Bayesian Linear Regression

BIOS 6618

CU Anschutz

Motivating Data and Frequentist MLR

Motivating Example: Lung Function in Children

Study Objective: To describe how lung function develops in children, and how smoking affects development.

Study Design: Cross-sectional survey. A random sample of children ages 3 to 19 from the East Boston area from which 654 had usable data.

Variables Measured: FEV (forced expiratory volume), age, sex, height, current smoking status. (FEV) measures how much air a person can exhale during a forced breath. Higher FEV indicates better lung function.

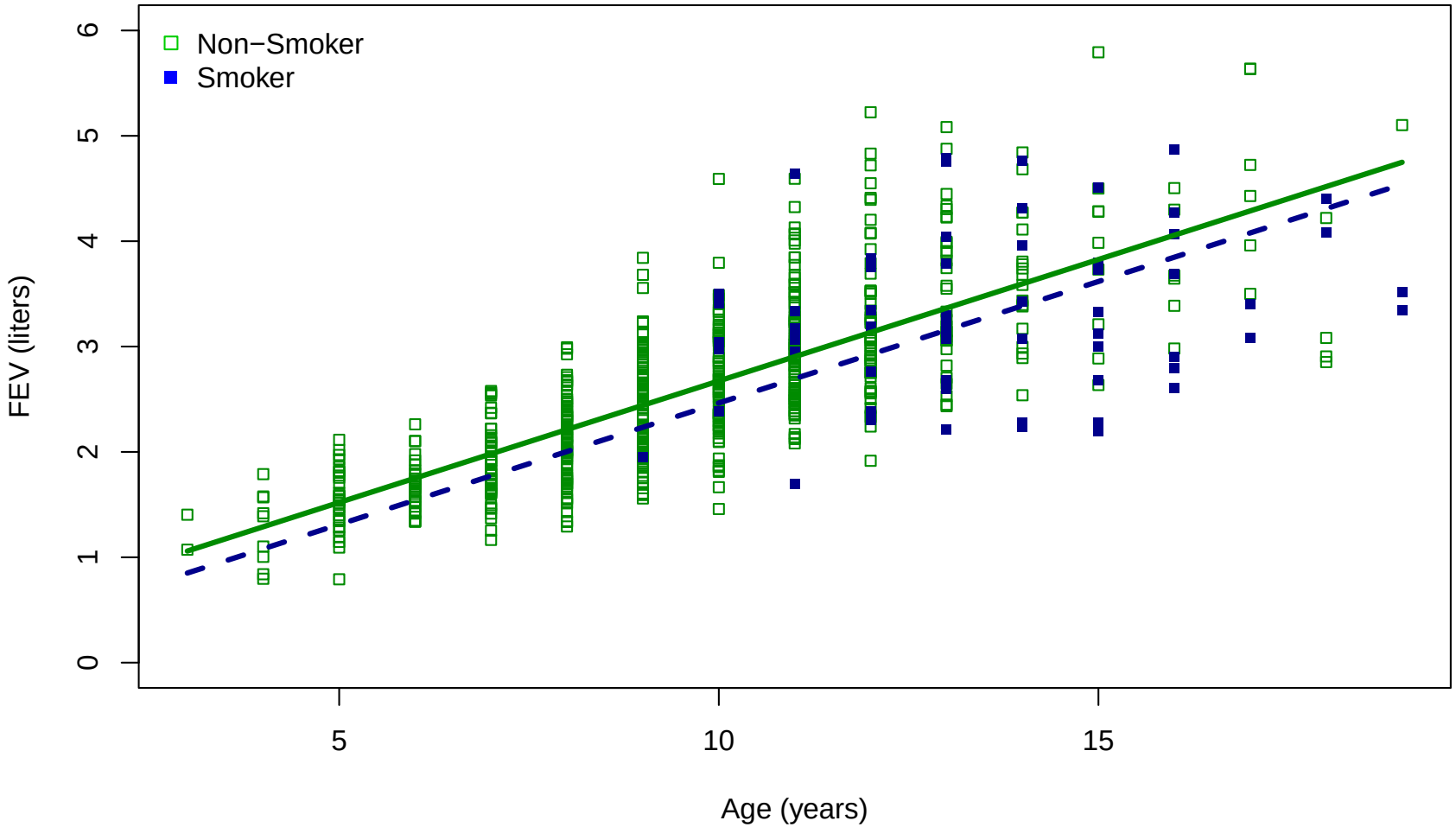
Outcome Variable (Y): FEV

Primary Explanatory Variable (X): age, sex, height, smoking status (depending on the question of interest)

Covariates (C): age, sex, height, smoking status (depending on the question of interest)

Source: Lung function in children (FEV data) [Am J Epidemiology, 110(1): 15-26, 1980.]

Motivating Example Figure



Frequentist MLR

We previously fit a MLR with an outcome of FEV and predictors of age and smoking status, which we will use for comparison to our Bayesian models:

```
fev <- read.csv('FEV_rosner.csv')
mlr <- lm( fev ~ smoke + age, data=fev )
summary(mlr)$coefficients
```

```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)  0.3673730 0.081435716  4.511203  7.647680e-06
## smokesmoker -0.2089949 0.080745337 -2.588321  9.859773e-03
## age          0.2306046 0.008184372 28.176209 8.279537e-115
```

```
confint(mlr)
```

```
##                   2.5 %      97.5 %
## (Intercept)  0.2074647  0.52728140
## smokesmoker -0.3675476 -0.05044215
## age          0.2145336  0.24667553
```

Bayesian Approaches

We will explore three different sets of priors to fit our MLR and walk through how we can use the `brms` package in R to estimate each Bayesian model. The three sets of priors will include:

1. "Non-informative" priors on our beta coefficients of $\beta_{age}, \beta_{smoke} \sim N(0, 1000)$
2. "Informative" priors on our beta coefficients for smoking and age of $\beta_{age} \sim N(0.2, 0.1)$ and $\beta_{smoke} \sim N(-0.33, 0.5)$
3. Poorly specified priors to illustrate why we still need to be thoughtful about our approach of $\beta_{age} \sim N(25, 0.1)$ and $\beta_{smoke} \sim N(-25, 0.1)$

"Non-Informative" Prior Bayesian MLR

Model Fit with `brms`

For our lecture we will use the `brms` package in R to fit and evaluate our Bayesian models. We will first start with our "noninformative" priors of $\beta_{age}, \beta_{smoke} \sim N(0, 1000)$ and assume default priors for everything else (e.g., the intercept, sigma):

```
library(brms) # load package

mod1 <- brm( fev ~ smoke + age, data=fev,
      prior = c(
        set_prior("normal(0,1000)", class="b", coef="age"),
        set_prior("normal(0,1000)", class="b", coef="smokesmoker")),
      seed = 123, # set seed for reproducibility
      chains = 4, # number of chains
      warmup = 1000, # burn-in length to discard from iter
      iter = 2000) # total number of iterations
```

Non-Informative Prior Model Summary

```
summary(mod1)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: fev ~ smoke + age
##    Data: fev (Number of observations: 654)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept       0.37      0.08     0.21     0.53 1.00     3911     2976
## smokesmoker    -0.21      0.08    -0.38    -0.05 1.00     3454     2745
## age             0.23      0.01     0.21     0.25 1.00     3719     2873
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.57      0.02     0.54     0.60 1.00     3878     2844
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

- Estimate is mean from pooled chains
- $\hat{\beta}$ have same interpretation as frequentist MLR
- `l-95% CI` and `u-95% CI` is our equal-tailed credible interval
- R-hat values of 1.00 suggest good convergence (rule of thumb is <1.05)
- `sigma` is our estimated $\sqrt{MSE} = \hat{\sigma}_{Y|X}$

Check Priors Used

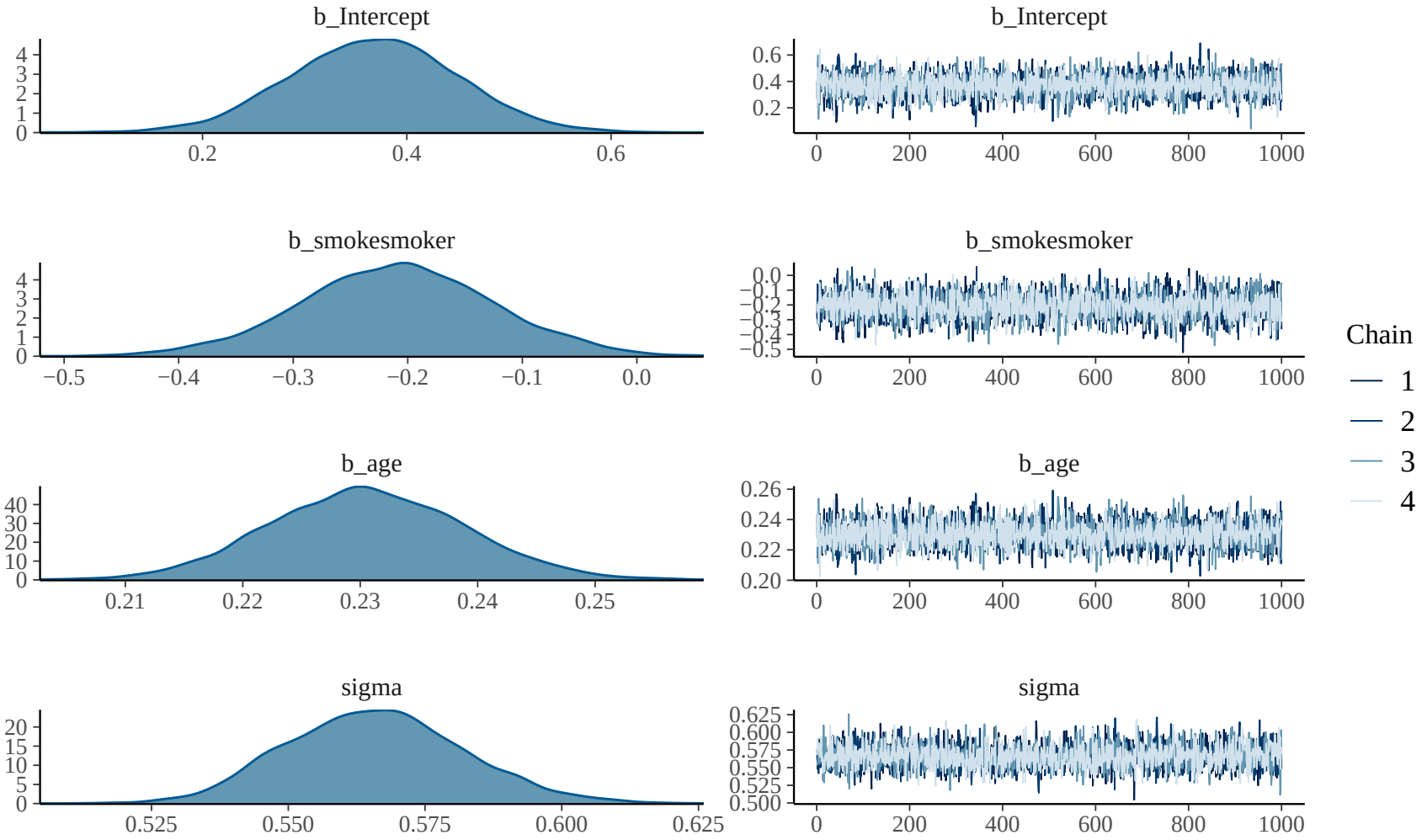
We can also easily check the priors we set (or the defaults used) using `prior_summary()`:

```
priors <- prior_summary(mod1)
priors[,c(1:3,10)] # removed columns not needed for example
```

```
##                   prior     class        coef  source
##                  (flat)         b             default
##          normal(0,1000)         b         age    user
##          normal(0,1000)         b smokesmoker    user
##  student_t(3, 2.5, 2.5) Intercept             default
##    student_t(3, 0, 2.5)     sigma             default
```

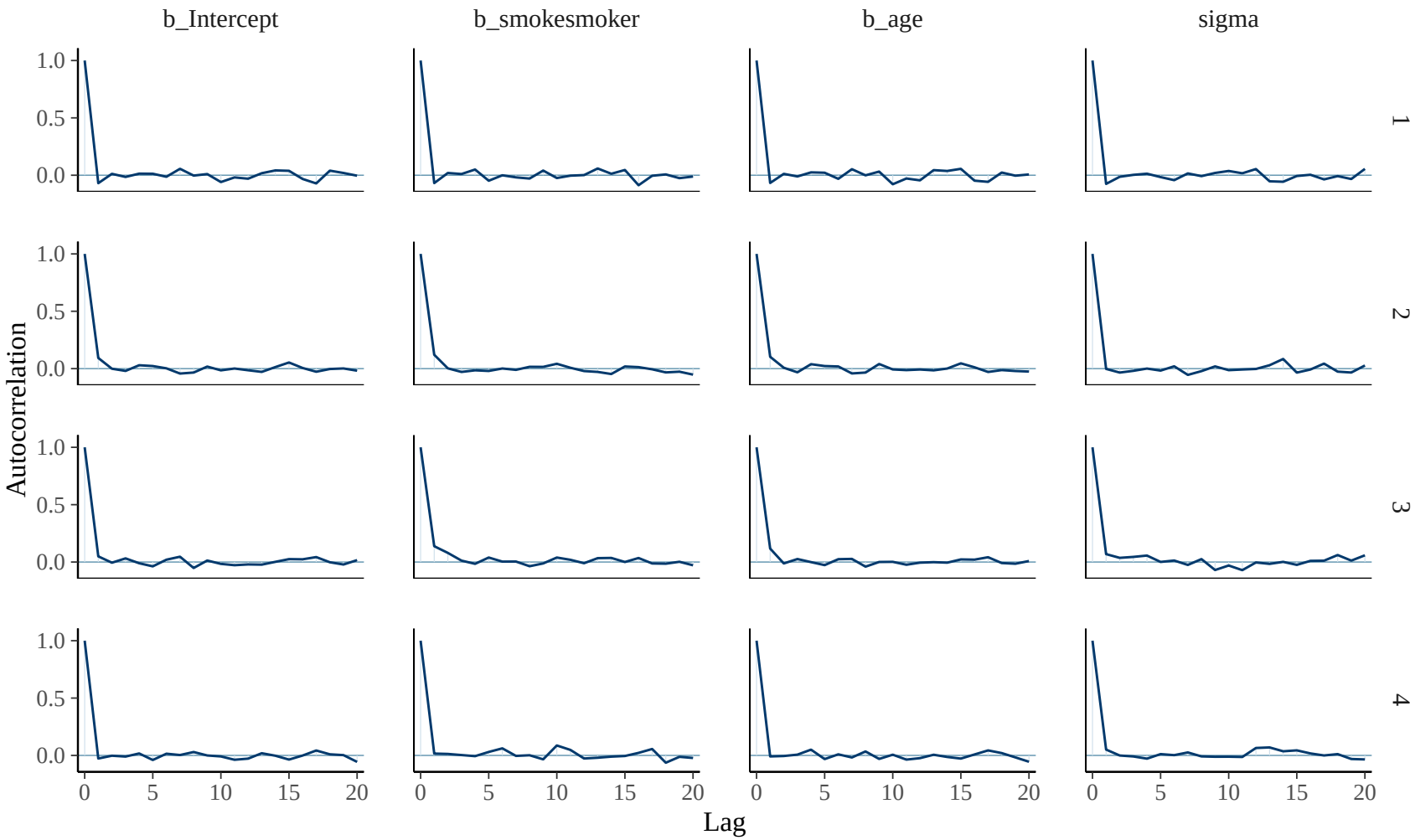
Diagnostic Plots: Density and Trace

```
plot(mod1)
```



Diagnostic Plots: Autocorrelation

```
mcmc_plot(mod1, type="acf")
```



HPD Intervals

While `brms` reports the 95% equal-tailed credible interval by default, we can also use other functions to estimate the highest posterior density (HPD) interval (i.e., the narrowest interval that includes 95% of the posterior):

```
bayestestR::hdi(mod1, ci=0.95)
```

```
## Highest Density Interval
##
## Parameter   |        95% HDI
## ------------------------------
## (Intercept) | [ 0.21,  0.53]
## smokesmoker | [-0.37, -0.04]
## age         | [ 0.21,  0.25]
```

Given our fairly symmetric trace plots, the equal-tailed CrI are similar:

- `(Intercept)`: (0.21, 0.53)
- `smokesmoker`: (-0.38, -0.05)
- `age`: (0.21, 0.25)

Extracting Posterior Chains

We can also extract all the posterior chain iterations:

```
post1 <- as_draws_df(mod1) # extract as data frame
nrow(post1) # check total number of iterations from pooled chains
```

```
## [1] 4000
```

```
post1[c(1:4, 3997:4000),]
```

```
## # A draws_df: 8 iterations, 2 chains, and 6 variables
##   b_Intercept b_smokesmoker b_age sigma lprior lp__
## 1        0.27         -0.283  0.24  0.58    -19 -574
## 2        0.50         -0.070  0.21  0.55    -19 -576
## 3        0.22         -0.327  0.24  0.57    -19 -575
## 4        0.30         -0.075  0.23  0.57    -19 -578
## 5        0.38         -0.325  0.23  0.59    -19 -576
## 6        0.29         -0.182  0.24  0.55    -19 -574
## 7        0.27         -0.269  0.24  0.58    -19 -574
## 8        0.30         -0.313  0.24  0.57    -19 -574
## # ... hidden reserved variables {'.chain', '.iteration', '.draw'}
```

Estimate Posterior Probabilities

Once we've extracted our posterior iterations, it is really easy to estimate any posterior probabilities (PP) we are interested in. For example, let's test the hypothesis $P(\beta_{smoke} > 0)$, $P(\beta_{smoke} < 0)$, and $P(\beta_{smoke} < -0.1)$:

```
mean( post1$b_smokesmoker > 0 ) # P(smoke > 0)
```

```
## [1] 0.00475
```

```
mean( post1$b_smokesmoker < 0 ) # P(smoke < 0)
```

```
## [1] 0.99525
```

```
mean( post1$b_smokesmoker < -0.1 ) # P(smoke < -0.1)
```

```
## [1] 0.90425
```

Estimate Posterior Probabilities

We can also easily calculate the probability that a 15-year old smoker has an average FEV less than 3.82[1] by estimating the FEV at each iteration (e.g., as if we have a regression equation):

```
mean( (post1$b_Intercept + post1$b_smokesmoker + 15*post1$b_age) < 3.82)
```

```
## [1] 0.999
```

[1] This is the estimated mean FEV for a 15-year old non-smoker from our frequentist MLR to use as an example.

Summarize Different Point Estimates

With our posterior iterations, we can also summarize the median instead of the mean:

```
apply(post1, 2, FUN=median)[1:4] # median
```

```
##   b_Intercept b_smokesmoker         b_age         sigma
##     0.3690756    -0.2080753     0.2304503     0.5656746
```

```
apply(post1, 2, FUN=mean)[1:4] # mean (Estimate in brms output)
```

```
##   b_Intercept b_smokesmoker         b_age         sigma
##     0.3676358    -0.2091936     0.2305654     0.5659766
```

"Informative" Prior Bayesian MLR

Model Fit with `brms`

Let's next examine "informative" priors of $\beta_{age} \sim N(0.2, 0.1)$ and $\beta_{smoke} \sim N(-0.33, 0.5)$ and assume default priors for everything else (e.g., the intercept, sigma):

```
mod2 <- brm( fev ~ smoke + age, data=fev,
      prior = c(
        set_prior("normal(0.2,0.1)", class="b", coef="age"),
        set_prior("normal(-0.33,0.5)", class="b", coef="smokesmoker")),
      seed = 123, # set seed for reproducibility
      chains = 4, # number of chains
      warmup = 1000, # burn-in length to discard from iter
      iter = 2000) # total number of iterations
```

Informative Prior Model Summary

```
summary(mod2)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: fev ~ smoke + age
##    Data: fev (Number of observations: 654)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept       0.37      0.08     0.21     0.53 1.00     3797     3359
## smokesmoker    -0.21      0.08    -0.37    -0.05 1.00     3559     3110
## age             0.23      0.01     0.21     0.25 1.00     3571     3387
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.57      0.02     0.54     0.60 1.00     3758     2993
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

- Results similar to "noninformative" priors
- In this case, not a major difference in results

Nonsensical, Poorly Specified Priors Bayesian MLR

Model Fit with `brms`

Let's next examine strong priors that have more extreme (and nonsensical) effects and small standard deviations of $\beta_{age} \sim N(25, 0.1)$ and $\beta_{smoke} \sim N(-25, 0.1)$ and assume default priors for everything else (e.g., the intercept, sigma):

```
mod3 <- brm( fev ~ smoke + age, data=fev,
      prior = c(
        set_prior("normal(25,0.1)", class="b", coef="age"),
        set_prior("normal(-25,0.1)", class="b", coef="smokesmoker")),
      seed = 123, # set seed for reproducibility
      chains = 4, # number of chains
      warmup = 1000, # burn-in length to discard from iter
      iter = 2000) # total number of iterations
```

Poorly Specified Prior Model Summary

```
summary(mod3)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: fev ~ smoke + age
##    Data: fev (Number of observations: 654)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -240.49      2.15  -244.77  -236.15 1.00     3999     2826
## smokesmoker   -25.00      0.10   -25.20   -24.81 1.00     4539     3453
## age            24.72      0.10    24.53    24.92 1.00     4435     2852
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma    69.56      1.96    65.88    73.44 1.00     4289     3242
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

- Results *very* different from our prior models
- Priors have overwhelmed our observed data and posterior estimates are more reflective of the prior
- Since priors include effects that are not plausible for our FEV outcome, we would definitely want to consider different models

Closing Summary

Bayesian approaches to linear regression give us the flexibility to incorporate prior information and to calculate summaries (e.g., posterior probabilities and credible intervals) with more intuitive interpretations.

Other Bayesian languages and packages exist, and can be used to fit regression models. Feel free to explore different options!

In practice, we should consider multiple (reasonably specified) priors. In this example, our final poorly specified example illustrates the dangers that can happen if we are not careful.