Bayesian Inference and Decision Theory

Unit 8: Bayesian Regression

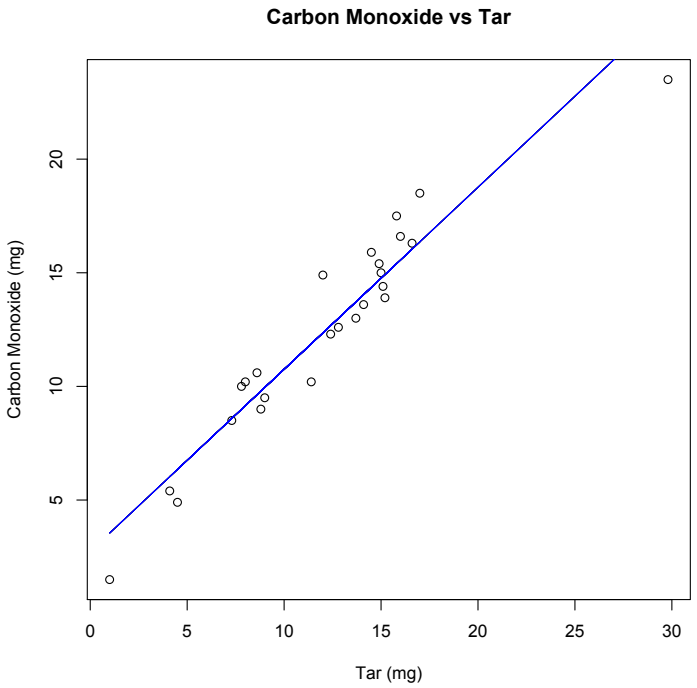
Learning Objectives for Unit 8

- Specify hierarchical models (aka multi-level models) for comparing samples from subgroups of a population
  - Each sample consists of iid observations from a subgroup of a larger population
  - Parameters for the groups are viewed as a sample from a larger population of groups
  - Information can be shared across groups
- Explain the benefits of Bayesian hierarchical models for complex multi-parameter problems
- Apply techniques from previous units to inference in hierarchical models
- Apply techniques from previous units to evaluate structural assumptions in hierarchical models

What is Regression?

- Regression models the relationship of one or more dependent (or response or outcome) variables to one or more independent (or explanatory) variables
  - Linear or nonlinear
  - One explanatory variable or many
- Objectives:
  - Understand relationship between explanatory variables and dependent variable(s)
  - Predict values of dependent variable(s) given particular values of independent variable(s)
  - Infer cause and effect relationships
  - Estimate systematic relationships and filter out noise

Example taken from http://www.mste.uiuc.edu/regression/cig.html

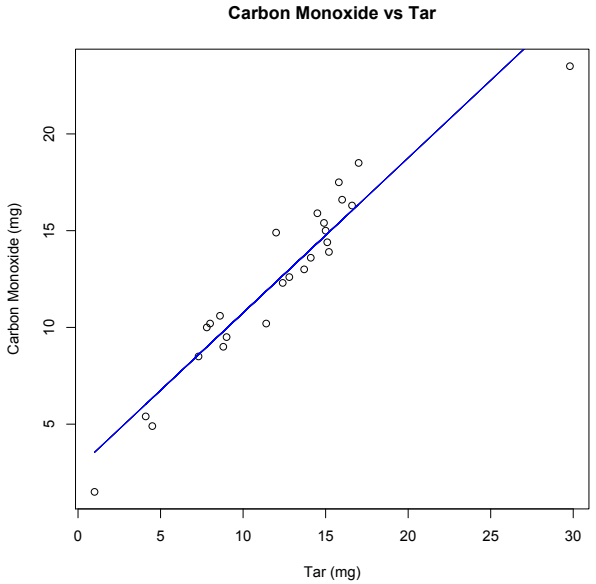


Example: Cigarette Data

- Data set consists of measurements of weight, tar, nicotine and CO content of 25 cigarette brands
- Graph shows scatterplot of data points and linear regression line with coefficients estimated by method of least squares

Simple Linear Regression

- Mean of dependent variable Y is linearly related to a single independent variable X
  - $E[Y \mid X] = \alpha + \beta X$
  - Equivalently, $Y = \alpha + \beta X + \varepsilon$ where $E[\varepsilon] = 0$
- Data consist of a sample of (Y_i, X_i) pairs
- Objectives of regression analysis:
  - Infer whether or not distribution of Y depends on X (whether slope of line is zero)
  - Estimate coefficients of relationship between Y and X (slope and intercept of line)
  - Find credible intervals for coefficients of the slope and intercept
  - Evaluate how much of the variability in Y is explained by X
  - Predict not-yet-observed Y for a given X value
  - Evaluate adequacy of model (are modeling assumptions met?)



Simple Linear Regression with Normal Homoscedastic* Errors

- Common assumptions for simple linear regression model:
  - Y given X is normally distributed
  - $E[Y \mid X] = \alpha + \beta X$
  - Observations $Y_{1:n}$ are independent given $X_{1:n}$
  - $Var[Y|X] = \sigma^2$ is independent of X
- The likelihood function: $f(\underline{y} \mid \underline{x}, \alpha, \beta, \sigma) = (2\pi)^{-n/2} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \alpha - \beta x_i)^2 \right\}$

How would you verify these assumptions?

*Homoscedastic (from ancient Greek "homo" (same) and "skedasis" (dispersion) means all observations have same variance

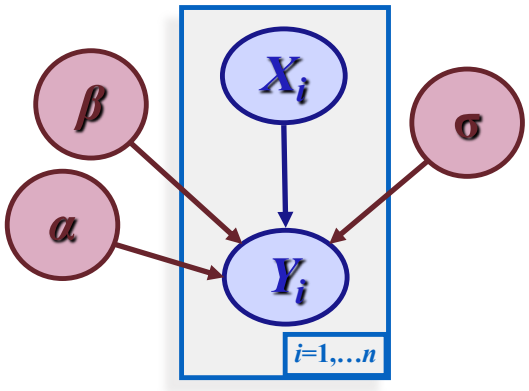


Plate representation of normal simple linear regression model: Goal is to find posterior distribution of α, β and σ

Least Squares Estimation of Regression Coefficients

- The standard estimates for α and β are:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \;\; \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- These are often called *least squares estimates* because they minimize the sum of squared deviations from the regression line (residuals):

$$(a,b) = \arg\min\{S_{ee}\}, \text{ where } S_{ee} = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

- If the error term is normally distributed with equal variances, then a and b are the maximum likelihood estimates of α and β

A Non-Informative Prior Distribution

- The model:
  - Observations Yi are independent and normally distributed
  - $E[Y \mid X] = \alpha + \beta X$
  - $\text{Var}[Y|X] = \sigma^2$ is independent of X
- Likelihood function:

$$f(\underline{y} \mid \underline{x}, \alpha, \beta, \sigma) = (2\pi)^{-n/2} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \alpha - \beta x_i)^2 \right\}$$

- To simplify the calculations we re-parameterize:
  - $\eta = \alpha + \beta \overline{X}$ where $\overline{X} = \frac{1}{n} \sum_i X_i$ is the sample mean
  - Then $E[Y \mid X] = \eta + \beta(X - \overline{X})$
  - β is called the slope of the regression line
  - η is the transformed intercept
  - $\rho = 1/\sigma^2$ is the precision

(be careful: η is easy to confuse with n)

- A commonly used non-informative prior distribution:
  - $g(\eta, \beta, \rho) \propto \rho^{-1}$ Uniform on η and β; weakly favoring smaller values of ρ
- Posterior distribution: η and β are independent and normal; precision $\rho = 1/\sigma^2$ has a gamma distribution

Finding the Posterior Distribution (1 of 2)

- The likelihood function:

$$f(\underline{y} \mid \underline{x},\eta,\beta,\rho) = \left(\frac{\rho}{2\pi}\right)^{n/2} \exp\left\{-\frac{\rho}{2}\sum_i (y_i - \eta - \beta(x_i - \bar{x}))^2\right\}$$

- Prior times likelihood:

$$f(\underline{y} \mid \underline{x},\eta,\beta,\rho)g(\rho) \propto \rho^{n/2-1} \exp\left\{-\frac{\rho}{2}\sum_i (y_i - \eta - \beta(x_i - \bar{x}))^2\right\}$$

- Algebraic manipulation of squared deviations:

$$\sum_i (y_i - \eta - \beta(x_i - \bar{x}))^2 = S_{ee} + n(\eta - \bar{y})^2 + S_{xx}(\beta - b)^2$$

- Re-express prior times likelihood:

$$f(\underline{y} \mid \underline{x},\eta,\beta,\rho)g(\rho) \propto \rho^{n/2-1} \exp\left\{-\frac{\rho}{2}\left(S_{ee} + n(\eta - \bar{y})^2 + S_{xx}(\beta - b)^2\right)\right\}$$

$$= \left(\rho^{(n-2)/2-1} \exp\left\{-\tfrac{1}{2}\rho S_{ee}\right\}\right)\left(\rho^{1/2} \exp\left\{-\tfrac{1}{2}\rho n(\eta - \bar{y})^2\right\}\right)\left(\rho^{1/2} \exp\left\{-\tfrac{1}{2}\rho S_{xx}(\beta - b)^2\right\}\right)$$

  - The first term is proportional to a gamma density function
  - Given ρ, the second and third parts are proportional to normal density functions

Notation:

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{ee} = \sum_{i=1}^{n} (y_i - \bar{y} - b(x_i - \bar{x}))^2$$

Finding the Posterior Distribution (2 of 2)

- Posterior distribution for ρ is Gamma with shape $(n-2)/2$ and scale $2/S_{ee}$
- η and β are independent and normally distributed conditional on ρ
  - η has mean $\bar{y}$ and precision $n\rho$
  - β has mean b and precision $S_{xx}\rho$
  - α has mean $\bar{y} - b\bar{x}$ and precision $\rho\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)^{-1}$
- Note: the untransformed intercept α is not independent of the slope β
- Distribution of (α, β) conditional on $\rho = \sigma^{-2}$:
  - (α, β) has a bivariate normal distribution
  - Mean is $(\bar{y} - b\bar{x}, b)$
  - Covariance matrix:
    - $Var(\alpha) = Var(\eta - \beta x) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$
    - $Var(\beta) = \sigma^2/S_{xx}$
    - $Cov(\alpha, \beta) = E[(\alpha - \bar{y} + b\bar{x})(\beta - b)] = E[\bar{x}(\beta - b)^2] = \bar{x}\sigma^2/S_{xx}$

Marginal Posterior Distribution for Parameters

Transformed parameters:

- η has nonstandard t distribution with:
  - Center $\bar{y}$
  - Spread $(n(n-2)/S_{ee})^{-1/2}$
  - Degrees of freedom $n-2$
- β has nonstandard t distribution with:
  - Center b
  - Spread $(S_{xx})(n-2)/S_{ee})^{-1/2}$
  - Degrees of freedom $n-2$
- η and β are:
  - Independent conditional on ρ
  - Uncorrelated but not independent when we marginalize out ρ

- Original parameters:
- (α, β) has a bivariate t distribution with:
  - Center $\begin{pmatrix} \bar{y} - b\bar{x} \\ b \end{pmatrix}$
  - Spread matrix $\frac{S_{ee}}{(n-2)} \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} & \frac{\bar{x}}{S_{xx}} \\ \frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix}$
  - Degrees of freedom $n-2$

Cigarette Example: Posterior Distribution

CO (mg)	Tar (mg)
13.60	14.10
16.60	16.00
23.50	29.80
10.20	8.00
5.40	4.10
15.00	15.00
9.00	8.80
12.30	12.40
16.30	16.60
15.40	14.90
13.00	13.70
14.40	15.10
10.00	7.80
10.20	11.40
9.50	9.00
1.50	1.00
18.50	17.00
12.60	12.80
17.50	15.80
4.90	4.50
15.90	14.50
8.50	7.30
10.60	8.60
13.90	15.20
14.90	12.00

- ρ has a Gamma distribution with
  - Shape $(n-2)/2 = 11.5$
  - Scale = $2/S_{ee} = 0.0446$
- Conditional on ρ
  - The transformed intercept η has a normal distribution with mean 12.53 and precision 25ρ
  - The slope β has a normal distribution with mean 0.80 and precision $\rho S_{xx} = 770.43\rho$
- Unconditionally
  - The transformed intercept η has a nonstandard t distribution with
    - Center $\bar{y}$ = 12.53
    - Spread $(n(n-2)/S_{ee})^{-1/2}$ = 0.28
    - Degrees of freedom $n-2$ = 23
  - The slope β has a a nonstandard t distribution with
    - Center b = 0.80
    - Spread $(S_{xx}(n-2)/S_{ee})^{-1/2}$= 0.05
    - Degrees of freedom $n-2 = 2$

$\bar{x}$	12.22
$\bar{y}$	12.53
S_{xy}	617.10
S_{xx}	770.43
S_{ee}	44.87
n	25
$\eta = \bar{y}$	12.53
$b = S_{xy}/S_{xx}$	0.80
$s^2 = S_{ee}/(n-2)$	1.95
s	1.40
a	2.74

Bayesian Interpretation of Standard Regression

- Most statistical software packages provide regression analysis
- We can use these software packages and give the result a Bayesian interpretation
- Results from cigarette regression in R:

```
Call:
lm(formula = mgCarbonMonoxide ~ mgTar, data = cigdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1124 -0.7167 -0.3754  1.0091  2.5450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74328    0.67521   4.063 0.000481 ***
mgTar        0.80098    0.05032  15.918 6.55e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.397 on 23 degrees of freedom
Multiple R-squared: 0.9168,  Adjusted R-squared: 0.9132
F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

Center of intercept and slope

Spread of intercept and slope

Approximate posterior expectation of observation std dev

Degrees of freedom

Bayesian Regression and Least Squares Regression

- We can give results from standard regression packages a Bayesian interpretation
- Assumptions:
  - Linear regression with normal error
  - Noninformative prior distribution on parameters
- Under these assumptions, confidence intervals for regression coefficients can be interpreted as Bayesian credible intervals
- If a t-test rejects the hypothesis that coefficient is equal to zero, then 0 does not belong to the corresponding posterior credible interval for the coefficient

Predictive Distribution for New Observation (Exact)

- We would like to predict the CO content of a cigarette with given tar content
- Given tar content x_{new}, slope β, transformed intercept η and precision ρ, the CO content is normally distributed with:
  - Mean $(x_{new} - \bar{x})\beta + \eta = x_{new}\beta + \alpha$
  - Variance $\sigma^2 = 1/\rho$
- Given the precision ρ, predictive distribution for x_{new} is normal with
  - Mean $(x_{new} - \bar{x})b + \bar{y} = x_{new}b + a$ *Fitted value*
  - Variance $\left(\frac{(x_{new}-\bar{x})^2}{S_{xx}} + \frac{1}{n} + 1\right)\sigma^2$
- Integrating out the precision, the predictive distribution has a non-standard t distribution with
  - Center $(x_{new} - \bar{x})b + \bar{y} = x_{new}b + a$ *Fitted value*
  - Spread $\sqrt{\left(\frac{(x_{new}-\bar{x})^2}{S_{xx}} + \frac{1}{n} + 1\right)\left(\frac{S_{ee}}{n-2}\right)}$
  - Degrees of freedom $(n-2)$

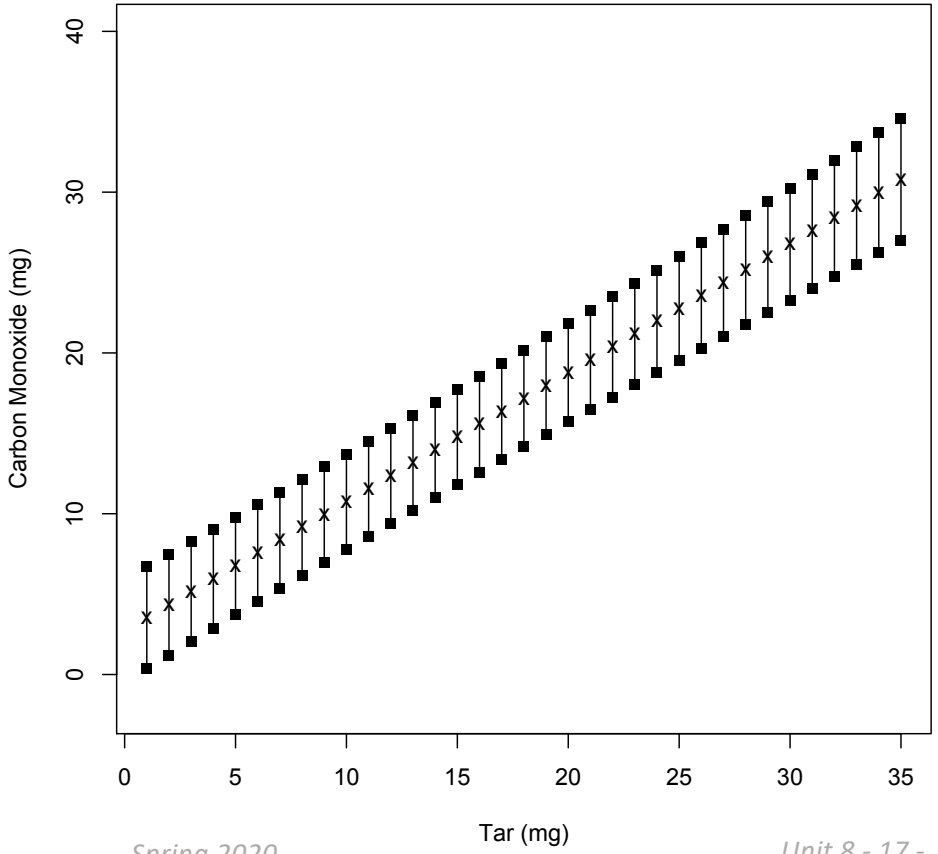
Predictive Distribution: Examples

- For tar content 14 mg the CO content $Y_{new}|(X_{new} = 14)$ has a nonstandard t distribution with
  - Center 13.96
  - Spread 1.43
  - Degrees of freedom 23
- For tar content 30 mg the CO content $Y_{new}|(X_{new} = 30)$ has a nonstandard t distribution with
  - Center 26.77
  - Spread 1.68
  - Degrees of freedom 23

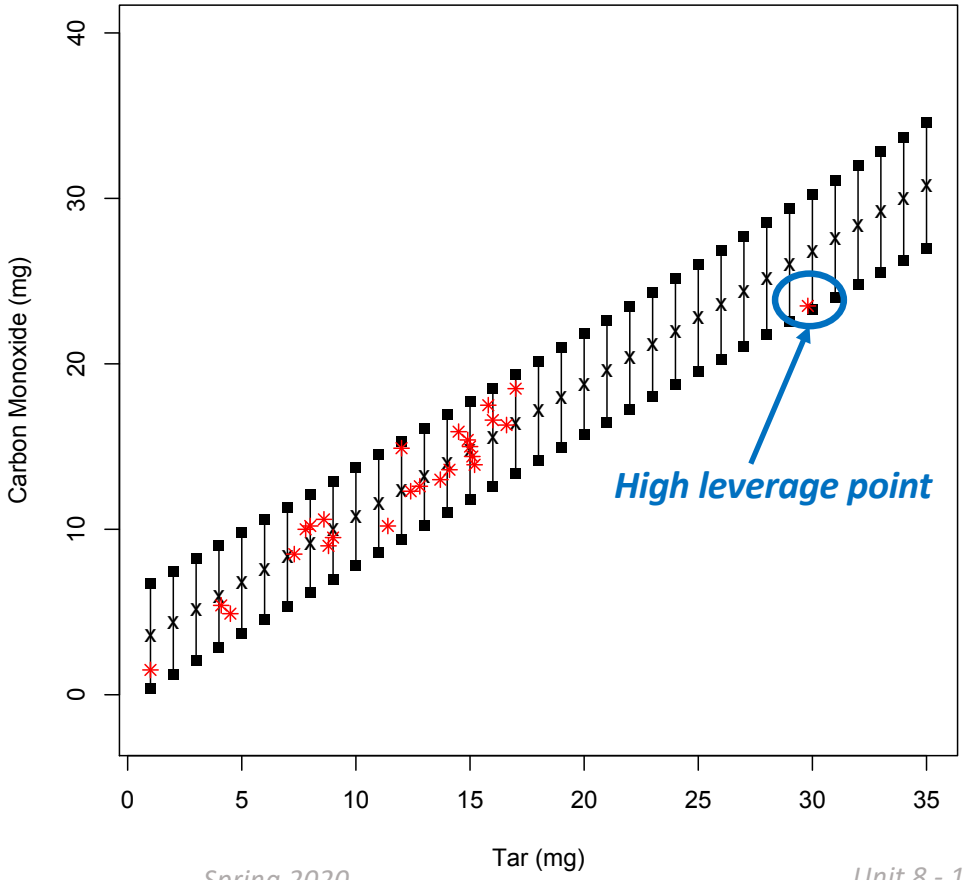
Predictive Distribution for New Observation (Direct MC)

- We can also estimate the predictive distribution by direct Monte Carlo
- The algorithm: For $k = 1, \ldots, K$ (K = desired sample size)
  - Simulate $\rho^{(k)}$ from a Gamma distribution with shape $(n-2)/2$ and scale $2/S_{ee}$
  - Simulate $\eta^{(k)}$ from a Normal distribution with mean $\bar{y}$ and precision $n\rho^{(l)}$
  - Simulate $\beta^{(k)}$ from a Normal distribution with mean b and precision $S_{xx}\ \rho^{(k)}$
  - Simulate $y_{new}^{(k)}$ from a Normal distribution with mean $\eta^{(k)} + \beta^{(k)}(x_{new} - \bar{x})$ and precision $\rho^{(k)}$
- The sample $y_{new}^{(1)}, \ldots, y_{new}^{(K)}$ approximates the posterior predictive distribution of Y given $X = x_{new}$

95% Intervals for Predictive Distribution of New Observation



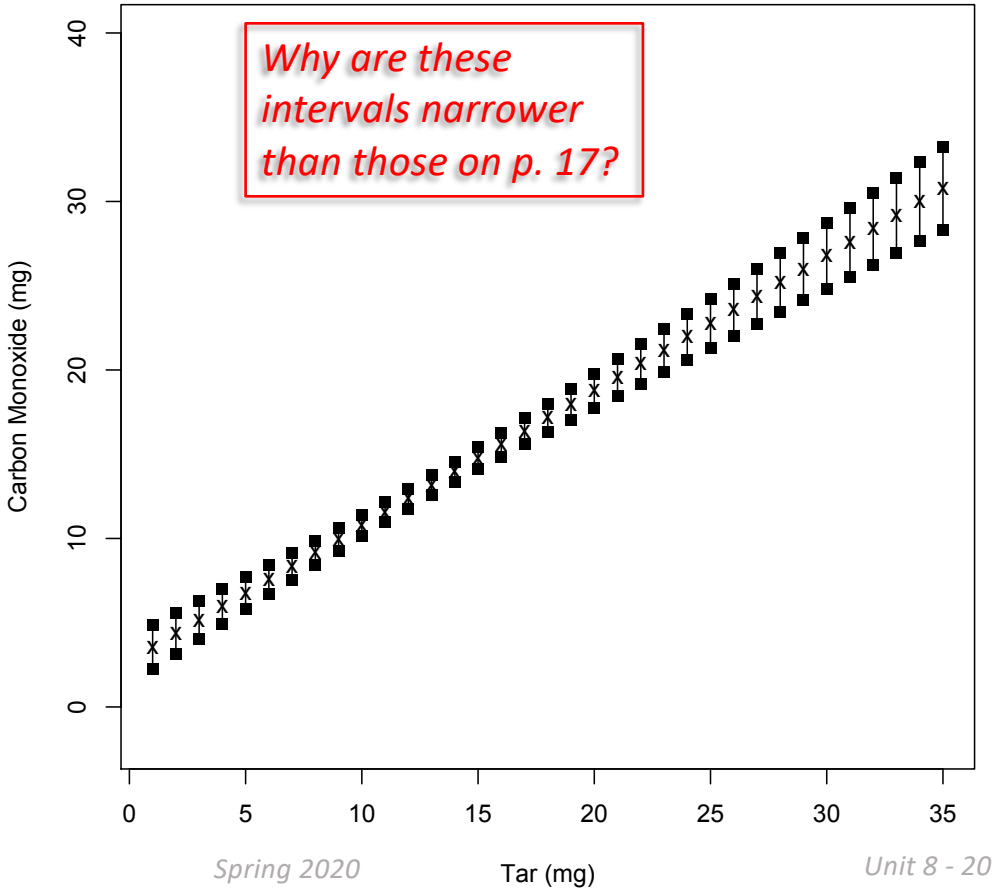
Posterior Predictive Model Check



Mean of Predictive Distribution for New Observation

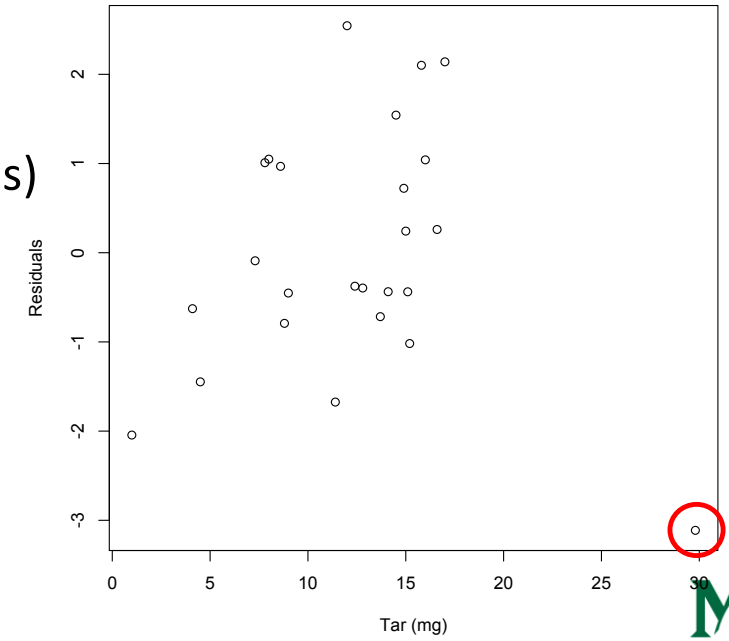
- Given tar content x_{new}, slope β, transformed intercept η and precision ρ, the tar content is normally distributed with:
  - Mean $x_{new}\beta + \alpha$
  - Variance $\sigma^2 = 1/\rho$
- Given the precision ρ, ***mean*** $x_{new}\beta + \alpha$ of the predictive distribution is normally distributed with
  - Mean $x_{new}b + \alpha$
  - Variance $\left(\frac{(x_{new}-\bar{x})^2}{S_{xx}} + \frac{1}{n}\right)\sigma^2$
- Integrating out the precision, ***mean*** $x_{new}\beta + \alpha$ of the predictive distribution has a non-standard t distribution with
  - Center $x_{new}b + \alpha$
  - Spread $\sqrt{\left(\frac{(x_{new}-\bar{x})^2}{S_{xx}} + \frac{1}{n}\right)\left(\frac{S_{ee}}{n-2}\right)}$
  - Degrees of freedom $(n-2)$

95% Credible Interval for Mean of Predictive Distribution for New Observation

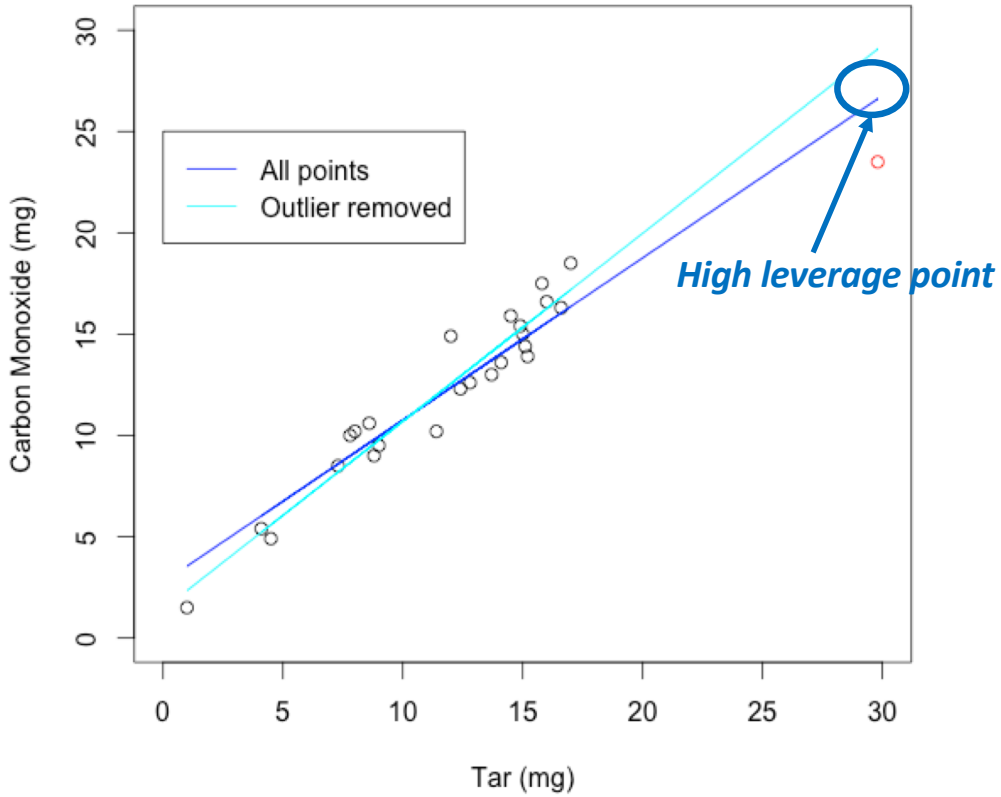


Diagnostics

- Residual plots are a good way to diagnose model inadequacy
- Residuals should show no systematic pattern ("white noise")
- Things to look for in residuals
  - U or bowl shape (indicates nonlinearity)
  - Outliers (very small or very large residuals)
  - Fan shape (indicates variance is not constant in X)
  - Unusual values

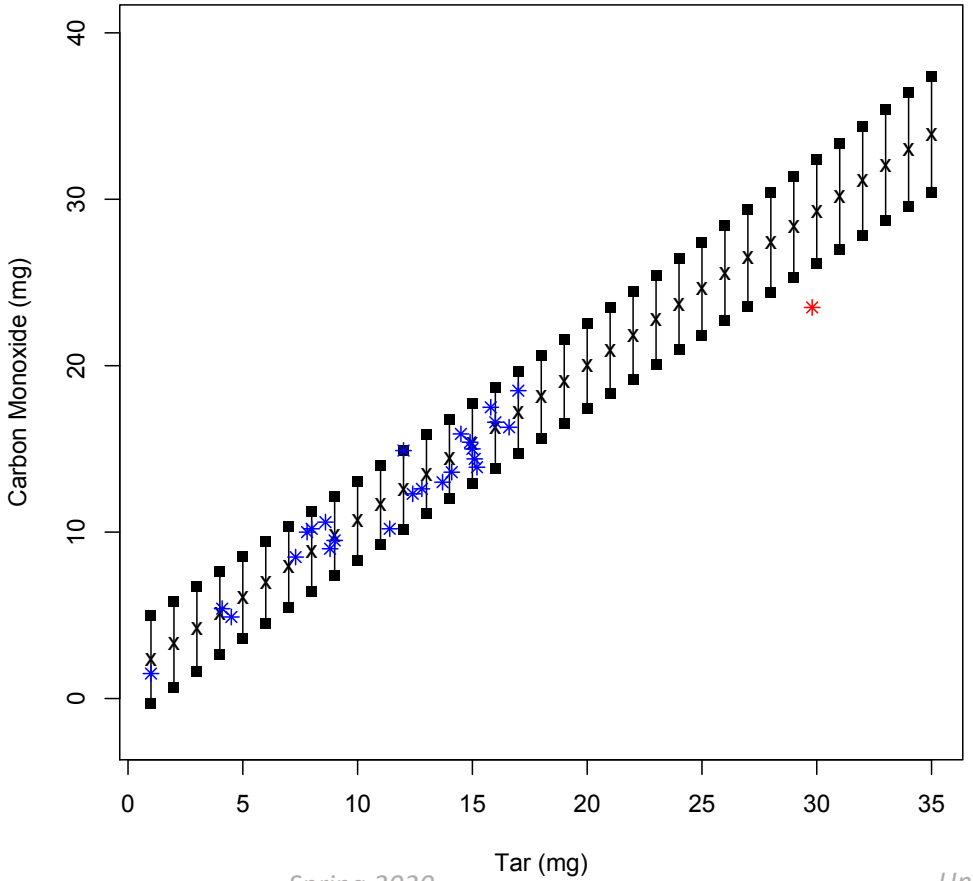


Cigarette Example with Outlier Removed



	All Data	Outlier removed
Slope est	0.80	0.93
Slope SE	0.05	0.05
Intrcp est	2.74	1.41
Intrcp SE	0.68	0.65
Regr SE	1.40	1.12

Predicting Outlier from Model without Outliers



An Informative Normal-Gamma Prior Distribution

- An informative conjugate prior distribution has the same form as the posterior distribution from a regression with a noninformative prior distribution:
  - Given the precision $\rho = 1/\sigma^2$, η and β follow a bivariate normal distribution
    - Mean of η is μ_η and precision of η is $k_\eta \rho$
    - Mean of β is μ_β and precision of β is $k_\beta \rho$
    - η and β are independent (a more general conjugate family allows dependence)
  - Precision $\rho = 1/\sigma^2$ has a Gamma(c, d) distribution
- The posterior distribution given the data $(y_{1:n}, x_{1:n})$ has the same form as the prior distribution:
  - Given the precision ρ, the coefficients η and β are independent and normal
    - Mean of η is $\mu_\eta^* = (k_\eta \mu_\eta + n\bar{y})/(k_\eta + n)$
    - Precision of η is $\rho(k_\eta + n)$
    - Mean of β is $\mu_\beta^* = \frac{k_\beta \mu_\beta + S_{xy}}{k_\beta + S_{xx}}$
    - Precision of β is $\rho(k_\beta + S_{xx})$
  - Precision ρ has a gamma distribution, with shape $c^* = c + \frac{n}{2}$ and scale

$$d^* = \left(d^{-1} + \frac{1}{2} S_{ee} + \frac{k_\eta n}{2(k_\eta + n)} (\bar{y} - \mu_\eta)^2 + \frac{k_\beta S_{xx}}{2(k_\beta + S_{xx})} (b - \mu_\beta)^2 \right)^{-1}$$

Interpreting the Parameters of the Conjugate Distribution

- μ_η : the mean of the transformed intercept term η
- k_η : the ratio of the precision ρk_η of the intercept η to the precision ρ of the deviations from the regression line
- μ_β : the mean of the slope term β
- k_β: the ratio of the precision ρk_β of the slope β to the precision ρ of the deviations from the regression line
- c : the shape parameter for the Gamma prior for the precision ρ of the deviations from the regression line
- d : the scale parameter for the Gamma prior for the precision ρ of the deviations from the regression line

Generalization: Multiple Regression

- In the multiple regression model, a dependent variable Y depends on several independent variables $X_1, \ldots, X_p$
- Multiple linear regression with normal errors:
  - Y given $X_1, \ldots, X_p$ is normally distributed
  - $E[Y|X_1, \ldots, X_p] = X_1\beta_1 + \cdots + X_p\beta_p$
  - Observations Y_i are independent
  - $Var[Y|X_1, \ldots, X_p] = \sigma^2$, and σ^2 is independent of $X_1, \ldots, X_p$
- If we have n observations, we write $y = X\beta + \varepsilon$ where:
  - $\underline{y}$ is a $n \times 1$ vector of observations
  - $\boldsymbol{X}$ is a $n \times p$ matrix of independent variable values
  - $\underline{\beta}$ is a $p \times 1$ vector of regression coefficients
    - Usually the first column consists of 1's, corresponding to a constant term in the regression equation
    - We can represent nonlinear equations by including polynomial terms like X_1^2 , interaction terms like X_1X_2, or other nonlinear terms like $\log(X_i)$ as predictors
  - $\underline{\varepsilon}$ is a $n \times 1$ vector of deviations, assumed normal with mean 0 and variance σ^2

Multivariate Normal Distribution

- The multivariate normal distribution $N(\underline{\theta}, \Sigma)$ is usually parameterized by a mean vector $\underline{\theta} = (\theta_1, \ldots, \theta_p)^T$ and covariance matrix $\Sigma = [\sigma_{ij}]$
- The density function for a random vector $X = (X_1, \ldots, X_p)^T$ with the $N(\underline{\theta}, \Sigma)$ distribution is:

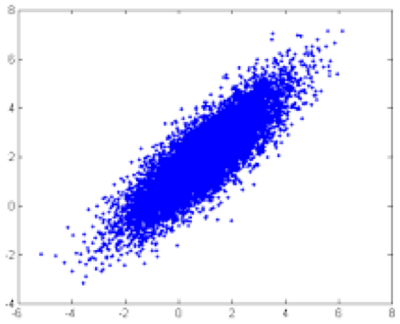
$$f(\underline{x} \mid \underline{\theta}, \Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}} \left(\sqrt{2\pi}\right)^k} \exp\left\{-\tfrac{1}{2}(\underline{x} - \underline{\theta})^{\mathrm{T}} \Sigma^{-1} (\underline{x} - \underline{\theta})\right\}$$

- Bayesian statisticians often use an alternate parameterization that is more convenient for Bayesian updating
- Bayesians often parameterize the multivariate normal distribution with mean vector $\underline{\theta} = (\theta_1, \ldots, \theta_p)^T$ and precision matrix $\mathrm{P} = [\rho_{ij}]$, where $\mathrm{P} = \Sigma^{-1}$
- The multivariate normal density function can be written in terms of $\underline{\theta}$ and P :

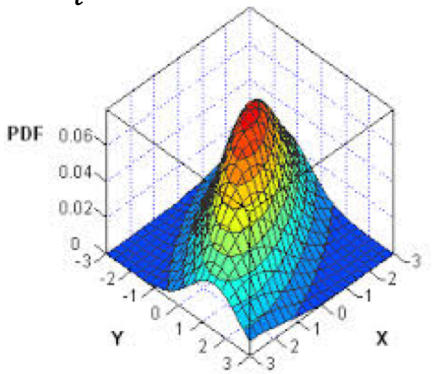
$$f(\underline{x} \mid \underline{\theta}, \mathrm{P}) = \frac{|\mathrm{P}|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^k} \exp\left\{-\tfrac{1}{2}(\underline{x} - \underline{\theta})^{\mathrm{T}} \mathrm{P} (\underline{x} - \underline{\theta})\right\}$$

Multivariate Normal Distribution

- If a random vector $X = (X_1, \dots, X_p)^T$ has a multivariate normal distribution with mean $\underline{\theta} = (\theta_1, \dots, \theta_p)^T$ and covariance matrix Σ, then
  - Each component X_i has a univariate normal distribution with mean θ_i and variance $\sigma_{ii} = \sigma_i^2$
  - The covariance $\sigma_{ij} = Cov(X_i, X_j) = \mathrm{E}[(X_i - \theta_i)(X_j - \theta_j)]$) measures how X_i and X_j vary together
  - X_i and X_j are independent if σ_{ij} = 0
  - The correlation $Cor(X_i, X_j)$ = $Cov(X_i, X_j)/\sigma_{ij}$, lies between -1 and 1.
  - A correlation of 1 or -1 (perfect correlation) occurs when X_i is a linear function of X_j



Scatterplot of Bivariate Normal Random Variables



Joint Density Function of Bivariate Normal Random Variables

Least Squares Estimation of Multiple Regression Coefficients

- The standard estimate for $\underline{\beta}$ is $\underline{b} = (X^T X)^{-1} X^T \underline{y}$
- This is often called the *least squares* estimate because it minimizes the sum of squared deviations from the regression line:

  $\underline{b} = \arg\min_{b'} \{\sum_{i=1}^{n} (y_i - X_i b')^2\}$
- If the error term is normally distributed with equal variances, then $\underline{b}$ is the maximum likelihood estimate of β
- Simple regression is a special case:
  - X is a $nx2$ matrix with first column a vector of 1's and second column the independent variable
  - It is straightforward (although tedious) to verify that the least squares estimates for simple regression are

    $\begin{pmatrix} a \\ b \end{pmatrix} = (X^T X)^{-1} X^T \underline{y}$

Conjugate Prior Distribution for Multiple Linear Regression with Normal Errors

- Observations $\underline{y} \sim_{indep} Normal\left(X\underline{\beta}, \sigma^2 I\right)$, where I is an $n \times n$ identity matrix
- Normal-Gamma conjugate prior for β and $\rho = 1/\sigma^2$
  - Given the precision ρ, the regression coefficients β are multivariate normal random variables with mean $\underline{\mu}$ and nonnegative definite precision matrix ρK
  - The precision ρ has a gamma distribution with shape c and scale d
- Posterior distribution for β and ρ
  - Given the precision ρ, the regression coefficients β are normally distributed with mean

    $\underline{\mu}^* = (K + X^T X)^{-1}(K\beta + X^T y)$
  - The precision ρ has a gamma distribution with shape $c^* = c + n/2$ and scale

$$d^* = \left(d^{-1} + \frac{1}{2}\left(\left(y - X\underline{b}\right)^T\left(y - X\underline{b}\right) + \frac{1}{2}\left(\underline{b} - \underline{\mu}\right)^T\left(K^{-1} + (X^T X)^{-1}\right)\left(\underline{b} - \underline{\mu}\right)\right)\right)^{-1}$$

- Marginal distribution for β is multivariate t
- Predictive distribution / marginal likelihood for observations is multivariate t

Marginal Likelihood for Normal-Gamma Conjugate Prior Distribution

- We are given values $\underline{x}$ for the independent variables and want to predict the dependent variable y
  - y given $\underline{x} = (x_1, \ldots, x_p)$ is normal with mean $\underline{x}^T\underline{\beta}$ and precision ρ
  - β is normal with mean μ and precision matrix ρK (a p×p matrix)
  - ρ is gamma with shape c and scale d
- Given ρ, K and μ
  - y is normally distributed
  - Mean is $\underline{x}^T\underline{\mu}$
  - Precision is $\rho\ (\underline{x}^T K^{-1}\underline{x} + 1)^{-1}$
- Integrate out ρ to obtain the predictive distribution:
  - y is nonstandard t
  - Center is $\underline{x}^T\underline{\mu}$
  - Spread is $\sqrt{\frac{1}{cd}\ (\underline{x}^T K^{-1}\underline{x} + 1)^{-1}}$
- Typically we use simulation to estimate marginal likelihoods
  - Simulate ρ and $\underline{\beta}$
  - Simulate y given $\underline{x}$, ρ and $\underline{\beta}$

Zellner's g-Prior Distribution

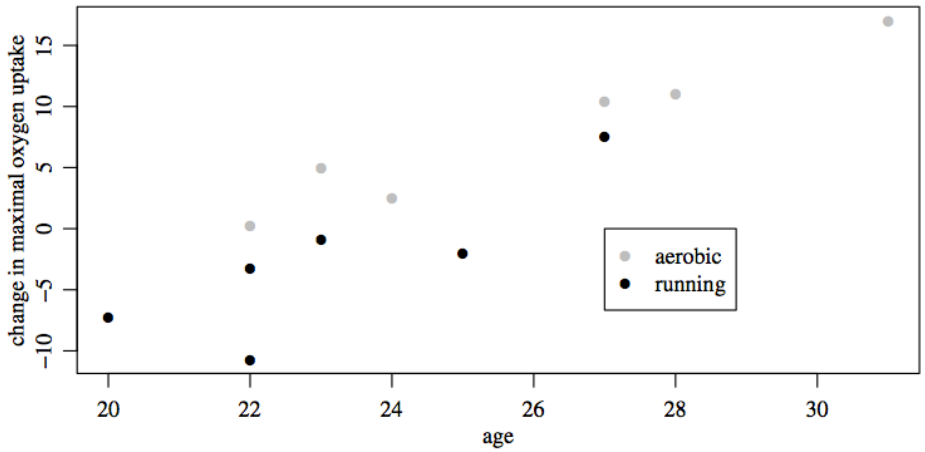
- It is often difficult to specify prior hyperparameters
- Zellner's g-prior is a commonly used "weakly informative" conjugate prior distribution
  - y given $\underline{x} = (x_1, \ldots, x_p)$ is normal with mean $\underline{x}^T\underline{\beta}$ and precision ρ
  - β is normal with mean $\underline{0}$ and precision matrix $\rho g^{-1}(X^TX)^{-1}$ (a $p \times p$ matrix)
  - ρ is gamma with shape c and scale d
- Posterior distribution for β and ρ
  - Given the precision ρ, the regression coefficients β are normally distributed with mean $\underline{\mu}^* = \frac{g}{g+1}(X^TX)^{-1}X^T\underline{y}$ and variance $v = \frac{g}{g+1}\sigma^2(X^TX)^{-1}$
  - The precision ρ has a gamma distribution with shape $c^* = c + n/2$ and scale $d^* = \left(d^{-1} + \frac{1}{2}\left(y - X\underline{\mu}^*\right)^T\left(y - X\underline{\mu}^*\right)\right)^{-1^{-1}}$
- Predictive distribution / marginal likelihood for observations is multivariate t

Remarks on the g-Prior

- Weakly informative
- Shrinks toward prior mean $\underline{\mu} = \underline{0}$
  - Precision matrix $\rho g^{-1}(X^T X)^{-1}$ for prior distribution of regression coefficient β depends on observed values of dependent variable X
    - Using this distribution implies that our prior knowledge about β before seeing the response variable y depends (weakly) on the values of the independent variables X
  - Parameter g controls amount of shrinkage
    - Larger g corresponds to less information and less shrinkage
  - Shrinkage toward $\underline{\mu} = \underline{0}$ helps prevent overfitting

Example: Oxygen Uptake (from Hoff Chapter 9)

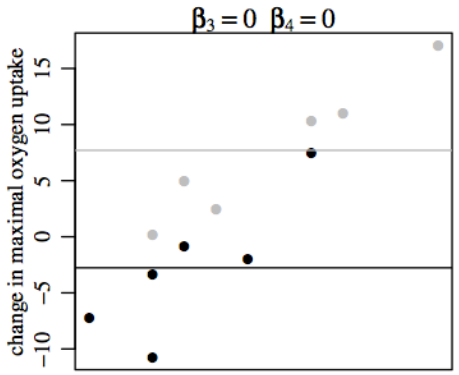
- Plot shows data from study on oxygen uptake during exercise
  - There are 12 subjects in the study
- Regression equation:
  $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon$
  - Y_i = change in maximal oxygen uptake after 12-week exercise program
  - x_{i1} = 1 for each subject i
  - x_{i2} = 0 if subject i is on running program; 1 if on aerobic exercise program
  - x_{i3} = age of subject i
  - $x_{i4} = x_{i2}\, x_{i3}$ (interaction between age and program)

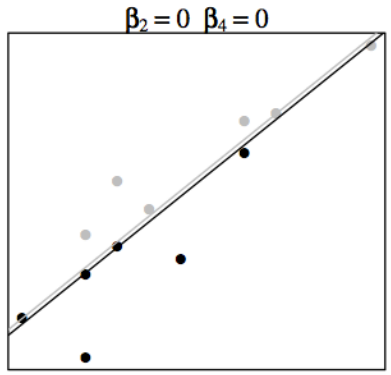


$$E[Y \mid \underline{x}] = \begin{cases} \beta_1 + \beta_3 \times \text{age} & \text{if running} \\ (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age} & \text{if aerobic} \end{cases}$$

Least Squares Estimates for Alternative Models

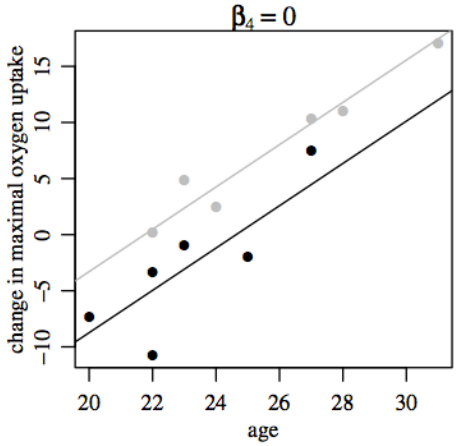
No dependence on age

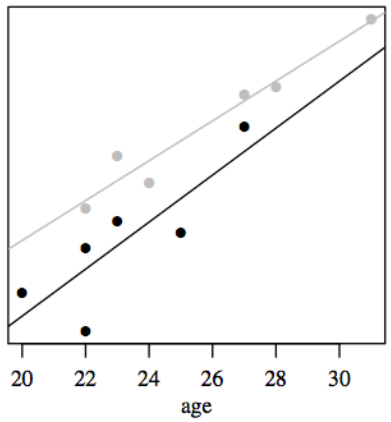




Running same as aerobic

Running slope same as aerobic slope but intercepts are different





Slopes and intercepts are different

Oxygen Uptake: Analysis

- Hoff text uses a weakly informative conjugate Zellner g-prior distribution with g=12 (about one observation of prior information)
  - Conditional on the precision ρ, the coefficients $\underline{\beta}$ are normally distributed with mean $\underline{\mu} = \underline{0}$ and precision ρK, where $K = \frac{1}{12} X^T X$
  - Precision ρ has gamma distribution with shape $c = 1/2$ and scale $d = 2/s^2$, where $s = S_{ee}/(n-p)$ is an unbiased estimate of the variance computed from the least squares sum of squared residuals
- Posterior distribution
  - Conditional on precision ρ, the coefficients β are normally distributed with mean $\underline{\mu}^*$ and precision matrix ρK^*

$$\underline{\mu}^* = \frac{12}{13}\left(X^\top X\right)^{-1} X^\top \underline{y} \qquad K^* = \frac{13}{12}\left(X^\top X\right)^{-1}$$

  - Precision ρ has gamma distribution with shape $c^* = (n+1)/2$ and scale d^*

$$d^* = \left(\frac{1}{d} + \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\mu^*)^2\right)^{-1} = \left(\frac{s^2 + SSR}{2}\right)^{-1}$$

- In this analysis, we use the observed values of the predictors X to specify the prior distribution on the response $\underline{y}$
- The relatively small virtual sample size limits the impact of the prior distribution

Posterior Distribution for β2 and β4

- Recall that β_2 and β_4 are the difference in intercept and slope for the two exercise groups
- These plots suggest lack of evidence for a difference between the groups

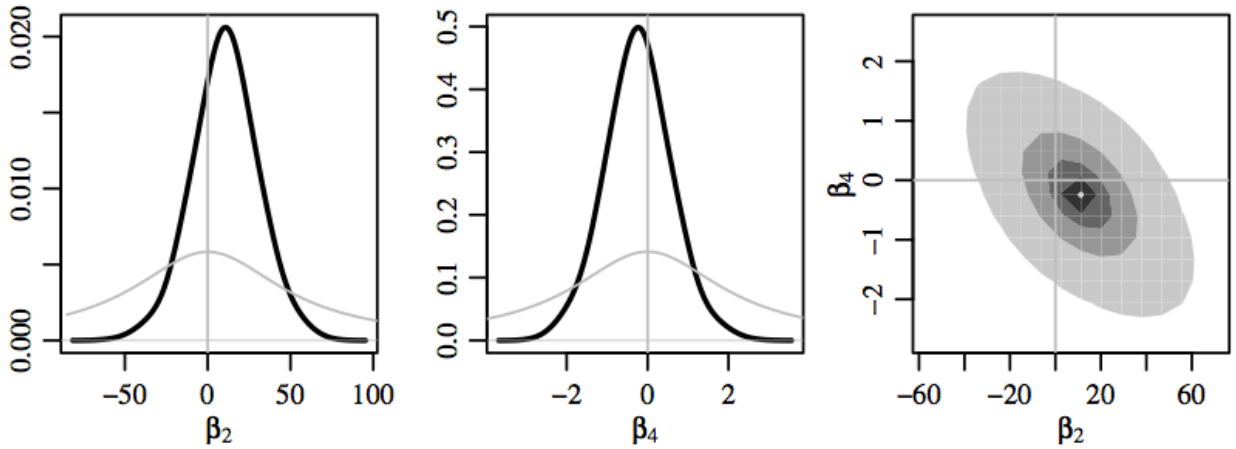


Fig. 9.3. Posterior distributions for β_2 and β_4, with marginal prior distributions in the first two plots for comparison.

Posterior Distribution for Aerobics Effect by Age

- The difference in expected oxygen uptake for an individual of age x in aerobics versus running program is is $\beta_2 + \beta_4\, x$
- The plot shows 95% intervals of of β2 + β4 x at each age x
- There is reasonably strong evidence of a difference between programs for younger ages

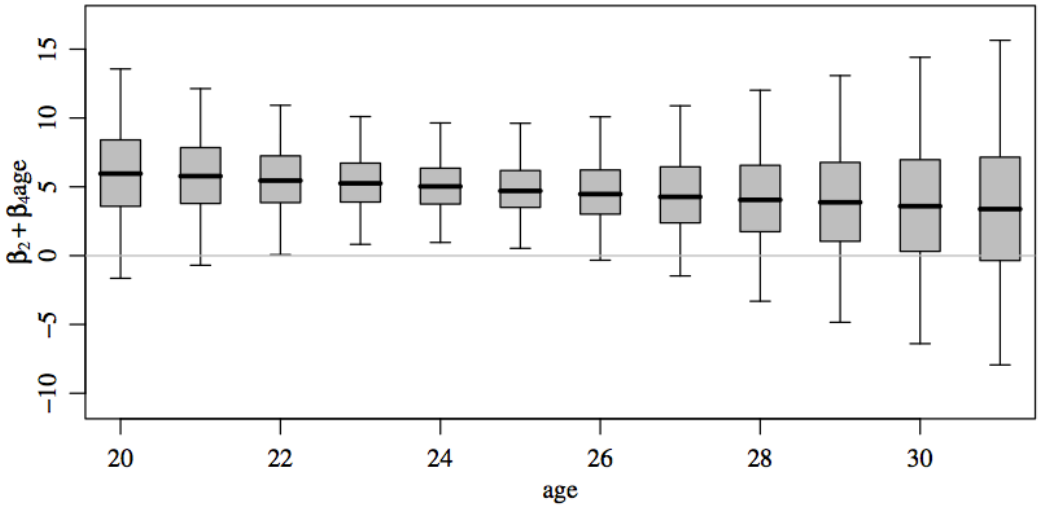
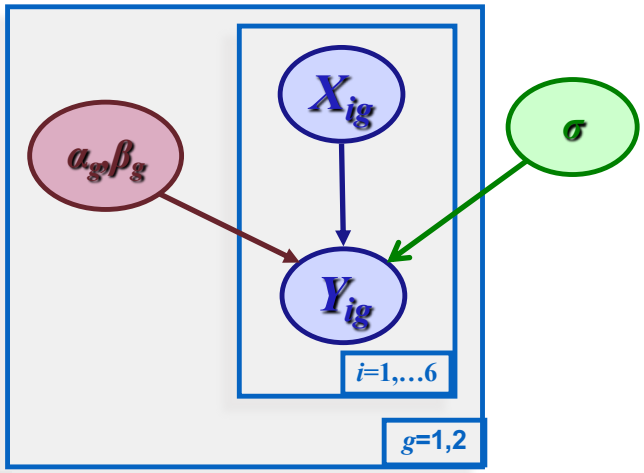


Fig. 9.4. Ninety-five percent confidence intervals for the difference in expected change scores between aerobics subjects and running subjects.

Reframing Oxygen Model as Hierarchical Model

- We could reframe the oxygen model as a hierarchical model with vague prior information on top-level hyperparameters:
  - $Y_{gi} \sim Normal(\alpha_g + \beta_g x_{ig}, \sigma^2)$
  - $\alpha_g \sim iid\ Normal(0, 100)$
  - $\beta_g \sim iid\ Normal(0, 100)$
  - $\rho = \sigma^{-2} \sim Gamma(0.5, 1000)$
- We can use JAGS to estimate the parameters of this model



- The model in JAGS:

```
model{
    for (i in 1:n) {
        y[i] ~ dnorm((a[2]+b[2]*age[i])*aerobic[i]+(a[1]+b[1]*age[i])*(1-aerobic[i]),rho)
            }
    for (i in 1:2) {
        a[i] ~ dnorm(0,0.0001)
        b[i] ~ dnorm(0,0.0001)
    }
    rho ~ dgamma(0.5,1/1000)
    }
```

Note that JAGS parameterizes normal distribution with precision (not standard deviation) and Gamma distribution with rate (not scale)

Comparison: Least Squares Regression with Hierarchical Model in JAGS

```
Call:
lm(formula = y ~ age * aerobic)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5295 -0.9610  0.3945  2.1717  2.2883

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -51.2939    12.2522  -4.187  0.00305 **
age           2.0947     0.5264   3.980  0.00406 **
aerobic      13.1071    15.7620   0.832  0.42978
age:aerobic  -0.3182     0.6498  -0.490  0.63746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.923 on 8 degrees of freedom
Multiple R-squared: 0.9049, Adjusted R-squared: 0.8692
F-statistic: 25.36 on 3 and 8 DF,  p-value: 0.0001938
```

Summary Comparison:

Effect	OLS	JAGS
a_{run}	-51.29	-50.51
a_{aer}	-38.19	-37.57
b_{run}	2.09	2.06
b_{aer}	1.78	1.75
s	2.92	2.80

```
Inference for Bugs model at "Oxygen.model.2.jags", fit using jags,
 3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 9
 n.sims = 3000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a[1]      -50.514  12.802 -75.516 -58.575 -50.715 -42.629 -24.519 1.001  2600
a[2]      -37.566  10.913 -59.713 -44.066 -37.473 -30.650 -16.305 1.001  3000
b[1]        2.061   0.550   0.921   1.720   2.065   2.412   3.145 1.002  1900
b[2]        1.753   0.420   0.926   1.500   1.749   2.003   2.608 1.001  3000
rho         0.132   0.061   0.039   0.086   0.124   0.167   0.265 1.004   690
deviance   60.746   3.870  55.888  57.952  59.926  62.629  70.124 1.001  3000
```


MASON

Example: Educational Testing (Gelman, et al. Section 5.5)

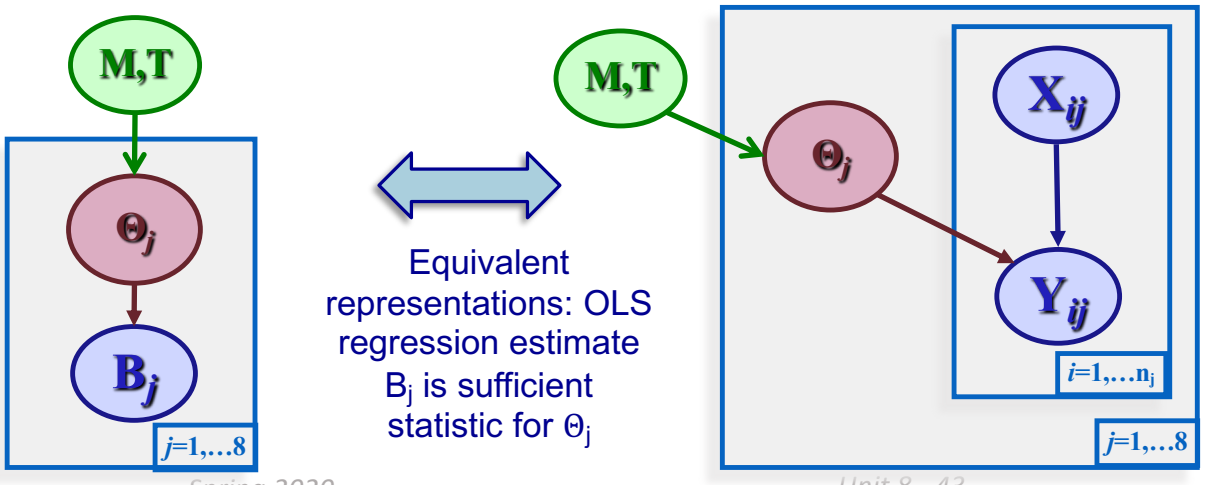
- Educational Testing Service conducted a study of SAT coaching
  - Eight schools with different coaching programs
  - All students had already taken PSAT
  - Some students were coached; some were not
  - Regression estimates were calculated for each school
    - Estimated coaching effects assumed independent for the eight schools
    - Sampling distributions are approximately normal
    - Sampling variances assumed known (over 30 students in each school provides a large sample for estimating variance if equal variances are assumed)
- Why not just compute 8 separate coaching effects estimates?
  - Confidence intervals for the coaching effects have a great deal of overlap
  - It is difficult to distinguish statistically between the experiments
- Why not just assume all coaching effects are equal and compute a pooled estimate?
  - Do we really believe the schools are all the same?
  - The largest school has estimated effect 28 under the separate model and 7 under the pooled model

Bayesian Hierarchical Model

- Estimated coaching effect B_j for school j was estimated by linear regression
  - B_j is the least squares regression coefficient
  - Hierarchical model treats B_j as drawn from a common distribution
- Compromises between the extremes of separate estimates and pooling
  - When schools are similar it behaves more like pooling
  - When schools are very different it behaves more like separate estimates
  - When schools are different but not too different it gives results between pooling and separate estimates

The Coaching Effects Model

- B_j has normal distribution with unknown mean Θ_j and known standard deviation σ_j
  - Θ_j is a school-specific mean difference in scores between coached and uncoached students in school j
  - $\sigma_j^2 = \sigma^2/n_j$ is the variance (assumed known) of B_j, equal to the common variance divided by the sample size for the jth school
- Prior distributions for expected coaching effect Θ_j are iid normal with mean M and standard deviation T
- Prior distribution for M and T are uniform

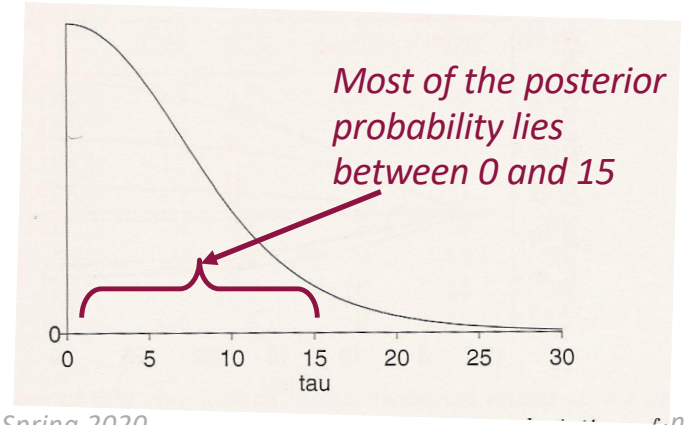


Comments on the Model

- This is the same kind of model as the one we used for the rat experiment
  - Data are from an exponential family
  - Parameter for each experiment has a conjugate prior distribution with experiment-specific parameter
  - Experiment-specific parameters are treated as drawn from a population
  - Distribution for experiment-specific parameter has no conjugate hyperprior
  - We have vague prior information about the parameters of the experiment-specific distribution
  - We can use numerical integration and simulation to estimate the posterior distribution for the hyperparameter
  - We have an analytic solution to the conditional posterior distribution for experiment-specific parameters given hyperparameter
  - We combine numerical integration and simulation to obtain an estimate of the joint posterior distribution
- Comments
  - This kind of model combines tractability and simplicity of conjugate updating with flexibility and power of hierarchical models

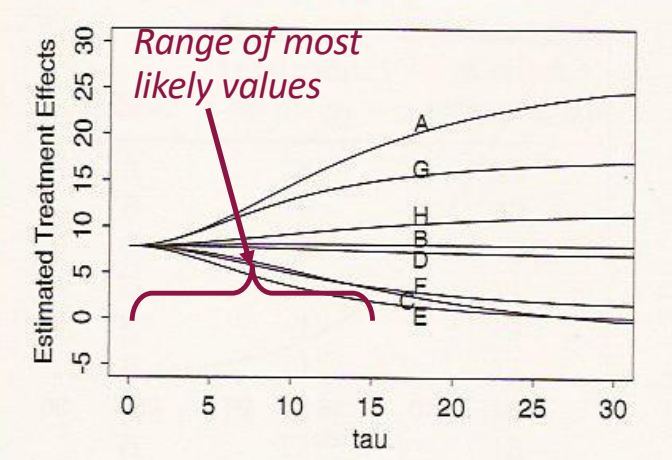
Inferences about T

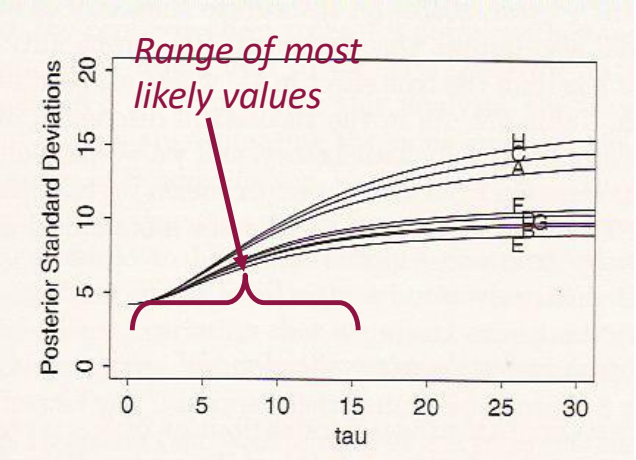
- The parameter T measures amount of school-to-school variability in coaching effect
- The mode of the posterior distribution is zero (no variation)
- There is considerable uncertainty in the posterior distribution
- It would be a mistake to use maximum *a posteriori* point estimate $T = 0$
- It would also be a mistake to analyze each school completely independently of the others



Posterior Distribution for School Coaching Effect Means

- Distribution of school-level coaching effects depend on coaching effect standard deviation T
  - Difference in school-level mean coaching effects increases with T
  - Standard deviations of school-level coaching effects increase and become more different from each other as T increases
- Conclusions
  - Coaching effects vary from almost none to as high as 15 points and probably differ among schools
  - Raw estimate of 28 points for School A is probably too high
    - Posterior median for School A is 10; 97.5% point is 23





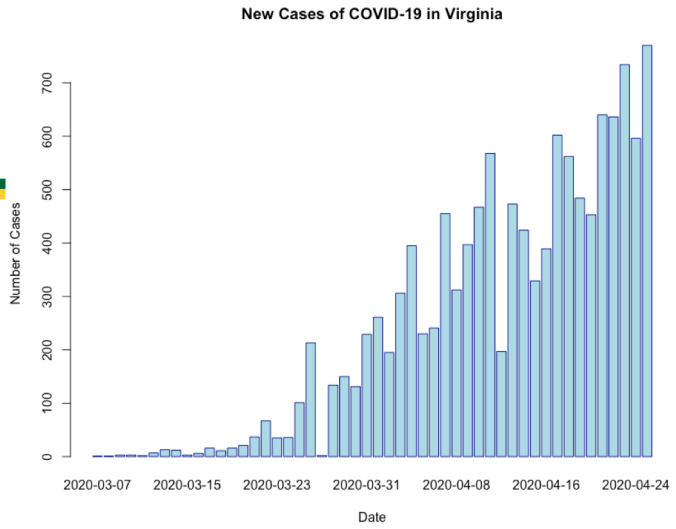
Generalizing Linear Regression with Normal Errors

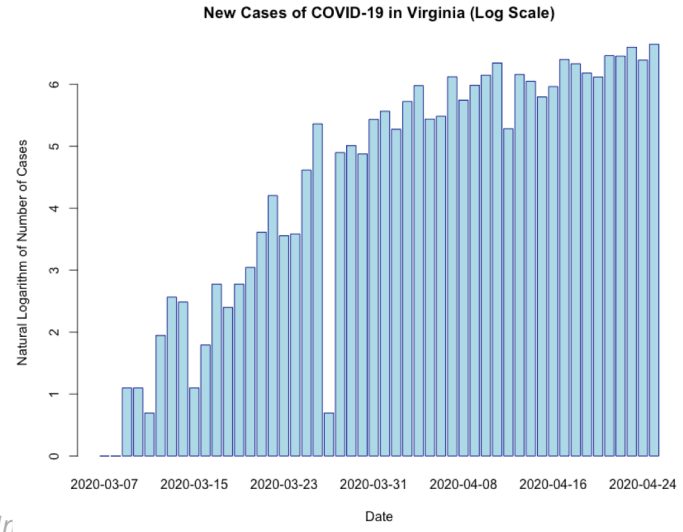
- Ordinary linear regression assumes:
  - Expected value of response variable is a linear function of predictor variables
  - Errors are normally distributed
- We can model many kinds of nonlinear relationship by transforming dependent and/or independent variables
- Generalized linear model (GLM) extends ordinary linear regression model to non-normal errors
  - Linear predictor is transformed by *link function:* $E[Y|X = g^{-1}(X_1\beta_1 + \cdots + X_p\beta_p)]$
  - Probability distribution of Y given X is a member of an exponential family of distributions
- Examples:
  - Logistic regression for 0/1 dependent variable
  - Poisson regression for count data
- R fits generalized linear model using `glm` function

Example: New COVID-19 Cases in Virginia

- Data: new cases from date of first case until April 25, 2020
- Question of interest: what is rate of growth?
  - If growth is exponential with constant rate, logarithm of new cases should follow approximately a straight line
  - Is rate of growth decreasing?
  - Will number of new cases start to decline?
- Constant growth rate: $E[\log(y)\,|d] = \alpha + \beta d$, where d is the number of days since 3/07/20 (date of first case)
- We will consider a quadratic model:

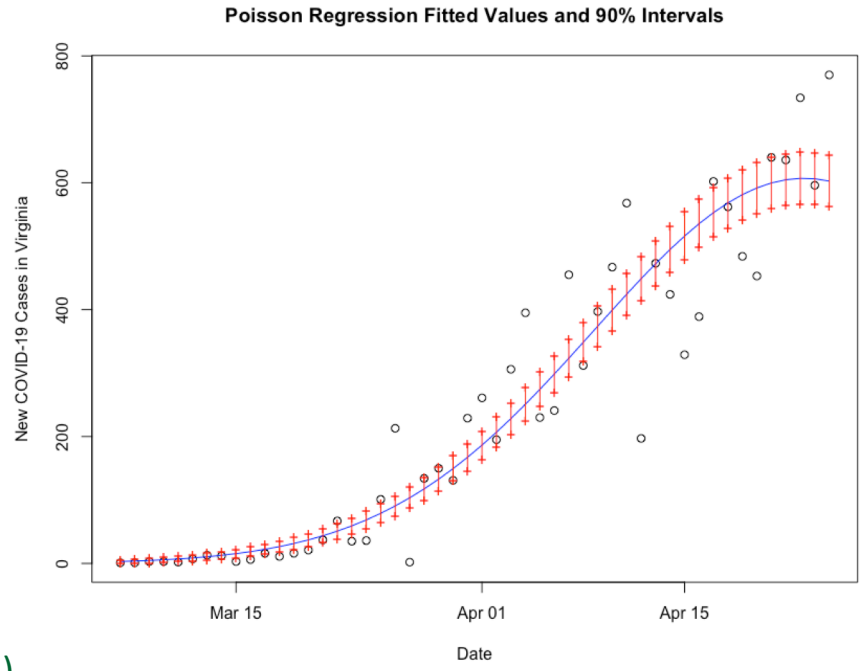
$$E[\log(y)\,|d] = \beta_0 + \beta_1 d + \beta_1 d^2$$





Poisson Regression using `glm` in R

```
Call:
glm(formula = new ~ day + I(day^2), family = "poisson", data = va.cases)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-13.6739   -3.2095   -0.8165    1.9502   10.9287

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.608e-01  9.826e-02    8.76   <2e-16 ***
day          2.299e-01  5.870e-03   39.17   <2e-16 ***
I(day^2)    -2.382e-03  8.486e-05  -28.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12446.3  on 49  degrees of freedom
Residual deviance:  1135.2  on 47  degrees of freedom
AIC: 1454.1

Number of Fisher Scoring iterations: 5
```



Observations have too high a variance for Poisson distribution (variance is much larger than mean)

Bayesian Model for New Case Counts

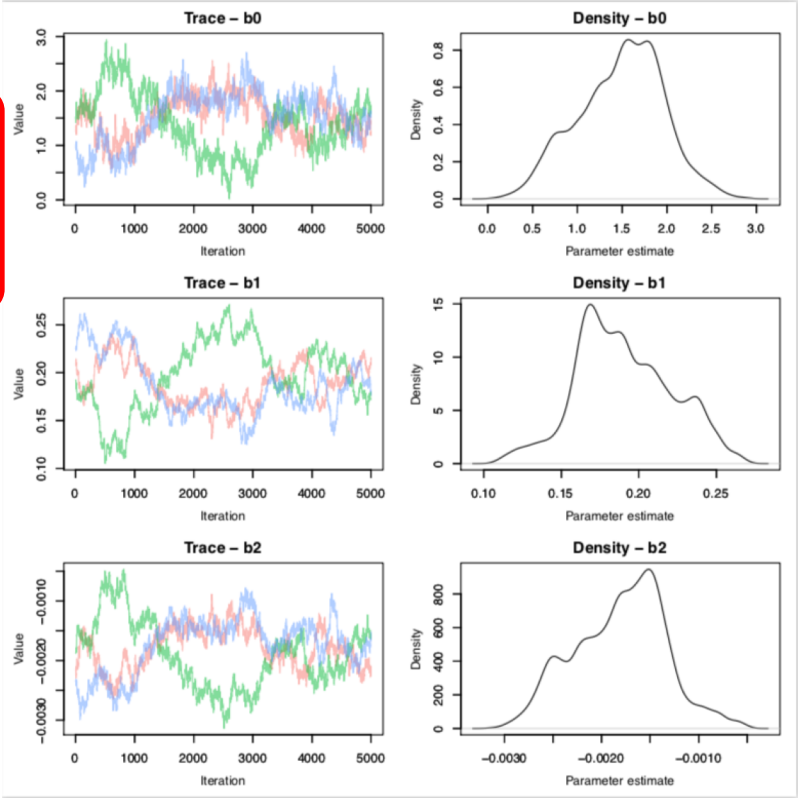
- Negative binomial observations (overdispersed Poisson – Poisson distribution with gamma-distributed mean)
- Expected log count is quadratic function of day
- `sfactor` controls overdispersion (small values mean large dispersion)

```
#Gibbs sample using JAGS

nc = va.cases$new
nd=dim(va.cases)[1]

covid.data <- list("nd","nc")  # data to pass to JAGS

covid.params <- c("b0","b1","b2","sfactor")  # parameters to monitor

covid.inits <- function(){
  list("b0"=0.9, "b1"=0.23, "b2"=-0.002, "sfactor"=0.1)
}

# The jags function takes data and starting values as input. It automatically writes
# a jags script, calls the model, and saves the simulations for easy access in R.
covid.fit <- jags(data=covid.data, inits=covid.inits, covid.params, n.chains=3,
                  n.iter=5500, n.burnin=500,model.file="pandemic.va.negbin.jags",
                  n.thin=1)
```

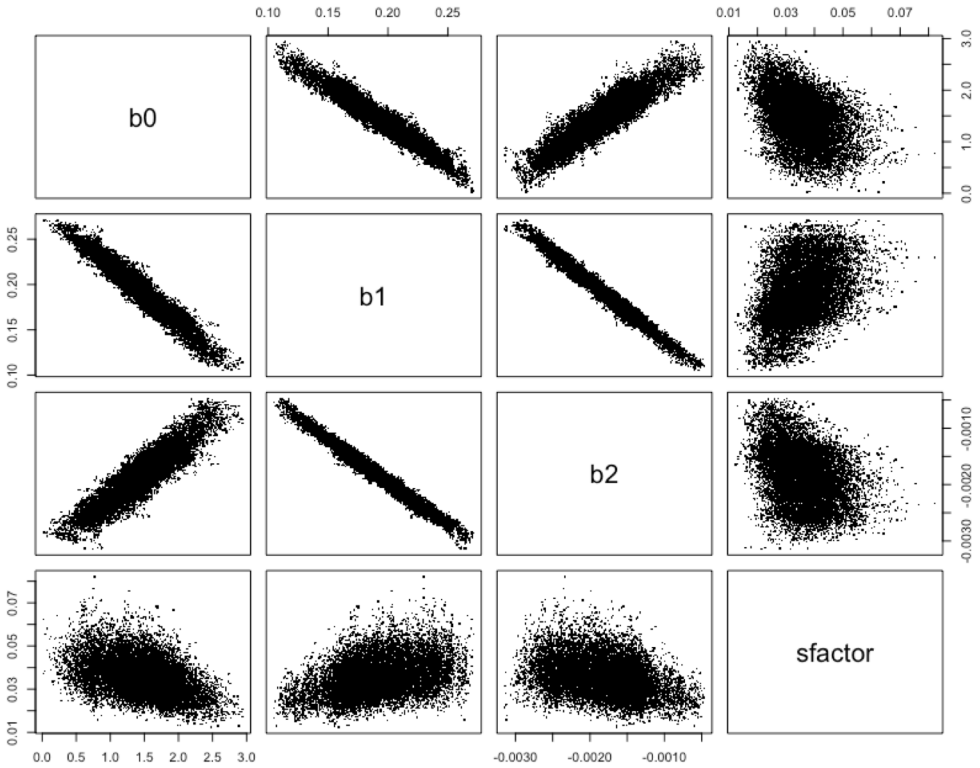
```
model {
  pr <- sfactor/(sfactor+1)
  for(d in 1:nd) {
    ll[d] <- b0 + b1*d + b2*d^2     # log mean number of cases
    elambda[d] <- exp(ll[d])        # mean number of cases
    sz[d] <- elambda[d]*sfactor     # Overdispersion parameter
    nc[d] ~ dnegbin(pr,sz[d])       # Negative binomial
    }
  sfactor ~ dbeta(1,1)T(0.01,0.2)   # controls overdispersion
  b0 ~ dnorm(0, .1)T(-0.3,4)      # constant term
  b1 ~ dnorm(0, .1)T(0.1,0.34)      # rate of increase at day 1
  b2 ~ dnorm(0, .1)T(-0.0045,0.0)   # rate of change of increase
  }
```

Results of Model Fit: We have a Problem!

```
Inference for Bugs model at "pandemic.va.negbin.jags", fit using jags,
 3 chains, each with 5500 iterations (first 500 discarded)
 n.sims = 15000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
b0         1.483   0.478   0.532   1.158   1.530   1.823   2.377 1.090    62
b1         0.190   0.031   0.129   0.168   0.187   0.212   0.250 1.056   140
b2        -0.002   0.000  -0.003  -0.002  -0.002  -0.001  -0.001 1.058    98
sfactor    0.036   0.008   0.021   0.030   0.035   0.041   0.054 1.002  2300
deviance 529.109   3.187 525.053 526.703 528.413 530.769 537.033 1.054    59

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 4.9 and DIC = 534.0
DIC is an estimate of expected predictive error (lower deviance is better).
```



The Problem: Correlated Parameters



- MCMC samplers have high autocorrelation when parameters are highly correlated ("mixing is poor")
- We can transform our variables to a less highly correlated representation

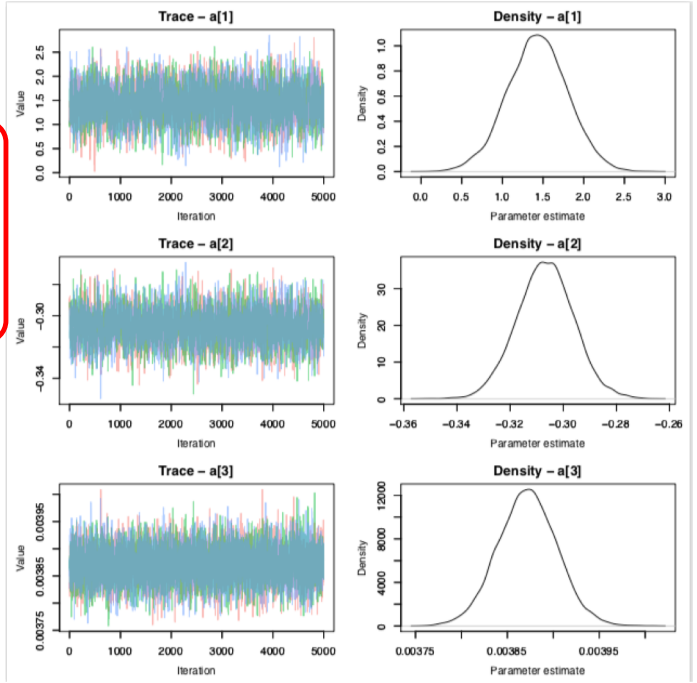
Transformed Observations

```
x = va.cases[c("day","d2")]     # day and day^2
x=cbind(rep(1,nd),x)            # vector of 1's
names(x)=c("const","day","d2")  # names for columns
x=as.matrix(x)                  # design matrix for regression
t(x) %*% x                      # X-transpose times X
bcov = solve(t(x) %*% x)        # Least squares covariance is proportional to this

v=eigen(bcov)$vectors           # eigenvectors of inverse(X-transpose times X)

a=c(b0,b1,b2)%*%v               # convert parameters to less correlated representation

nc = va.cases$new
nd=dim(va.cases)[1]

covid.data <- list("nd","nc","v")  # data to pass to JAGS

covid.params <- c("a","sfactor")  # parameters to monitor

covid.inits <- function(){
  list("a"=c(0.4, -0.3, 0.004), "sfactor"=0.1)
}


covid.fit.a <- jags(data=covid.data, inits=covid.inits, covid.params, n.chains=3,
                    n.iter=5500, n.burnin=500,model.file="pandemic.va.negbina.jags",
                    n.thin=1)
```

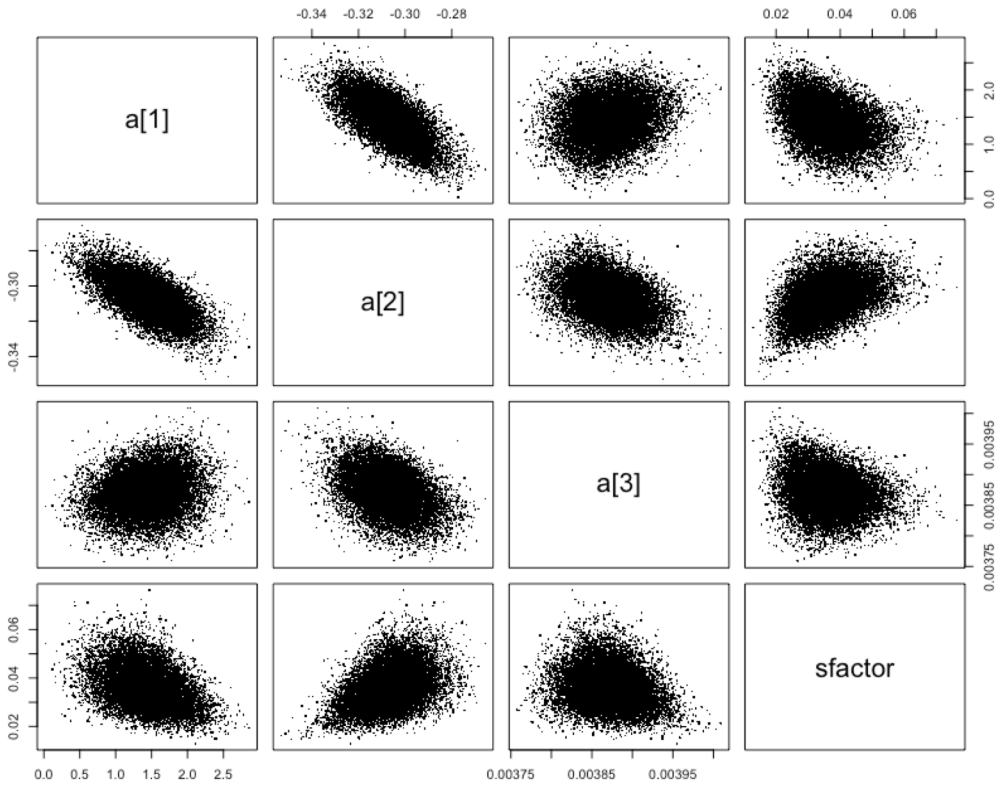
```
model {
  b<- a %*% t(v)
  pr <- sfactor/(sfactor+1)   # Probability for negative binomial
  for(d in 1:nd) {
    ll[d] <-  b[1] + b[2]*d + b[3]*d^2
    elambda[d] <- exp(ll[d])        # mean number of cases
    sz[d] <- elambda[d]*sfactor     # Overdispersion parameter
    nc[d] ~ dnegbin(pr,sz[d])       # Negative binomial
    }
  sfactor ~ dbeta(1,1)T(0.01,0.2)    # controls overdispersion
  a[1] ~ dnorm(1, 1)T(-0.83,3.13)    # 1st transformed coefficient
  a[2] ~ dnorm(0, .1)T(-0.45,-0.2)  # 2nd transformed coefficient
  a[3] ~ dnorm(0, .1)T(0.001,0.006) # 3rd transformed coefficient
  }
```

Transformation Improves Mixing

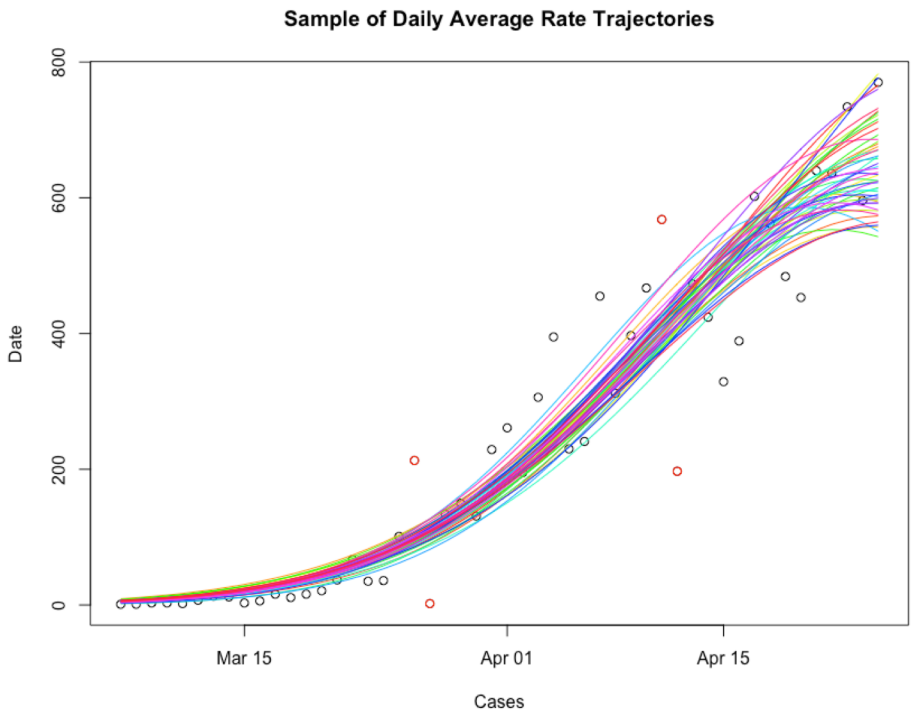
```
Inference for Bugs model at "pandemic.va.negbina.jags", fit using jags,
 3 chains, each with 5500 iterations (first 500 discarded)
 n.sims = 15000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a[1]       1.432   0.370   0.688   1.185   1.432   1.678   2.156 1.001 15000
a[2]      -0.307   0.011  -0.328  -0.314  -0.307  -0.300  -0.286 1.001 11000
a[3]       0.004   0.000   0.004   0.004   0.004   0.004   0.004 1.001 15000
sfactor    0.036   0.008   0.022   0.030   0.035   0.041   0.053 1.002  2900
deviance 528.378   2.749 524.972 526.381 527.740 529.673 535.539 1.001  5600

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 3.8 and DIC = 532.2
DIC is an estimate of expected predictive error (lower deviance is better).
```

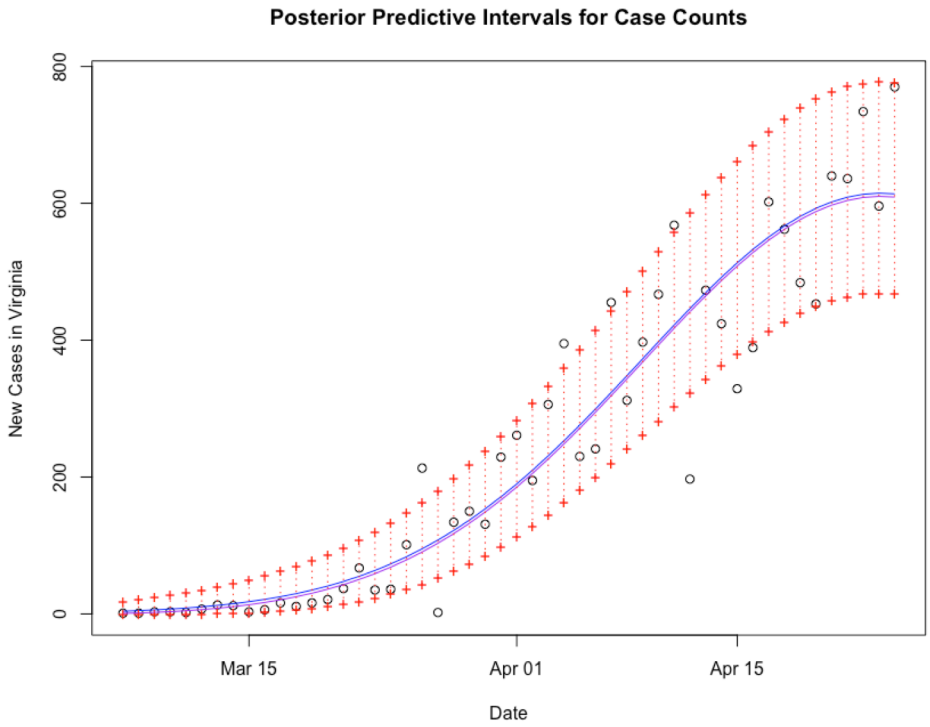


Transformation Reduces Corelation



Some Results





- Expected peak in new cases is 4/28/20
- Estimated 72.2% probability peak is after 4/25/20 (last data point in sample)
- Estimated 89.3% probability peak is 05/04 or earlier

Missing Data

- Missing observations are common in real-world data sets
- Common approaches to missing data
  - Remove all cases with missing data
    - Reduces power
    - Can bias parameter estimates (depending on why observations are missing)
  - Impute missing values
    - Many statistical packages have methods for imputing (filling in) missing values
    - Simple ad hoc imputation approaches (e.g., use mean of non-missing values) can bias analysis, especially of relationships among variables
    - Regression imputation (use regression model to predict missing observations) gives unbiased estimates of mean but underestimates variability
    - Multiple imputation (use observed data to learn about relationships and randomly generate multiple completions) gives unbiased estimates of both means and variances
  - Model-based methods
    - EM algorithm (iterative approach to find maximum likelihood or maximum a posteriori estimates in presence of missing data)
    - Fully Bayesian model of observed and unobserved data
- It is important to account for mechanism for missing observations

Missing Data: Some Terminology

- Missing Completely at Random (MCAR)
  - The events leading to an observation being missing are not related to any of the random variables being observed or any of the parameters of interest
  - Example – Respondent randomly fails to fill in some items in a questionnaire in a manner unrelated to any of the data being collected or any of the variables being studied
- Missing at Random (MAR)
  - Also known as ignorable non-response
  - Missingness is not random but can be accounted for by the observed data
  - Example – Female respondents are less likely to answer question on job satisfaction, but responses do not depend on job satisfaction after accounting for gender
  - Accounting for gender difference in estimates of job satisfaction can lead to sound statistical estimates
- Missing Not at Random (MNAR)
  - Also known as non-ignorable non-response
  - Whether or not value is missing is related to the reason it is missing in a way that cannot be accounted for by other measurements
  - Example – People who are not satisfied in their job are less likely to answer question on job satisfaction
  - Estimates will be biased unless we can measure something related to reason observations are missing

Missingness Mechanism

- Notation:
  - Full data set: $\underline{Y} = (\underline{Y}_O, \underline{Y}_M)$
  - Observed data: $\underline{Y}_O$
  - Missing data: $\underline{Y}_M$
  - Response indicator: $\underline{R}$ (1 if observed, 0 if missing)
    - $\underline{R}$ is a data structure parallel to $\underline{Y}$, with one entry for each entry in $\underline{Y}$ indicating whether that entry is missing or not
- Missing completely at random (MCAR)
  - $\mathrm{P}(r \mid \underline{Y}_O, \underline{Y}_M) = P(r)$
  - Probability of an observation being missing does not depend on either the observed or unobserved data
  - Deleting all cases with missing values yields unbiased estimates if data are MCAR
- Missing at random (MAR)
  - $\mathrm{P}\left(r \mid \underline{Y}_O, \underline{Y}_M\right) = \mathrm{P}\left(r \mid \underline{Y}_O\right)$
  - Units with same observed attributes have same distribution for missing attributes
  - EM algorithm and simple Bayesian models assume MAR
- Missing not at random (MNAR)
  - Missingness mechanism is not ignorable
  - Requires joint model of $\underline{Y}$ and $\underline{R}$

Bayesian Models with Missing Data

- The Bayesian approach handles missing data naturally
  - Specify joint probability distribution $P(\underline{Y}, \underline{\Theta}, \underline{R})$ on full data, parameters, and missing data mechanism
  - Use Bayesian inference to find posterior distribution $P(\underline{\Theta}, \underline{Y}_M | \underline{Y}_O, \underline{R})$ of parameters $\underline{\Theta}$ and missing data YM given observed data YO and indicator R of which observations are missing
  - Calculate and/or approximate desired features of posterior distribution
- JAGS handles missing data
  - NA values are allowed in data variables
  - JAGS will sample values for the NA variables
  - Full probability model $P(\underline{Y}, \underline{\Theta}, \underline{R})$ must be specified
- Examples:
  - Predict future dependent variables for regression given independent variable values
  - Estimate parameters of a regression model when some values of independent variables are missing

JAGS and Missing Data

- Predict future dependent variables for regression given independent variable values
  - Specify distribution for observed $\underline{Y}$ given observed $\underline{X}$ and $\underline{\beta}$
  - Specify prior distribution for $\underline{\beta}$
  - Specify independent variables X_{new} for cases to be predicted
  - Set Y_{new} = NA for cases to be predicted
  - JAGS will simultaneously sample parameters and future observations from $P(\underline{\beta}, Y_{new} | \underline{Y}, \underline{X}, X_{new})$
- Estimate parameters of a regression model when some values of independent variables are missing (assume MAR)
  - Specify distribution for observed $\underline{Y}$ given full data $\underline{X}$ and $\underline{\beta}$
  - Specify prior distribution for $\underline{\beta}$
  - Specify prior distribution for $\underline{X}$
  - Set $\underline{Y}_M$= NA for missing observations
  - JAGS will simultaneously sample parameters and missing data from P($\underline{\beta}, \underline{Y}_M | \underline{Y}, \underline{X}_O$) under assumption that observations are missing at random

Example: Oxygen Uptake with Missing Data

Missing Data:

```
# Add new observation with missing age
age[13]=NA       # Add observation with missing age
aerobic[13]=1    # Person with unknown age was on aerobic program
y[13]=5.5        # Oxygen update for person with unknown age
```

JAGS Model:

```
model{
  for (i in 1:n) {
    age[i] ~ dunif(18,35)
        y[i] ~ dnorm((a[2]+b[2]*age[i])*aerobic[i]+(a[1]+b[1]*age[i])*(1-aerobic[i]),rho)
           }
    for (i in 1:2) {
        a[i] ~ dnorm(0,0.0001)
        b[i] ~ dnorm(0,0.0001)
    }
    rho ~ dgamma(0.5,1/1000)
  age13 <- age[13]
    }
```

JAGS Output:

```
Inference for Bugs model at "Oxygen.model.missing.jags", fit using jags,
 3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 9
 n.sims = 3000 iterations saved
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a[1]      -49.633  13.566 -76.502 -58.389 -49.579 -40.767 -22.209 1.016   170
a[2]      -35.408  11.652 -59.582 -42.680 -35.159 -27.705 -13.500 1.009   270
age13      24.632   2.231  20.063  23.318  24.646  25.907  29.190 1.001  2600
b[1]        2.024   0.584   0.864   1.642   2.025   2.394   3.189 1.016   170
b[2]        1.668   0.449   0.829   1.372   1.655   1.952   2.598 1.007   320
rho         0.128   0.063   0.038   0.082   0.116   0.162   0.283 1.001  3000
deviance  134.241   4.227 128.039 131.112 133.627 136.572 144.145 1.001  3000
```

Collinearity and Identifiability

- Collinearity means $X^T X$ is singular
  - One of the explanatory variables is a linear combination of other explanatory variables
  - The classical regression estimate $\underline{b} = (X^T X)^{-1} X^T \underline{y}$ does not exist
  - The posterior distribution exists if the prior distribution is informative, but the data provide no information about some linear combinations of the β's
  - Inferences depend on prior assumptions about these linear combinations
- When $X^T X$ is nearly singular we say the data are nearly collinear
  - Data are not very informative about some linear combinations of the β's
  - Inferences about these linear combinations are sensitive to the prior distribution
  - Standard errors for some regression coefficients may be very large
  - Posterior distributions for some regression coefficients are highly correlated
- If some β's cannot be estimated uniquely we say they are not identifiable
- Regression modules in many statistical packages contain diagnostics for collinearity

Indicator Variables

- Sometimes one of the regressors is a categorical variable
- To use a categorical variable with k categories as an explanatory variable, we transform it into k-1 indicator variables
- Example: Region variable has possible values North, South, East and West
  - Transform into 3 indicator variables X_1, X_2, X_3
  - North: $X_1 = 1, X_2 = 0, X_3 = 0$
  - South: $X_1 = 0, X_2 = 1, X_3 = 0$
  - East: $X_1 = 0, X_2 = 0, X_3 = 1$
  - West: $X_1 = 0, X_2 = 0, X_3 = 0$

    (If we added a fourth categorical variable to have value 1 for West and 0 for other regions, and the regression had a constant term, the data would be collinear)
- Regression coefficient for X_i represents the effect of being in the corresponding region relative to being in the West

ANOVA

- ANalysis Of VAriance (ANOVA) refers to a method for summarizing the results of a normal regression model
- ANOVA partitions the sum of squares into components:
  - $S_{yy} = S_{reg} + S_{ee}$
  - $S_{yy} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares
  - $S_{reg} = \sum_i (a + bx_i - \bar{y})^2$ is the regression sum of squares
  - $S_{ee} = \frac{1}{n}\sum_{i=1}^{n} (y_i - a - b(x_i - \bar{x}))^2$ is the error (or residual) sum of squares
- Interpretation:
  - S_{reg} measures the variation in the data explained by the regression model
  - S_{ee} measures the unexplained variation
- Most regression software includes ANOVA table
  - F test evaluates whether S_{reg} is larger than expected by chance if α and β are zero

More on ANOVA

- ANOVA is also used to refer to a type of regression model in which the predictor variables are categorical variables
- One Way ANOVA refers to a model in which the data fall into groups with a separate mean for each group and a common variance (e.g., the reaction time data)
- Multi factor ANOVA refers to a model in which
  - There are p factors, each with a finite number of levels
  - The predictor variables consist of the levels of each factor
  - The response variable is normally distributed with mean depending on the factor levels and common variance
- No interaction model: category mean is equal to sum of a coefficient for each factor
- Example: 2-way ANOVA, no interaction
  - Y_{ijk} denotes the kth observation having the ith level on the first factor and the jth level on the second factor
  - $E[Y_{ijk}] = \alpha_i + \beta_j$
  - The ordinary least squares estimates of α_i and β_j are the regression estimates obtained by forming categorical variables for the factors

ANOVA and Regression

- Example: evaluate effect of new teaching method
  - Categorize students by new/old teaching method T, ability level A (categorical), socioeconomic status S (categorical)
  - Measure performance score of ith student who uses method t and has ability level a and socioeconomic status s Xtasi
- Model:
  - Observation X_{itas} is normally distributed with mean μ_{tas} and standard deviation σ
  - Mean score$\mu_{tas} = \mu + \mu_t + \mu_a + \mu_s$
- We can represent this as a linear regression model with "dummy variables" corresponding to the categories
- Because this kind of model is so common, standard statistical packages provide special-purpose methods for this kind of regression model
- Theory for ANOVA is a special case of linear model theory

Transformation of Variables

- We often need to transform variables to achieve a linear relationship
- Common transformations are power, logarithm, and logit (log(x/(1-x))
- We can transform the dependent variable and/or the independent variables
- We can include more than one transformation of the same independent variable (e.g., polynomial regression)
- A transformation changes the interpretation of the regression coefficient

Summary and Synthesis

- We studied regression models in which a dependent variable y was related to an independent variable x by a linear equation with normal errors
- We considered the posterior distribution for the coefficients under a noninformative prior distribution
  - Relation to classical estimators
  - Credible intervals for parameters
- We generalized to an informative Normal-Gamma conjugate prior distribution
- We studied the marginal likelihood
  - Used in predicting as-yet-unseen observations
  - Used in hypothesis testing (next unit)
- We discussed how to generalize these results to multiple regression
- We considered Bayesian hierarchical regression models
- We looked at regression models for count data
- We explored regression models with missing data

References for Unit 8

- Draper, N.R., and Smith, H. (1981) Applied Regression Analysis, 2nd Edition, Wiley: New York
- Lee, Peter, Bayesian Statistics: An Introduction (2nd edition), Arnold, 1997
- Mosteller, F. and Tukey, J. Data Analysis and Regression: A Second Course in Statistics, Addison Wesley, 1977
- Press, J Bayesian Statistics, Wiley, 1989