Final Project - Bayesian Analysis

```r
library(readxl)
library(dplyr)
library(tidyr)
library(lubridate)
library(stringr)
library(lme4)
library(tidymodels)
library(purrr)
library(tidyr)
library(pROC)
library(groupdata2)
```

```r
load("beatspy.RData")
m3_df = m3_df |>
  mutate(
    beat_spy = factor(
      beat_spy,
      levels = c(1, 0),       # 1 = event
      labels = c("yes", "no")
    ),
    gics_sector_name = factor(gics_sector_name),
    Ticker = factor(Ticker)
  )
set.seed(123)
```

Frequentist

Models

Sector Models

Firm Mixed Effects

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

```
## 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
```

	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	−398.89
RMSE	0.49	0.49	0.47	0.49	0.49	0.48	0.50

```
##    4632.9    4688.0   -2307.4    4614.9      3391
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9505 -1.0067  0.6846  0.9363  1.1874
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  Ticker (Intercept) 0.289497 0.53805
##         log_pe      0.012843 0.11333  -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      -0.02220    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')
```

Sector Fixed Effects

```
##
## Call:
## glm(formula = beat_spy ~ log_pe + div_yield + factor(gics_sector_name),
##     family = binomial(link = "logit"), data = m3_df)
##
## Coefficients:
```

```
##                                              Estimate Std. Error z value
## (Intercept)                                   0.51439    0.31086   1.655
## log_pe                                       -0.16173    0.06488  -2.493
## div_yield                                     0.10129    0.02318   4.370
## factor(gics_sector_name)Consumer Discretionary -0.25790    0.24538  -1.051
## factor(gics_sector_name)Consumer Staples      0.26459    0.24560   1.077
## factor(gics_sector_name)Energy                0.02855    0.26369   0.108
## factor(gics_sector_name)Financials           -0.42514    0.22947  -1.853
## factor(gics_sector_name)Health Care           0.12594    0.24359   0.517
## factor(gics_sector_name)Industrials          -0.31252    0.23072  -1.355
## factor(gics_sector_name)Information Technology -0.67038    0.24352  -2.753
## factor(gics_sector_name)Materials             0.01109    0.25882   0.043
## factor(gics_sector_name)Real Estate           0.39848    0.26779   1.488
## factor(gics_sector_name)Utilities            -0.12672    0.24909  -0.509
##                                              Pr(>|z|)
## (Intercept)                                   0.09798 .
## log_pe                                        0.01267 *
## div_yield                                    1.24e-05 ***
## factor(gics_sector_name)Consumer Discretionary  0.29324
## factor(gics_sector_name)Consumer Staples      0.28134
## factor(gics_sector_name)Energy                0.91379
## factor(gics_sector_name)Financials            0.06392 .
## factor(gics_sector_name)Health Care           0.60513
## factor(gics_sector_name)Industrials           0.17555
## factor(gics_sector_name)Information Technology  0.00591 **
## factor(gics_sector_name)Materials             0.96582
## factor(gics_sector_name)Real Estate           0.13675
## factor(gics_sector_name)Utilities             0.61092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4692.6  on 3399  degrees of freedom
## Residual deviance: 4553.0  on 3387  degrees of freedom
## AIC: 4579
##
## Number of Fisher Scoring iterations: 4
```

Sector Fixed Effects w/ Interactions

```
##
## Call:
## glm(formula = beat_spy ~ log_pe + div_yield * factor(gics_sector_name),
##     family = binomial(link = "logit"), data = m3_df)
##
## Coefficients:
##                                                         Estimate Std. Error
## (Intercept)                                             0.598294   0.382658
## log_pe                                                 -0.141407   0.065402
## div_yield                                               0.056334   0.072561
## factor(gics_sector_name)Consumer Discretionary         -0.651732   0.386631
## factor(gics_sector_name)Consumer Staples                0.061950   0.397369
```

```
## factor(gics_sector_name)Energy                               0.269860   0.406570
## factor(gics_sector_name)Financials                          -0.724126   0.350106
## factor(gics_sector_name)Health Care                         -0.132709   0.362233
## factor(gics_sector_name)Industrials                         -0.364062   0.355297
## factor(gics_sector_name)Information Technology              -1.078496   0.379784
## factor(gics_sector_name)Materials                           -0.046934   0.386036
## factor(gics_sector_name)Real Estate                          0.436358   0.490961
## factor(gics_sector_name)Utilities                           -0.190843   0.517561
## div_yield:factor(gics_sector_name)Consumer Discretionary     0.149613   0.107978
## div_yield:factor(gics_sector_name)Consumer Staples           0.064631   0.100157
## div_yield:factor(gics_sector_name)Energy                    -0.050386   0.088010
## div_yield:factor(gics_sector_name)Financials                 0.105227   0.087464
## div_yield:factor(gics_sector_name)Health Care                0.107264   0.105975
## div_yield:factor(gics_sector_name)Industrials               -0.002932   0.103243
## div_yield:factor(gics_sector_name)Information Technology     0.161282   0.107027
## div_yield:factor(gics_sector_name)Materials                  0.012621   0.096039
## div_yield:factor(gics_sector_name)Real Estate                0.000487   0.103716
## div_yield:factor(gics_sector_name)Utilities                  0.026034   0.120654
##                                                             z value Pr(>|z|)
## (Intercept)                                                   1.564  0.11793
## log_pe                                                       -2.162  0.03061 *
## div_yield                                                     0.776  0.43754
## factor(gics_sector_name)Consumer Discretionary               -1.686  0.09186 .
## factor(gics_sector_name)Consumer Staples                      0.156  0.87611
## factor(gics_sector_name)Energy                                0.664  0.50685
## factor(gics_sector_name)Financials                           -2.068  0.03861 *
## factor(gics_sector_name)Health Care                          -0.366  0.71409
## factor(gics_sector_name)Industrials                          -1.025  0.30552
## factor(gics_sector_name)Information Technology               -2.840  0.00451 **
## factor(gics_sector_name)Materials                            -0.122  0.90323
## factor(gics_sector_name)Real Estate                           0.889  0.37412
## factor(gics_sector_name)Utilities                            -0.369  0.71232
## div_yield:factor(gics_sector_name)Consumer Discretionary       1.386  0.16587
## div_yield:factor(gics_sector_name)Consumer Staples             0.645  0.51874
## div_yield:factor(gics_sector_name)Energy                      -0.573  0.56698
## div_yield:factor(gics_sector_name)Financials                   1.203  0.22894
## div_yield:factor(gics_sector_name)Health Care                  1.012  0.31146
## div_yield:factor(gics_sector_name)Industrials                 -0.028  0.97734
## div_yield:factor(gics_sector_name)Information Technology       1.507  0.13183
## div_yield:factor(gics_sector_name)Materials                    0.131  0.89544
## div_yield:factor(gics_sector_name)Real Estate                  0.005  0.99625
## div_yield:factor(gics_sector_name)Utilities                    0.216  0.82916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4692.6  on 3399  degrees of freedom
## Residual deviance: 4542.1  on 3377  degrees of freedom
## AIC: 4588.1
##
## Number of Fisher Scoring iterations: 4
```

Model Comparison

Information Criterion

```
##                 Model df      AIC
## 1           pooled_fe 13 4579.022
## 2  pooled_fe_interact 23 4588.136
## 3       sector_models 33 4590.064
## 4 mixed_random_slopes  9 4632.865
```

```
##                 Model df      BIC
## 1           pooled_fe 13 4658.732
## 2 mixed_random_slopes  9 4688.048
## 3  pooled_fe_interact 23 4729.162
## 4       sector_models 33 4792.405
```

Predictive Performance

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

```
## Setting levels: control = yes, case = no
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = yes, case = no
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = yes, case = no
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = yes, case = no
```

```
## Setting direction: controls < cases
```

```
## # A tibble: 4 x 2
##   model              auc
##   <chr>            <dbl>
## 1 pooled           0.610
## 2 sector_by_sector 0.606
## 3 pooled_int       0.606
## 4 mixed            0.590
```

Bayesian

Model 1 (Logistic)

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

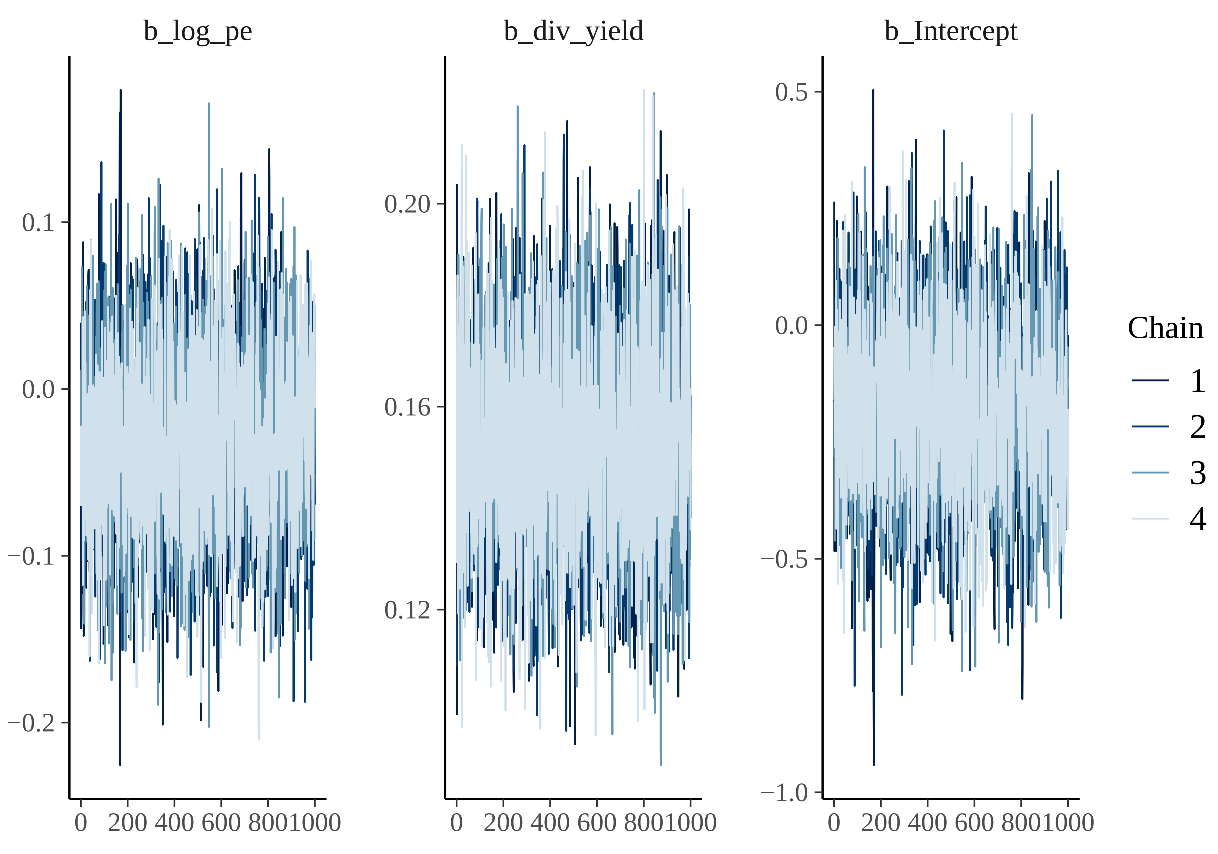
```
## Compiling Stan program...
```

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

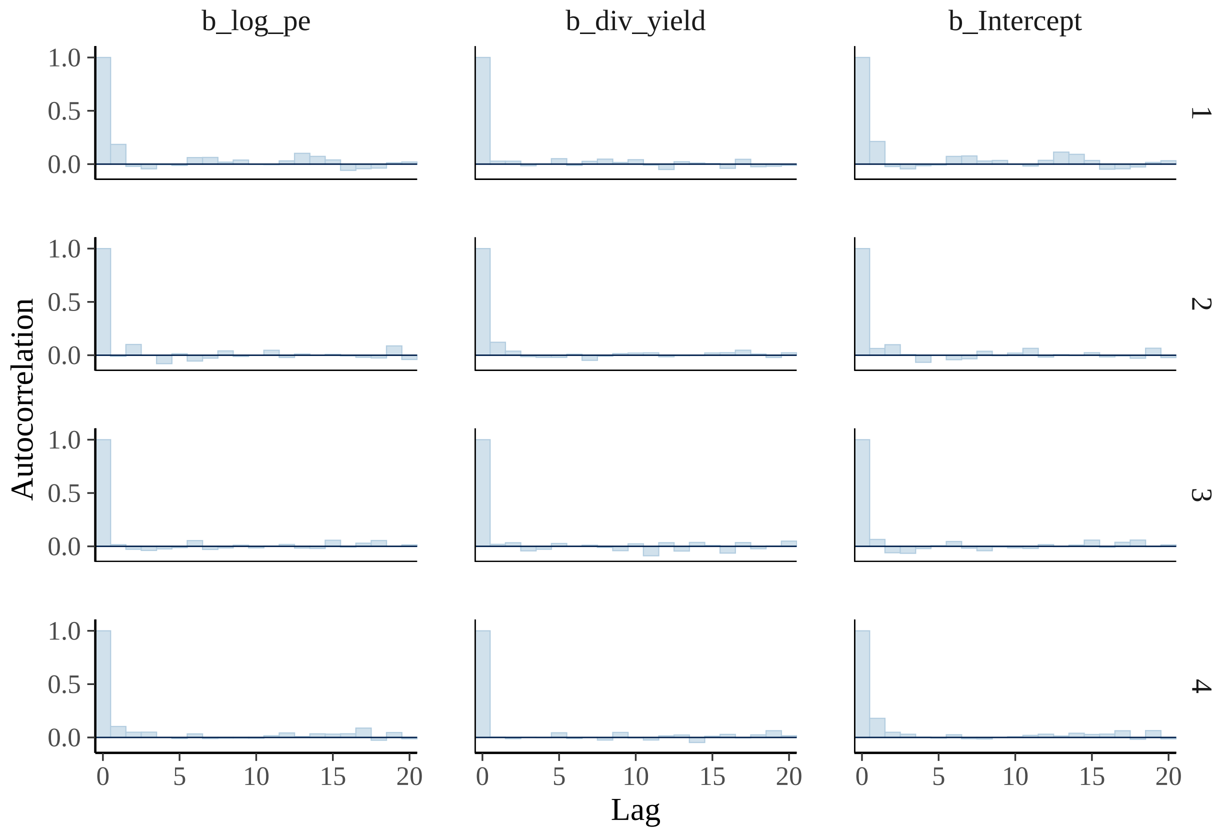
```
## Start sampling
```

```
##              Estimate  Est.Error   l-95% CI   u-95% CI     Rhat Bulk_ESS
## Intercept -0.17759681 0.18714168 -0.5503537 0.18279735 1.000445 3151.577
## log_pe    -0.03052351 0.05440108 -0.1344534 0.07678756 1.000324 3418.781
## div_yield  0.15239539 0.01982122  0.1144111 0.19214307 1.001008 3661.874
##           Tail_ESS
## Intercept 2686.697
## log_pe    2877.849
## div_yield 2709.490
```

Rhats are ~1 and effective sample sizes » 100



No discernable pattern from trace plots



acfs fall off quickly

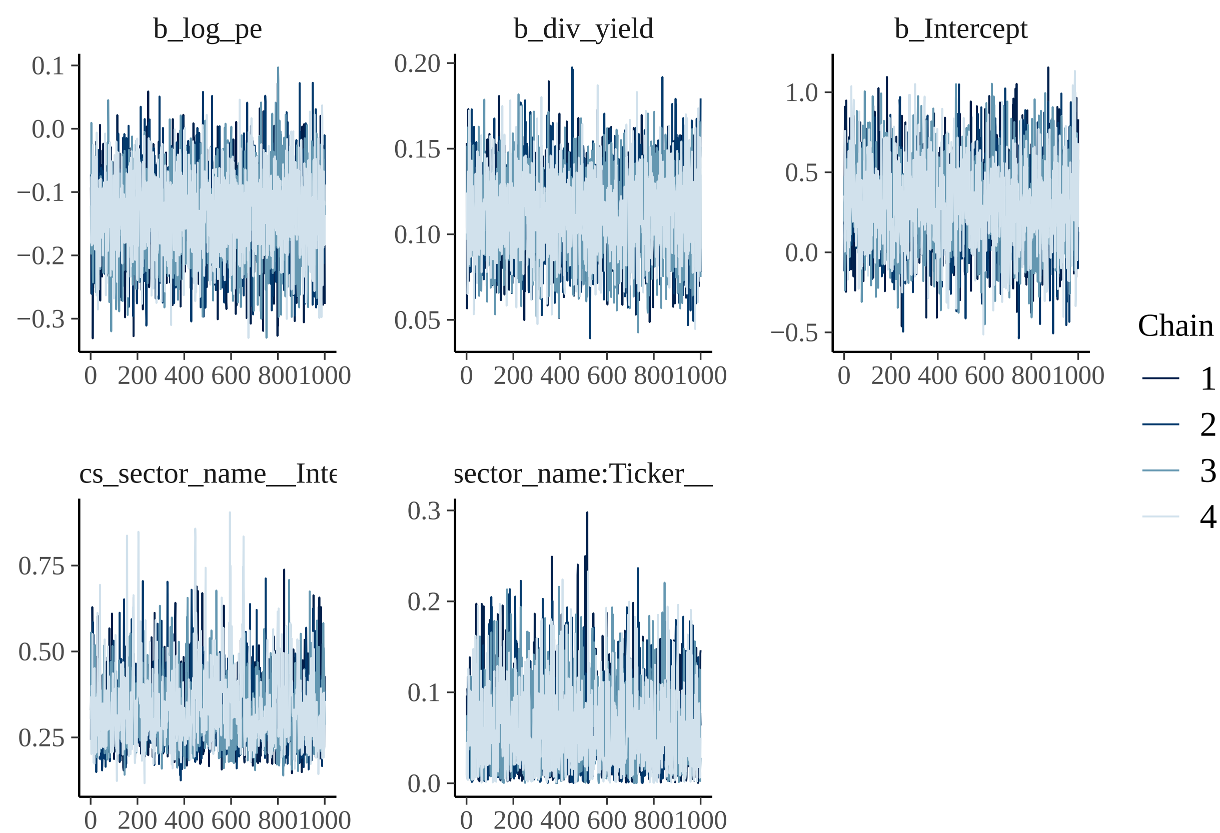
Model 2 (Nested random intercepts)

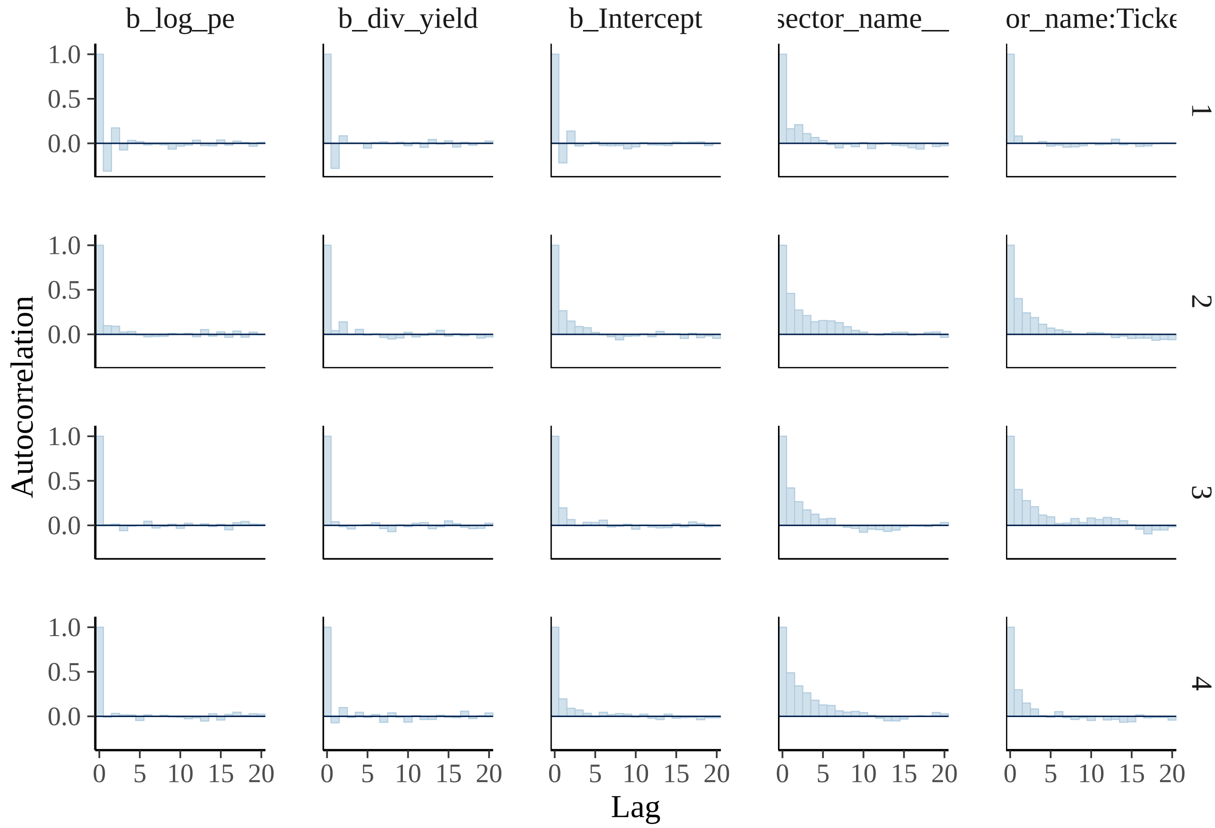
$$logit(p_{i,t}) = \beta_0 + \beta_1 log(PE_{i,t}) + \beta_2 DivYield_{i,t} + u_j + v_i$$

```
## Compiling Stan program...

## Trying to compile a simple C file

## Start sampling

##  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)     0.33      0.10     0.18     0.56 1.00     1141     1522
##
## ~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      0.04     0.00     0.17 1.00     1435     1544
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.30      0.26    -0.19     0.84 1.00     2505     2799
## log_pe       -0.13      0.06    -0.27    -0.01 1.00     3961     3077
## div_yield     0.11      0.02     0.07     0.16 1.00     3786     3272
##
## 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).
```

Rhats are all ~1, effective sample sizes » 100





acfs fall off quickly

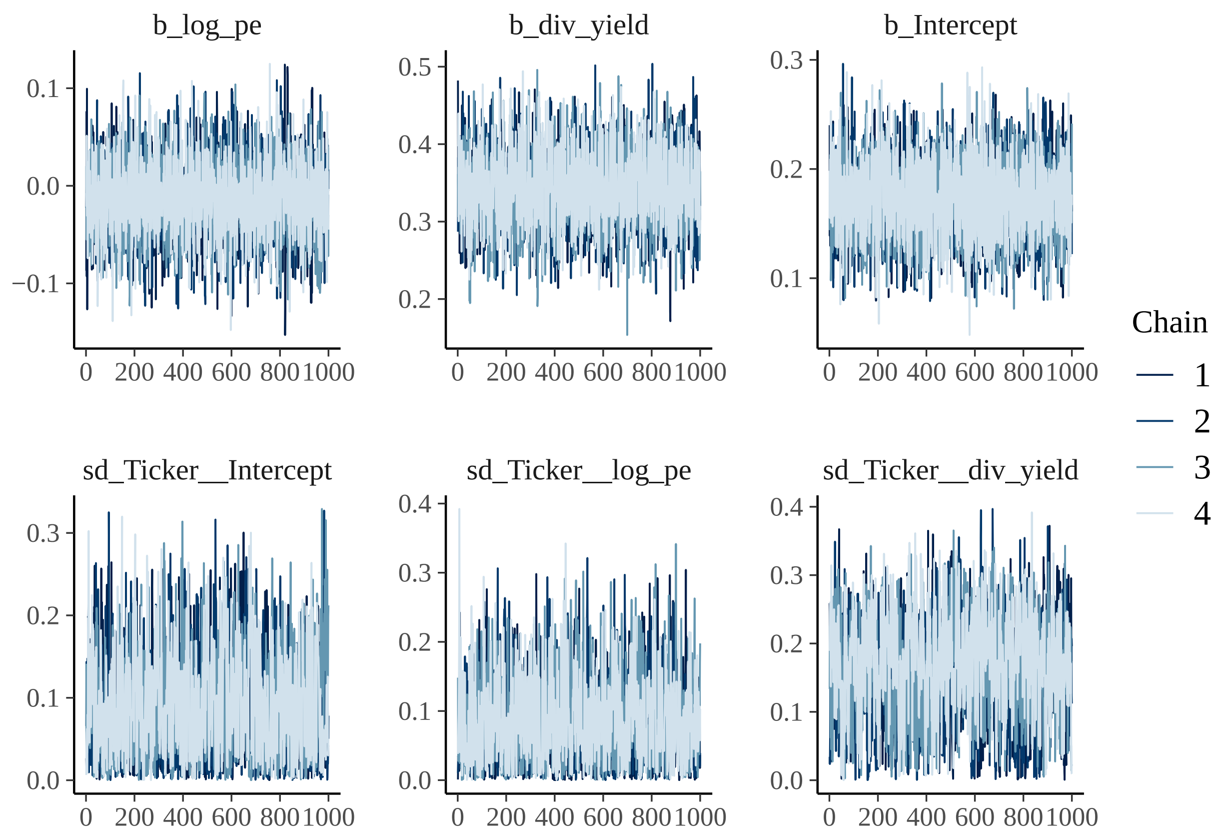
Model 3 (Mixed Effects + firm-level random slopes/intercepts)

Covariates are centered

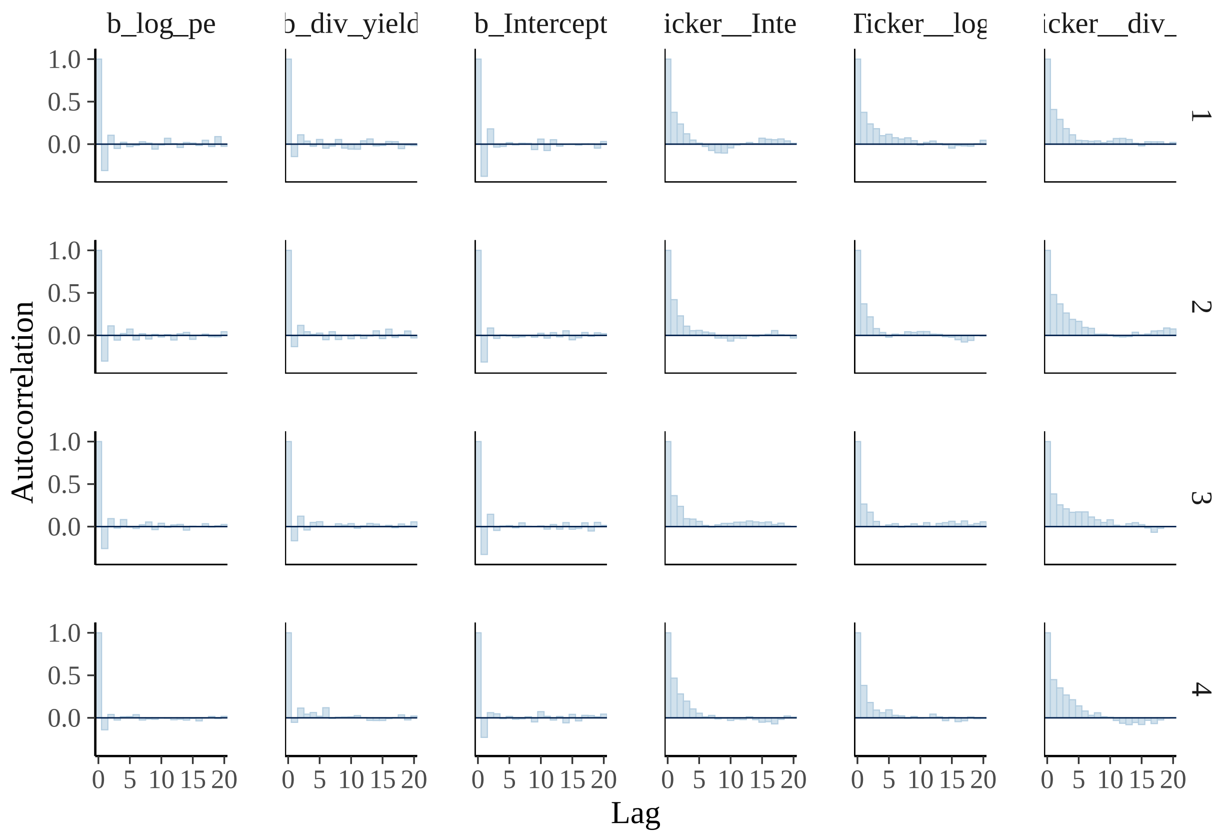
```
## Compiling Stan program...

## Trying to compile a simple C file

## Start sampling

##  Family: bernoulli
##   Links: mu = logit
## Formula: beat_spy ~ log_pe + div_yield + (1 + log_pe + div_yield | Ticker)
##    Data: mutate(m3_df, log_pe = scale(log_pe), div_yield =  (Number of observations: 3400)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~Ticker (Number of levels: 406)
##                             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)                   0.09      0.06     0.00     0.23 1.00     1377
## sd(log_pe)                      0.08      0.06     0.00     0.22 1.00     1258
## sd(div_yield)                   0.16      0.07     0.02     0.30 1.01      891
```

```
## cor(Intercept,log_pe)        0.03      0.50     -0.86      0.88 1.00       2565
## cor(Intercept,div_yield)     0.11      0.48     -0.84      0.90 1.00       1112
## cor(log_pe,div_yield)        0.20      0.49     -0.80      0.93 1.00       1266
##                          Tail_ESS
## sd(Intercept)                2140
## sd(log_pe)                   1740
## sd(div_yield)                1644
## cor(Intercept,log_pe)        2375
## cor(Intercept,div_yield)     2187
## cor(log_pe,div_yield)        2621
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.17      0.04     0.10     0.24 1.00     6974     2983
## log_pe       -0.01      0.04    -0.09     0.07 1.00     5986     2722
## div_yield     0.34      0.05     0.25     0.44 1.00     3244     2842
##
## 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).
```

Rhats are ~1 and effective sample sizes » 100



No discernable pattern from trace plots



Model 4 (Pooled + sector FE)

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

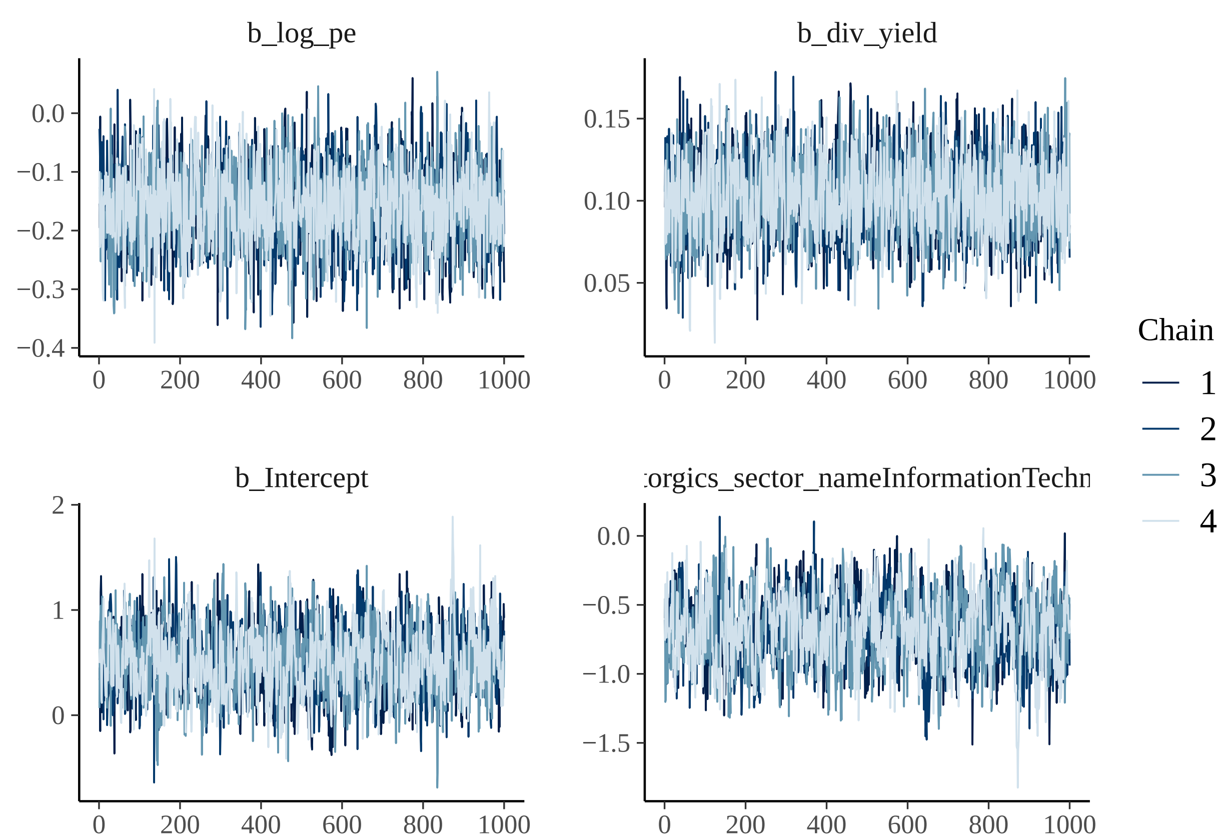
```
## Trying to compile a simple C file
```

```
## Start sampling
```

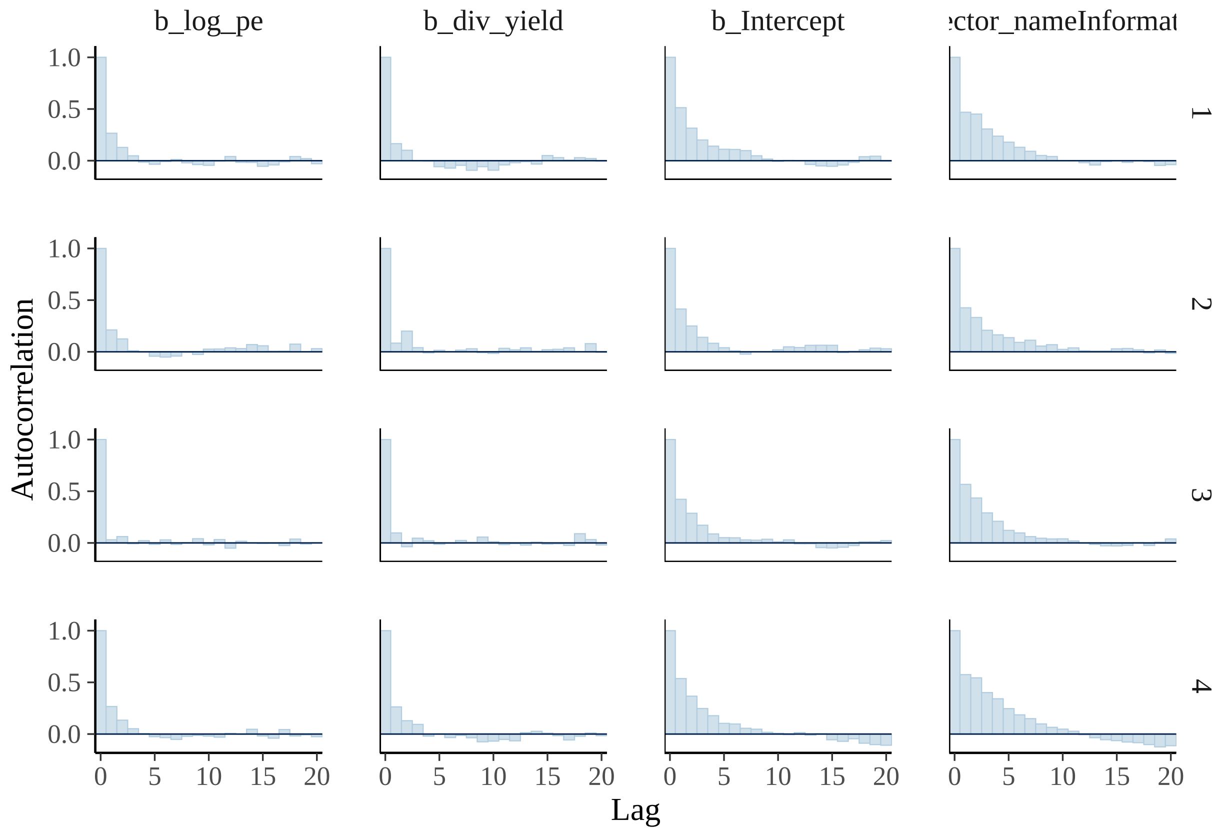
```
##  Family: bernoulli
##   Links: mu = logit
## Formula: beat_spy ~ log_pe + div_yield + factor(gics_sector_name)
##    Data: m3_df (Number of observations: 3400)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                                               Estimate Est.Error l-95% CI
## Intercept                                         0.51      0.31    -0.10
## log_pe                                           -0.16      0.07    -0.29
## div_yield                                         0.10      0.02     0.06
## factorgics_sector_nameConsumerDiscretionary      -0.27      0.25    -0.77
## factorgics_sector_nameConsumerStaples             0.26      0.25    -0.23
## factorgics_sector_nameEnergy                      0.02      0.27    -0.52
## factorgics_sector_nameFinancials                 -0.43      0.23    -0.90
```

```
## factorgics_sector_nameHealthCare                0.12      0.25    -0.38
## factorgics_sector_nameIndustrials              -0.32      0.23    -0.79
## factorgics_sector_nameInformationTechnology    -0.68      0.25    -1.17
## factorgics_sector_nameMaterials                 0.00      0.26    -0.51
## factorgics_sector_nameRealEstate                0.39      0.27    -0.16
## factorgics_sector_nameUtilities                -0.14      0.25    -0.64
##                                             u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept                                       1.14 1.00     1100     1778
## log_pe                                         -0.03 1.00     2425     2578
## div_yield                                       0.15 1.00     2519     2880
## factorgics_sector_nameConsumerDiscretionary     0.22 1.00      795     1173
## factorgics_sector_nameConsumerStaples           0.75 1.00      814     1443
## factorgics_sector_nameEnergy                    0.53 1.00      913     1456
## factorgics_sector_nameFinancials                0.00 1.00      751     1248
## factorgics_sector_nameHealthCare                0.59 1.00      796     1406
## factorgics_sector_nameIndustrials               0.12 1.01      726     1349
## factorgics_sector_nameInformationTechnology    -0.21 1.00      793     1265
## factorgics_sector_nameMaterials                 0.53 1.00      907     1454
## factorgics_sector_nameRealEstate                0.91 1.00      871     1414
## factorgics_sector_nameUtilities                 0.34 1.00      808     1185
##
## 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).
```

Rhats are ~1 and effective sample sizes » 100



No discernable pattern from trace plots



Model Comparison

```
##                elpd_diff se_diff
## bayes_fe         0.0       0.0
## bayes_model2    -0.6       1.2
## bayes_model1   -22.4       8.0
## bayes_model3   -24.9       7.8
```

LOOCV favors the `pooled model with sector fixed effects` and the `model with nested random intercepts` over the `baseline pooled model` and `mixed effects model w/ firm-level random intercepts and slopes`.

The sector FE and nested random intercepts models are generally comparable (firm-level variation may be small)