Pew Simulation Results

Valerie Bradley

September 5, 2019

The purpose of this simulation is to test the efficacy of distribution regression for predicting individual-level political support from political polling data.

1 the data

Pew Research conducts regular public opinion research polls measuring poltical attitudes in the US, like this political survey from September 2018. We use 4 surveys fielded over the 6 months leading up to the 2018 US miderm elections that all ask respondents which party they plan to support in the upcoming election, in addition to a selection of demographic variables (e.g. age, income bracket, education, race, etc.).

```
pew_data = fread('~/github/bdr/data/data_recoded.csv')
```

The data contains `pew_data[, .N]` total responses collected from live interviews on landlines and cell phones.

2 results

2.1 performance relative to benchmarks

2.2 optimal kernel parameters?

2.3 does weighting help?

```
mse_files = list.files('~/github/bdr/pew-experiment/results/sim_randparams/', pattern = 'mse_', full.nam
mses = rbindlist(lapply(mse_files, function(f) fread(f)))
mses[, match_rate_bkt := floor(match_rate * 5)]

mses[, length(unique(results_id))]
```

```
## [1] 529
```

The models considered here are variations of basic distribution regression models.

```
model_desc_tab = data.frame(model_name = unique(mses$model))
model_desc_tab$kernel = ifelse(grepl('cust',model_desc_tab$model_name), "custom"
                               , ifelse(grepl('linear',model_desc_tab$model_name), "linear"
                                        , ifelse(grepl('dr',model_desc_tab$model_name), "rbf", "")))
model_desc_tab$weighted = ifelse(grepl('wdr',model_desc_tab$model_name), "X", "")
model_desc_tab$separate_bags = ifelse(grepl('sepbags',model_desc_tab$model_name), "X", "")

kable(model_desc_tab)
```

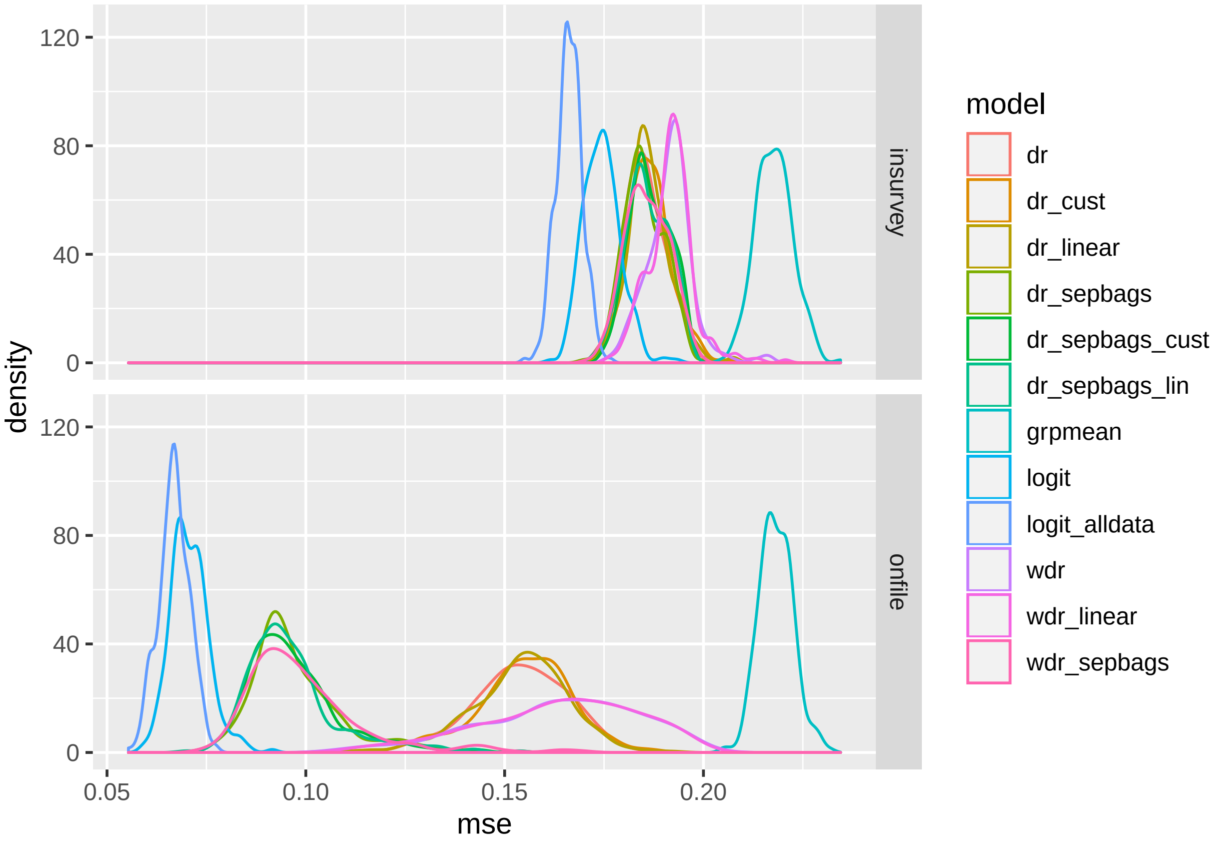
model_name	kernel	weighted	separate_bags
logit			
logit_alldata			

model_name	kernel	weighted	separate_bags
dr_linear	linear		
wdr_linear	linear	X	
dr	rbf		
wdr	rbf	X	
dr_cust	custom		
dr_sepbags	rbf		X
wdr_sepbags	rbf	X	X
dr_sepbags_lin	rbf		X
dr_sepbags_cust	custom		X
grpmean			

```r
pred_files = list.files('~/github/bdr/pew-experiment/results/sim_randparams', pattern = '^party', full.

holdout_error = rbindlist(lapply(pred_files, function(f){
  temp = fread(f)
  holdout_ind = which(temp[model == 'logit',]$holdout == 1)

  temp$act_class = rep(pew_data$support, length(unique(temp$model)))
  temp[, pred_class := c('1-Dem', '2-Rep', '3-Oth')[apply(temp[, .(y_hat_dem, y_hat_rep, y_hat_oth)], 1
  temp[, correct_class := as.numeric(act_class == pred_class)]

  holdout_error = cbind(temp[holdout == 1, .(y_hat_dem = mean(y_hat_dem)
                                      , y_hat_rep = mean(y_hat_rep)
                                      , y_hat_oth = mean(y_hat_oth)
                                      , class_rate = mean(correct_class)
                                      ), by = .(model, results_id, match_rate, n_bags, n_landmarks, refi
  , pew_data[holdout_ind, .(y_dem = mean(y_dem)
                                , y_rep = mean(y_rep)
                                , y_oth = mean(y_oth)
                                )]
        )
  holdout_error[, y_hat_dem_2way := y_hat_dem/(1 - y_hat_oth)]

  holdout_error[, error_dem := y_hat_dem - y_dem]
  holdout_error[, error_rep := y_hat_rep - y_rep]
  holdout_error[, error_oth := y_hat_oth - y_oth]
  holdout_error[, error_dem_2way := y_hat_dem_2way - (y_dem/(1-y_oth))]

  holdout_error
}))

holdout_error
```

```
##                  model                                           results_id
##    1:            logit partyinsurvey_match100_bags136_lmks26_refitbagsFALSE
##    2:    logit_alldata partyinsurvey_match100_bags136_lmks26_refitbagsFALSE
##    3:        dr_linear partyinsurvey_match100_bags136_lmks26_refitbagsFALSE
##    4:       wdr_linear partyinsurvey_match100_bags136_lmks26_refitbagsFALSE
##    5:               dr partyinsurvey_match100_bags136_lmks26_refitbagsFALSE
##   ---
## 6104:       dr_sepbags     partyonfile_match99_bags80_lmks68_refitbagsFALSE
## 6105:      wdr_sepbags     partyonfile_match99_bags80_lmks68_refitbagsFALSE
## 6106:  dr_sepbags_lin     partyonfile_match99_bags80_lmks68_refitbagsFALSE
## 6107: dr_sepbags_cust     partyonfile_match99_bags80_lmks68_refitbagsFALSE
## 6108:         grpmean     partyonfile_match99_bags80_lmks68_refitbagsFALSE
##       match_rate n_bags n_landmarks refit_bags    party y_hat_dem
##    1:  0.9974763    136           26      FALSE insurvey 0.4915911
##    2:  0.9974763    136           26      FALSE insurvey 0.4753241
##    3:  0.9974763    136           26      FALSE insurvey 0.5038523
##    4:  0.9974763    136           26      FALSE insurvey 0.5189263
##    5:  0.9974763    136           26      FALSE insurvey 0.5035291
##   ---
## 6104:  0.9896966     80           68      FALSE   onfile 0.5205693
## 6105:  0.9896966     80           68      FALSE   onfile 0.4988696
## 6106:  0.9896966     80           68      FALSE   onfile 0.5153288
## 6107:  0.9896966     80           68      FALSE   onfile 0.5201237
## 6108:  0.9896966     80           68      FALSE   onfile 0.5173217
##       y_hat_rep  y_hat_oth class_rate y_dem y_rep y_oth y_hat_dem_2way
##    1: 0.4405297 0.06787916      0.609 0.487 0.415 0.098      0.5273899
##    2: 0.4320492 0.09262665      0.613 0.487 0.415 0.098      0.5238462
##    3: 0.4246997 0.07144791      0.539 0.487 0.415 0.098      0.5426215
##    4: 0.3988787 0.08219503      0.507 0.487 0.415 0.098      0.5653993
##    5: 0.4242549 0.07221601      0.536 0.487 0.415 0.098      0.5427223
##   ---
## 6104: 0.4054501 0.07398059      0.823 0.503 0.399 0.098      0.5621581
## 6105: 0.4169994 0.08413109      0.822 0.503 0.399 0.098      0.5446954
## 6106: 0.4084171 0.07625406      0.826 0.503 0.399 0.098      0.5578686
## 6107: 0.4121137 0.06776265      0.827 0.503 0.399 0.098      0.5579305
## 6108: 0.4144835 0.06819483      0.459 0.503 0.399 0.098      0.5551823
##          error_dem    error_rep    error_oth error_dem_2way
##    1:  0.004591092  0.025529743 -0.030120836  -0.0125214317
##    2: -0.011675877  0.017049231 -0.005373354  -0.0160650647
##    3:  0.016852341  0.009699747 -0.026552088   0.0027102069
##    4:  0.031926285 -0.016121310 -0.015804975   0.0254879861
##    5:  0.016529056  0.009254939 -0.025783994   0.0028109834
##   ---
## 6104:  0.017569298  0.006450114 -0.024019412   0.0045084163
## 6105: -0.004130443  0.017999355 -0.013868912  -0.0129542962
## 6106:  0.012328811  0.009417127 -0.021745938   0.0002188873
## 6107:  0.017123683  0.013113667 -0.030237350   0.0002808669
## 6108:  0.014321697  0.015483470 -0.029805168  -0.0024674089
```

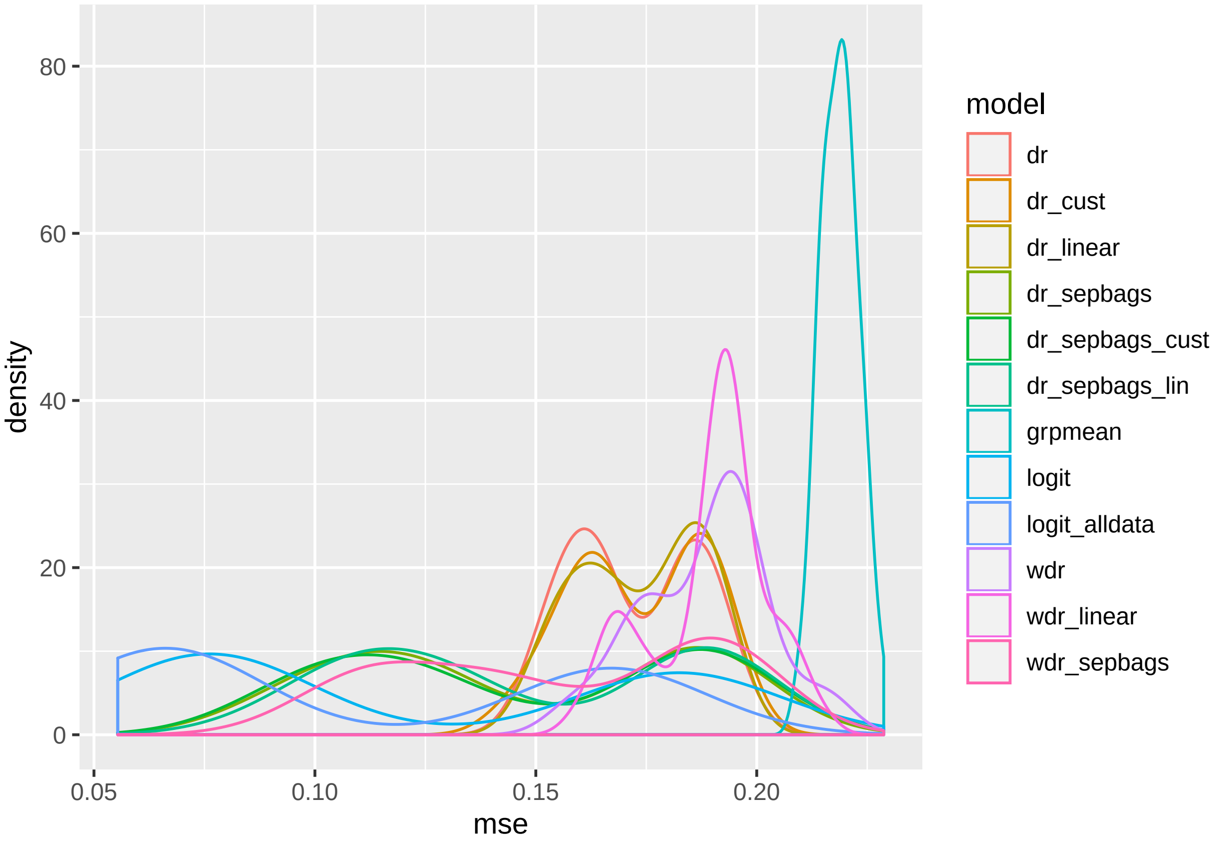
```r
ggplot(mses, aes(x = mse, color = model)) + geom_density() + facet_grid(party~.)
```



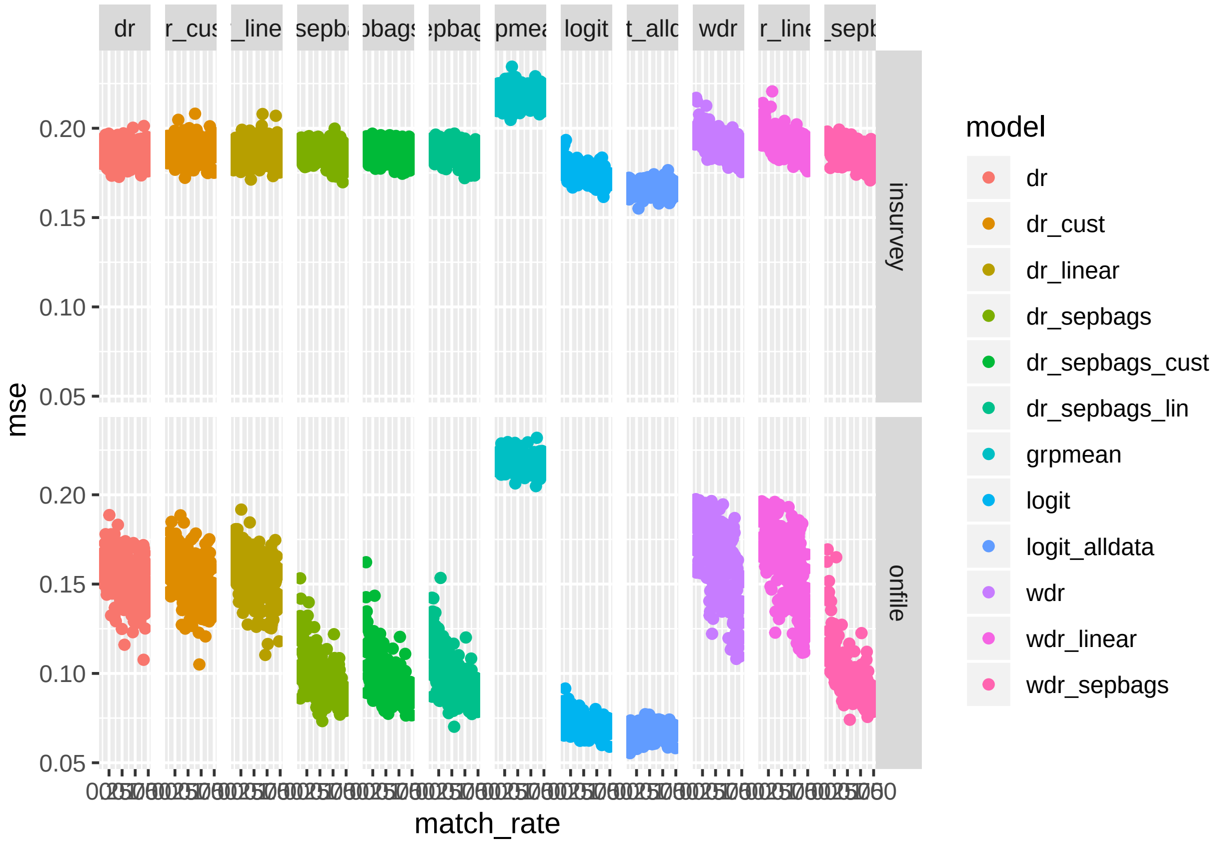
```r
mses[mse_rellogit < 1 & model != 'logit_alldata']
```

```
##                                            results_id match_rate n_bags
##  1:  partyinsurvey_match11_bags104_lmks41_refitbagsFALSE  0.1107272    104
##  2:  partyinsurvey_match11_bags104_lmks41_refitbagsFALSE  0.1107272    104
##  3:  partyinsurvey_match11_bags104_lmks41_refitbagsFALSE  0.1107272    104
##  4:   partyinsurvey_match11_bags137_lmks35_refitbagsTRUE  0.1073929    137
##  5:   partyinsurvey_match11_bags137_lmks35_refitbagsTRUE  0.1073929    137
##  6:   partyinsurvey_match11_bags137_lmks35_refitbagsTRUE  0.1073929    137
##  7:   partyinsurvey_match11_bags137_lmks35_refitbagsTRUE  0.1073929    137
##  8:  partyinsurvey_match11_bags66_lmks104_refitbagsFALSE  0.1072078     66
##  9:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 10:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 11:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 12:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 13:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 14:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 15:  partyinsurvey_match13_bags43_lmks168_refitbagsFALSE  0.1305678     43
## 16:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 17:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 18:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 19:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 20:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 21:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 22:   partyinsurvey_match16_bags81_lmks138_refitbagsTRUE  0.1606479     81
## 23:  partyinsurvey_match17_bags147_lmks211_refitbagsTRUE  0.1725560    147
## 24:  partyinsurvey_match21_bags83_lmks146_refitbagsFALSE  0.2137370     83
## 25: partyinsurvey_match25_bags122_lmks117_refitbagsFALSE  0.2519162    122
## 26: partyinsurvey_match25_bags122_lmks117_refitbagsFALSE  0.2519162    122
## 27: partyinsurvey_match25_bags122_lmks117_refitbagsFALSE  0.2519162    122
## 28:   partyinsurvey_match27_bags63_lmks228_refitbagsTRUE  0.2687904     63
## 29:   partyinsurvey_match28_bags94_lmks96_refitbagsFALSE  0.2797190     94
## 30:   partyinsurvey_match28_bags94_lmks96_refitbagsFALSE  0.2797190     94
## 31:  partyinsurvey_match37_bags63_lmks115_refitbagsFALSE  0.3728651     63
## 32:  partyinsurvey_match37_bags63_lmks115_refitbagsFALSE  0.3728651     63
## 33:  partyinsurvey_match37_bags63_lmks115_refitbagsFALSE  0.3728651     63
## 34:  partyinsurvey_match51_bags113_lmks94_refitbagsFALSE  0.5107417    113
## 35:  partyinsurvey_match51_bags113_lmks94_refitbagsFALSE  0.5107417    113
## 36:  partyinsurvey_match51_bags113_lmks94_refitbagsFALSE  0.5107417    113
## 37:    partyinsurvey_match51_bags91_lmks62_refitbagsTRUE  0.5061082     91
## 38:    partyinsurvey_match94_bags57_lmks89_refitbagsTRUE  0.9356278     57
##                                                results_id match_rate n_bags
##     n_landmarks refit_bags    party           model       mse mse_relall
##  1:          41      FALSE insurvey       dr_linear 0.1764472   1.057238
##  2:          41      FALSE insurvey              dr 0.1800281   1.078694
##  3:          41      FALSE insurvey         dr_cust 0.1803816   1.080812
##  4:          35       TRUE insurvey              dr 0.1813353   1.132252
##  5:          35       TRUE insurvey         dr_cust 0.1828413   1.141655
##  6:          35       TRUE insurvey      dr_sepbags 0.1824568   1.139255
##  7:          35       TRUE insurvey dr_sepbags_cust 0.1842207   1.150268
##  8:         104      FALSE insurvey dr_sepbags_cust 0.1882692   1.121439
##  9:         168      FALSE insurvey       dr_linear 0.1811343   1.084551
## 10:         168      FALSE insurvey      wdr_linear 0.1902896   1.139369
## 11:         168      FALSE insurvey              dr 0.1804500   1.080454
## 12:         168      FALSE insurvey         dr_cust 0.1859140   1.113170
## 13:         168      FALSE insurvey      dr_sepbags 0.1856473   1.111573
## 14:         168      FALSE insurvey  dr_sepbags_lin 0.1833073   1.097562
## 15:         168      FALSE insurvey dr_sepbags_cust 0.1821746   1.090780
## 16:         138       TRUE insurvey       dr_linear 0.1834689   1.101961
## 17:         138       TRUE insurvey              dr 0.1818670   1.092339
## 18:         138       TRUE insurvey         dr_cust 0.1817845   1.091844
## 19:         138       TRUE insurvey      dr_sepbags 0.1895380   1.138413
## 20:         138       TRUE insurvey     wdr_sepbags 0.1863657   1.119360
## 21:         138       TRUE insurvey  dr_sepbags_lin 0.1915414   1.150446
## 22:         138       TRUE insurvey dr_sepbags_cust 0.1843398   1.107191
## 23:         211       TRUE insurvey     wdr_sepbags 0.1777601   1.070404
## 24:         146      FALSE insurvey         dr_cust 0.1820750   1.094845
## 25:         117      FALSE insurvey       dr_linear 0.1786261   1.078012
## 26:         117      FALSE insurvey              dr 0.1774758   1.071070
## 27:         117      FALSE insurvey         dr_cust 0.1767657   1.066785
## 28:         228       TRUE insurvey      dr_sepbags 0.1800516   1.087599
## 29:          96      FALSE insurvey         dr_cust 0.1810981   1.109021
## 30:          96      FALSE insurvey  dr_sepbags_lin 0.1811810   1.109528
## 31:         115      FALSE insurvey       dr_linear 0.1806476   1.097175
## 32:         115      FALSE insurvey              dr 0.1798089   1.092081
## 33:         115      FALSE insurvey         dr_cust 0.1796536   1.091138
## 34:          94      FALSE insurvey       dr_linear 0.1775567   1.065671
## 35:          94      FALSE insurvey              dr 0.1763791   1.058604
## 36:          94      FALSE insurvey         dr_cust 0.1761152   1.057020
## 37:          62       TRUE insurvey     wdr_sepbags 0.1793064   1.051247
## 38:          89       TRUE insurvey      dr_sepbags 0.1697721   1.042805
##     n_landmarks refit_bags    party           model       mse mse_relall
##     mse_rellogit match_rate_bkt
##  1:    0.9604413              0
##  2:    0.9799333              0
##  3:    0.9818573              0
##  4:    0.9781333              0
##  5:    0.9862565              0
##  6:    0.9841827              0
##  7:    0.9936970              0
##  8:    0.9957487              0
##  9:    0.9490358              0
## 10:    0.9970039              0
## 11:    0.9454506              0
## 12:    0.9740787              0
## 13:    0.9726813              0
## 14:    0.9604211              0
## 15:    0.9544865              0
## 16:    0.9485904              0
## 17:    0.9403084              0
## 18:    0.9398816              0
## 19:    0.9799698              0
## 20:    0.9635681              0
## 21:    0.9903277              0
## 22:    0.9530932              0
## 23:    0.9925986              0
## 24:    0.9993024              1
## 25:    0.9886046              1
## 26:    0.9822381              1
## 27:    0.9783082              1
## 28:    0.9991811              1
## 29:    0.9949866              1
## 30:    0.9954423              1
## 31:    0.9847430              1
## 32:    0.9801707              1
## 33:    0.9793243              1
## 34:    0.9946604              2
## 35:    0.9880636              2
## 36:    0.9865852              2
## 37:    0.9937872              2
## 38:    0.9970783              4
##     mse_rellogit match_rate_bkt
```

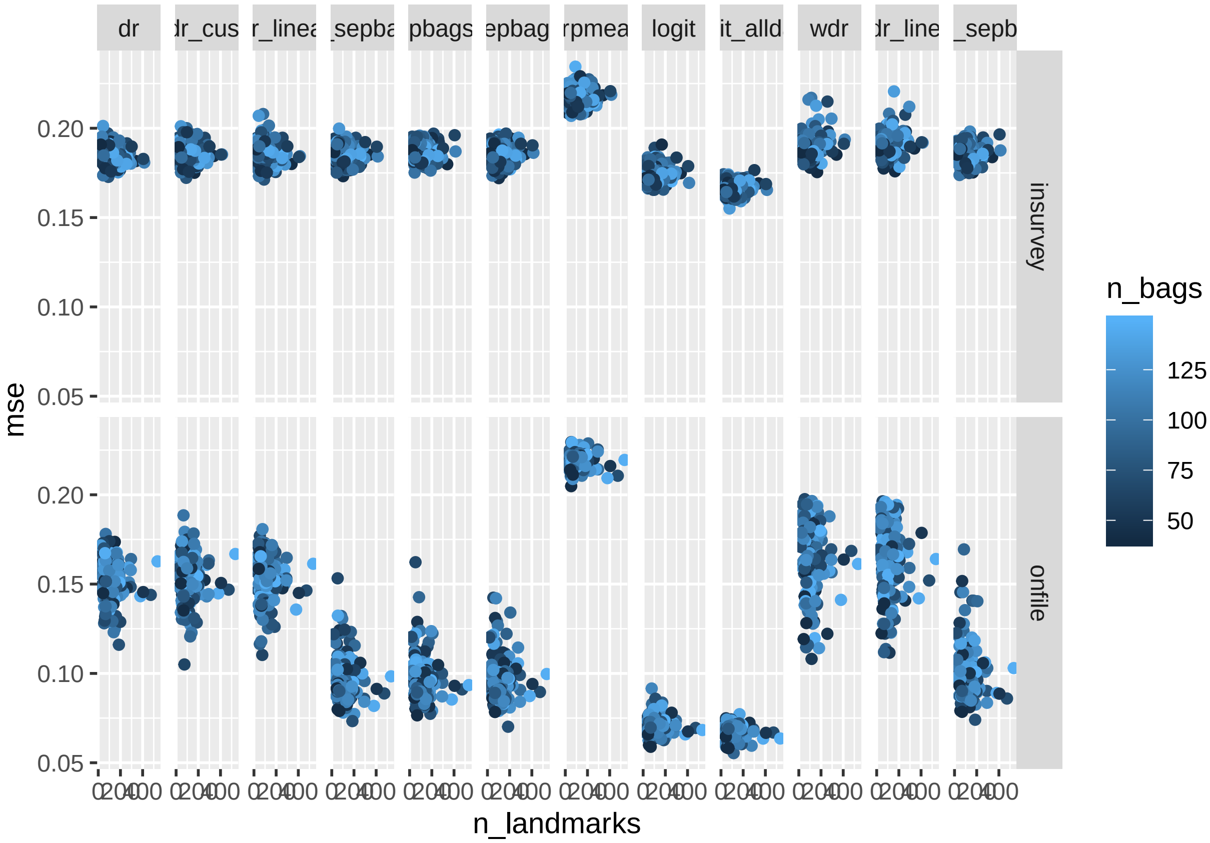
```r
ggplot(mses[match_rate < 0.2 ], aes(x = mse, color = model)) + geom_density()
```



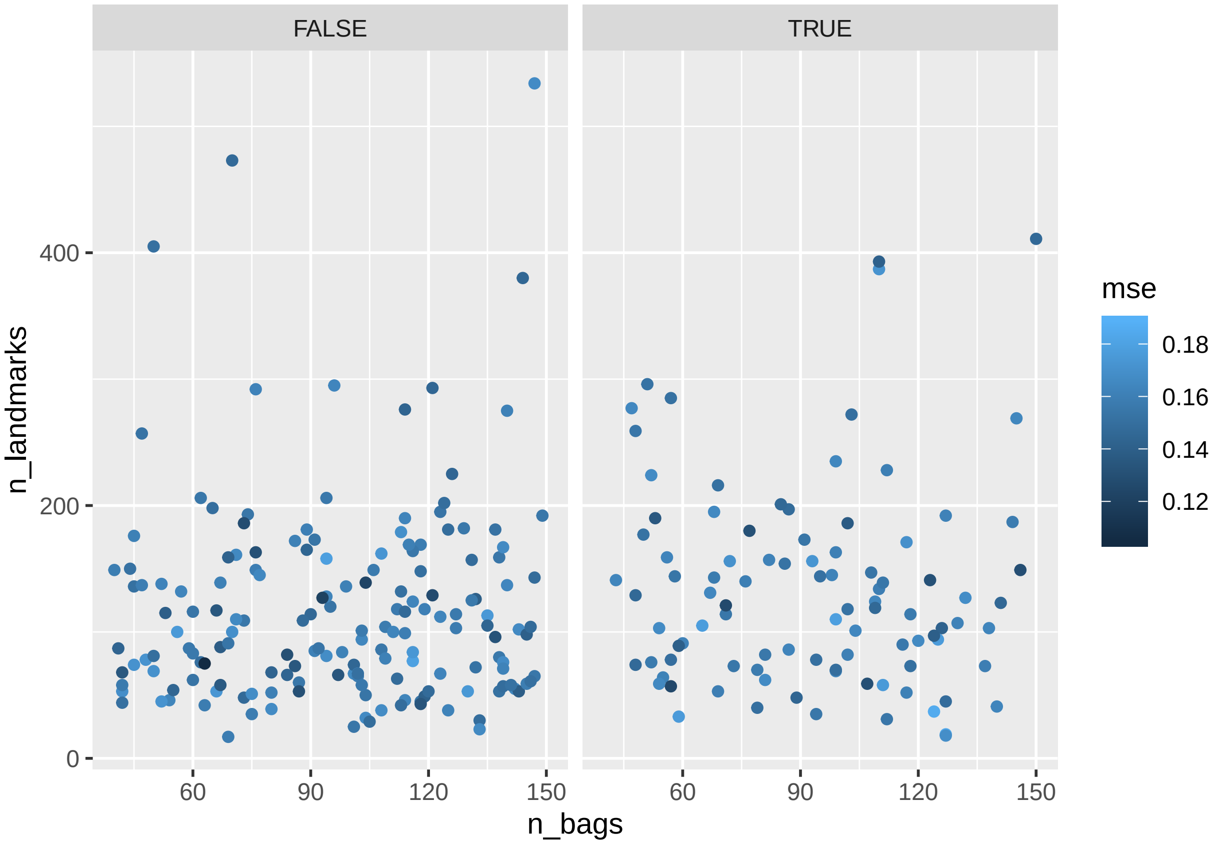
```
ggplot(mses, aes(x = match_rate, y = mse, color = model)) +
  geom_point() +
  #geom_smooth() +
  facet_grid(party~model)
```



```
ggplot(mses[refit_bags == F], aes(x = n_landmarks, y = mse, color = n_bags)) +
  geom_point() +
  #geom_smooth() +
  facet_grid(party~model)
```

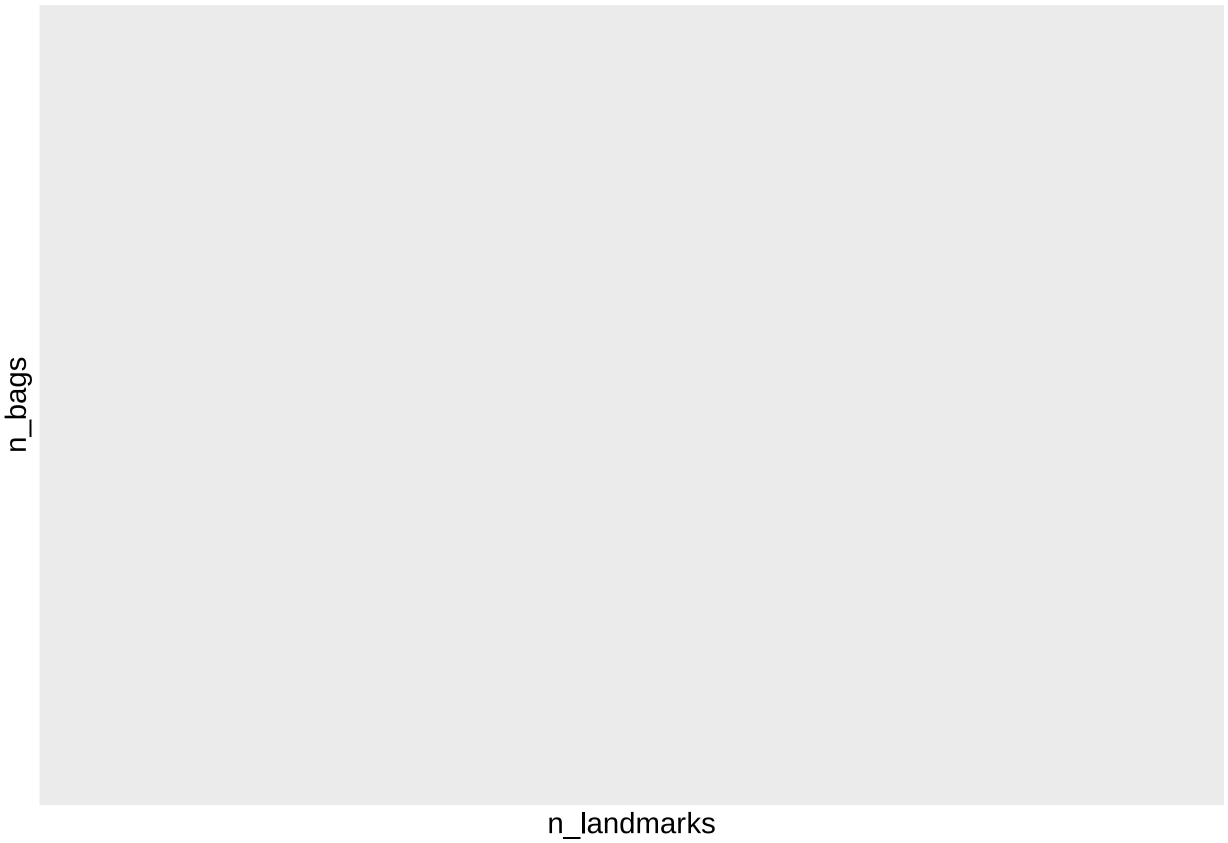


```r
ggplot(mses[model == 'dr_cust' & party == 'onfile']) +
  geom_point(aes(x = n_bags, y = n_landmarks, color = mse)) +
  facet_grid(~refit_bags)
```

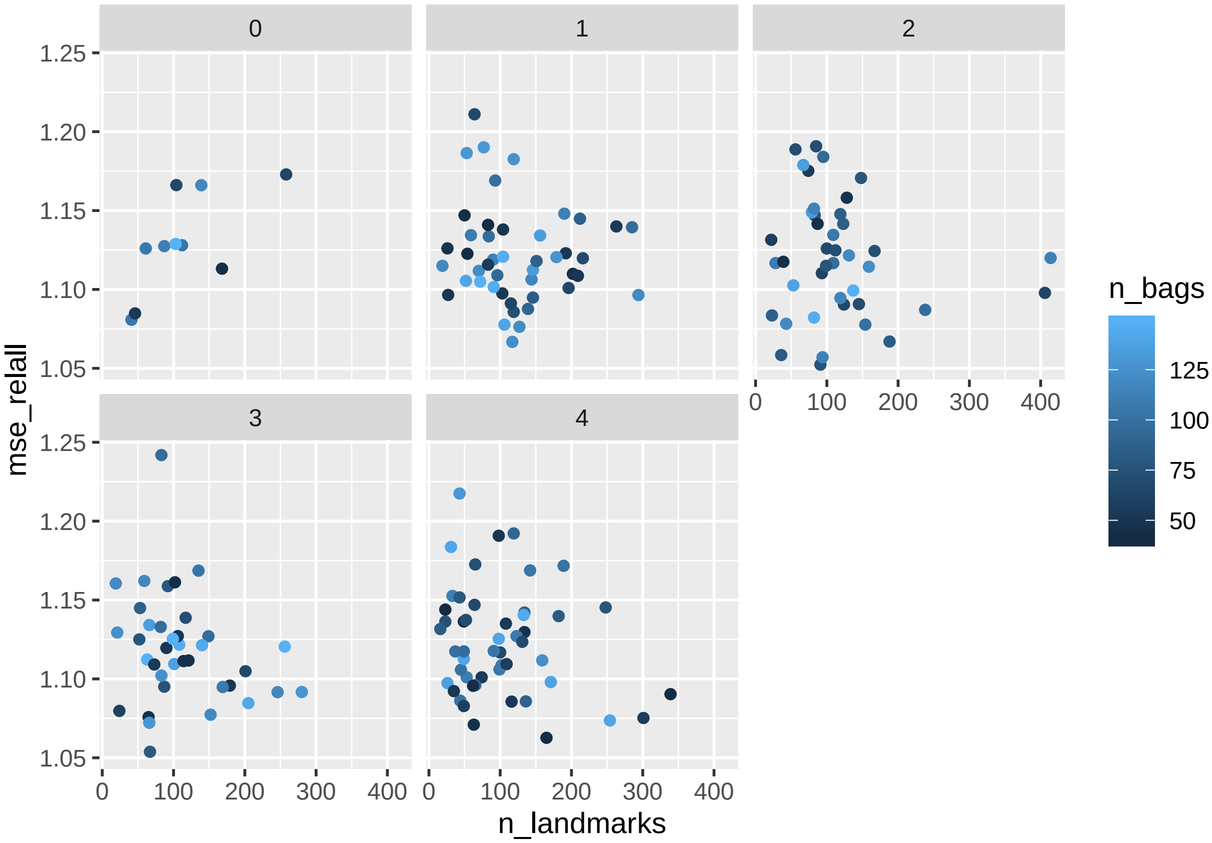


```
ggplot(mses[model == 'dr_cust']) + geom_contour(aes(x = n_landmarks, y = n_bags, z = mse), bins = 2)
```

```
## Warning: Computation failed in `stat_contour()`:
## Contour requires single `z` at each combination of `x` and `y`.
```



```r
ggplot(mses[model == 'dr_cust' & party == 'insurvey' & refit_bags == F]) +
  geom_point(aes(x = n_landmarks, y = mse_relall, color = n_bags)) + facet_wrap(~match_rate_bkt)
```



```
#
#
# ggplot(holdout_error[model %in% c('logit', 'logit_alldata', 'dr', 'dr_sepbags', 'wdr')]) +
#   geom_density(aes(x = error_dem, color= model)) +
#
#   facet_grid(~party)
#
# ggplot(holdout_error) +
#   geom_density(aes(x = error_rep, color= model)) +
#   facet_grid(~party)
#
# ggplot(holdout_error) +
#   geom_density(aes(x = error_oth, color= model)) +
#   facet_grid(~party)
#
#
# ggplot(holdout_error) +
#   geom_density(aes(x = error_dem_2way, color= model)) +
#   facet_grid(~party)
#
# # classification rate
# ggplot(holdout_error[model %in% c('logit', 'logit_alldata', 'dr', 'dr_sepbags', 'wdr')]) +
#   geom_density(aes(x = class_rate, color = model)) +
#   facet_grid(~party)
```