantive knowledge, either as a formal prior distribution or more informally, in choosing which variable to include.