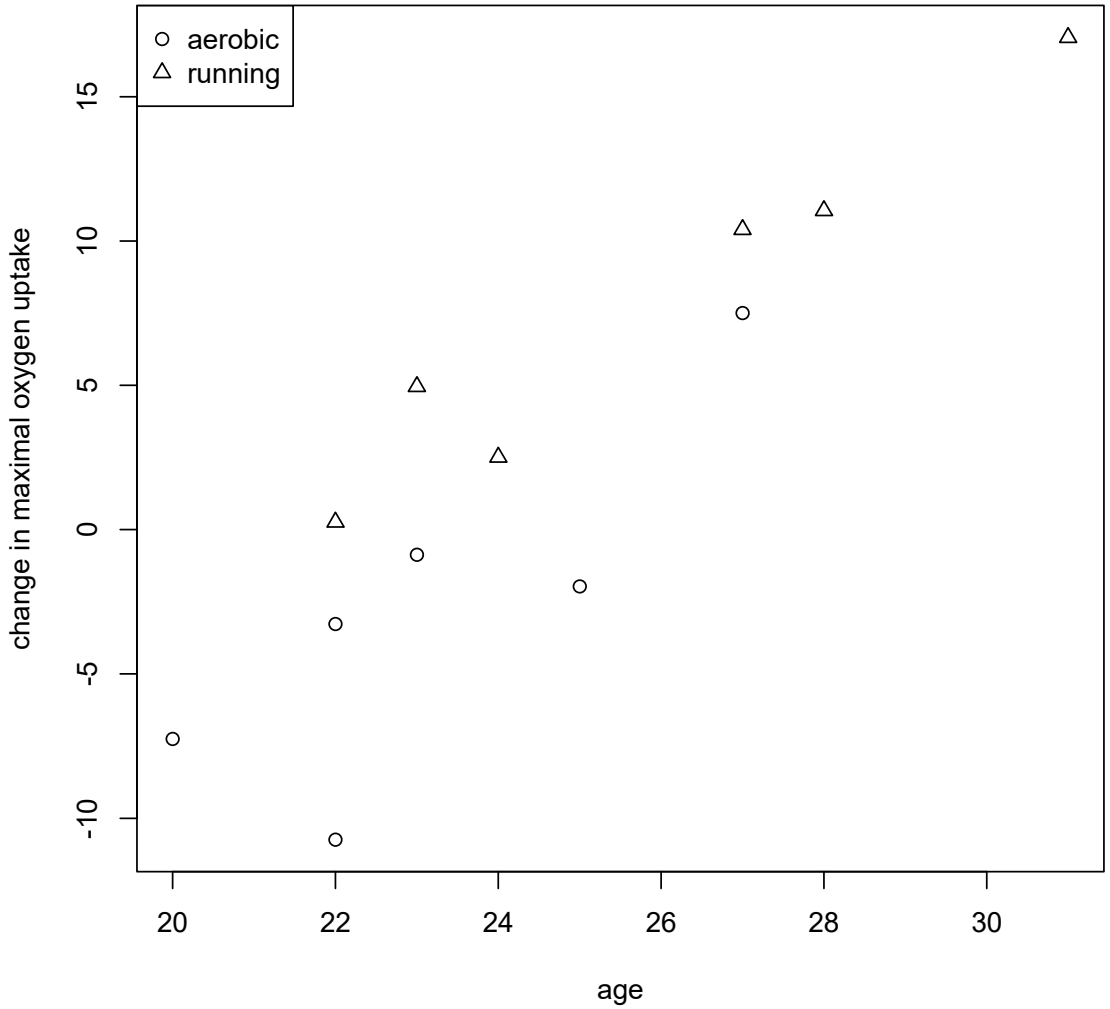
Normal Error Regression Models

Ch 5 BMUW

Example: Oxygen uptake (Kuehl 2000)

Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each subject was measured (in liters per minutes) while running on an inclined treadmill, both before and after the 12-week program. A graph of the change in maximal oxygen uptake vs age by treatment is shown below.


aerobic
running
change in maximal oxygen uptake
-10
-5
0
5
10
15
20
22
24
26
28
30
age

Regression Basics

Normal error regression models are some of the most popular statistical models.

Context:

- The response variable Y is continuous.
- We have p explanatory variables, $X_1, X_2, \ldots, X_p$ which we believe give us insight into the typical behavior of the response Y.
- We have a coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.
- $y_i | x_i, \beta, \sigma^2 \sim N(x_i^T \beta, \sigma^2)$ and $y_1, y_2, \ldots, y_n$ are independent.
    - x_i is the vector of explanatory values for observation i.

The normal regression model can also be represented as

$$Y \mid X, \beta, \sigma^2 = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \epsilon,$$
$$\epsilon \mid \sigma^2 \sim N(0, \sigma^2).$$

Let $y = (y_1, \ldots, y_n)$ denote an n-dimensional vector of response and X be an $n \times (p + 1)$ matrix whose ith row is $x_i = (1, x_{i1}, \ldots, x_{ip})$.

The normal error regression model is

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

where I is the $n \times n$ identity matrix.

Least-Squares Estimation

In the frequentist paradigm, the most common technique used to estimate regression parameters is ordinary least-squares.

In ordinary least-squares estimation, we estimate the values of β using the values that minimize the residual sums of squares

$$RSS = \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2.$$

The solution to this problem (assuming the X matrix has linearly independent columns) is

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

and an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}\right)^2}{n - (p+1)}.$$

The sampling distribution of the estimated regression coefficients is

$$\hat{\beta} | X, \beta, \sigma^2 \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

Bayesian Estimation of a Regression Model

The data distribution is

$$y|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I).$$

To do Bayesian analysis of the regression model, we only need to specify a prior distribution for our parameters.

Conjugate Prior Distributions

Conjugate prior distributions for σ^2 and β are

$$\sigma^2 \sim \text{Inv-Gamma}(a, b)$$

and

$$\beta|\sigma^2 \sim N\left(\mu_\beta, \sigma^2 V_\beta\right),$$

(taken together to form a Normal-Inverse-Gamma prior distribution).

Under these prior distributions, the posterior distribution for σ^2 is $\sigma^2|y \sim \text{Inv-Gamma}(a^*, b^*)$ with

$$a^* = a + n/2,$$

$$b^* = b + \frac{1}{2}(\mu_\beta^T V_\beta^{-1} \mu_\beta + y^T y - \mu_*^T V_* \mu_*),$$

$$\mu_* = \left(V_\beta^{-1} + X^T X\right)^{-1}\left(V_\beta^{-1}\mu_\beta + X^T X y\right),$$

and

$$V_* = \left(V_\beta^{-1} + X^T X\right)^{-1}.$$

The posterior distribution for β is

$$\beta|y \sim t_{\nu^*}(\mu_*, \Sigma^*),$$

with $\nu^* = 2a^*$ and $\Sigma^* = \left(\frac{b^*}{a^*}\right) V_*$.

A Simple Independent Prior Distribution

The conjugate prior distribution for (β, σ^2) is convenient for posterior inference and simulations, but other prior distributions are also commonly used.

A very simple approach for prior distributions in a regression setting is to assume that all parameters are independent with

$$\beta_j \sim N\left(\mu_{\beta_j}, c_j^2\right), \quad j = 0,1, \ldots, p,$$

and

$$\sigma^2 \sim \text{Inv-Gamma}(a, b)$$

which is equivalent to $\tau = 1/\sigma^2 \sim \text{Gamma}(a, b)$.

A low information prior for μ uses something like $\mu_{\beta_j} = 0$ and the variance c_j^2 to a large value like 10^4.

- This centers our prior belief around zero, which corresponds to the assumption that X_j has no effect on Y.
- This known as a "skeptical" prior.
- The large variance corresponds to high uncertainty or high ignorance.
- Conversely, this corresponds to a small precision like 10^{-4}.

A low information prior for the τ is to use $a = b = 0.01$.

- The prior mean is $E(\tau) = 1$ and $\text{var}(\tau) = 100$.
- Interestingly, the prior mean and variance of σ^2 are undefined.

Interpretation of Regression Coefficients

The regression coefficients describe the mean effect of the explanatory variables on the response variable Y after adjusting for the other explanatory variables.

We may ask three questions:

1. Is the effect of X_j important for the prediction or description of Y?
2. What is the association between Y and X_j (positive, negative, or other)?
3. What is the magnitude of the effect of X_j on Y?

In answering these questions, we note:

- A posterior interval farther from zero suggests that X_j has more impact on the response variable.
  - If the entire posterior interval is more than zero, the relationship is positive (as the explanatory variable increases, the response increases).
  - If the entire posterior interval is negative, the relationship is negative (as one increases, the other decreases).
  - If the interval straddles zero, we must assess whether the relationship is more likely to be positive or negative.
- β_j represents the change in the mean value of Y when X_j increases by one unit and the other explanatory variables remain constant.

- The precision τ indicates the precision of the model. If the precision is large, it means that we can accurately predict the expected value of Y.
- The R_B^2 statistic given by

$$R_B^2 = 1 - \frac{\sigma^2}{s_Y^2} = 1 - \frac{1}{\tau\, s_Y^2},$$

  represents the proportional reduction of uncertainty concerning the response variable Y achieved by incorporating the explanatory variables in the model, with s_Y^2 being the sample variance of the responses.
- The R_B^2 statistic is the Bayesian analogue of the adjusted R^2 statistic (R_a^2). We prefer models with larger values of R_B^2.

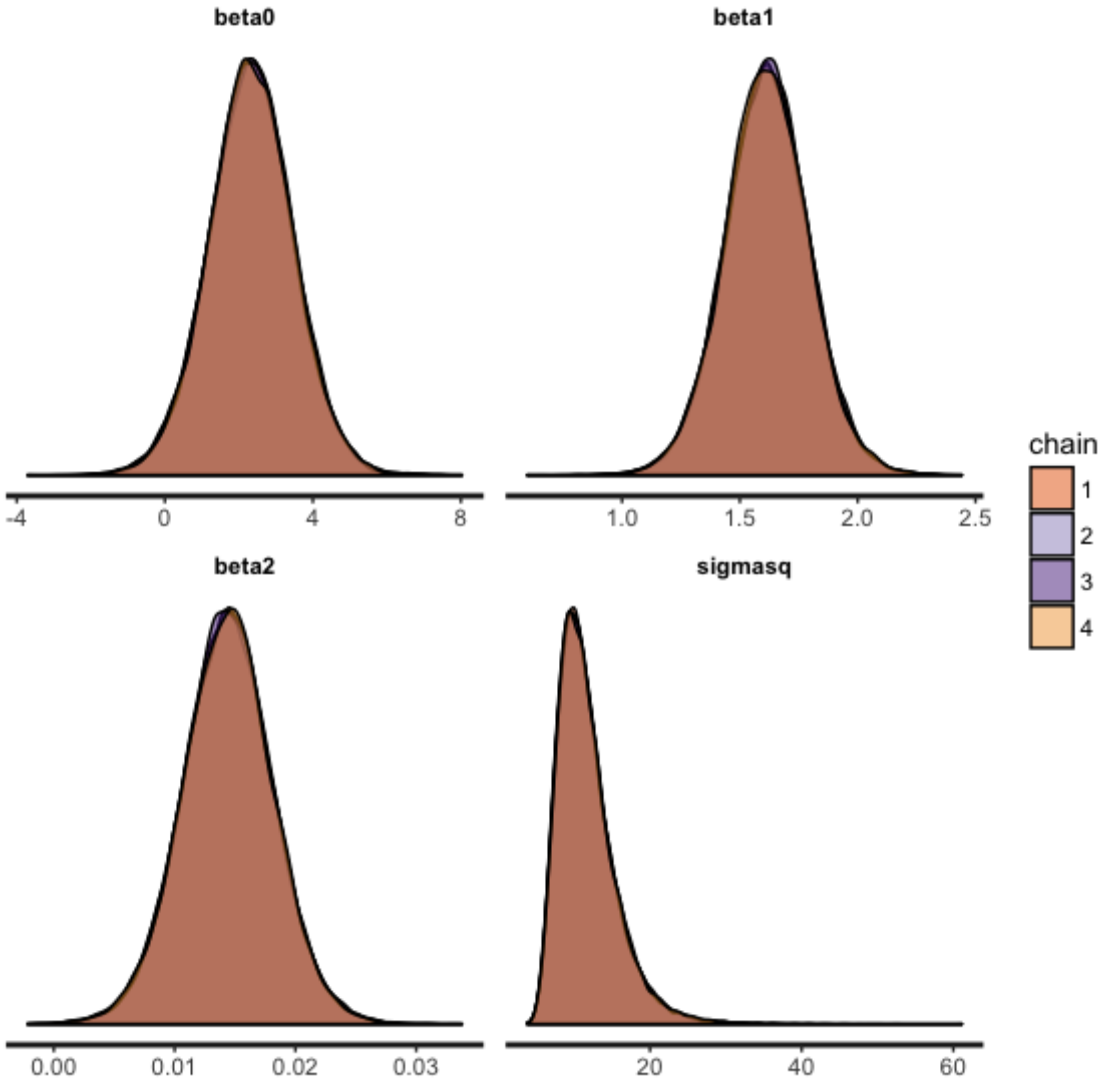
Example: Soft drink delivery times

We are interested in estimation of the required time needed by each employee in a delivery system network to refill an automatic vending machine. For this reason, a small quality assurance study was set up by an industrial engineer of the company. As the response variable, the engineer considered the total service time (measured in minutes) of each machine, including its stocking with beverage products and any required maintenance or housekeeping. After examining the problem, the industrial engineer recommended two important variables that affect delivery time: the number of cases of stocked products and the distance walked by the employee (measured in feet).

Data distribution: $y_i|x_i, \beta, \tau \sim N\left(x_i^T\beta, \tau^{-1}\right), i = 1,2, \ldots, n.$

Prior distributions:
$\beta_j \overset{i.i.d.}{\sim} N(0,10^4), j = 0,1,2.$
$\tau \sim \text{Gamma}(0.01, 0.01).$

Approximate posterior densities for the parameters of interest are shown below.


beta0
beta1
beta2
sigmasq
-4
0
4
8
1.0
1.5
2.0
2.5
0.00
0.01
0.02
0.03
20
40
60
chain
1
2
3
4

Check convergence using the Gelman-Rubin statistics:

```
> summary(fit)$summary[,"Rhat"]
     prec     beta0     beta1     beta2
1.0000913 0.9999862 0.9999840 0.9999754
    sigma   sigmasq      Rbsq       lp__
1.0001161 1.0001225 1.0001225 1.0000503
```

95% central posterior intervals:

```
> summary(fit)$summary[,c("2.5%",
"97.5%")]
                    2.5%         97.5%
prec        0.047204700   0.15743966
beta0       0.065099050   4.59749804
beta1       1.262671566   1.96951975
beta2       0.006891852   0.02188431
sigma       2.520245930   4.60264391
sigmasq     6.351639549  21.18433101
Rbsq        0.912106457   0.97364712
lp__     -46.522413637 -40.72450795
```

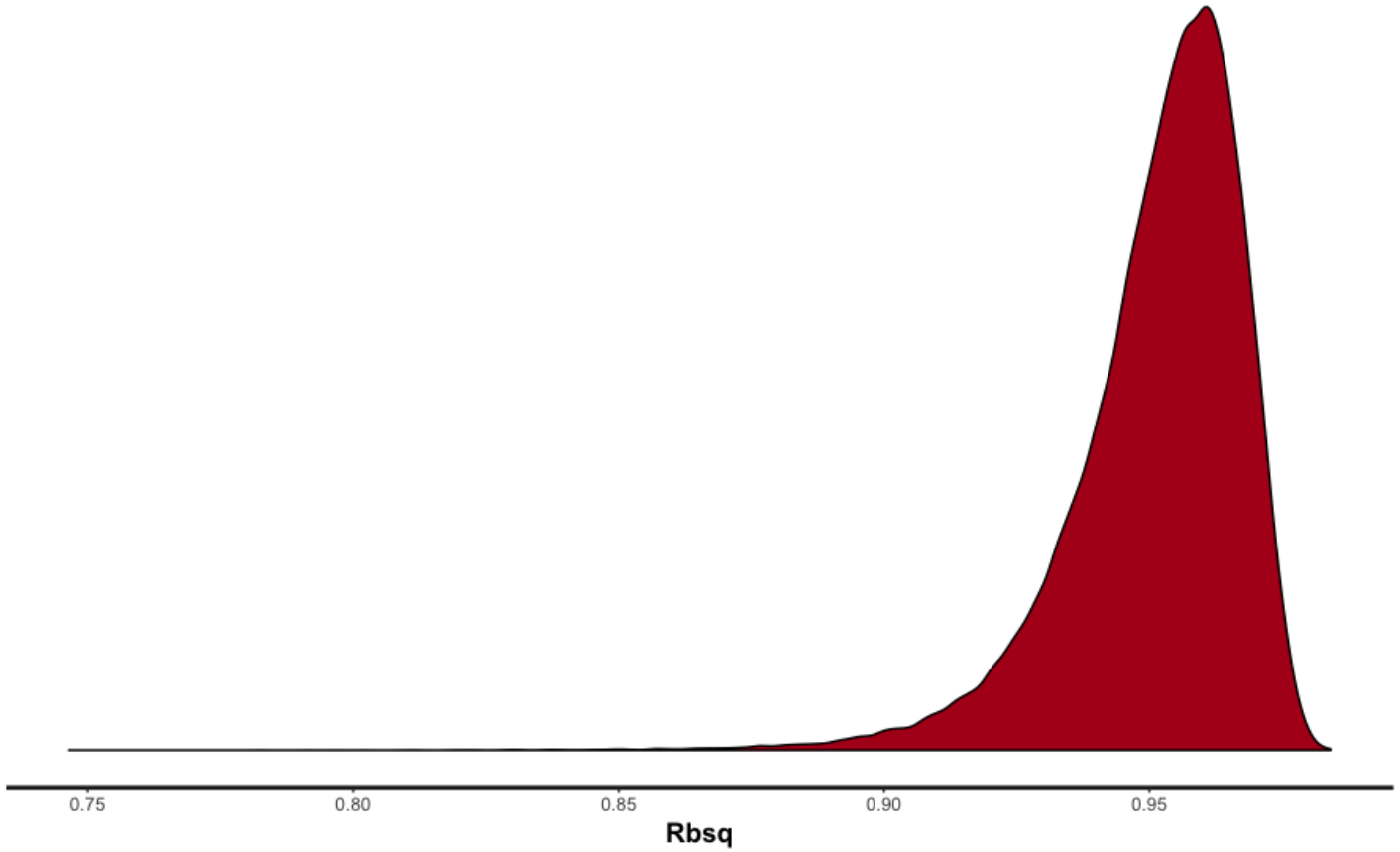
A point estimate of the regression model (using the means) is

$$E[time|cases, dist] = 2.34 + 1.62 \times cases + 0.14 \times dist.$$

The expected delivery time a posteriori (i.e., after the data) increases by about 1.6 [1.23, 1.97] minutes for each additional case stocked (assuming the distance walked remains the same).

The expected delivery time a posteriori (i.e., after the data) increases by about 0.014 [0.007, 0.022] minutes for each additional foot of distance walked (assuming the number of cases to be delivered remains the same).

These explanatory variables seem to explain a great deal of the variation in the responses. Consider the posterior distribution of R_B^2.



Other prior distributions

Recall that conjugate prior distributions for σ^2 and β are $\sigma^2 \sim \text{Inv-Gamma}(a, b)$and $\beta|\sigma^2 \sim N\left(\mu_\beta, \sigma^2 V_\beta\right)$.

A special case of the conjugate prior distributions is the popular Zellner (1986) g-prior, in which $V_\beta = c^2(X^T X)^{-1}$.

- It is called the g-prior because $c^2 = g$ in the original publication.
- The default choice for c^2 is $c^2 = n$ since it has an interpretation of adding prior information equivalent to one data point.

- It is said to be equivalent to the information of one data point because the precision of $\hat{\beta}$ (in a frequentist context) using n observations is $(X^T X)/\sigma^2$, the information of a single data point is $(X^T X)/n\sigma^2$.
- If no information is available, one might have $V_\beta = I_{p+1}$ and set c^2 equal to some large value like 100^2.
- The conjugate prior setup described above is very convenient for implementing Bayesian variable selection (deciding which variables should be included in the model).
- Zellner's g-prior is a popular choice among these because it allows for a sensible "default" choice of prior distribution.
- Other distributions such as the t distribution and the Cauchy distribution have been suggested as prior distributions for β, but they are less popular.

Example: Soda (continued)

We will examine the same model as before using Zellner's g-prior ($c^2 = n$ and $c^2 = 100^2$).

Analysis of Variance Models

Suppose we have:

- Y: a continuous response variable
- A: a categorical explanatory variable L levels 1, 2, …, L.

Assuming the response variable is normally distributed, we are assuming

$$Y|A = j \sim N(\mu_j, \sigma^2), \quad j = 1, \ldots, L,$$

where:

- j indicates the category a response belongs to.
- μ_j is the mean of Y for group j.

We believe that Y has (potentially) different means for each level of A.

Equivalently, we can write this model as

$$Y|A = j \sim N\left(\mu_j', \sigma^2\right), \quad j = 1, \ldots, L,$$

with $\mu_j' = \mu_0 + \alpha_j$, where:

- μ_0 is an overall mean effect common to all levels of *A.*
- $\alpha_j, j = 1, \ldots, L$ are group-specific parameters indicatinog the association of level *j* on the mean of the response variable *Y*.

These types of models are known as a **one-way analysis of variance models**.

- The key point is that there is only one categorical variable.
- With two categorical variables we could fit a two-way analysis of variance (ANOVA) model, etc.

We will use the alternative expression of the one-way ANOVA model because it:

1. Separates the constant overall effect from the effect of the variable A.
2. It allows for generalization of the ANOVA formulation when additional categorical explanatory variables are involved in the model.

Consider a random sample of n individuals resulting in n_j observations for each level j of variable A, and let Y_{jk} denote the kth observation of the jth level of variable A.

Thus,

$$Y_{jk}|\mu'_j, \sigma^2 \sim N\left(\mu'_j, \sigma^2\right), \quad k = 1,2, \ldots, n_j, \qquad j = 1,2, \ldots, L.$$

We must impose a constraint on the μ'_j parameters to make them identifiable (estimable).

We use the corner (CR) constraint, though the sum-to-zero (STZ) constraint is also quite popular.

For the corner constraint, the effect of a level $r \in \{1,2, \ldots, L\}$ is set equal to zero, i.e., $\alpha_r = 0$.

- The level r is referred to as the **baseline** or **reference category** of factor A.
- Typically, this is the first or last level (in order) of the factor A, though this is simply for convenience.
- For simplicity, we will assume $\alpha_1 = 0$ in what follows.

Under this parameterization, we have

$$\mu_1' = E(Y|A = 1) = \mu_0$$

and

$$\mu_j' = E(Y|A = j) = \mu_0 + \alpha_j, \quad j = 2, \ldots, L.$$

The interpretation of α_j is thus the expected difference in Y for an individual having level j of A in comparison to an individual in the reference group/level.

To model the mean effects in Stan, we only need to estimate $\mu_0, \alpha_2, \ldots, \alpha_L$, and will simply set $\alpha_1 = 0$.

- We also estimate the variance σ^2, which has only an indirect effect on the other parameters.

The one-way ANOVA model is a special case of the typical multiple regression model with specially crafted explanatory variables called **indicator variables**.

An indicator variable (dummy variable) can be used to represent the various levels of the categorical variable.

An indicator variable, D, takes the value 1 when an observation belongs to a certain group and 0 when it does not.

Indicator variables can be used to fit one-way, two-way, or higher ANOVA models, as well as multiple linear regression models that have different intercepts for each group, different slopes for each group, or both, etc.

To represent a categorical variable with L levels, we need $L - 1$ indicator variables.

- If all the indicator variables are zero, you know an observation must have the level that was left out.

One-way ANOVA and Multiple Regression

We want to fit the model

$$Y_{jk} \sim N\left(\mu'_j, \sigma^2\right), \quad k = 1,2, \dots, n_j, \qquad j = 1,2,3.$$

Since we have 3 levels, we need two indicator variables, D_2 and D_3.

- Let level 1 be the reference level.
- Let $D_{j2} = 1$ if observation j has level 2 (and 0 otherwise).
- Let $D_{j3} = 1$ if observation j has level 3 (and 0 otherwise).

Then

$$\begin{aligned} Y_{jk} &\sim N\left(\beta_0 + \beta_1 D_{j2} + \beta_2 D_{j3}, \sigma^2\right) \\ &= N\left(\mu_0 + \alpha_2 D_{j2} + \alpha_3 D_{j3}, \sigma^2\right). \end{aligned}$$

Consequently:

- An observation having level 1 (meaning both D_{j2} and D_{j3} are zero) will have mean β_0.
  - i.e., $\beta_0 \equiv \mu_0$.
- An observation having level 2 (meaning $D_{j2} = 1$ and $D_{j3} = 0$) will have mean $\beta_0 + \beta_1$.
  - i.e., $\beta_1 = \alpha_2$ (the effect of level 2 on the mean response of level 1).
- An observation having level 3 (meaning $D_{j2} = 0$ and $D_{j3} = 1$) will have mean $\beta_0 + \beta_2$.
  - i.e., $\beta_2 = \alpha_3$ (the effect of level 3 on the mean response of level 1).
- In the Stan models, we will use the notation α_i instead of β_{i-1} to be clearer.

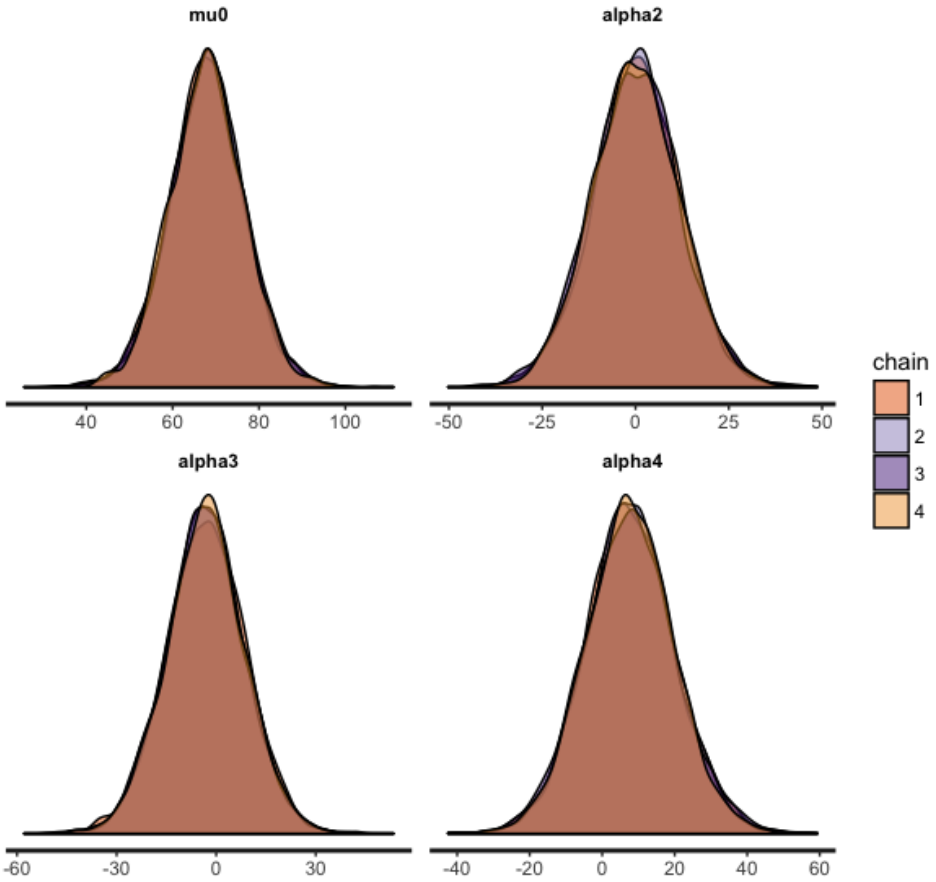
Example: Evaluation of Tutors

The director of a private school wishes to employ a new mathematics tutor. The ability of four candidates is examined using a small study. A group of 26 students was randomly divided into four classes. In all classes, the same mathematical topic was taught for 2 hours per day for 1 week. After completing the short course, all students had to take the same test. Their grades were recorded and compared. The administrator wishes to employ the tutor whose students attained the higher performance for the given test.

Data distribution: $Y_{jk}|\mu_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \sigma^2 \sim N\left(\mu_0 + \alpha_j, \sigma^2\right)$, $j = 1,2,3,4$, $k = 1,2, \ldots, n_j$.

Prior distributions:
$\mu_0, \alpha_j \sim N(0,100^2)$, $j = 2,3,4$ (recall that $\alpha_1 = 0$).
$\sigma^2 \sim$ Inv-Gamma(0.01, 0.01)

Posterior densities:



95% central posterior intervals:

```
> summary(fit)$summary[,c("2.5%", "97.5%")]
                  2.5%          97.5%
prec       0.00114236   0.003846241
mu0       50.55932615  85.158470522
alpha2   -23.50098854  23.707383028
alpha3   -26.71757916  20.181389711
alpha4   -16.70795679  32.484900760
sigma     16.12433320  29.586838757
sigmasq  259.99412107 875.381027631
lp__     -97.18268387 -90.650538269
```

Posterior means:

```
> summary(fit)$summary[,"mean"]
         prec            mu0         alpha2         alpha3
 0.002288687  67.886810894   0.109731835  -3.230837636
      alpha4          sigma       sigmasq            lp__
 7.643409242  21.661444458 481.233051609 -92.949617943
```

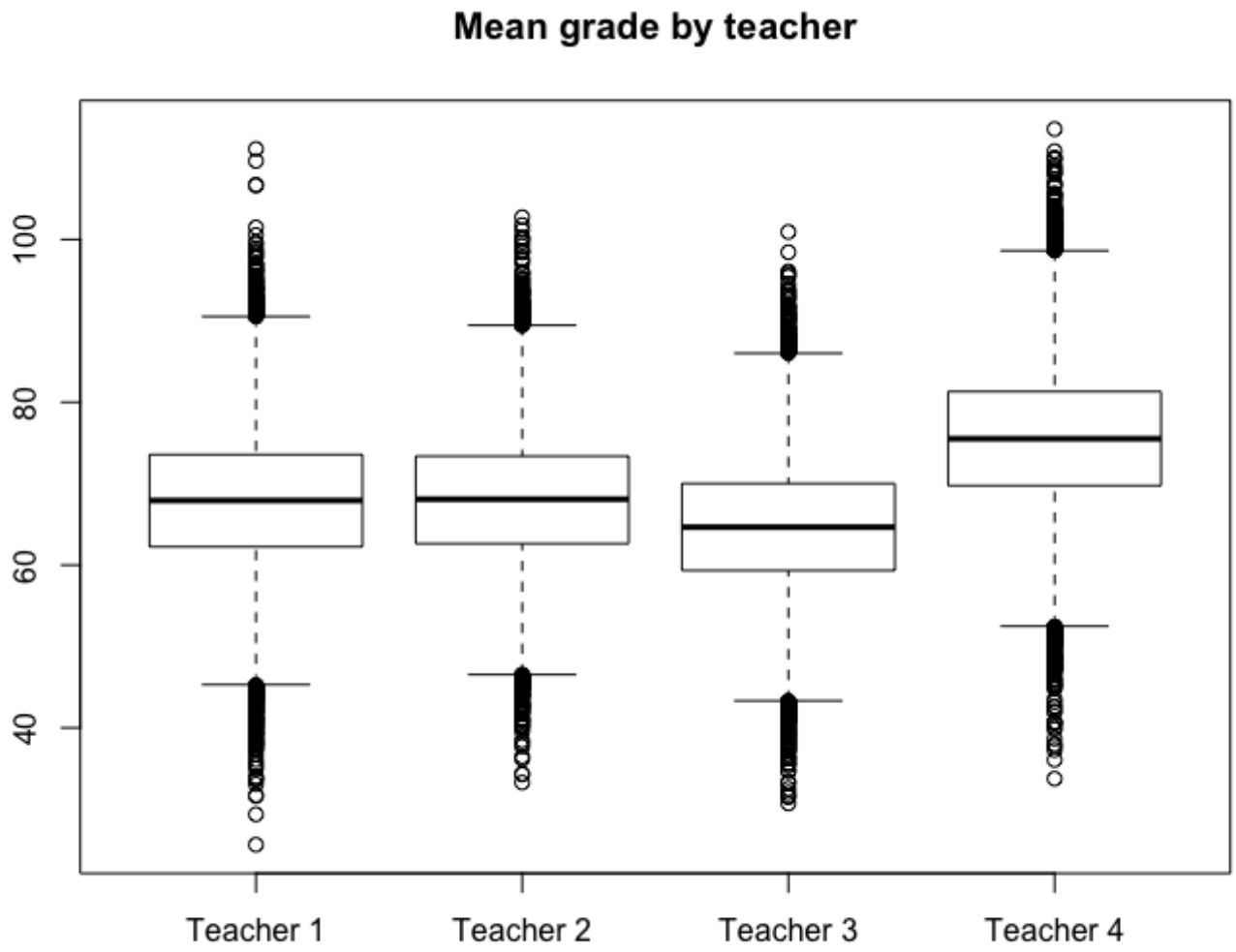
Using the mean of the posteriors as point estimates:

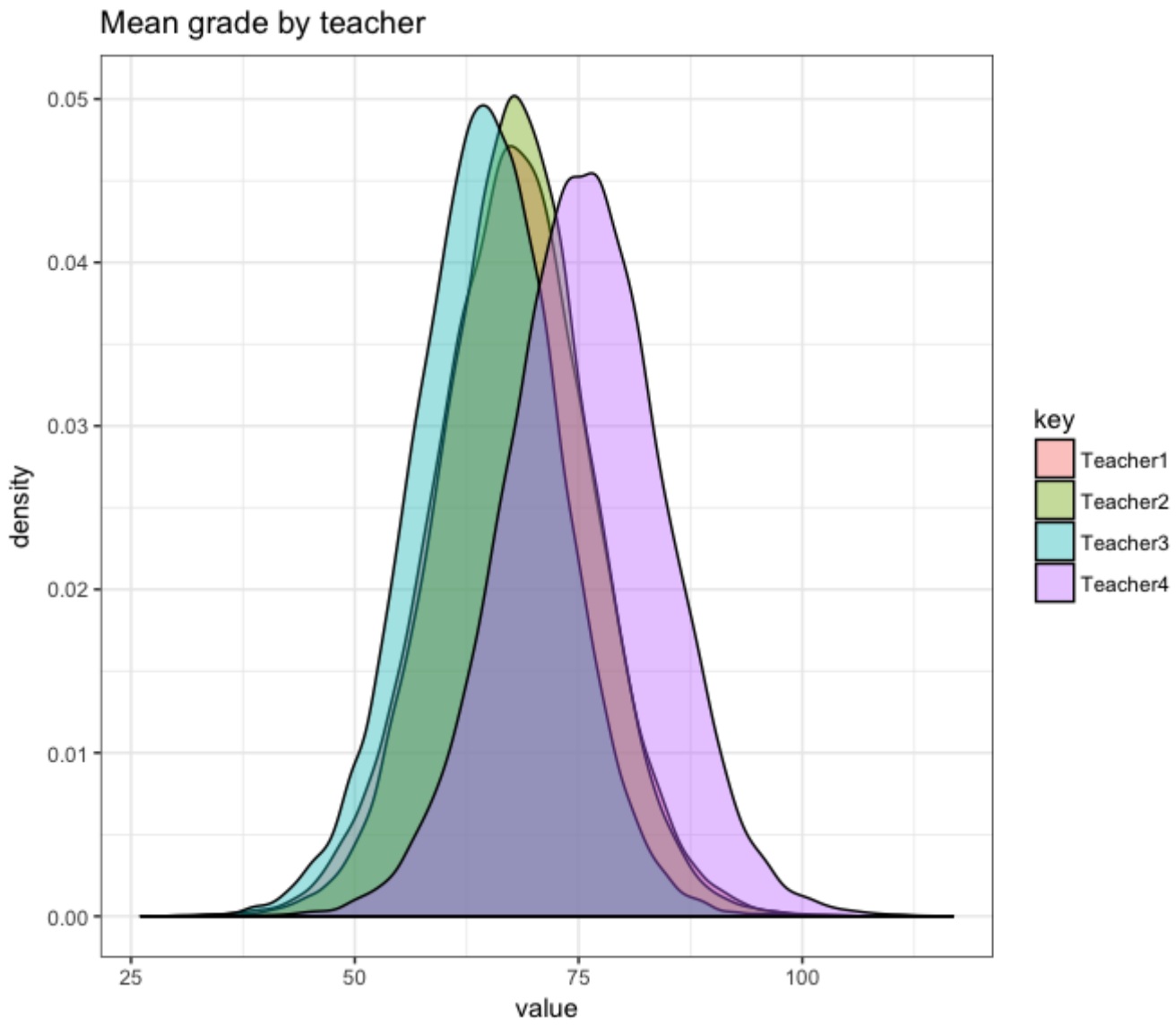
The overall mean of the grades (and the mean grade for the first teacher) is 67.89

The effect of the second teacher on the overall mean grade (or in comparison to the first teacher) is 0.11.

The effect of the third teacher on the overall mean grade is -3.23.

The effect of the fourth teacher on the overall mean grade is 7.64.


Mean grade by teacher
100
80
60
40
Teacher 1
Teacher 2
Teacher 3
Teacher 4


Mean grade by teacher
density
0.05
0.04
0.03
0.02
0.01
0.00
25
50
75
100
value
key
Teacher1
Teacher2
Teacher3
Teacher4

Looking at the posterior distributions of the mean grades for each teacher, there is substantial overlap.

It is not clear that one teacher is clearly better, but if we had to pick one, we would probably pick the fourth teacher.