Bayesian Final Project

2026-01-23

```
library(readxl)
library(dplyr)
library(tidyr)
library(lubridate)
library(stringr)
```

```
load("beatspy.RData")
```

Revised Frequentist Models

The mixed-effects models are likely overfitting (`boundary (singular) fit` issue)

```
#logistic regression models fitted within each sector
library(purrr)
library(modelsummary)

sector_models = m3_df |>
  group_split(gics_sector_name) |>
  setNames(unique(m3_df$gics_sector_name)) |>
  map(~ glm(
    beat_spy ~ log_pe + div_yield,
    data = .x,
    family = binomial(link = "logit")
  ))

modelsummary(sector_models)
```

```
#mixed-effects w/ firm-level random intercepts
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

	Health Care	Information Technology	Consumer Staples	Industrials	Utilities	Financials	Materials
(Intercept)	−0.846	1.524	0.684	−0.559	0.512	−2.651	0.231
	(1.388)	(0.741)	(1.268)	(0.571)	(0.566)	(1.001)	(0.726)
log_pe	0.215	−0.304	−0.262	0.049	0.022	0.791	0.000
	(0.403)	(0.212)	(0.373)	(0.153)	(0.168)	(0.293)	(0.214)
div_yield	−0.044	−0.279	−0.181	−0.012	−0.184	−0.038	−0.074
	(0.097)	(0.089)	(0.090)	(0.052)	(0.060)	(0.096)	(0.081)
Num.Obs.	94	283	306	189	603	321	577
AIC	131.3	387.2	395.5	257.9	826.0	427.6	803.8
BIC	138.9	398.2	406.7	267.7	839.2	438.9	816.9
Log.Lik.	−62.641	−190.622	−194.752	−125.971	−410.016	−210.811	−398.899
RMSE	0.49	0.49	0.47	0.49	0.49	0.48	0.50

```r
mixed_intercept = glmer(
  beat_spy ~ log_pe + div_yield +
    (1 | Ticker),
  data = m3_df,
  family = binomial(link = "logit")
)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
mixed_intercept |> summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: beat_spy ~ log_pe + div_yield + (1 | Ticker)
##    Data: m3_df
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    4626.4    4650.9   -2309.2    4618.4      3396
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1735 -0.9373 -0.7145  1.0129  4.7846
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Ticker (Intercept) 0        0
## Number of obs: 3400, groups:  Ticker, 406
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.17522    0.19108   0.917    0.359
## log_pe       0.03057    0.05524   0.553    0.580
## div_yield   -0.15169    0.01974  -7.684 1.55e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) log_pe
## log_pe    -0.946
## div_yield -0.539  0.299
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
#mixed effects w/ firm-level random slopes
mixed_random_slopes = glmer(
  beat_spy ~ log_pe + div_yield +
    (1 + log_pe + div_yield | Ticker),
  data = m3_df,
  family = binomial(link = "logit"),
  control = glmerControl(optimizer = "bobyqa")
)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
mixed_random_slopes |> summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: beat_spy ~ log_pe + div_yield + (1 + log_pe + div_yield | Ticker)
##    Data: m3_df
## Control: glmerControl(optimizer = "bobyqa")
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    4632.9    4688.0   -2307.4    4614.9      3391
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1874 -0.9363 -0.6846  1.0067  2.9505
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  Ticker (Intercept) 0.289481 0.53803
##         log_pe      0.012842 0.11332  -1.00
##         div_yield   0.007972 0.08928  -1.00  1.00
## Number of obs: 3400, groups:  Ticker, 406
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.24602    0.20147   1.221    0.222
## log_pe       0.02221    0.05797   0.383    0.702
## div_yield   -0.17073    0.02119  -8.057 7.84e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) log_pe
```

```
## log_pe    -0.949
## div_yield -0.565  0.335
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```r
#mixed-effects model w/ sector-level random slopes
mixed_sector_slopes = glmer(
  beat_spy ~ log_pe + div_yield +
    (1 + log_pe + div_yield | gics_sector_name),
  data = m3_df,
  family = binomial(link = "logit")
)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
mixed_sector_slopes |> summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## beat_spy ~ log_pe + div_yield + (1 + log_pe + div_yield | gics_sector_name)
##    Data: m3_df
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    4597.2    4652.4   -2289.6    4579.2      3391
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4695 -0.8955 -0.6795  0.9989  3.4987
##
## Random effects:
##  Groups           Name        Variance Std.Dev. Corr
##  gics_sector_name (Intercept) 0.400691 0.63300
##                   log_pe      0.004766 0.06904  -0.99
##                   div_yield   0.002169 0.04658  -0.99  0.96
## Number of obs: 3400, groups:  gics_sector_name, 11
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.32205    0.29108  -1.106   0.2686
## log_pe       0.12805    0.06343   2.019   0.0435 *
## div_yield   -0.11209    0.02587  -4.333 1.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) log_pe
## log_pe    -0.885
## div_yield -0.743  0.472
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```r
#mixed-effects model with sector-level variation + firm-level variation within sector
mixed_nested = glmer(
  beat_spy ~ log_pe + div_yield +
    (1 | gics_sector_name/Ticker),
  data = m3_df,
  family = binomial(link = "logit")
)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
mixed_nested |> summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: beat_spy ~ log_pe + div_yield + (1 | gics_sector_name/Ticker)
##    Data: m3_df
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    4593.4    4624.0   -2291.7    4583.4      3395
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4931 -0.9133 -0.6750  1.0000  3.8468
##
## Random effects:
##  Groups                  Name        Variance  Std.Dev.
##  Ticker:gics_sector_name (Intercept) 2.908e-10 1.705e-05
##  gics_sector_name        (Intercept) 7.238e-02 2.690e-01
## Number of obs: 3400, groups:
## Ticker:gics_sector_name, 406; gics_sector_name, 11
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29217    0.24022  -1.216   0.2239
## log_pe       0.13164    0.06247   2.107   0.0351 *
## div_yield   -0.11290    0.02243  -5.034 4.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) log_pe
## log_pe    -0.894
## div_yield -0.595  0.414
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Bayesian Analysis

```r
library(brms)
library(tidybayes)
```

```r
library(bayesplot)
library(posterior)
```

Model 1 (Logistic)

$$
\begin{aligned}
Y_{i,t} &\sim Bernoulli(p_{i,t}) \\
logit(p_{i,t}) &= \beta_0 + \beta_1 + log(PE_{i,t}) + \beta_2 DivYield_{i,t}
\end{aligned}
$$

```r
bayes_model1 <- brm(
  beat_spy ~ log_pe + div_yield,
  data = m3_df,
  family = bernoulli(link = "logit"),
  seed = 123
)
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

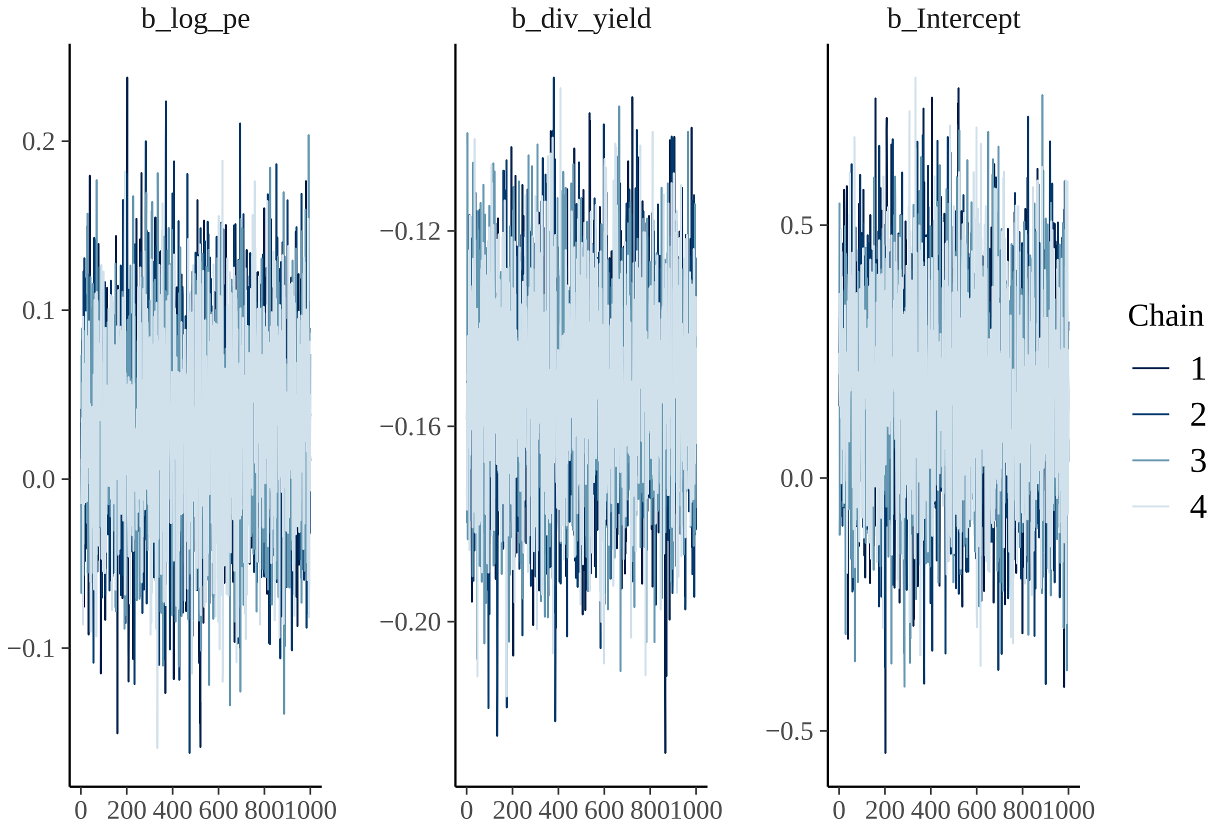
```
## Start sampling
```

```r
summary(bayes_model1)$fixed
```

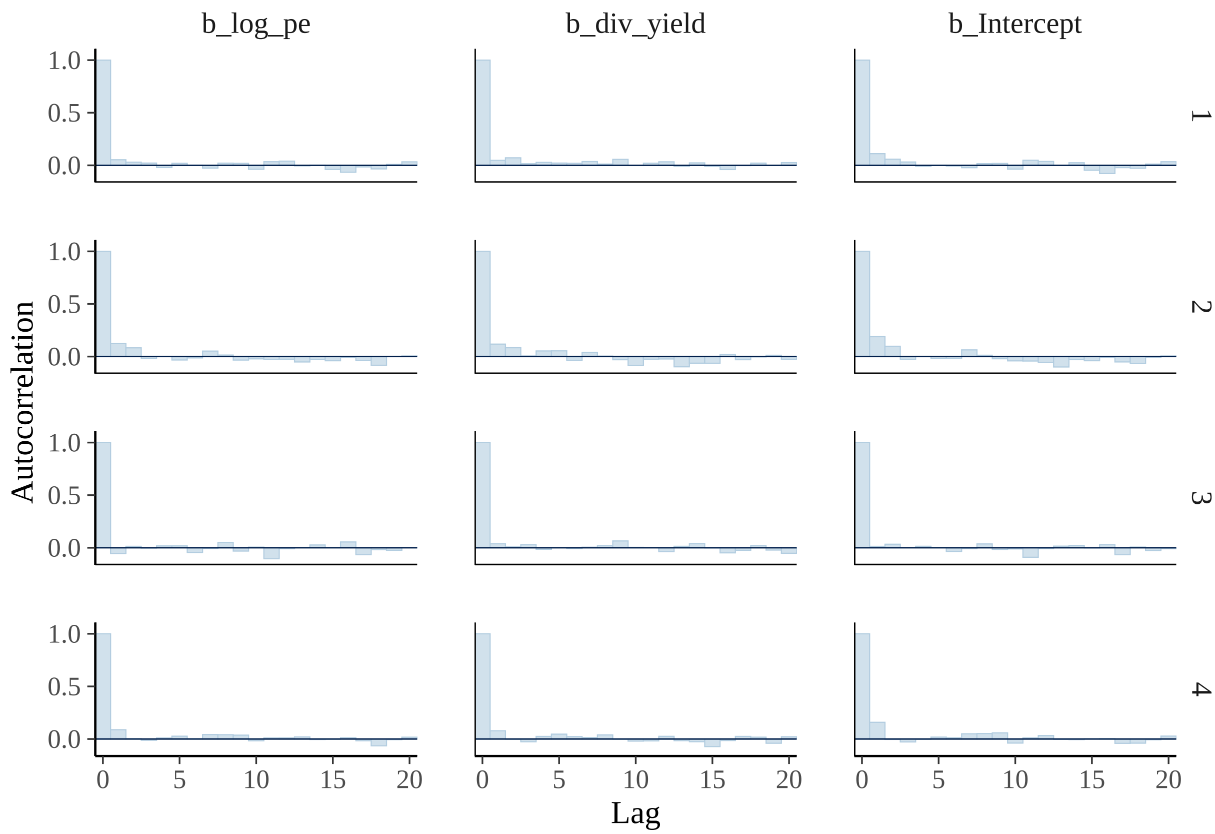
```
##              Estimate  Est.Error   l-95% CI   u-95% CI     Rhat Bulk_ESS
## Intercept  0.17273278 0.18791917 -0.1830719  0.5376506 1.001122 3007.931
## log_pe     0.03149262 0.05432937 -0.0760189  0.1343352 1.001685 3368.702
## div_yield -0.15189077 0.01951898 -0.1906835 -0.1138507 1.000259 2819.099
##           Tail_ESS
## Intercept 2906.451
## log_pe    3373.531
## div_yield 2796.484
```

```r
draws_model1 = bayes_model1 |>
  as_draws_array()

mcmc_trace(draws_model1,
           pars = c("b_log_pe", "b_div_yield", "b_Intercept"))
```




```
mcmc_acf_bar(
  draws_model1,
  pars = c("b_log_pe", "b_div_yield", "b_Intercept")
)
```



Model 2 (Hierarchical)

```
bayes_model2 <- brm(
  beat_spy ~ log_pe + div_yield + (1 | gics_sector_name/Ticker),
  data = m3_df,
  family = bernoulli(link = "logit"),
  seed = 123
)
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Start sampling
```

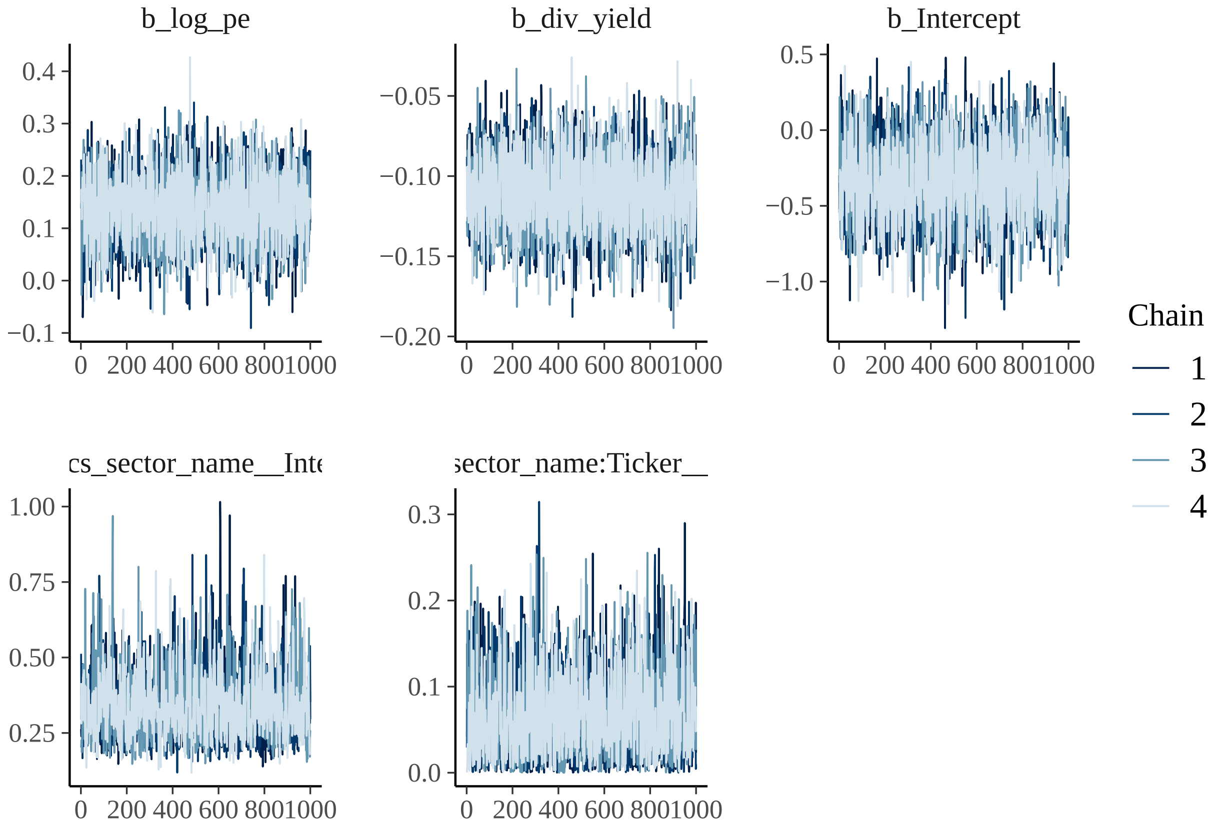
```
summary(bayes_model2)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: beat_spy ~ log_pe + div_yield + (1 | gics_sector_name/Ticker)
##    Data: m3_df (Number of observations: 3400)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
```

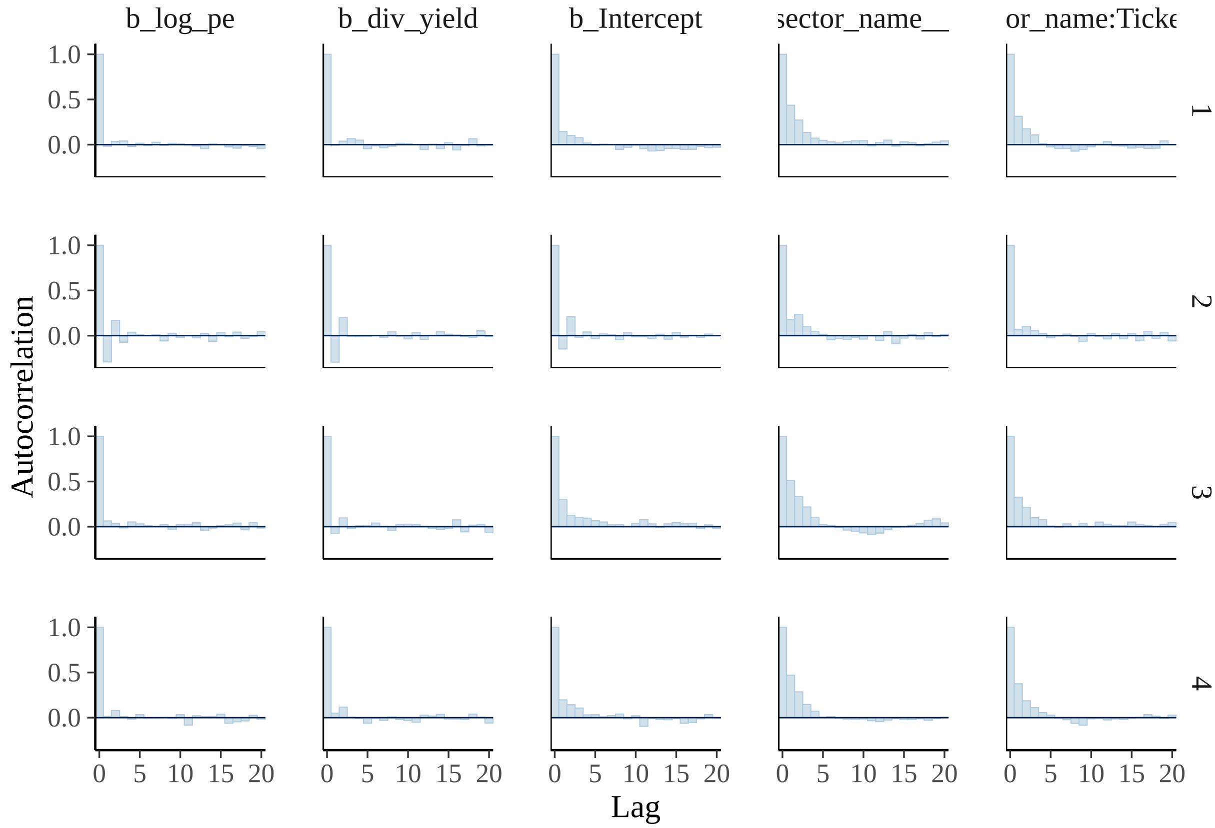
```
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~gics_sector_name (Number of levels: 11)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.34      0.10     0.19     0.59 1.00     1348     2025
##
## ~gics_sector_name:Ticker (Number of levels: 406)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.06      0.05     0.00     0.17 1.00     1607     1590
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -0.32      0.26    -0.82     0.17 1.00     2110     2593
## log_pe        0.14      0.06     0.02     0.26 1.00     3616     2987
## div_yield    -0.11      0.02    -0.15    -0.07 1.00     3638     2948
##
## 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
draws_model2 = bayes_model2 |>
  as_draws_array()

mcmc_trace(
  draws_model2,
  pars = c("b_log_pe",
           "b_div_yield",
           "b_Intercept",
           "sd_gics_sector_name__Intercept",
           "sd_gics_sector_name:Ticker__Intercept")
)
```




```r
mcmc_acf_bar(
  draws_model2,
  pars = c("b_log_pe", "b_div_yield", "b_Intercept", "sd_gics_sector_name__Intercept", "sd_gics_sector_n
)
```



Model Comparison

```r
loo_compare(loo(bayes_model1),
            loo(bayes_model2))
```

```
##              elpd_diff se_diff
## bayes_model2   0.0       0.0
## bayes_model1 -22.2       7.0
```

The difference in expected log predictive density is 22.2 (SE = 7.0). LOOCV provides strong evidence that the multilevel specification improves out-of-sample predictive performance.