Lecture 11/16/2023 (in class)

Alex Kale

2023-11-16

Load student absences data

```
df = read_csv("../data/students.csv", show_col_types = FALSE)
head(df)
```

```
## # A tibble: 6 x 10
##     age address travel_time study_time failures internet absences g_edu g_job
##   <dbl> <chr>         <dbl>      <dbl>    <dbl> <chr>       <dbl> <dbl> <chr>
## 1    18 urban         27.2        3.03        0 no              6     4 at_home
## 2    17 urban         11.0        4.15        0 yes             4     1 other
## 3    15 urban          6.57       2.02        3 yes            10     1 at_home
## 4    15 urban          9.98       6.47        0 yes             2     4 health
## 5    16 urban         12.0        4.32        0 no              4     3 other
## 6    16 urban         14.3        3.11        0 yes            10     4 services
## # i 1 more variable: alcohol <dbl>
```

Preprocessing

```
model_df = df |>
  mutate(
    # factors
    address = as.factor(address),
    failures = as.factor(failures),
    internet = as.factor(internet),
    g_edu = as.factor(g_edu),
    g_job = as.factor(g_job),
    # centered continuous predictors
    c_age = age - mean(age),
    c_tt = travel_time - mean(travel_time),
    c_st = study_time - mean(study_time),
    c_alc = alcohol - mean(alcohol)
  )
```

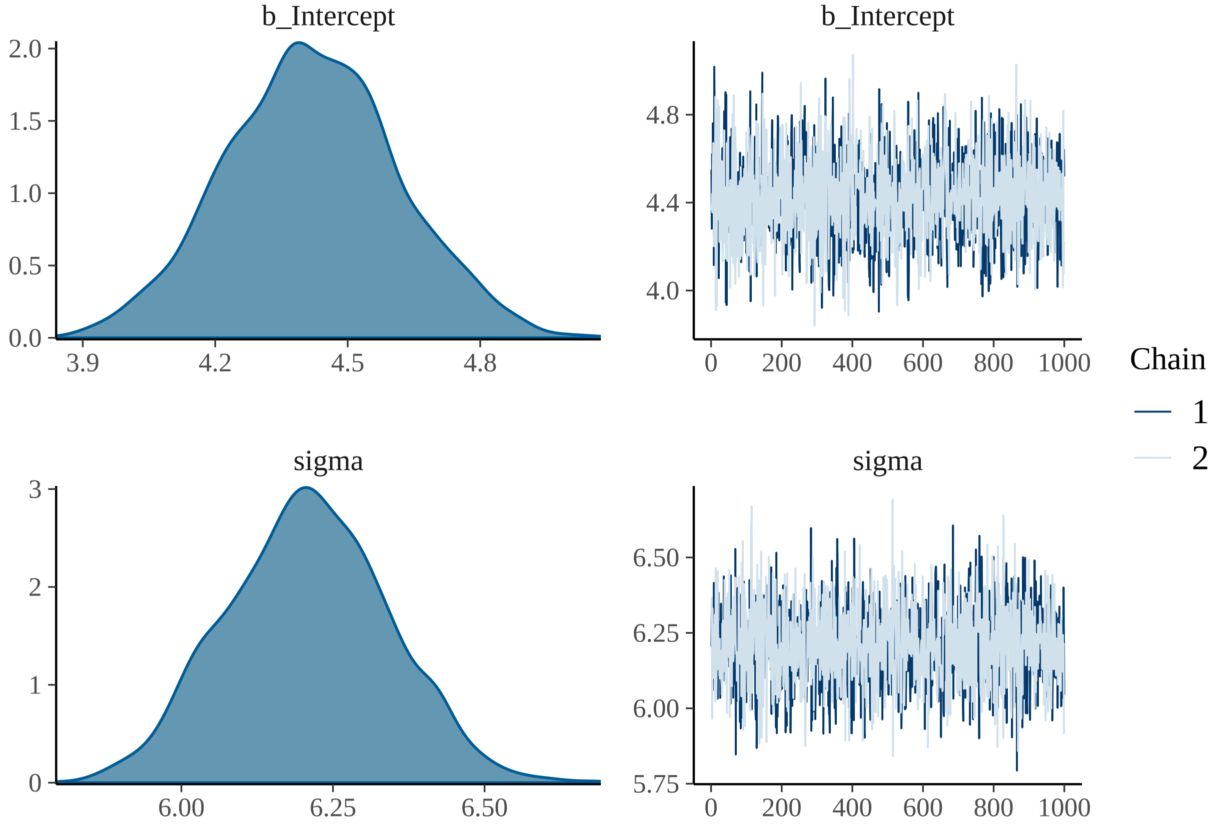
Find a model family

```
m_norm = brm(
  bf("absences ~ 1"),
  family = "normal",
  data = model_df,
  iter = 2000, warmup = 1000, chains = 2,
  file = "m0.rds"
)
```

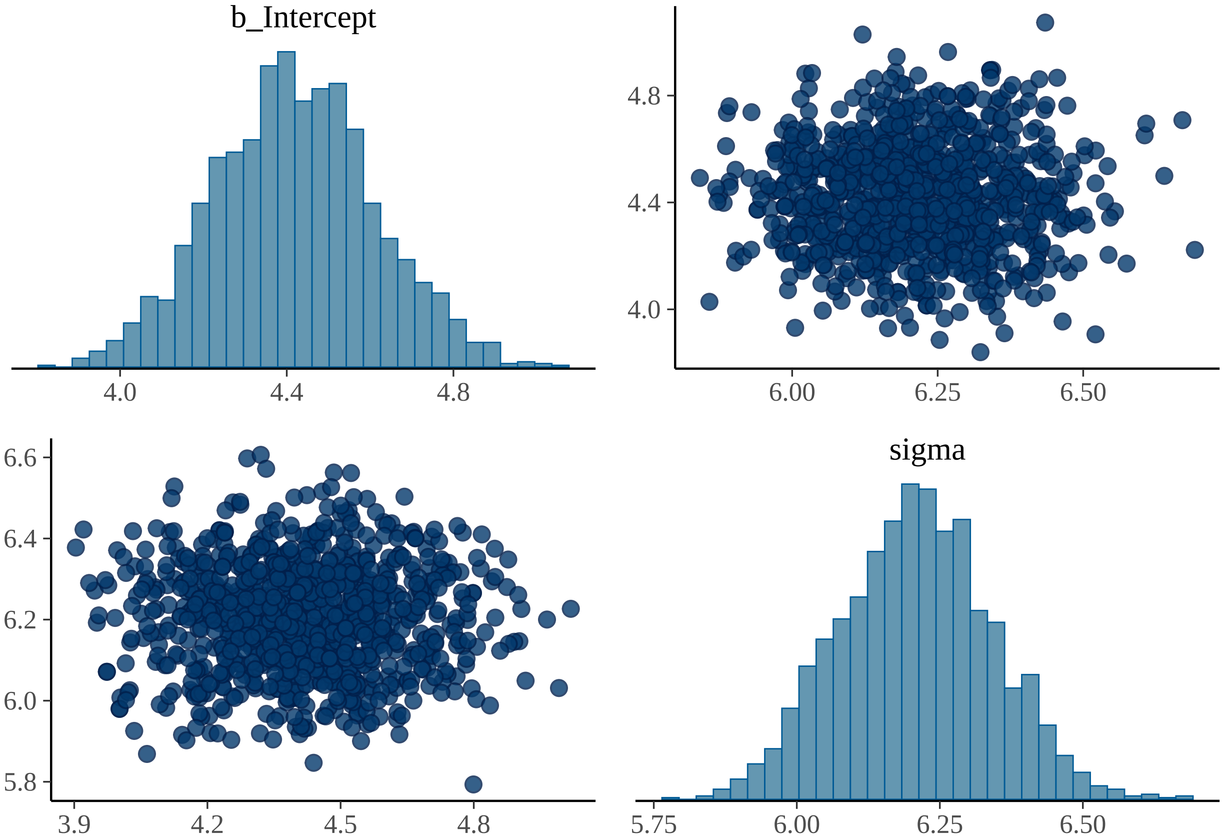
```
summary(m_norm)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: absences ~ 1
##    Data: model_df (Number of observations: 1044)
##   Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 2000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     4.42      0.19     4.04     4.81 1.00     2071     1232
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     6.21      0.13     5.96     6.47 1.00     1890     1476
##
## 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).
```

```
plot(m_norm)
```

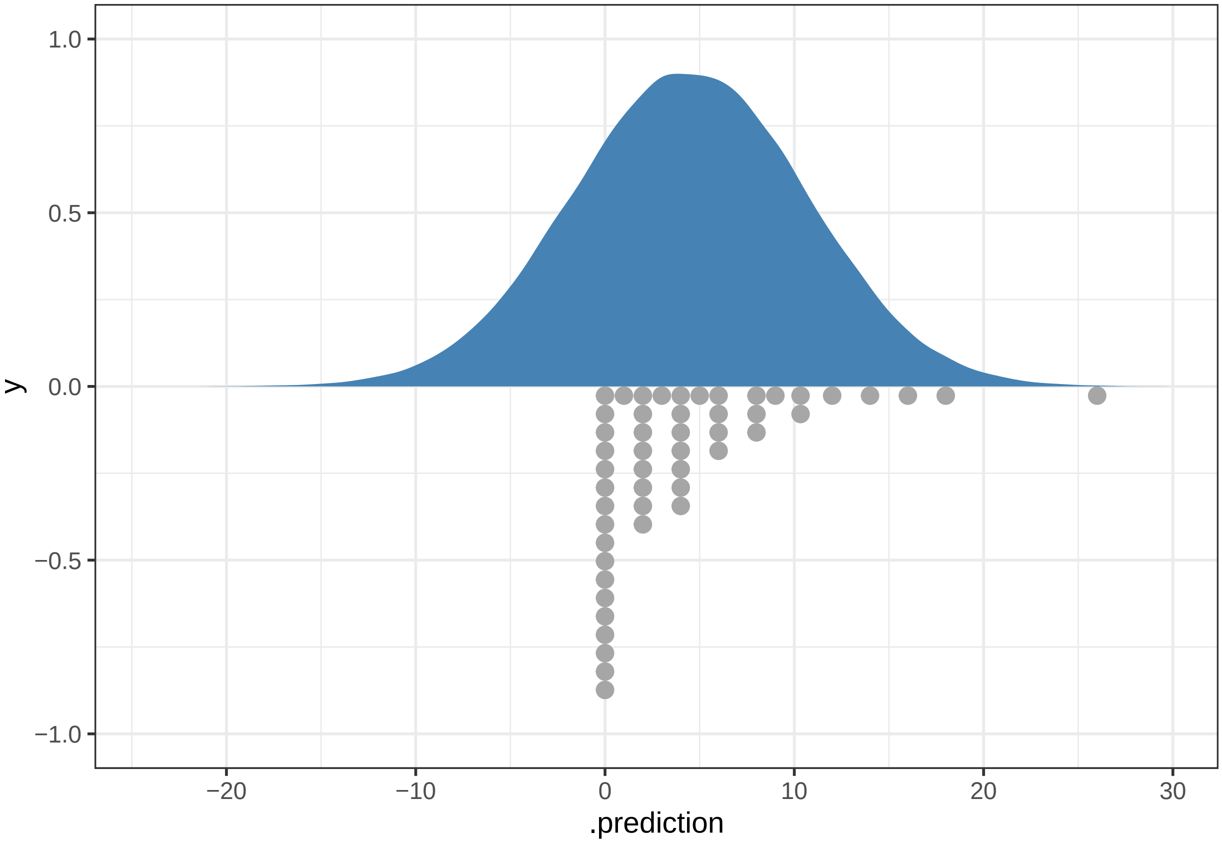


```
pairs(m_norm)
```



PP check

```r
model_df |>
  select(absences) |>
  add_predicted_draws(m_norm, ndraws = 100) |>
  ggplot(aes(x = .prediction)) +
  stat_slab(fill = "steelblue") +
  stat_dots(aes(x = absences), quantiles = 50, side = "bottom", data = model_df) +
  theme_bw()
```



Now try lognormal.

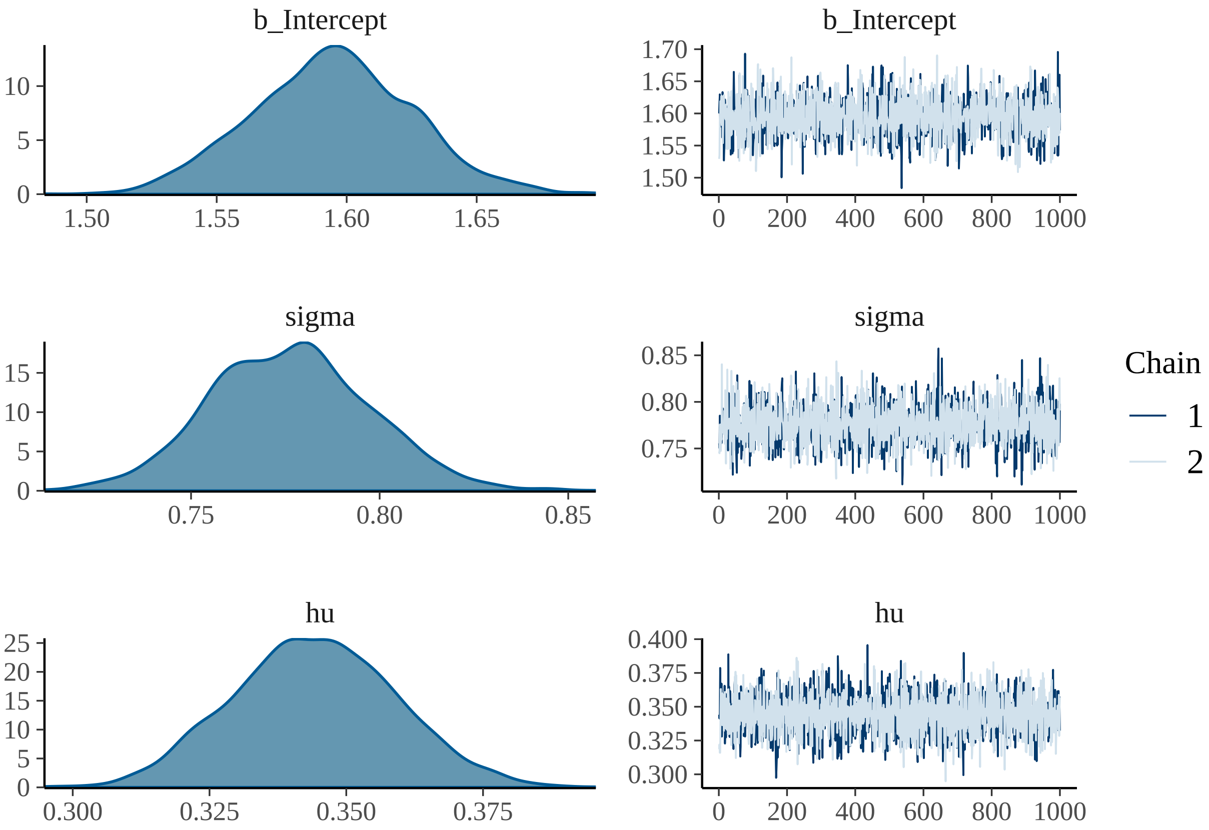
```
m_hlogn = brm(
  bf("absences ~ 1"),
  family = hurdle_lognormal(),
  data = model_df,
  iter = 2000, warmup = 1000, chains = 2,
  file = "m1.rds"
)
```

```
summary(m_hlogn)
```

```
##  Family: hurdle_lognormal
##   Links: mu = identity; sigma = identity; hu = identity
## Formula: absences ~ 1
##    Data: model_df (Number of observations: 1044)
##   Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 2000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     1.59      0.03     1.53     1.66 1.00     1832     1211
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.78      0.02     0.74     0.82 1.00     2490     1518
## hu        0.34      0.01     0.32     0.37 1.00     2246     1398
##
```

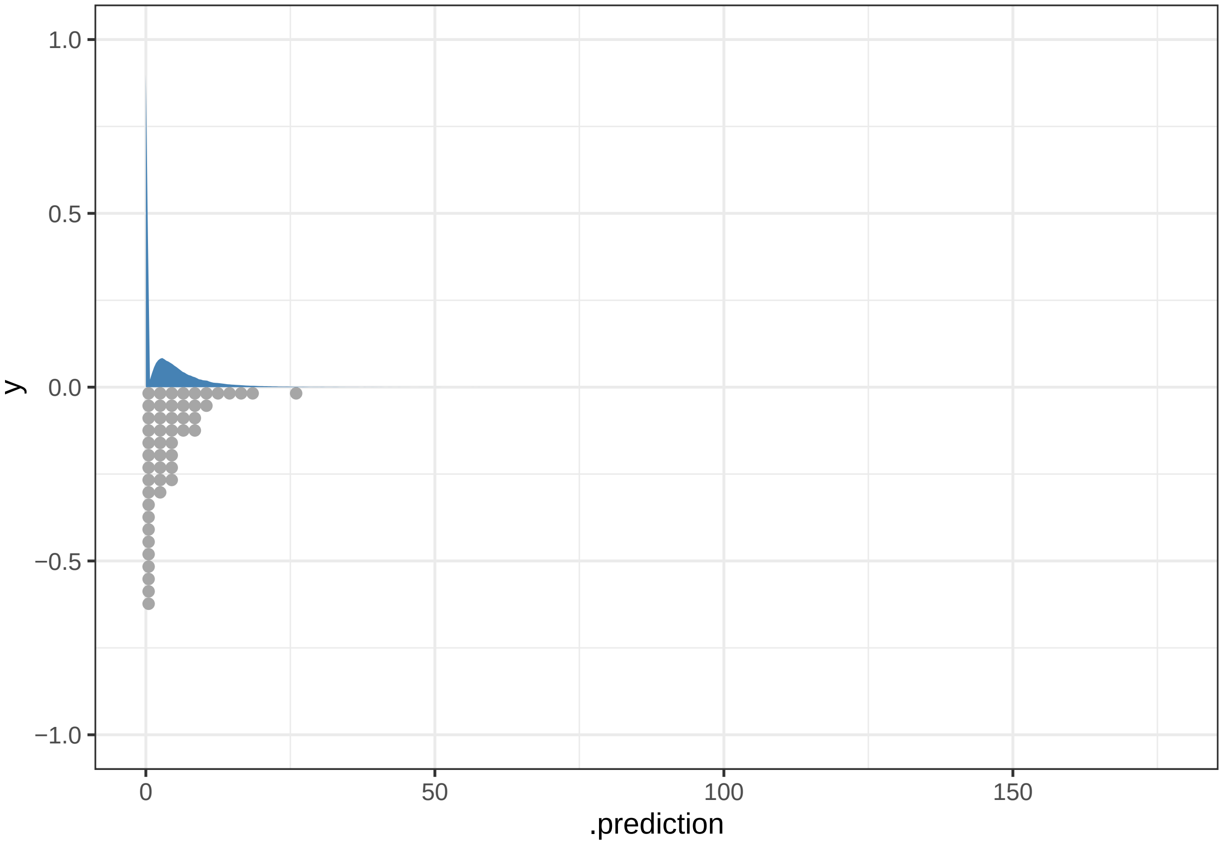
```
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
plot(m_hlogn)
```



PP check

```
model_df |>
  select(absences) |>
  add_predicted_draws(m_hlogn, ndraws = 100) |>
  ggplot(aes(x = .prediction)) +
  stat_slab(fill = "steelblue") +
  stat_dots(aes(x = absences), quantiles = 50, side = "bottom", data = model_df) +
  theme_bw()
```



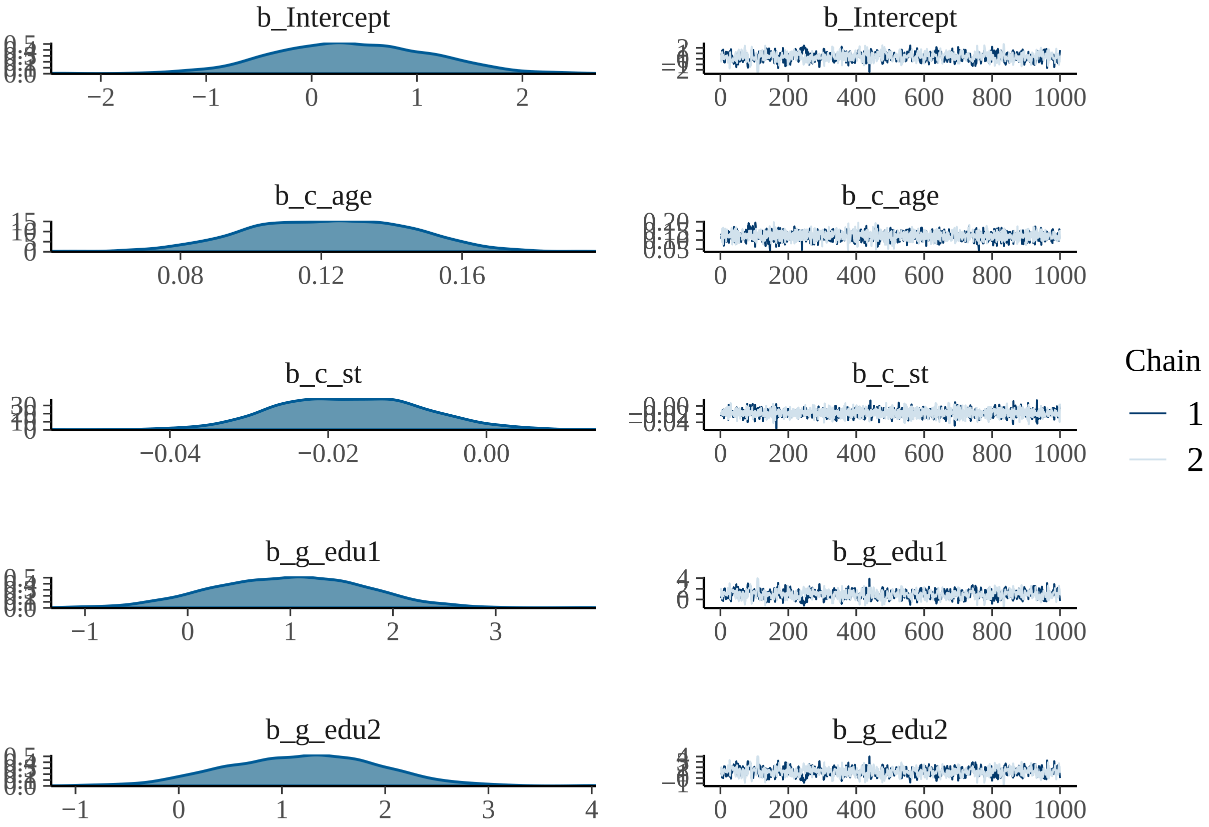
Add predictors

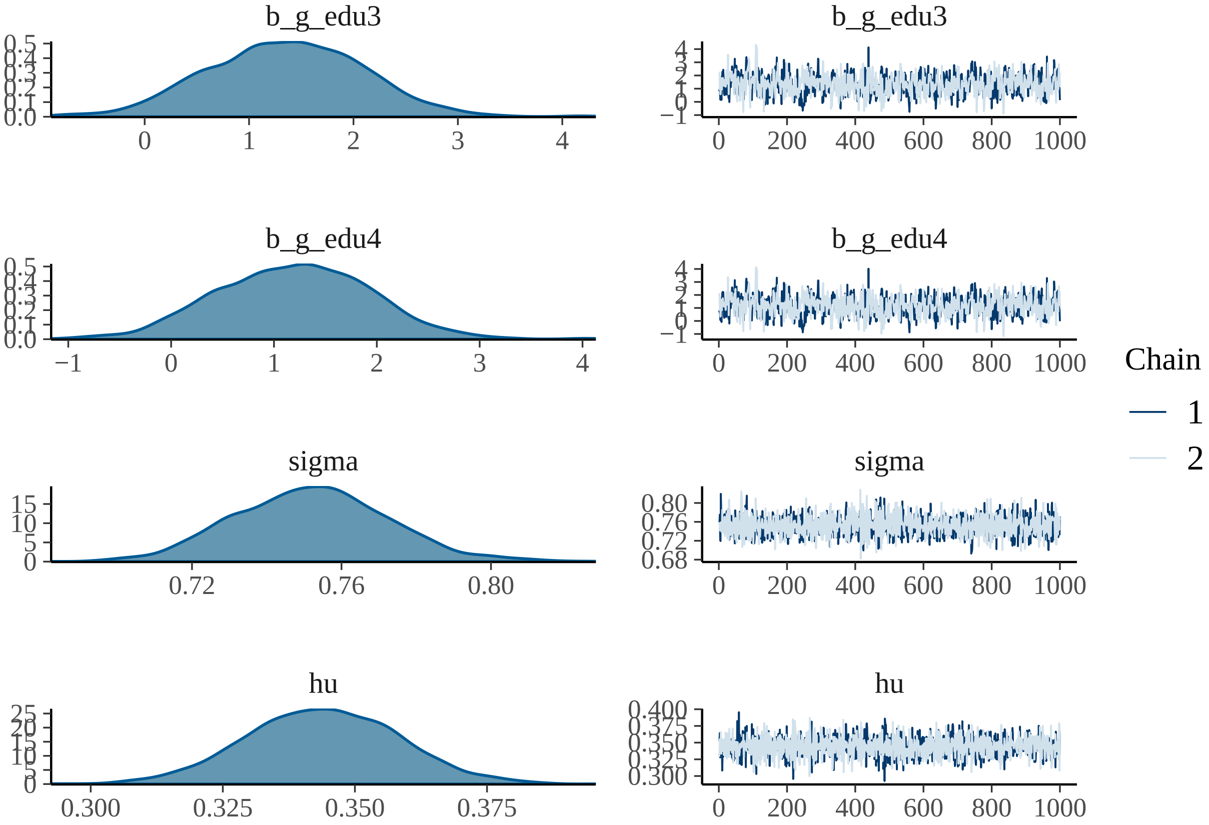
```
m_mains = brm(
  bf("absences ~ c_age + c_st + g_edu"),
  family = hurdle_lognormal(),
  data = model_df,
  iter = 2000, warmup = 1000, chains = 2,
  file = "m2.rds"
)
```

```
summary(m_mains)
```

```
##  Family: hurdle_lognormal
##   Links: mu = identity; sigma = identity; hu = identity
## Formula: absences ~ c_age + c_st + g_edu
##    Data: model_df (Number of observations: 1044)
##   Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 2000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.39      0.74    -1.08     1.81 1.00      570      843
## c_age         0.12      0.02     0.08     0.17 1.00     2079     1193
## c_st         -0.02      0.01    -0.03     0.00 1.00     2878     1306
## g_edu1        0.99      0.75    -0.43     2.44 1.00      571      818
## g_edu2        1.21      0.75    -0.20     2.63 1.00      569      853
## g_edu3        1.35      0.75    -0.07     2.79 1.00      569      826
## g_edu4        1.19      0.75    -0.21      2.64 1.00      565      863
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.75      0.02     0.71     0.79 1.00     2280     1449
## hu        0.34      0.01     0.32     0.37 1.00     2233     1482
##
## 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).
```

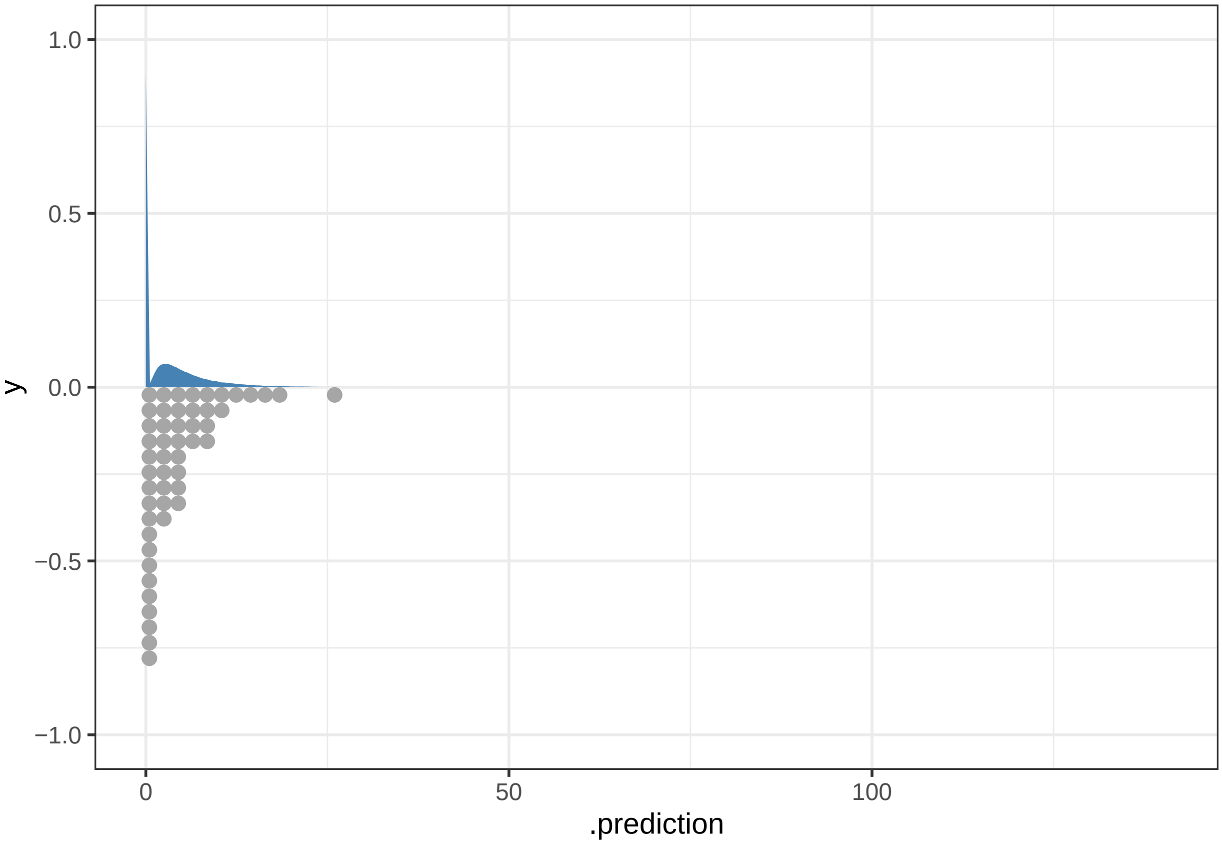
```r
plot(m_mains)
```



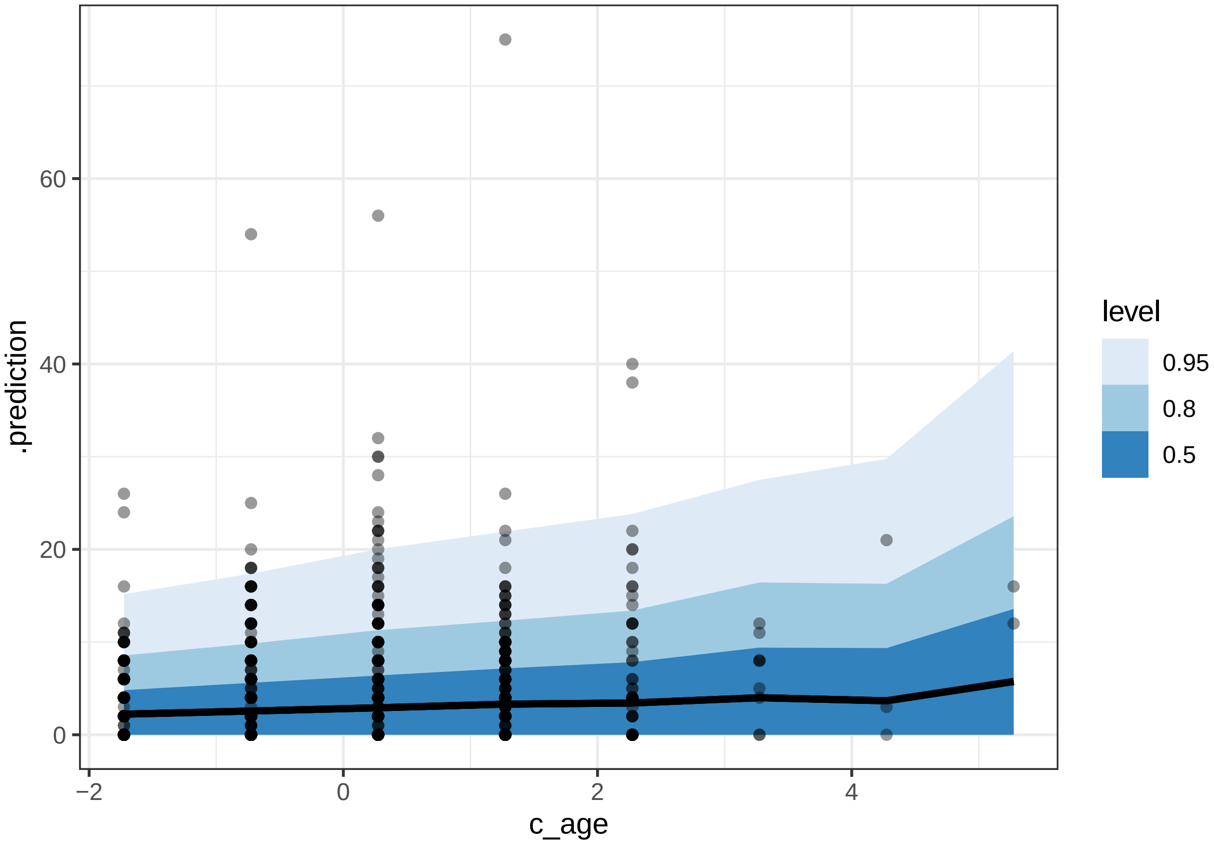


PP check

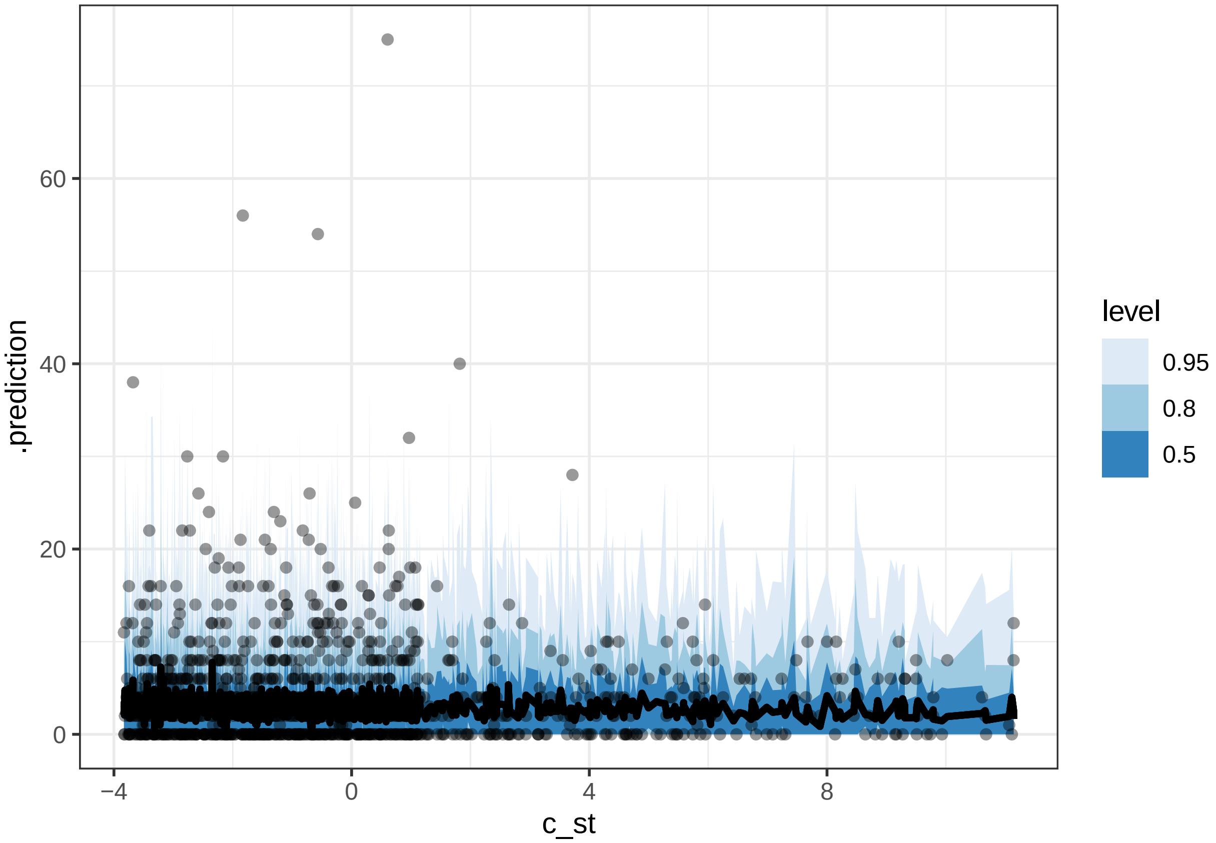
```
model_df |>
  select(absences, c_age, c_st, g_edu) |>
  add_predicted_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = .prediction)) +
  stat_slab(fill = "steelblue") +
  stat_dots(aes(x = absences), quantiles = 50, side = "bottom", data = model_df) +
  theme_bw()
```



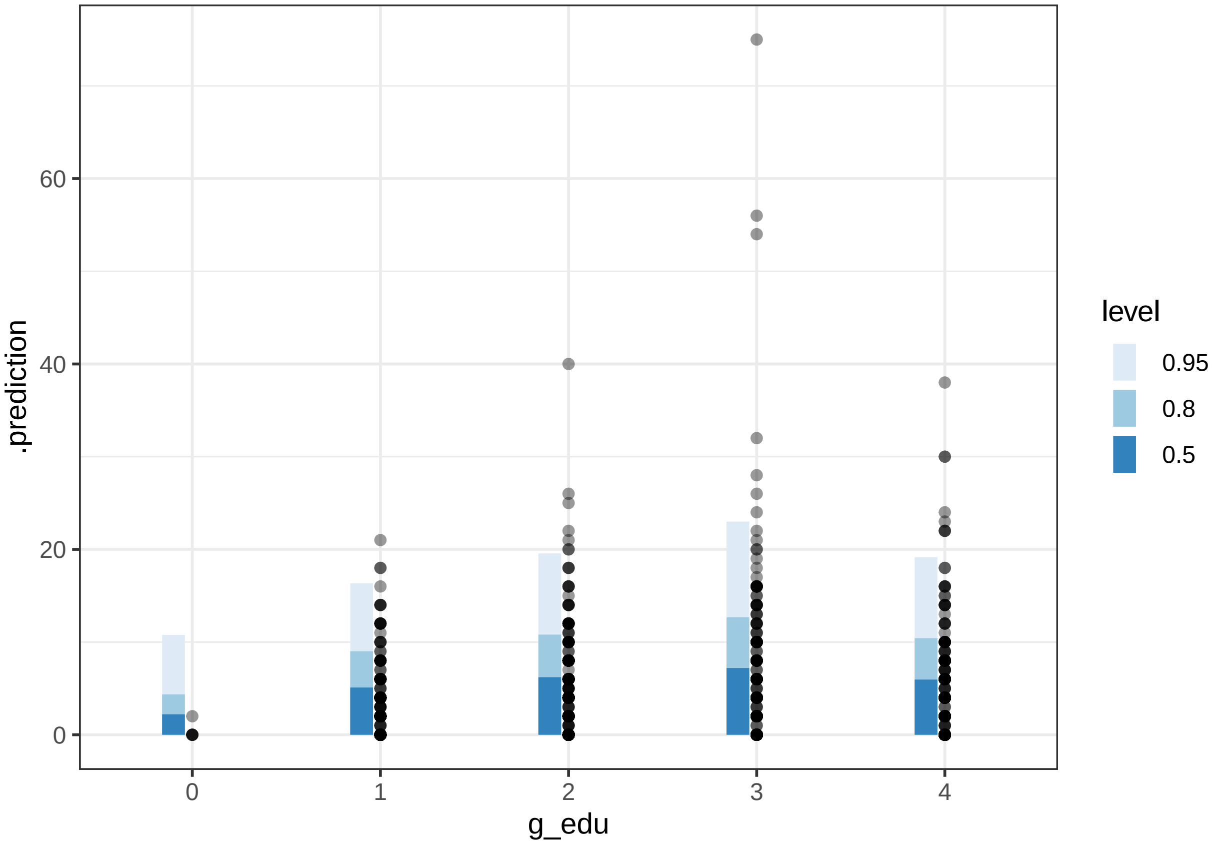
```
model_df |>
  select(absences, c_age, c_st, g_edu) |>
  add_predicted_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = c_age, y = .prediction)) +
  stat_lineribbon(.width = c(0.5, 0.8, 0.95)) +
  scale_fill_brewer() +
  geom_point(aes(y = absences), alpha = 0.4, data = model_df) +
  theme_bw()
```



```
model_df |>
  select(absences, c_age, c_st, g_edu) |>
  add_predicted_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = c_st, y = .prediction)) +
  stat_lineribbon(.width = c(0.5, 0.8, 0.95)) +
  scale_fill_brewer() +
  geom_point(aes(y = absences), alpha = 0.4, data = model_df) +
  theme_bw()
```



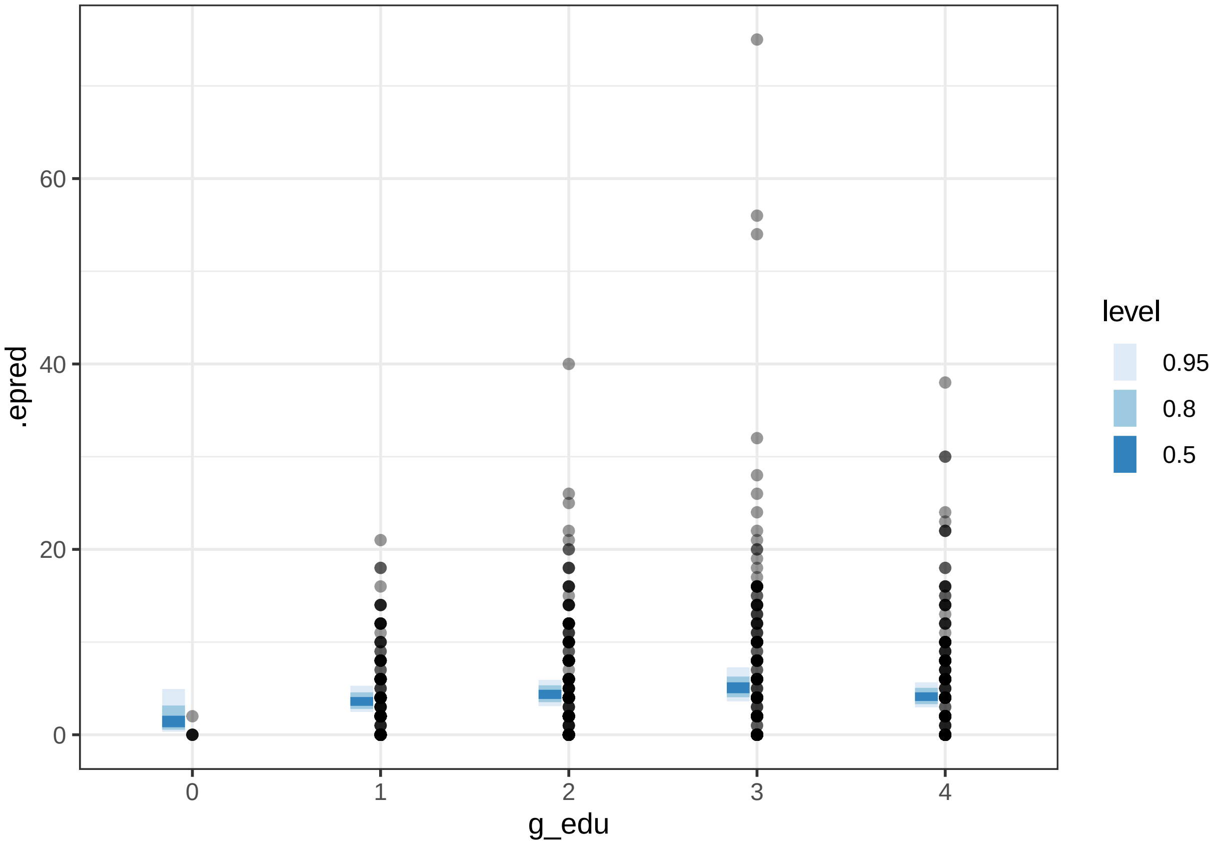
```
model_df |>
  select(absences, c_age, c_st, g_edu) |>
  add_predicted_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = g_edu, y = .prediction)) +
  stat_interval(.width = c(0.5, 0.8, 0.95), position = position_nudge(x = -0.1)) +
  scale_color_brewer() +
  geom_point(aes(y = absences), alpha = 0.4, data = model_df) +
  theme_bw()
```



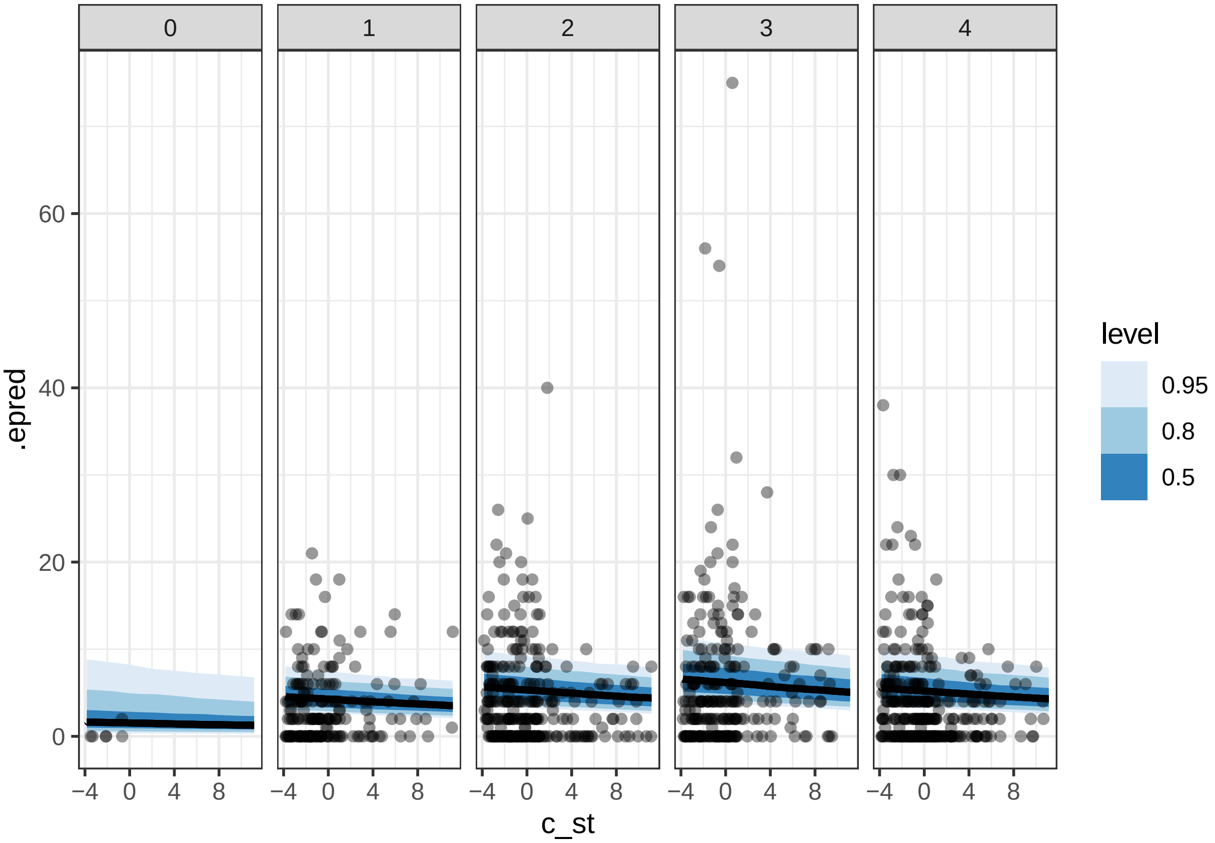
Inferential uncertainty

```
predictor_grid = model_df |>
  data_grid(c_age, c_st, g_edu) |>
  ungroup()
```

```
model_df |>
  select(absences, c_age, c_st, g_edu) |>
  add_epred_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = g_edu, y = .epred)) +
  stat_interval(.width = c(0.5, 0.8, 0.95), position = position_nudge(x = -0.1)) +
  scale_color_brewer() +
  geom_point(aes(y = absences), alpha = 0.4, data = model_df) +
  theme_bw()
```

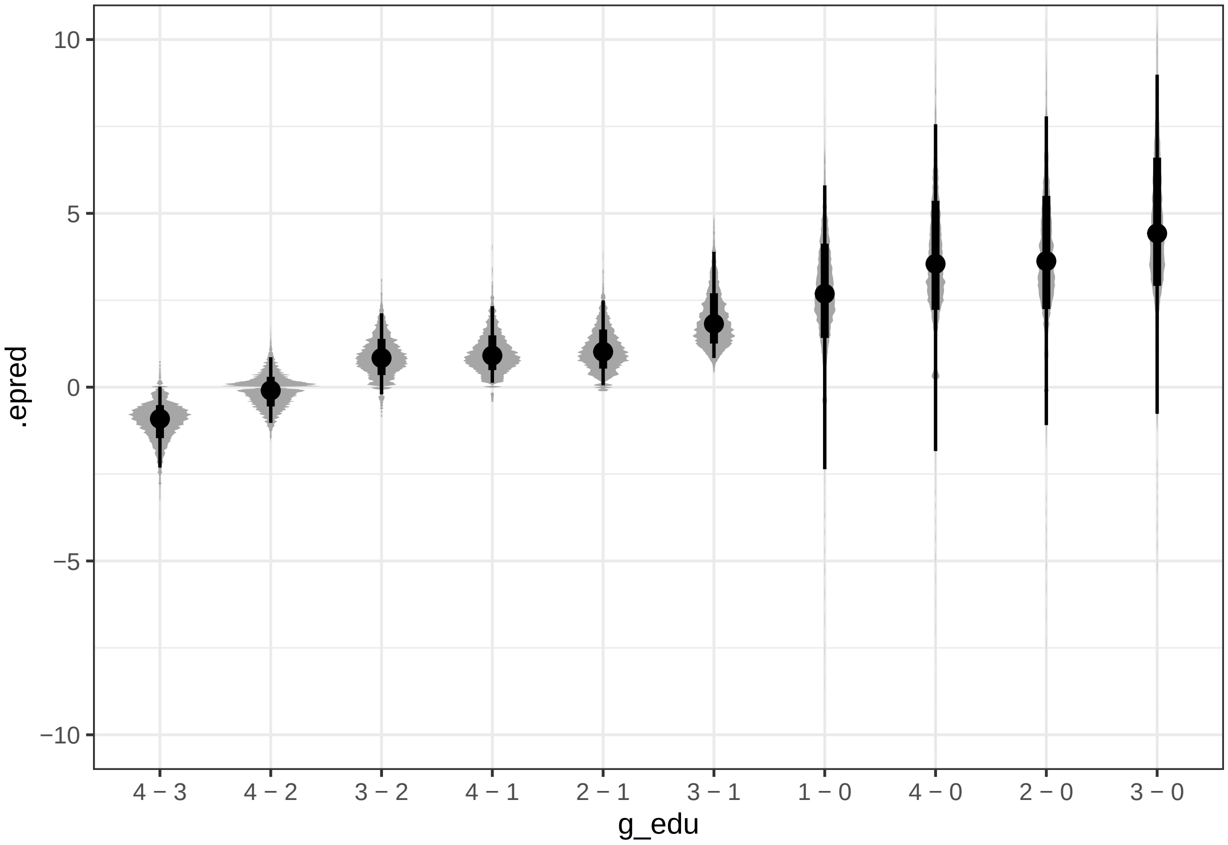


```
predictor_grid |>
  add_epred_draws(m_mains, ndraws = 100) |>
  ggplot(aes(x = c_st, y = .epred)) +
  stat_lineribbon(.width = c(0.5, 0.8, 0.95)) +
  scale_fill_brewer() +
  geom_point(aes(y = absences), alpha = 0.4, data = model_df) +
  theme_bw() +
  facet_grid(. ~ g_edu)
```



```
countrasts_g_edu = predictor_grid |>
  add_epred_draws(m_mains, ndraws = 100) |>
  compare_levels(.epred, by = g_edu) |>
  ungroup() |>
  mutate(g_edu = reorder(g_edu, .epred))
```

```
countrasts_g_edu |>
  ggplot(aes(x = g_edu, y = .epred)) +
  stat_eye() +
  # geom_hline(yintercept = 0, linetype = "dashed") +
  coord_cartesian(ylim = c(-10, 10)) +
  theme_bw()
```


10
5
0
−5
−10
.epred
4 − 3
4 − 2
3 − 2
4 − 1
2 − 1
3 − 1
1 − 0
4 − 0
2 − 0
3 − 0
g_edu