More Bayesian Linear Regression

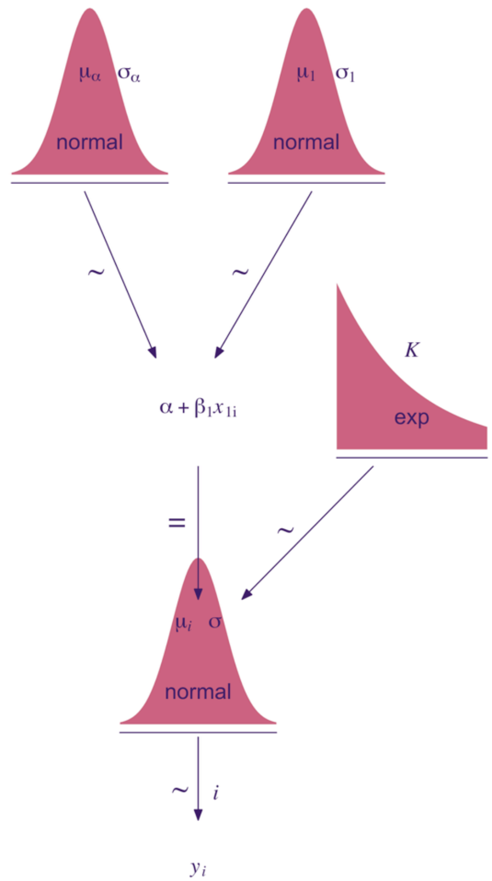
Bayesian Data Analysis

Steve Buyske

Bayesian Framework for the Linear Model

- Remember our model for linear regression in the Bayesian setting:
- Model:
  - $y_i \sim \text{Normal}(\mu_i, \sigma)$
  - $\mu_i = \alpha + \beta x_i$ *this is sometimes called the link*
  - together, *these will give us the likelihood*.
- The priors:
  - $\alpha \sim \text{Normal}(\text{something}, \text{something})$,
  - $\beta \sim \text{Normal}(0, \text{something})$,
  - $\sigma \sim \text{Exponential}(\text{something})$.
  - (These are not the only possible priors, just a common set.)

Kruschke Diagram



Slide added after I made the video

- In the Kruschke diagram, it is worth noting something about the distributions of the parameters.
- As the diagram indicates, in the Bayesian framework we consider that all parameters have distributions.
- Before we get data, we label the distributions “prior distributions”, since they are prior to our evidence.
  - The diagram still holds—it represents a model of the data, regardless of the specifics of the distributions.
- After we get data, we update the distributions of the parameters which we now label “posterior distributions”, since they are following our updates with the new evidence.
  - Again, the diagram still holds—it still represents a model of the data, but now we have updated the specifics of the distributions.
- By “specifics of the distributions,” I mean the value of K, μ_α, σ_α, and μ_1, σ_1.

Quick summary of our model

- Remember we fit a model `gini_stan` with `stan_glm(life_expectancy ~ gini, data = mini_gapminder)`.
- We can get estimates from the posterior of the median, mean, and mode a posteriori with

```
point_estimate(gini_stan)
```

```
## # Point Estimates
##
## # Fixed Effects (Conditional Model)
##
## Parameter   | Median |  Mean |   MAP
## ------------------------------------
## (Intercept) |  87.99 | 87.96 | 88.43
## gini        |  -0.39 | -0.39 | -0.39
##
## # Sigma (fixed effects)
##
## Parameter | Median | Mean |  MAP
## --------------------------------
## sigma     |   6.57 | 6.58 | 6.56
```

- We can get the highest density interval version of a 90% credible interval with

```
hdi(gini_stan, ci = .9)
```

```
## # Highest Density Interval
##
## # Fixed Effects (Conditional Model)
##
## Parameter   |        90% HDI
## ----------------------------
## (Intercept) | [83.62, 92.20]
## gini        | [-0.49, -0.27]
##
## # Sigma (fixed effects)
##
## Parameter |        90% HDI
## --------------------------
## sigma     | [ 5.99,  7.14]
```

- There is also a nice function `model_parameters()` that summarizes the model parameters.
  - "CI" refers to credible interval (the HDI by default).
  - We will talk about pd, ROPE, Rhat, and ESS in the future.

```
model_parameters(gini_stan, ci = .9)
```

```
## # Fixed effects
##
## Parameter   | Median |          90% CI |   pd | % in ROPE |  Rhat |     ESS |                   Prior
## ----------------------------------------------------------------------------------------------------
## (Intercept) |  87.99 | [83.62, 92.20]  | 100% |        0% | 1.001 | 3894.82 | Normal (72.92 +- 17.93)
## gini        |  -0.39 | [-0.49, -0.27]  | 100% |      100% | 1.001 | 3981.03 |   Normal (0.00 +- 2.34)
##
## # Fixed effects sigma
##
## Parameter | Median |       90% CI |   pd | % in ROPE |  Rhat |     ESS |     Prior
## ----------------------------------------------------------------------------------
## sigma     |   6.57 | [5.99, 7.14] | 100% |        0% | 0.999 | 3859.26 | NA ( +- )
```

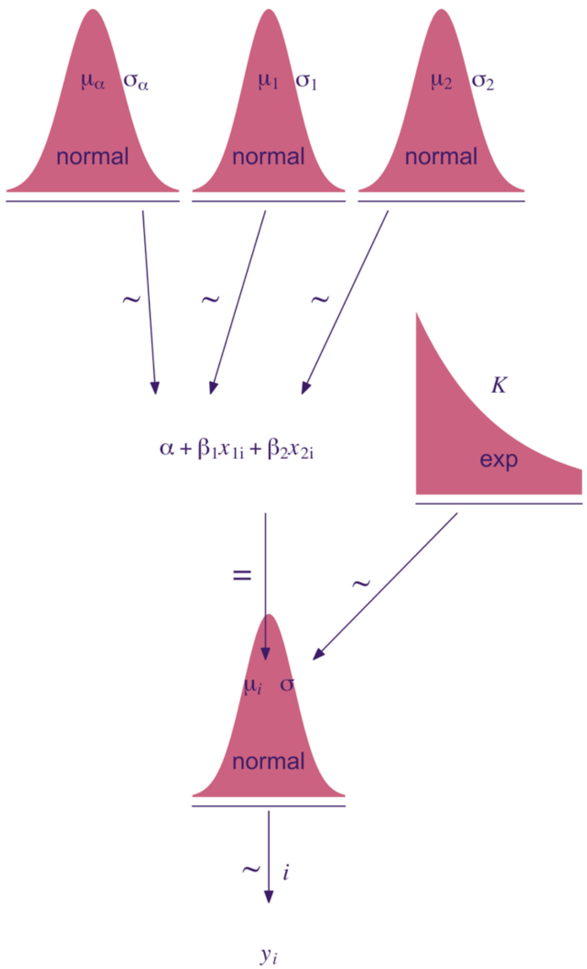
- (The `model_parameters()` function is part of the `parameters` package, something that is not installed in the Introduction workspace, so don't try using it yourself there.)

Bayesian Multiple Linear Regression

- To fit include multiple predictor variables (with no interaction), just using the `+` sign on the right hand side of the formula. For example, let's regress `life_expectancy` on both `gini` and `log10_gdp_per_capita`.

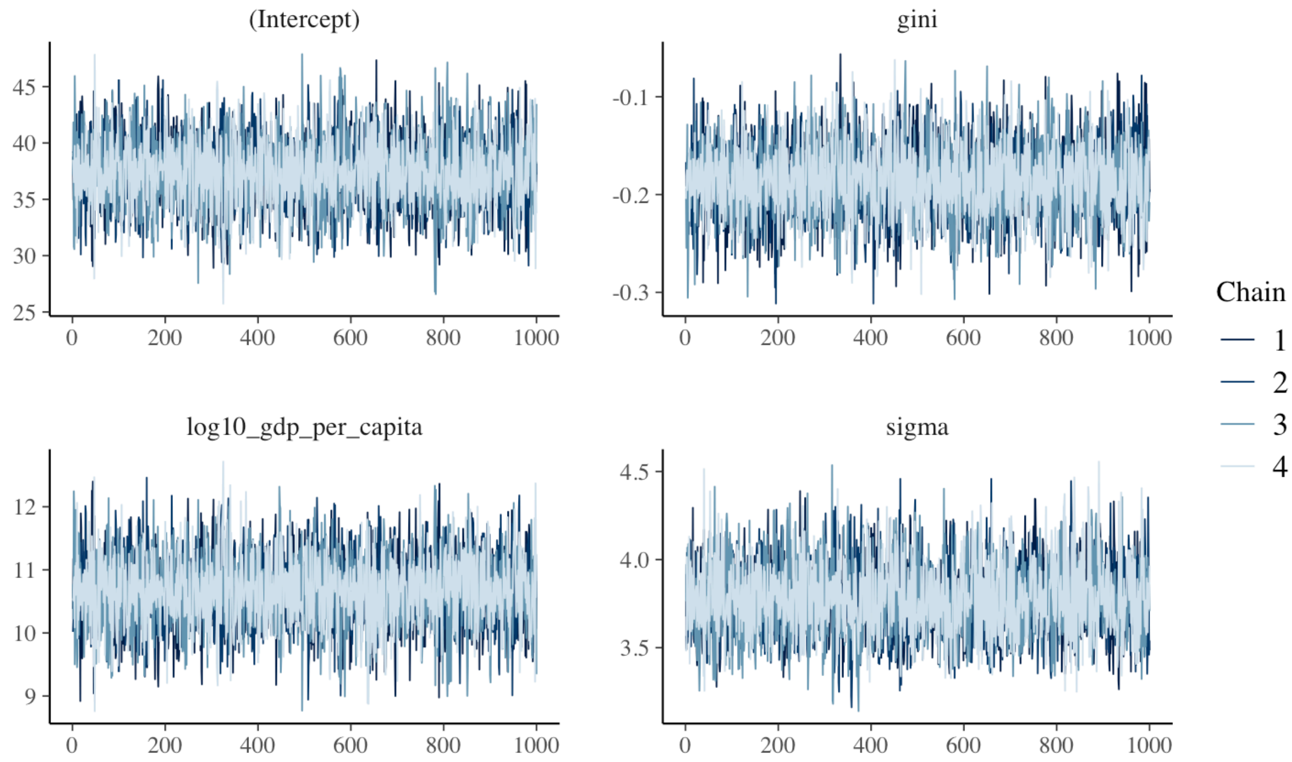
```
gini_stan_multiple <-
  stan_glm(life_expectancy ~ gini + log10_gdp_per_capita, data = mini_gapminder)
```

Graphical representation of our model



The trace plots look fine

```
plot(gini_stan_multiple, plotfun = "trace")
```



- And here are point estimates …

```
point_estimate(gini_stan_multiple)
```

```
## # Point Estimates
##
## # Fixed Effects (Conditional Model)
##
## Parameter            | Median |  Mean |   MAP
## ---------------------------------------------
## (Intercept)          |  37.40 | 37.38 | 37.45
## gini                 |  -0.19 | -0.19 | -0.19
## log10_gdp_per_capita |  10.67 | 10.67 | 10.65
##
## # Sigma (fixed effects)
##
## Parameter | Median | Mean |  MAP
## --------------------------------
## sigma     |   3.77 | 3.77 | 3.77
```

- … and credible intervals.

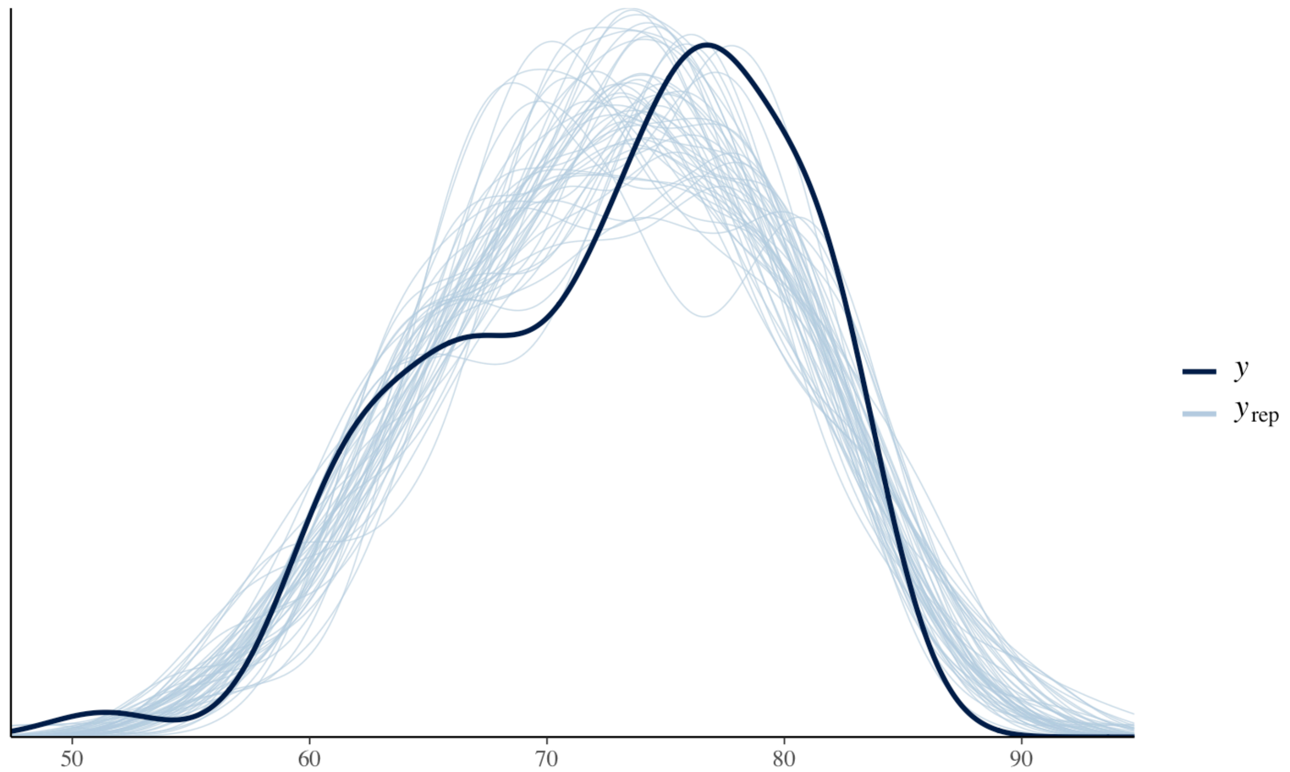
```
hdi(gini_stan_multiple, ci = .9)
```

```
## # Highest Density Interval
##
## # Fixed Effects (Conditional Model)
##
## Parameter            |       90% HDI
## -------------------------------------
## (Intercept)          | [32.33, 42.53]
## gini                 | [-0.25, -0.13]
## log10_gdp_per_capita | [ 9.74, 11.61]
##
## # Sigma (fixed effects)
##
## Parameter |       90% HDI
## --------------------------
## sigma     | [ 3.45,  4.12]
```

Posterior Predictive Check

- The posterior predictive check is still not good, unfortunately.

```
pp_check(gini_stan_multiple)
```



Interaction Terms

- Sometimes with several predictor variables, we are interested in their interaction.
- We will illustrate it by adding a new variable to our data set, an indicator for the gini index is 40.
  - The code below shows how you can add a variable to a data frame

```
mini_gapminder <-
  mini_gapminder %>%
  mutate(high_gini = (gini > 40))
```

```
## Rows: 177
## Columns: 3
## $ country   <chr> "Albania", "Algeria", "Angola", "Antigua and Barbuda", "Arg…
## $ gini      <dbl> 29.0, 27.6, 42.6, 40.0, 42.4, 32.6, 32.3, 30.5, 32.4, 43.7,…
## $ high_gini <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE…
```

Interaction Terms cont.

- We could fit a model like
  - `life_expectancy ~ log10_gdp_per_capita + high_gini`
  - This would have a coefficient for `log10_gdp_per_capita` and a coefficient for `high_gini`.
    - Since `log10_gdp_per_capita` is continuous, the coefficient represents a slope
    - Since `high_gini` is binary (think of it as 0 or 1), the coefficient represents the shift in `life_expectancy` for a country with high gini index as opposed to low.

R code for interactions

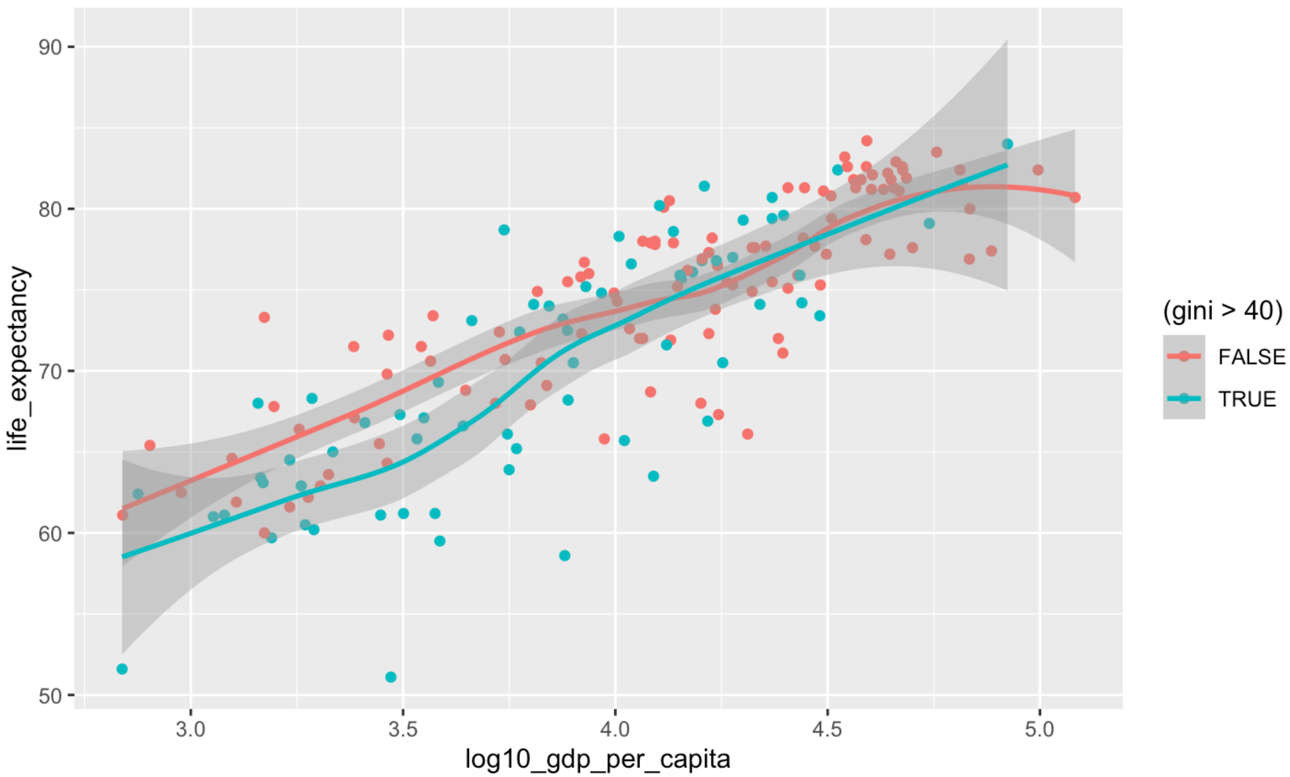
- But what if we want to model that the slope for `log10_gdp_per_capita` is different for low gini index countries than for high?
  - We would want to add an interaction term.
  - The R code would be
    - `life_expectancy ~ log10_gdp_per_capita + high_gini + log10_gdp_per_capita : high_gini`, or more compactly,
    - `life_expectancy ~ log10_gdp_per_capita * high_gini`.
    - One way to write the frequentist model for this would be

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

$$\text{where } \epsilon_i \sim \text{Normal}(0, \sigma) \text{ and independent.}$$

R code for interactions cont.

- The plot indicates that any interaction effect is fairly weak

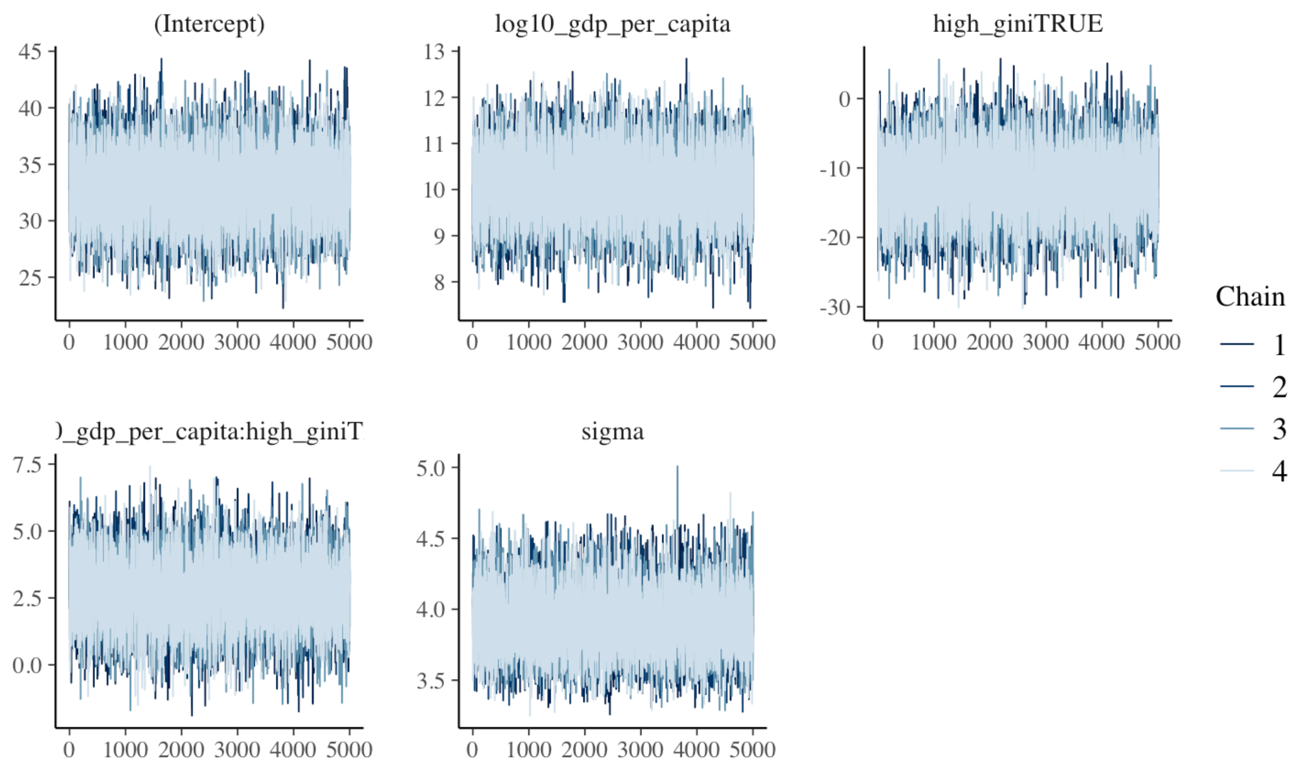


- but let’s fit a model.

```
gini_interaction <-
  stan_glm(life_expectancy ~ log10_gdp_per_capita * high_gini, data = mini_gapminder)
```

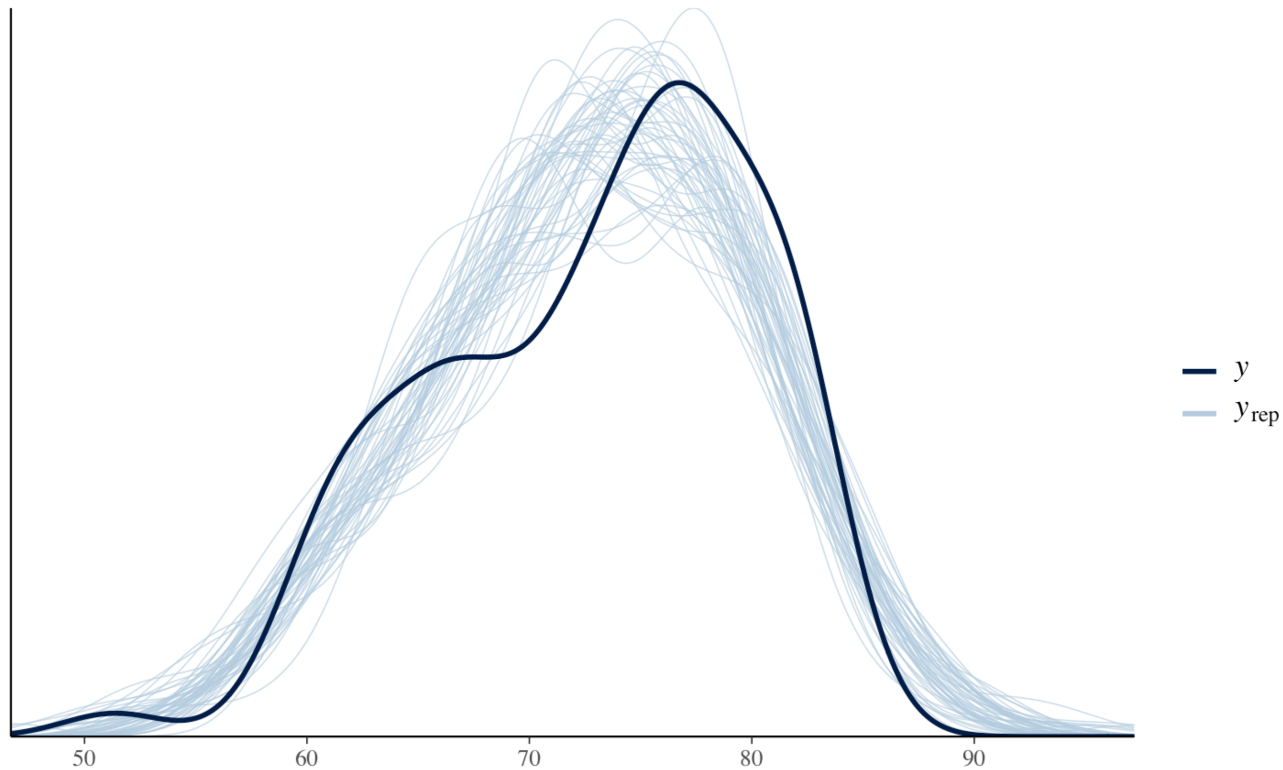
- The trace plots look fine

```
plot(gini_interaction, plotfun = "trace")
```



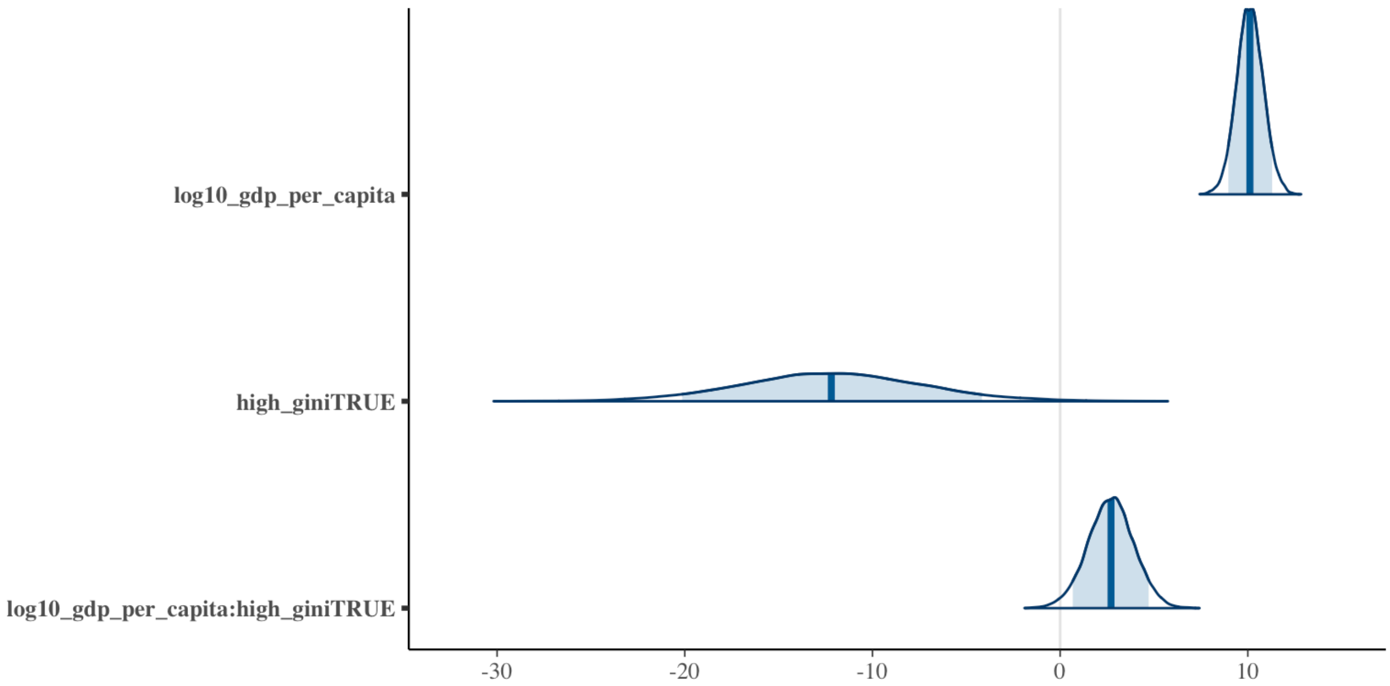
- The posterior predictive plot still has a problem.
    - We will continue to ignore that for now.

```
pp_check(gini_interaction)
```



- And here are the posteriors with 90% equal tails credible intervals shown.

```
plot(gini_interaction,
     plotfun = "areas",
     pars = c("log10_gdp_per_capita", "high_giniTRUE", "log10_gdp_per_capita:high_giniTRUE"),
     prob = .9
  )
```



```
point_estimate(gini_interaction)
```

```
## # Point Estimates
##
## # Fixed Effects (Conditional Model)
##
## Parameter                          | Median |   Mean |    MAP
## --------------------------------------------------------------
## (Intercept)                        |  33.08 |  33.06 |  33.34
## log10_gdp_per_capita               |  10.11 |  10.12 |  10.19
## high_giniTRUE                      | -12.20 | -12.17 | -11.96
## log10_gdp_per_capita:high_giniTRUE |   2.71 |   2.70 |   2.89
##
## # Sigma (fixed effects)
##
## Parameter | Median | Mean |  MAP
## -------------------------------
## sigma     |   3.90 | 3.91 | 3.89
```

Interpretation

```
## # Point Estimates
##
## Parameter                               |   Mean
## --------------------------------------------------
## (Intercept)                             |  33.06
## log10_gdp_per_capita                    |  10.12
## high_giniTRUE                           | -12.17
## log10_gdp_per_capita:high_giniTRUE      |   2.70
## sigma                                   |   3.91
```

- The median log10_gdp_per_capita is a little over 4—a value of 4 is $10^4 = \$10,000$.
- In a low gini index country with a `log10_gdp_per_capita` of 4, we would predict a life expectency of (using the Mean column)
  - 33.06 + 10.12 * 4 = 73.54.
- If the gini index is high, however, then both the intercept and slope are different. With `log10_gdp_per_capita` equal to 4 again, we would predict a life expectency of of
  - 33.06 - 12.17 + (10.12 + 2.70) * 4 = 72.17.

Interpretation cont.

```
## # Point Estimates
##
## Parameter                                |   Mean
## -------------------------------------------------
## (Intercept)                              |  33.06
## log10_gdp_per_capita                     |  10.12
## high_giniTRUE                            | -12.17
## log10_gdp_per_capita:high_giniTRUE       |   2.70
## sigma                                    |   3.91
```

- more formally in the first case we would have
  - 33.06 + 10.12 * 4 + -12.17 * 0 + 2.70 * 4 * 0 = 73.54
- and for the second case we would have
  - 33.06 + 10.12 * 4 + -12.17 * 1 + 2.70 * 4 * 1 = 72.17

Help with the concept of interactions

- If interactions are new to you and giving you a headache, you might want to look at
  - https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/mlr-with-interactions.html
  - https://moderndive.com/6-multiple-regression.html#model4 or the video at
  - https://youtu.be/ScKL40dp8M4
- Of course, you should also post questions on Piazza.