# Modeling Workspace

*William Morgan*

*10 May, 2018*

## Problem Outline

What do we want to answer?

- *Q1:* Does faculty activity on discussion forums affect student engagement on forums?
- *Q2:* How do student outcomes change with the level of engagement in online forums? If there is a change, it is independent of who is doing the posting?

---

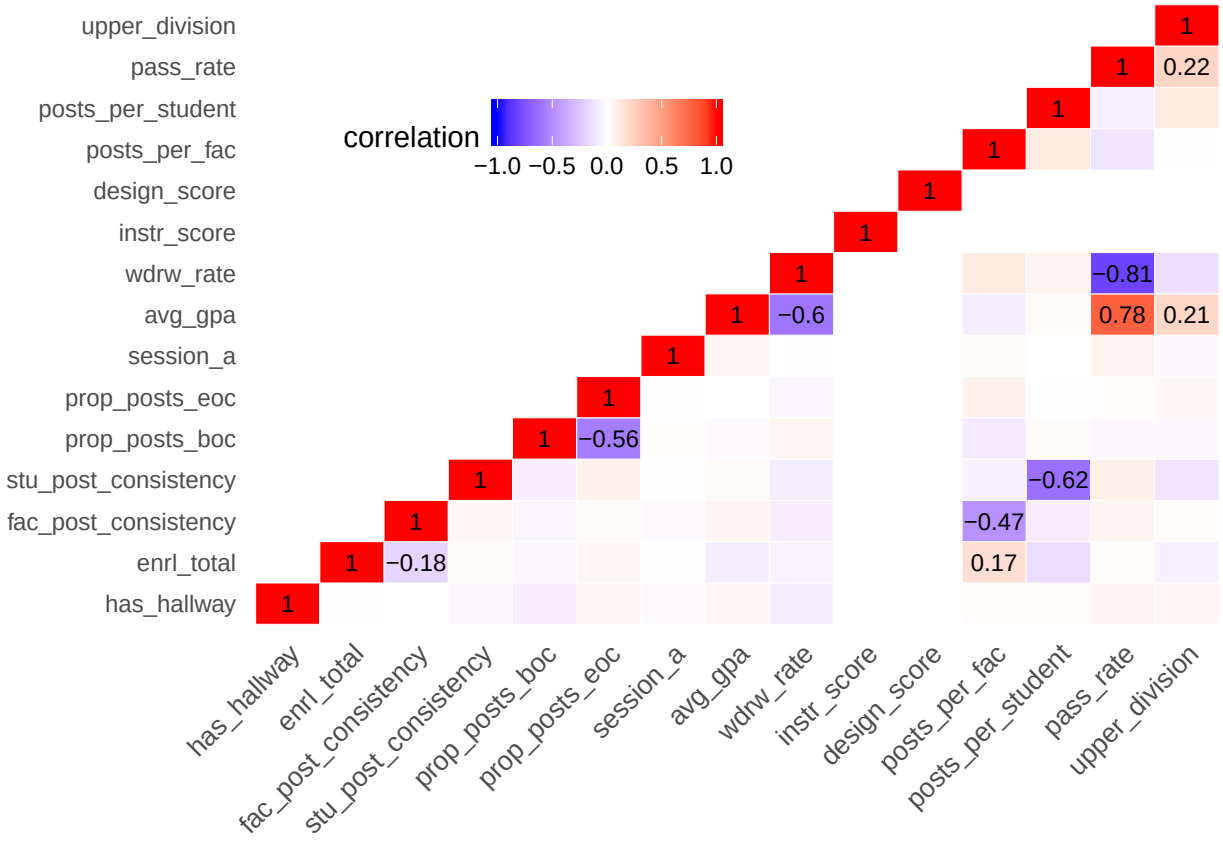
## Preliminary Feature Search

This section will be relatively brief, as we prefer to avoid a deep dive into the potential feature space. Instead, we will just use the Lasso on our original set of chosen variables and then run a plain linear model (without regularization) to get at the standard errors of the estimates.

We have expectations on which variables might carry the most signal, but it is worth doing some investigation in case there is something we overlooked. Our first objective will be creating a correlation heat map of all the variables that we anticipate using, including the responses we wish to test.

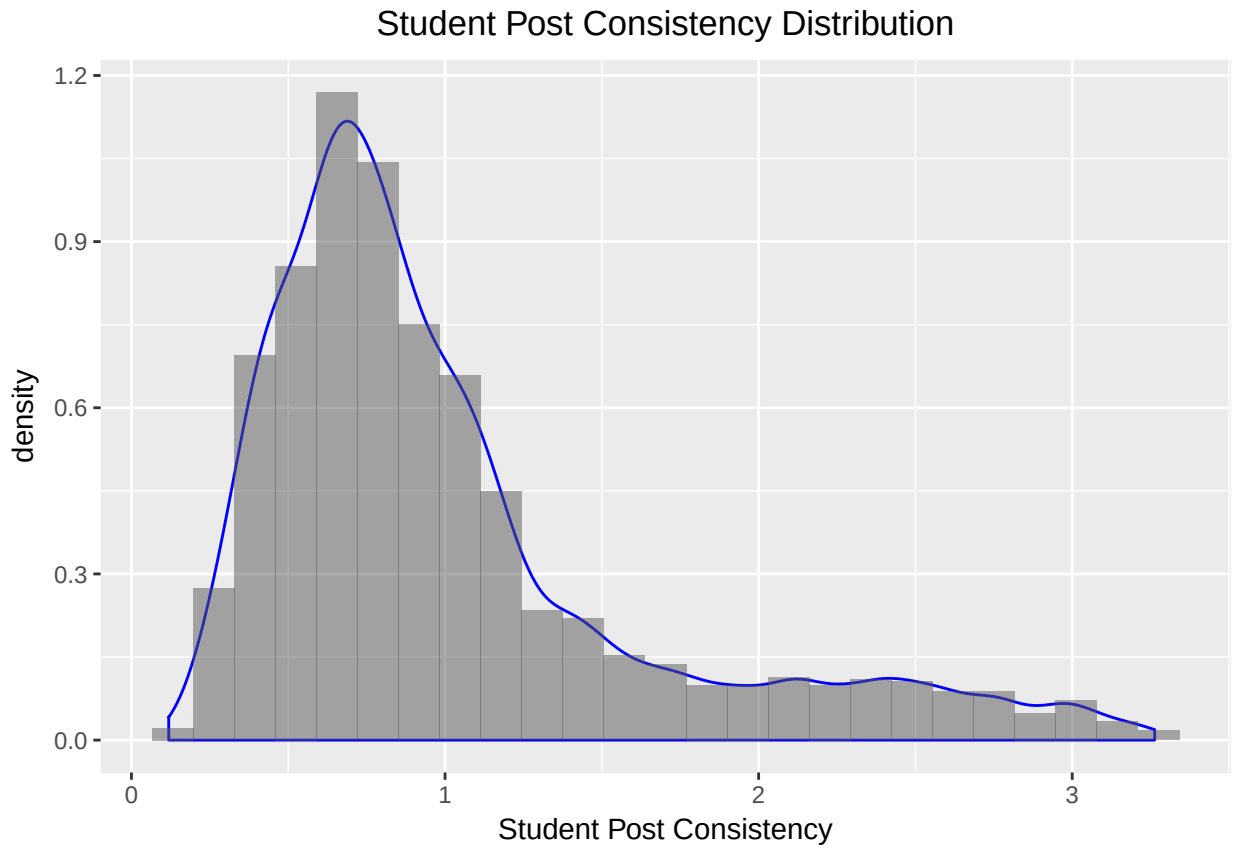
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```




Nothing really jumps out that wasn't already expected. Obviously, positive grade outcomes are negatively associated with poor grade outcomes. For the most part, almost none of the variables we anticipate using in the model have a strong correlation with the responses.

(Not sure where to put the plot below so I'm keeping it here for now)




```
## Saving 6.5 x 4.5 in image
```

---

## Research Question 1

Does faculty activity on discussion forums affect student posting behavior?

Based on the available data, we can use several metrics for student behavior on discussion forums:

- number of posts per student (over the entire course)
- consistency of student posts (from week to week)
- variance of the number of posts from week to week

Notes:

We need to be careful about what we choose to include in these models because some things might not be interpretable in a meaningful way. If we were to use `pass_rate` as a predictor and find that it has a positive effect on say the number of posts per student, what would our interpretation be? That increasing the pass rate of a course creates more student discussion? That's silly and feels like some sort of logical fallacy.

- generally the problem is: A is caused by B even though B occurs after A

With that said, we can still use things that are fixed or occur during the course. Class sizes, existence of hallway forums, faculty activity, and others can all be a part of this question.

```
## [1] "Lasso Estimates of Posts per Student Model"
```

```
##                   row       value
## 1         (Intercept) 11.83683169
## 2           enrl_total -1.81599771
## 3      prop_posts_boc  0.20382084
## 4      prop_posts_eoc -0.05467978
## 5      upper_division  2.14751713
## 6 fac_post_consistency -0.37389256
## 7       posts_per_fac  1.32200497
```

```
## [1] "Lasso Estimates of Student Post Consistency Model"
```

```
##                   row       value
## 1         (Intercept)  1.73087611
## 2           session_a -0.01975066
## 3           enrl_total  0.03042758
## 4         has_hallway -0.06765607
## 5      prop_posts_boc -0.04762364
## 6      prop_posts_eoc  0.04911342
## 7      upper_division -0.24485037
## 8 fac_post_consistency  0.01553571
## 9       posts_per_fac -0.07021609
```

Because we are unable to directly get at the standard errors and p-values for the lasso estimates, we use these results to inform a second model; just a standard linear model.

```
## [1] "Coefficient estimates from linear model with response: Posts per Student"
```

```
##                   term    estimate   std.error       p.value
## 1         (Intercept) 14.99542508 0.665036266 4.734937e-107
## 2          session_a1 -0.17195563 0.302066557  5.692038e-01
## 3          enrl_total -0.04807624 0.003952791  1.539644e-33
## 4        has_hallway1  0.14862505 0.328889908  6.513623e-01
## 5      prop_posts_boc  0.84321659 0.626986292  1.787305e-01
## 6      prop_posts_eoc -0.50829205 0.965844367  5.987272e-01
## 7     upper_division1  2.37159343 0.317004319  8.717060e-14
## 8 fac_post_consistency -0.64975769 0.270348864  1.628123e-02
## 9       posts_per_fac  0.02761948 0.003463004  1.884684e-15
```

```
## [1] "Coefficient estimates from linear model with response: Student Post Consistency"
```

```
##                   term      estimate    std.error       p.value
## 1         (Intercept)  1.5317899128 0.0553113972 1.201397e-156
## 2          session_a1 -0.0202960718 0.0284783901  4.760767e-01
## 3          enrl_total  0.0008058002 0.0003725530  3.059739e-02
## 4        has_hallway1 -0.0663964556 0.0310047956  3.228518e-02
## 5      prop_posts_boc -0.2565715229 0.0493874732  2.131826e-07
## 6     upper_division1 -0.2419909960 0.0298706246  6.850101e-16
## 7 fac_post_consistency  0.0261472390 0.0254811945  3.048784e-01
## 8       posts_per_fac -0.0013775219 0.0003263203  2.473375e-05
```

---

## Research Question 2:

How do student outcomes change with the level of engagement in online forums? If there is a change, it is independent of who is doing the posting?

keep it simple; just use pass_rate and avg_gpa

```
## [1] "Lasso estimates of GPA model"
```

```
##                     row       value
## 1           (Intercept)  2.45728850
## 2           has_hallway  0.03641036
## 3            enrl_total -0.02550088
## 4  fac_post_consistency  0.01170358
## 5  stu_post_consistency  0.03879823
## 6        prop_posts_boc -0.01126247
## 7        prop_posts_eoc -0.01028979
## 8             session_a  0.06431982
## 9         posts_per_fac -0.02906266
## 10    posts_per_student  0.02387290
## 11       upper_division  0.22066685
```

We repeat the same procedure from the previous research question to get more interpretable results.

```
## [1] "Coefficient Estimates for linear GPA model: "
```

```
##                     term      estimate    std.error      p.value
## 1           (Intercept)  2.7665857937 0.0388001784 0.000000e+00
## 2          has_hallway1  0.0367374384 0.0153012169 1.639078e-02
## 3            enrl_total -0.0006530863 0.0001875252 5.009297e-04
## 4  fac_post_consistency  0.0184326998 0.0125778631 1.428541e-01
## 5  stu_post_consistency  0.0392618155 0.0091089069 1.663528e-05
## 6        prop_posts_boc -0.0391888099 0.0291753286 1.792657e-01
## 7        prop_posts_eoc -0.0555474947 0.0449567213 2.166768e-01
## 8            session_a1  0.0646242575 0.0140474196 4.325621e-06
## 9         posts_per_fac -0.0005840284 0.0001620924 3.177092e-04
## 10    posts_per_student  0.0022467004 0.0008581433 8.870252e-03
## 11      upper_division1  0.2209745734 0.0148569922 6.078420e-49
```

I don't exactly understand how to interpret the coefficients of `prop_posts_boc` and `prop_posts_eoc` when they are included in the same model, so it is worth running the model two more times using one variable at a time. The reason I think this is difficult to understand is because the two aren't entirely but pretty strongly dependent on one another. I didn't think this was a problem when I was looking at the correlation matrix, but now that the models are ran I realize this is kind of wonky.

```
## [1] "Coefficient Estimates for linear GPA model with only `prop_posts_boc`: "
```

```
##                     term      estimate    std.error      p.value
## 1           (Intercept)  2.7497421527 0.0363285408 0.000000e+00
## 2          has_hallway1  0.0364366770 0.0153001272 1.728315e-02
## 3            enrl_total -0.0006595687 0.0001874622 4.382013e-04
## 4  fac_post_consistency  0.0180544064 0.0125748321 1.511372e-01
## 5  stu_post_consistency  0.0387347637 0.0090994172 2.113292e-05
## 6        prop_posts_boc -0.0194838344 0.0244320969 4.252186e-01
## 7            session_a1  0.0647247006 0.0140479617 4.183723e-06
## 8         posts_per_fac -0.0005905808 0.0001620145 2.700026e-04
## 9     posts_per_student  0.0022223100 0.0008579637 9.620871e-03
## 10      upper_division1  0.2202947343 0.0148476219 1.020961e-48
```

```
## [1] "Coefficient Estimates for linear GPA model with only `prop_posts_eoc`: "
```

```
##                     term      estimate    std.error      p.value
## 1           (Intercept)  2.7376617265 0.0322791215 0.000000e+00
## 2          has_hallway1  0.0378050095 0.0152818526 1.340123e-02
## 3            enrl_total -0.0006512780 0.0001875362 5.196227e-04
```

```
## 4  fac_post_consistency  0.0195045918 0.0125535838 1.203208e-01
## 5  stu_post_consistency  0.0396987313 0.0091038667 1.324191e-05
## 6        prop_posts_eoc -0.0225385798 0.0376488993 5.494336e-01
## 7            session_a1  0.0645806002 0.0140485691 4.398206e-06
## 8         posts_per_fac -0.0005657566 0.0001615342 4.653337e-04
## 9     posts_per_student  0.0022545459 0.0008581960 8.639837e-03
## 10      upper_division1  0.2210931448 0.0148579855 5.498506e-49
```

---

## Future Options

- can we add data about instructors making announcements?
- what peoplesoft data can be used?

---

## Working Notes

- research questions should be reordered; train of thought should be:
- does faculty activity on forums affect grade outcomes?
  - what characteristics of faculty activity have greatest effect? (early course posts, late course posts, consistency of posting)
- does faculty activity on forums affect student activity on forums?
  - what student activity do we want to measure?

---

## Extras: