# Final Project - Bayesian Analysis

```r
library(readxl)
library(dplyr)
library(tidyr)
library(lubridate)
library(stringr)
library(lme4)
```

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

## Frequentist

```
## 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
##    4632.9    4688.0   -2307.4    4614.9      3391
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

| | 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 |

```
## -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')
```

```
##
## 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
```

```
##
## 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

```
##                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
```

(to-do – LOOCV)

## Bayesian

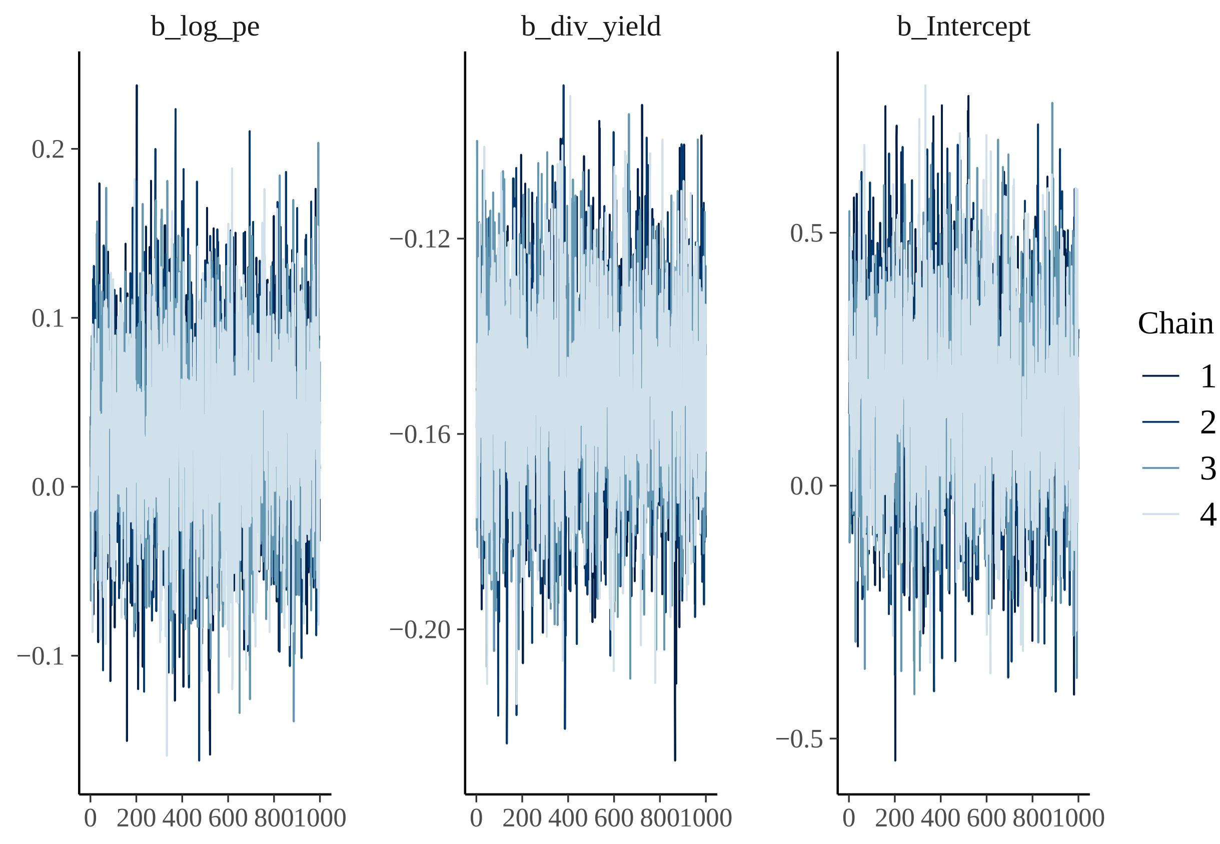
### Model 1 (Logistic)

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

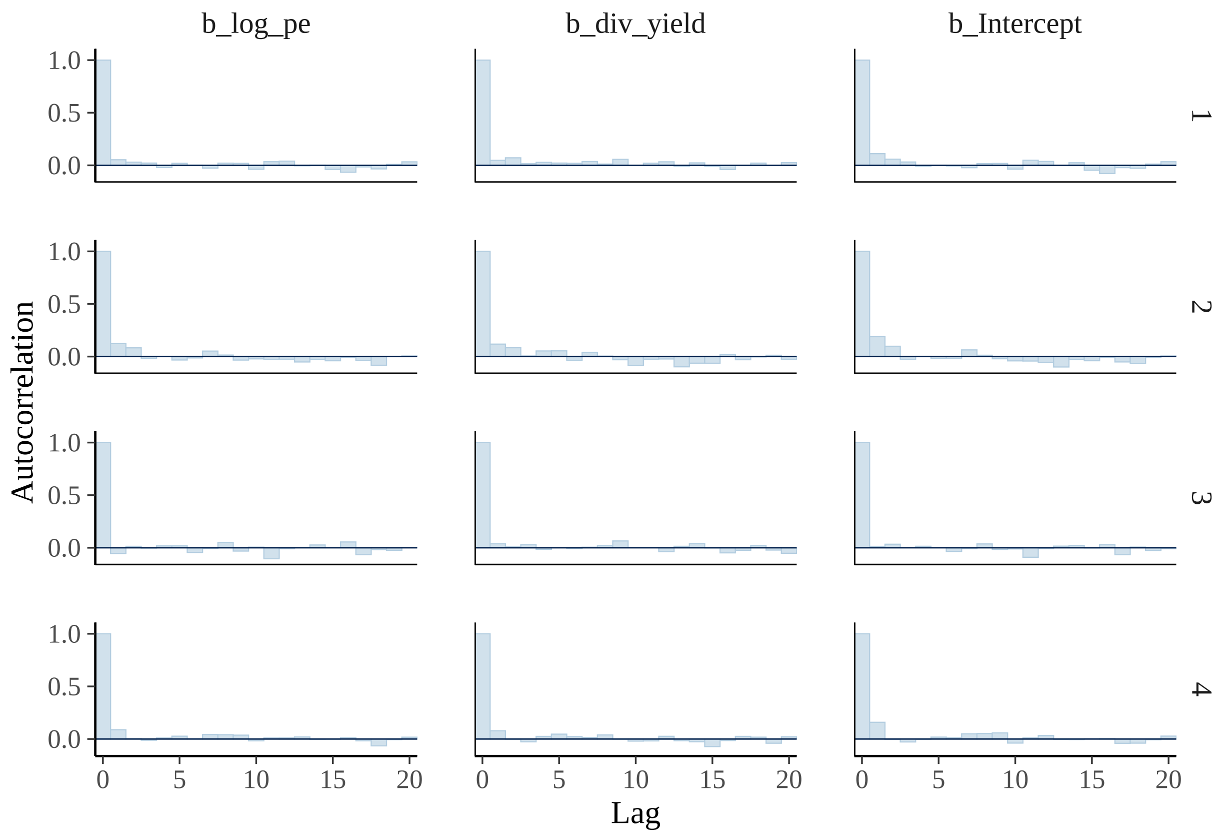
```
## Compiling Stan program...
```

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

## Start sampling

##               Estimate  Est.Error   l-95% CI   u-95% CI     Rhat Bulk_ESS
## Intercept  0.17273278 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
```

Rhats are ~1 and effective sample sizes » 100




No discernable pattern from trace plots




acfs fall off quickly

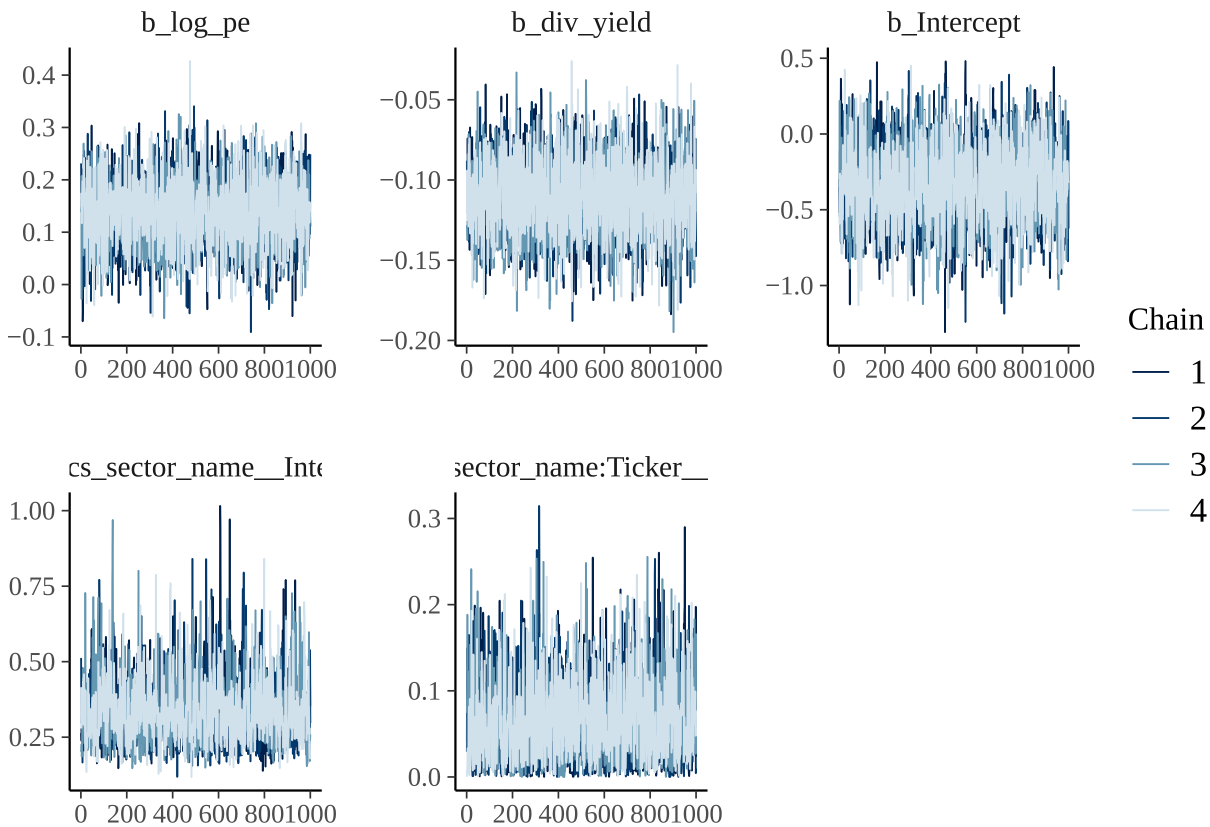
## Model 2 (Nested random intercepts)

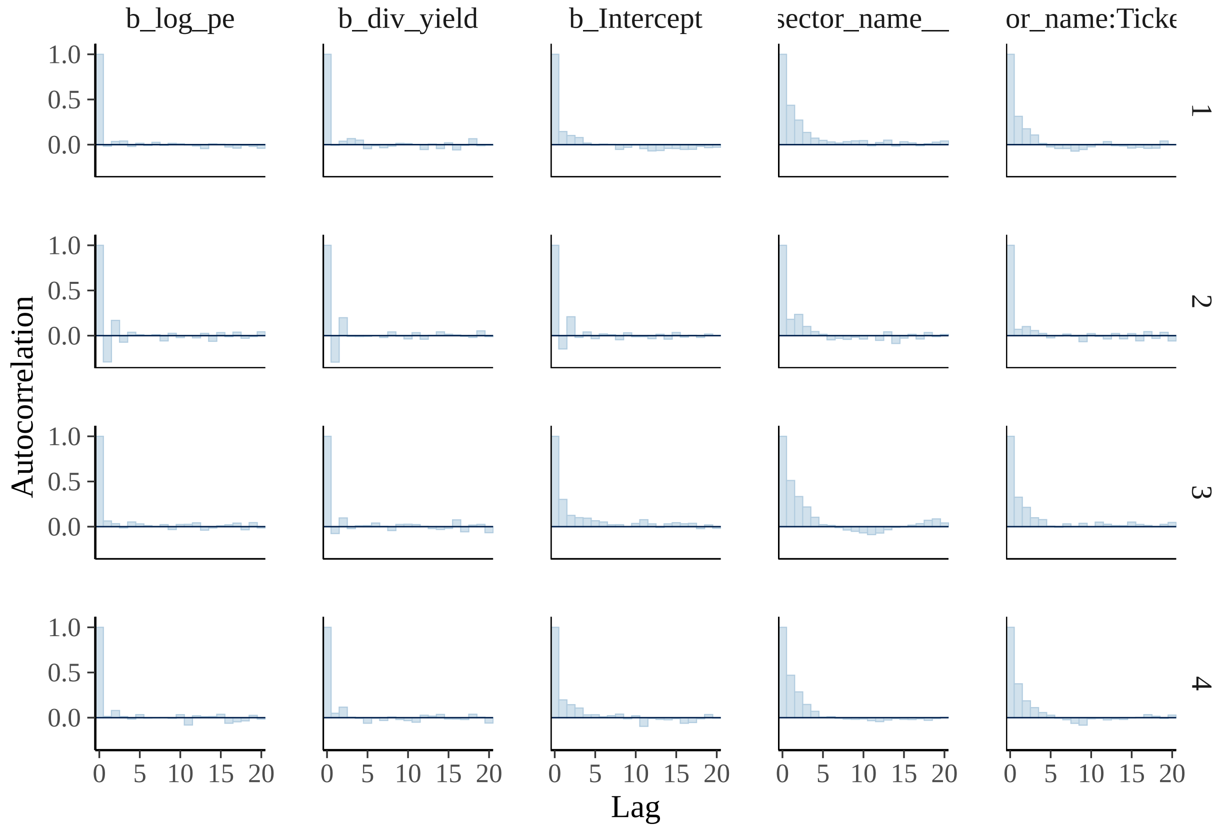
$$logit(p_{i,t}) = \beta_0 + \beta_1 log(PE_{i,t}) + \beta_2 DivYield_{i,t} + u_j + v_i$$

```
## 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)     0.34      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).
```

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...
```

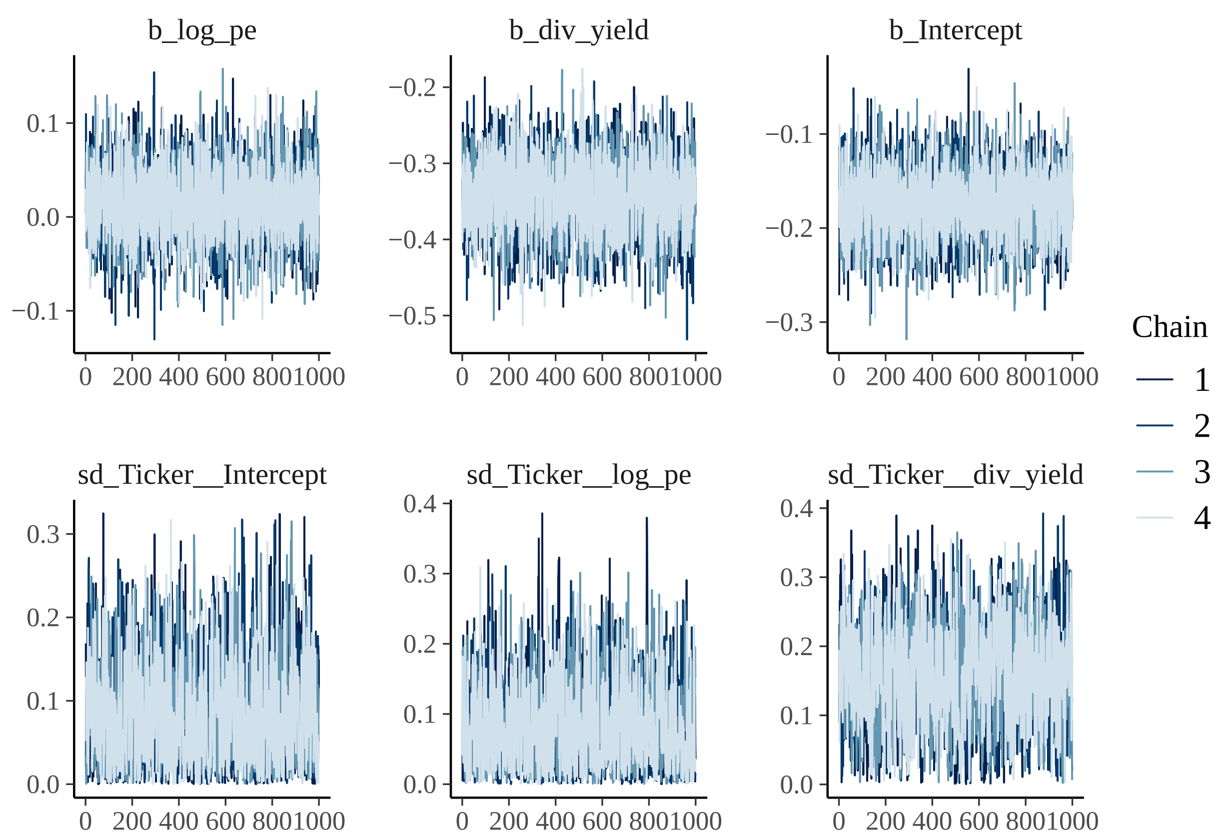
```
## Trying to compile a simple C file
```

```
## Start sampling
```

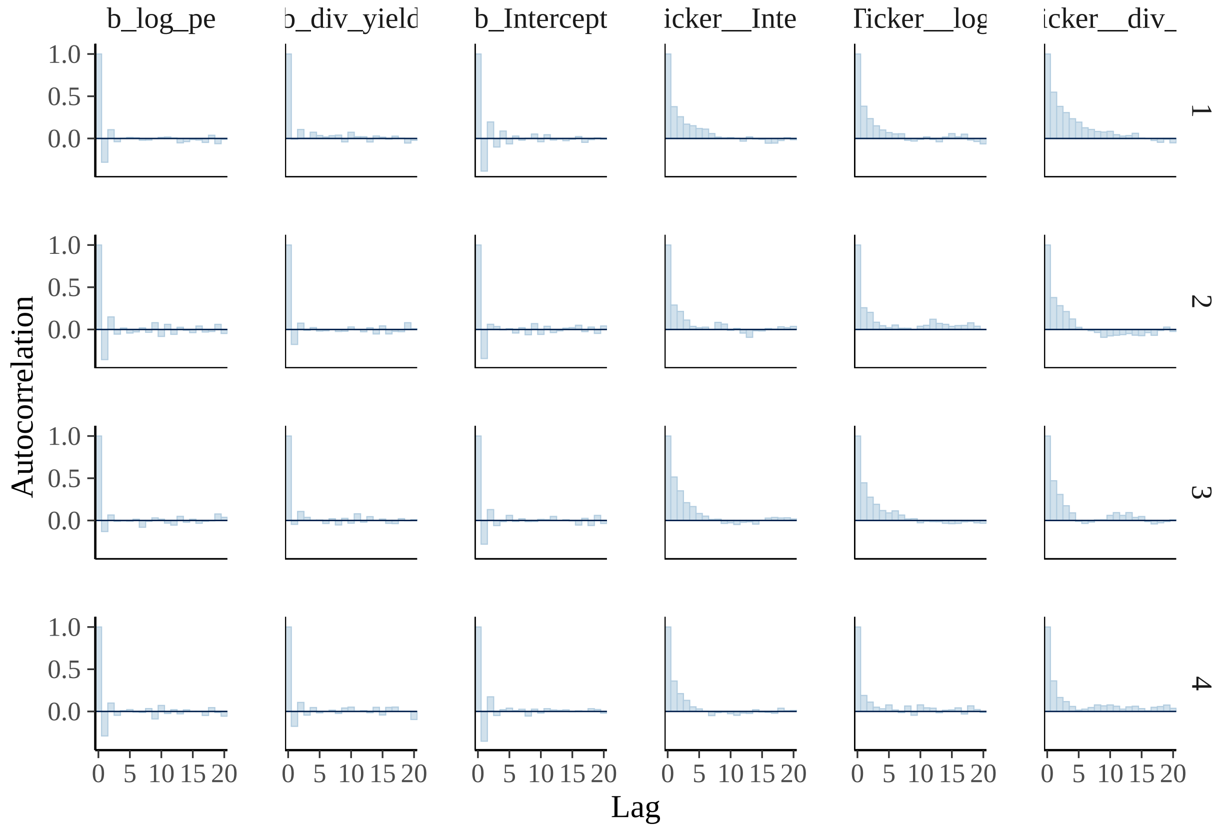
```
##  Family: bernoulli
##   Links: mu = logit
## Formula: beat_spy ~ log_pe + div_yield + (1 + log_pe + div_yield | Ticker)
##    Data: mutate(m3_df, log_pe = scale(log_pe), div_yield =  (Number of observations: 3400)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          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     1265
## sd(log_pe)                     0.08      0.06     0.00     0.23 1.00     1412
## sd(div_yield)                  0.16      0.07     0.02     0.30 1.00     1070
```

```
## cor(Intercept,log_pe)        0.03      0.50     -0.87      0.89 1.00      2685
## cor(Intercept,div_yield)     0.10      0.48     -0.83      0.90 1.00      1388
## cor(log_pe,div_yield)        0.20      0.48     -0.78      0.92 1.00      1513
##                          Tail_ESS
## sd(Intercept)                1754
## sd(log_pe)                   2025
## sd(div_yield)                1397
## cor(Intercept,log_pe)        2449
## cor(Intercept,div_yield)     2368
## cor(log_pe,div_yield)        2618
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -0.17      0.04    -0.24    -0.10 1.00     6658     3278
## log_pe        0.01      0.04    -0.06     0.09 1.00     6592     3033
## div_yield    -0.34      0.05    -0.44    -0.25 1.00     3832     3085
##
## 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

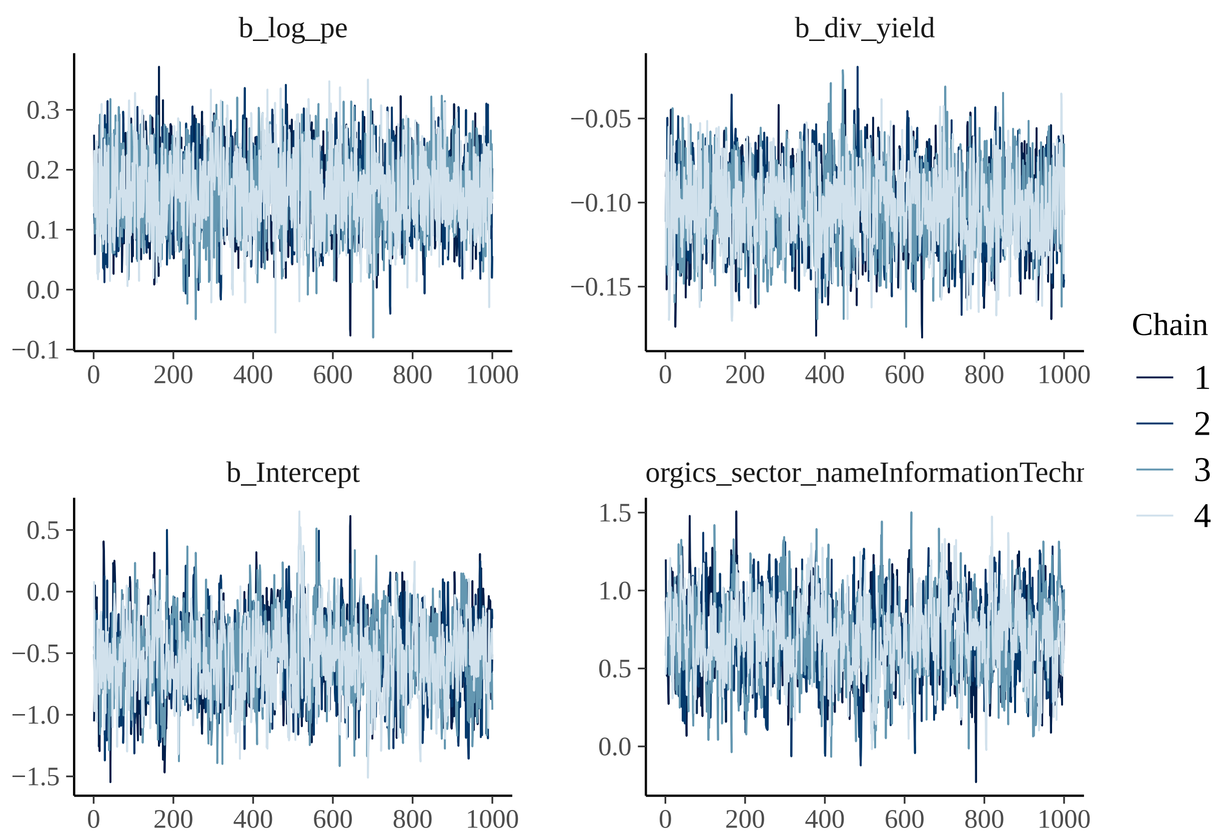



**Model 4 (Pooled + sector FE)**

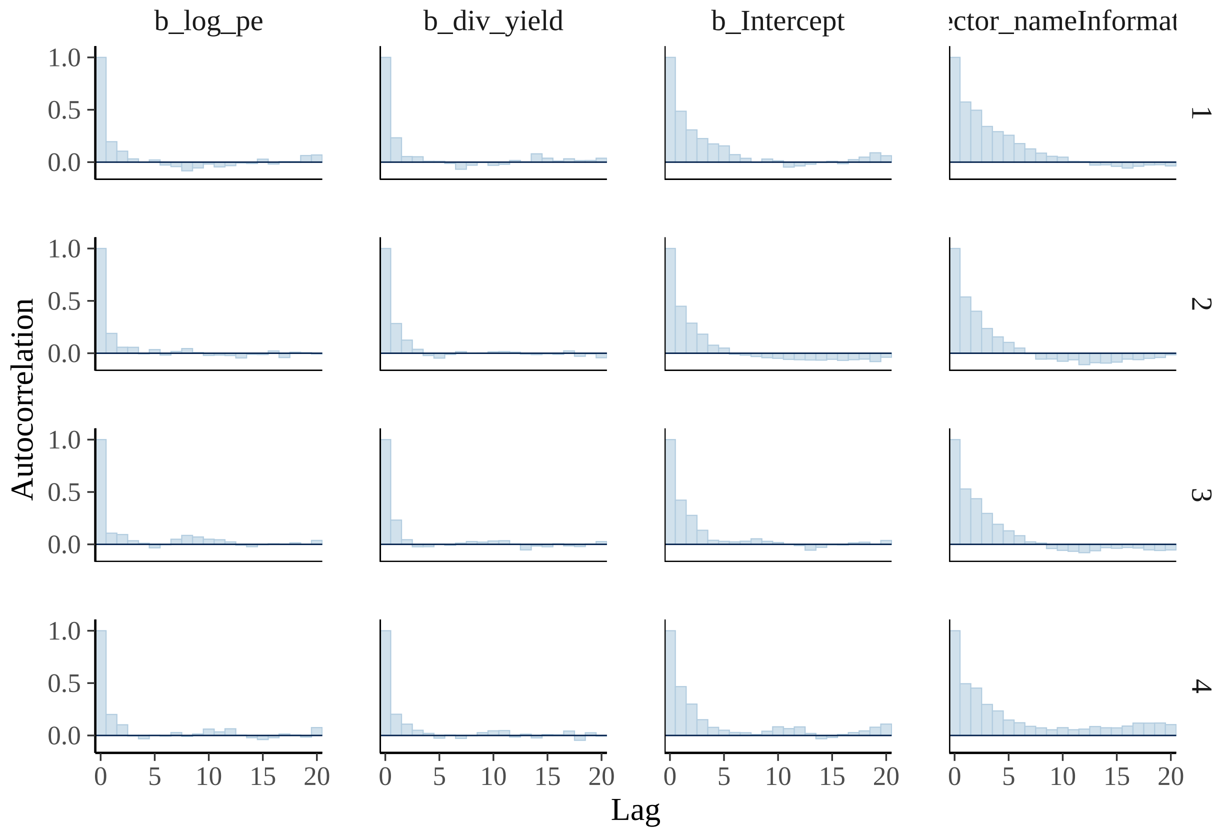
```
## Compiling Stan program...

## 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                                       -0.53      0.31    -1.15
## log_pe                                           0.16      0.07     0.04
## div_yield                                       -0.10      0.02    -0.15
## factorgics_sector_nameConsumerDiscretionary      0.28      0.25    -0.23
## factorgics_sector_nameConsumerStaples           -0.25      0.25    -0.75
## factorgics_sector_nameEnergy                    -0.01      0.27    -0.53
## factorgics_sector_nameFinancials                 0.45      0.23    -0.00
```

```
## factorgics_sector_nameHealthCare                -0.11      0.25    -0.61
## factorgics_sector_nameIndustrials                0.33      0.23    -0.13
## factorgics_sector_nameInformationTechnology      0.69      0.25     0.21
## factorgics_sector_nameMaterials                  0.01      0.26    -0.52
## factorgics_sector_nameRealEstate                -0.38      0.27    -0.91
## factorgics_sector_nameUtilities                  0.15      0.25    -0.35
##                                              u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept                                        0.08 1.00     1206     1782
## log_pe                                           0.29 1.00     2550     2373
## div_yield                                       -0.06 1.00     2367     2615
## factorgics_sector_nameConsumerDiscretionary      0.78 1.00      889     1279
## factorgics_sector_nameConsumerStaples            0.24 1.00      837     1187
## factorgics_sector_nameEnergy                     0.52 1.00      945     1424
## factorgics_sector_nameFinancials                 0.92 1.00      784     1064
## factorgics_sector_nameHealthCare                 0.39 1.00      877     1298
## factorgics_sector_nameIndustrials                0.79 1.00      764     1071
## factorgics_sector_nameInformationTechnology      1.18 1.00      855     1431
## factorgics_sector_nameMaterials                  0.52 1.00      915     1599
## factorgics_sector_nameRealEstate                 0.16 1.00      913     1482
## factorgics_sector_nameUtilities                  0.65 1.00      871     1256
##
## 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     -0.3       1.2
## bayes_model1    -22.5       8.1
## bayes_model3    -24.9       7.9
```

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)