Regression study: Graduation rates of PhysTEC Legacy Sites and Model Development

Phase I Study
April 26, 2021

Dr. Remy Dou
Executive Overview

The purpose of this report is to present statistical findings from data collected from 43 PhysTEC Legacy Sites teacher graduation rates, exploring possible contributing variables to teacher graduation outcomes. In particular, the analyses carried out were designed to confirm and build upon prior work (“2019-20 Funded Site Graduation Rate Evaluation” including analyses by Joe Taylor) regarding factors believed to contribute teacher graduation rates of PhysTEC sites during and post funding cycles. This report does not address PTEPA data, but rather identifies a set of variables that can be expected to contribute to graduation rates in order to account for their effect when testing the contribution of PTEPA measures to the graduation rates of PhysTEC sites (which will be undertaken in Summer 2021; i.e., Phase II).

Evaluation questions:

1. What factors predict whether a site experiences a change in teacher graduates during funding? Post-funding? Total increase pre-to-post?
2. What factors predict the size of the change in teacher graduates during funding? Post-funding? Total increase pre-to-post?
3. What factors predict the number of teacher graduates during funding? Post-funding?

Findings from this study include:

Model parameters: Average teacher graduation rate, teacher graduates prior to funding, whether a site is Comprehensive site, and whether the highest degree granted is a Bachelor’s. Of the four parsimonious regression models identified, the best regression model (defined as the one that explains the most change in the outcome variable) is the one where the outcome variable is the average graduation rate during the first three years of funding. All four models include the same three input variables: the average teacher graduates prior to funding, whether a site is a Comprehensive site, and whether the highest degree granted is Bachelor’s. These models used the following as the following output variables (in order or model strength):

- The average teacher graduation rate during the first three years of funding,
- The average teacher graduation rate during the entirety of the funding period\(^1\),
- The average teacher graduation rate during the post-funding period\(^2\),
- Whether there was a change in teacher graduation rates, represented by a binary outcome variable indicating whether the change in average teacher graduate rates from prior to funding\(^3\) to post-funding increased by 0.50 teacher graduates or more.

From these models we find:

1. Change in teacher graduation rates as an outcome (i.e., dependent variable) is not predicted by input variables. Thus, with the data given, the change in teacher graduation rates is not able to be modelled; only the actual number of graduates (represented as an average; see model above).
2. The average number of teacher graduates prior to funding is the strongest predictor of teacher graduation rates during funding. This variable explains 45% of the variance in during-funding graduation rates when looking only at the first three years of funding and 37% of the variance when looking at the entire funding period.

---

\(^1\) The funding period is defined as the first year of funding through and including the first year after funding ended.

\(^2\) The post-funding period is defined as the second year after funding ends and beyond.

\(^3\) The prior to funding period is defined as the three years immediately prior to when funding starts.
3. However, whether a site is a Comprehensive site is the only predictor of post-funding teacher graduation rates even when accounting for the number of teacher graduates prior to funding. In other words, although teacher graduation rates prior to funding predict graduation rates during funding, only Comprehensive sites were associated with higher post-funding teacher graduation rates regardless of prior performance. This finding suggests that Comprehensive funding is associated with more sustainable increases in teacher graduation rates.

4. Additionally, whether a site is a Comprehensive site is the only predictor of whether teacher graduation rates increased pre-to-post funding. Other background variables considered did not predict this outcome. The model shows there is a 98% probability that a Comprehensive Site will exhibit an increase of 0.50 or more teacher graduates when looking at the average change from pre to post funding.

5. Department size and change in department size are not significant predictors of teacher graduation rates during or post-funding.

6. Additionally, site maturity and highest degree granted were not significant predictors of teacher graduation rates. However, whether a site grants Bachelor’s degrees as the highest degree has a potential to rise to significance in future analyses.

7. The least successful sites in terms of teacher graduate rates are Recruiting sites.

Together, these results provide a model that can be tested in the future. In Summer 2021 we plan to add PTEPA Rubric results to the input variables to identify any additional contributions of that measurement. The addition of a control group (e.g. sites without PhysTEC funding) would help to establish a stronger case for causality.

**Method Overview**

Prior data collection activities in relation to 43 PhysTEC Legacy Sites documented the number of total graduates and teacher graduates for specific years (Bachelor’s degrees) in relation to PhysTEC funding (prior to, during, and post). These and related data were used to calculate the following outcome (i.e., dependent) variables of interest:

**Table I. Outcomes variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AvgDuring</strong> (continuous)</td>
<td>Average number of teacher graduates during the funding period.</td>
</tr>
<tr>
<td><strong>AvgPost</strong> (continuous)</td>
<td>Average number of teacher graduates during the post funding period.</td>
</tr>
<tr>
<td><strong>ChngDuringPost</strong> (continuous)</td>
<td>The difference AvgPost and AvgDuring was used to represent the change in teacher graduate rates from during to after funding.</td>
</tr>
<tr>
<td><strong>ChngDuringPre</strong> (continuous)</td>
<td>The difference between AvgDuring and the average number of teacher graduates during the three years prior to funding (see AvgGradsPre).</td>
</tr>
<tr>
<td><strong>ChngPostPre</strong> (continuous)</td>
<td>The difference between Avgpost and the average number of teacher graduates during the three years prior to funding (see AvgGradsPre).</td>
</tr>
<tr>
<td><strong>DidChngPostPre</strong> (binary)</td>
<td>Whether or not ChngPostPre increased by more than 0.50.</td>
</tr>
<tr>
<td><strong>AvgGradF1F3</strong> (continuous)</td>
<td>Average number of teacher graduates across the first three years of funding [see “Exploratory” below].</td>
</tr>
<tr>
<td><strong>AvgGradPF1PF2</strong> (continuous)</td>
<td>Average number of teacher graduates across the first two years post funding [see “Exploratory” below].</td>
</tr>
</tbody>
</table>
These six variables represented the three primary outcome variables of the models. An additional two variables were also tested as potential outcome variables: the average teacher graduation rate during the first three years of funding (AvgGradF1F3) and during the first two years after funding (AvgGradPF1PF2). Note: Here and throughout the rest of the report we refer to the average number of teacher graduates across a specific set of years as synonymous with the average teacher graduation rate.

Additionally, several predictors were tested against the possible outcomes above in stepwise fashion. These predictors included the following:

Table II. Predictor variables.

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgDptSizePre (continuous)</td>
<td>The average number of total graduates across the three years before funding.</td>
</tr>
<tr>
<td>AvgDptSizePost (continuous)</td>
<td>The average number of total graduates post-funding.</td>
</tr>
<tr>
<td>ChngDptSizePreDuring (continuous)</td>
<td>The average number of total graduates during funding minus AvgDptSizePre</td>
</tr>
<tr>
<td>DidDptSizePreDuring (binary)</td>
<td>Whether or not ChngDptSizePreDuring is greater than 0.50.</td>
</tr>
<tr>
<td>NmbrYrsSnceFndng (continuous)</td>
<td>2020 minus the year funding began.</td>
</tr>
<tr>
<td>AvgGradsPre (continuous)</td>
<td>The average number of teacher graduates the three years prior to funding.</td>
</tr>
</tbody>
</table>

Table II Continued. Predictor Variables.

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDG (categorical)</td>
<td>The highest degree granted by the department.</td>
</tr>
<tr>
<td>Site.Code (categorical)</td>
<td>Type of funding site.</td>
</tr>
<tr>
<td>PhDDum (binary)</td>
<td>Whether the highest degree granted by a site was PhD degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>MastDum (binary)</td>
<td>Whether the highest degree granted by a site was Master’s degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>BachDum (binary)</td>
<td>Whether the highest degree granted by a site was Bachelor’s degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>CompDum (binary)</td>
<td>Whether a site is a Comprehensive Site.</td>
</tr>
<tr>
<td>TargDum (binary)</td>
<td>Whether a site is a Targeted Site.</td>
</tr>
<tr>
<td>RecDum (binary)</td>
<td>Whether a site is a Recruiting site.</td>
</tr>
</tbody>
</table>

Again, to reiterate, the purpose of testing these variables in relation to one another was to identify the most parsimonious model(s) (e.g., greatest variance explained, non-significant control variables removed) that can be used to carry out additional tests of the contribution of PTEPA data to select outcome variables. In other words, these variables were identified because they allow evaluators to a) gauge the impact of funding as a factor of teacher graduate rates, b) determine whether prior site characteristics not related to
funding contribute to teacher graduate rates (e.g., AvgDptSizePre, HDG), and c) determine the extent to which site characteristics during funding contribute to post-funding outcomes.

The continuous variables listed in Table III below were analyzed to determine kurtosis (measure of distribution tails) and skewness (measure of asymmetry in distribution). Kurtosis values greater than or less than +/- 2 respectively are generally considered as evidence that the distribution of errors violates assumptions of normality and skewness values greater than or less than +/- 1 respectively suggest the same.

*Table III.* Descriptives table for continuous variables. Note that skewness values for AvgDuring and AvgGradF1F3 suggest violations of normality may exist.

<table>
<thead>
<tr>
<th>vars</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>skew</th>
<th>kurtosis</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgDptSizePre</td>
<td>15.85</td>
<td>12.8</td>
<td>12.04</td>
<td>0.67</td>
<td>57.67</td>
<td>57</td>
<td>1.75</td>
<td>2.91</td>
<td>1.97</td>
</tr>
<tr>
<td>AvgDptSizePost</td>
<td>21.76</td>
<td>15.7</td>
<td>17.38</td>
<td>2.4</td>
<td>71.6</td>
<td>69.2</td>
<td>1.34</td>
<td>1.39</td>
<td>2.42</td>
</tr>
<tr>
<td>ChngDptSizePreDuring</td>
<td>5.91</td>
<td>7.14</td>
<td>5.21</td>
<td>-5</td>
<td>40.05</td>
<td>45.05</td>
<td>2.45</td>
<td>9.92</td>
<td>1.1</td>
</tr>
<tr>
<td>NmbYrsSnceFndng</td>
<td>9.48</td>
<td>4.27</td>
<td>8</td>
<td>5</td>
<td>19</td>
<td>14</td>
<td>1.06</td>
<td>-0.07</td>
<td>0.65</td>
</tr>
<tr>
<td>AvgGradsPre</td>
<td>1.08</td>
<td>1.13</td>
<td>0.67</td>
<td>-0.39</td>
<td>3.67</td>
<td>4.06</td>
<td>0.78</td>
<td>-0.65</td>
<td>0.17</td>
</tr>
<tr>
<td>AvgDuringPre</td>
<td>1.84</td>
<td>1.33</td>
<td>1.67</td>
<td>0</td>
<td>6.14</td>
<td>6.14</td>
<td>1.24</td>
<td>1.29</td>
<td>0.2</td>
</tr>
<tr>
<td>AvgPost</td>
<td>1.92</td>
<td>1.34</td>
<td>1.67</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0.45</td>
<td>-0.75</td>
<td>0.2</td>
</tr>
<tr>
<td>ChngDuringPost</td>
<td>0.13</td>
<td>1.13</td>
<td>0</td>
<td>-2.23</td>
<td>3.67</td>
<td>5.9</td>
<td>0.77</td>
<td>1</td>
<td>0.18</td>
</tr>
<tr>
<td>AvgGradF1F3</td>
<td>1.71</td>
<td>1.29</td>
<td>1.67</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>1.41</td>
<td>1.87</td>
<td>0.2</td>
</tr>
<tr>
<td>AvgGradPF1PF2</td>
<td>2.5</td>
<td>1.7</td>
<td>2.5</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0.9</td>
<td>0.48</td>
<td>0.28</td>
</tr>
</tbody>
</table>

*Figure 1.* Histograms of AvgDptSizePre, AvgDptSizePost, and ChngDptSizePreDuring, respectively.

*Figure 2.* Histograms of AvgGradF1F3, AvgDuring, and NmbYrsSnceFndng, respectively.
Analyses: Initial, Exploratory, & Confirmatory

Model analyses took place in three parts: initial, exploratory, and confirmatory. Each part informed which variables were kept (and in some cases created) for the subsequent part.

Initial
Stepwise linear regressions were run to explore the associations between the input variables and the output variables. The initial set of models only tested six outcome variables that had been used in prior evaluations (see 2019 – 2020 Funded Site Graduation Rate Evaluation) or agreed upon during discussions, and nine input variables (i.e., AvgDptSizePre, ChngDptSizePreDuring, DidDptSizePreDuring, NmbrYrsSnceFndng, HDG, Site.Code, CompDum, RecDum, TargDum). 54 simple regressions were run first to look associations between individual input variables and each of the three outcome variables. An additional 54 linear regressions were run to explore the same relationships while accounting (or controlling) for AvgGradsPre. Models were compared based on estimated effect sizes and R-squared values.

Findings from initial analyses provided evidence to keep the following control variables:
- RecDum (whether a site was a Recruiting site)
- CompDum (whether a site was a Comprehensive site)
- AvgGradsPre (average teacher graduates prior to funding)

Findings from initial analyses provided evidence to drop the following control variables:
- AvgDptSizePre (average number of total graduates across the three years before funding)
- ChngDptSizePreDuring (average number of total graduates during funding minus AvgDptSizePre)
- DidDptSizePreDuring (Whether or not ChngDptSizePreDuring is greater than 0.50)
- NmbrYrsSnceFndng (2020 minus the year funding began)
- HDG (see “Exploratory”) (highest degree granted by the department)
- Site.Code (type of funding site)
- TargDum (whether a site is a Targeted site)

No statistical relationship was found between any of the control (or input) variables and the outcome variables listed below. As such, the following outcome variables were dropped:
- ChngDuringPost (difference between AvgPost and AvgDuring)
- ChngDuringPre (difference between AvgDuring and AvgGradsPre)
- ChngPostPre (difference between AvgPost and AvgGradsPre)

Analyses of the remaining variables are described below.

Exploratory
Given outcomes that supported further exploration of the value of HDG as a predictor, measures of this variable were converted into three dummy variables (i.e., PhDum, MastDum, BachDum). Two additional outcome variables were also created to test for short term outcomes (i.e., AvgGradF1F3, AvgGradPF1PF2). Similar analyses as the ones described previously were carried out with these variables.
Findings from exploratory analyses provided evidence to keep only the following control variables:
- AvgGradsPre (average teacher graduates prior to funding)
- CompDum (whether a site was a Comprehensive site)
- BachDum (whether the highest degree granted by a site was Bachelor’s degrees)

Findings from exploratory analyses provided evidence to keep only the following outcome variables:
- AvgDuring (average number of teacher graduates during the funding period)
- AvgPost (average number of teacher graduates post-funding period)
- DidChngPostPre (whether or not ChngPostPre was more than 0.50)
- AvgGradF1F3 (average number of teacher graduates across the first three years of funding)

Confirmatory
A dataframe consisting of only the final variables of interest was created and imputed once to account for missingness. The three final linear models were tested once more and test results were recorded. The models were run using a permutation algorithm to account for deviations from normality and Type II error from testing multiple models.

Box 1: Philosophy of Model Building
I typically approach model building by leaning on theoretical frameworks or prior evidence of effect when determining what variables to include or exclude. At times, that may mean keeping variables that are statistically insignificant in order to adhere to specific model parameters or when I want to make a specific claim about a variable. In this evaluative work, a particular theoretical framework did not guide variable selection or model building, but rather qualitative information gathered by Dr. Stephanie Chasteen and findings from prior evaluations. As such, my approach took on a less theoretical tone. Instead, I tested each variable against each outcome variable, eliminating those which exhibited no relationship. Those that exhibited a statistically significant relationship, I kept in the model and tested alongside others that were also statistically significant. In many of those cases, some of those variables were removed when no longer significant based on the assumption that the variance of other variables better explained the variances in the outcome variables.
Summary of Model Results

### Model A: Predicting average teacher graduates during funding

\[
\text{AvgDuring} \sim \beta_1 \text{AvgGradsPre} + \beta_2 \text{CompDum} + \beta_3 \text{BachDum} + e
\]

<table>
<thead>
<tr>
<th>F-statistic</th>
<th>R-squared</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(3,39) = 7.60***</td>
<td>0.37</td>
<td>0.54***</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Model B: Predicting the average teacher graduates post-funding

\[
\text{AvgPost} \sim \beta_1 \text{AvgGradsPre} + \beta_2 \text{CompDum} + \beta_3 \text{BachDum} + e
\]

<table>
<thead>
<tr>
<th>F-statistic</th>
<th>R-squared</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(3,39) = 7.60***</td>
<td>0.37</td>
<td>0.25</td>
<td>0.54**</td>
<td>0.07</td>
</tr>
</tbody>
</table>

### Model C: Predicting the average teacher graduates during just the first three years of funding

\[
\text{AvgGradF1F3} \sim \beta_1 \text{AvgGradsPre} + \beta_2 \text{CompDum} + \beta_3 \text{BachDum} + e
\]

<table>
<thead>
<tr>
<th>F-statistic</th>
<th>R-squared</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(3,39) = 10.62***</td>
<td>0.45</td>
<td>0.63***</td>
<td>0.14</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Model D: Likelihood of predicting a change from AvgGradsPre to AvgPost of 0.50 or more

\[
\text{DidChngPostPre} \sim \beta_1 \text{AvgGradsPre} + \beta_2 \text{CompDum} + \beta_3 \text{BachDum} + e
\]

<table>
<thead>
<tr>
<th>AIC</th>
<th>AvgGradsPre; Odds Ratio</th>
<th>CompDum; Odds Ratio</th>
<th>BachDum; Odds Ratio</th>
</tr>
</thead>
</table>
| 42.17 | 6.25
or
14% Probability | 0.02
or
98% Probability | 1.07
or
48% Probability |
Findings

The summary of analyses presented above suggest there are at least two ways of considering the outcomes of PhysTEC funding as a factor of teacher graduate rates: a short-term approach and a long-term approach. A short-term approach would focus more closely on maximizing the number of teacher graduates during the first three years of funding by almost exclusively considering the number of teacher graduates during the three years prior to funding. Findings supporting a short-term approach:

1. Model A suggested that the best and strongest predictor of teacher graduate rates during funding was the three-year average teacher graduation rate prior to funding (AvgGradsPre). This variable exhibited a large effect size such that for every increase in AvgGradsPre by one standard deviation, one could expect a 0.54 standard deviation increase in the average number of teacher graduates during funding (AvgDuring).

2. This is most clear seen in the model that only looked at the average teacher graduate rate during the first three years of funding (Model C), where a one standard deviation shift in AvgGradsPre roughly equates to an expected 0.63 standard deviation shift in the number of teacher graduates during those first three years (AvgGradF1F3). This one variable, AvgGradsPre, explained 45% percent of the variance in average teacher graduation rates during the first three years, which is quite large. Given that most sites only had data for the first three years of funding, and few had data beyond that, it makes sense that not only would this model exhibit the largest effect size but would also serve as the most appropriate baseline model of short-term outcomes moving forward.

A long-term approach would focus on supporting Comprehensive Sites to increase the number of teacher graduates after funding ends. Findings supporting a long-term approach:

3. Whether a site was a Comprehensive site (or not) was the best predictor of post-funding teacher graduation rates. These findings suggest a stronger association between long-term outcomes post funding and the type of funding an institution receives.

4. Other factors were not significant predictors of graduation rates. The size of the department (as judged by the number of majors averaged across 3 years pre-funding) was not a significant predictor of teacher graduation rates during or post-funding. It is worth noting that a site’s designation as PhD granting, Bachelor’s degree granting, Recruiting, or Targeted was not ultimately predictive of teacher graduation rates when accounting for other factors. Future analyses may consider the value of the BachDum variable as part of the baseline models.

5. Model D suggests that there is a 98% probability that a Comprehensive Site will exhibit an increase of 0.50 or more teacher graduates when looking at the average change from pre to post funding. Targeted and Recruiting sites were not predictive of teacher graduate rates.

Note: Average department sizes, changes in department sizes over time, and years since funding were not predictive of teacher graduation rates across any period.

Findings from Comparisons of Means: Recruiting Sites

In addition to the analyses described above, several comparisons of means were run using independent samples t-tests. These analyses support the findings described above, but also highlight some important things to consider regarding how funding is being allocated:

Phase I: Regression Study 2020 #1
A) Recruiting sites are more likely to have lower teacher graduation rates during the three years prior to funding.

B) Recruiting sites are more likely to have lower total graduation rates (number of Bachelor’s degrees granted) during the three years prior to funding.

C) Recruiting sites are more likely to see department sizes decrease between the funding period and the pre-funding period.

D) Recruiting sites are more likely to have been funded recently compared to other sites.

Limitations
These analyses are corroborated by prior examinations though one must take care when making causality statements in relation to the variables. While some longitudinal aspects of the design and analyses allow for causal interpretations of the results, the research design was at best quasi-experimental in that sites were not randomly assigned a funding type and that no true negative control group was identified.

Recommendations

Brief:

1. If the aim is to maximize the number of teacher graduates during the funding period, then funding should go to sites that have the highest number of teacher graduates prior to funding (regardless of site type or highest degree granted by the institution).

2. If the aim is to maximize the number of teacher graduates after the funding period, then funding should go to increase the number of Comprehensive sites (regardless of the number of teacher graduates prior to funding or highest degree granted by the institution).

Summary of outcomes:
Future decision-making around funding of new PhysTEC sites should consider the value of both short-term and long-term gains in their goals. Considerations of short-term gains should focus on teacher graduation rates during the first three years of funding and lean on average teacher graduation rates during the prior three years of funding as a very strong predictor of success.

Consideration of long-term gains should focus on average teacher graduation rates starting from the second year after funding ends. Comprehensive sites are much more likely than other types of sites to exhibit positive long-term outcomes. The characteristics of what makes these sites better at sustaining long-term outcomes could be an aspect of Phase II analyses.

The PhysTEC team should revisit the goals, objectives, and approaches, surrounding Recruiting Sites given that they not only exhibit no statistical likelihood of increasing teacher graduation rates, but also tend to grant fewer Bachelor’s degrees overall and/or see a decrease in the number of Bachelor’s degrees granted.

Furthermore, more nuanced articulation around interpretations of positive correlations with means (i.e., averages) of teacher graduation rates should be explored.
A Note on Phase II

Phase II of these analyses will focus on identifying whether aspects of PTEPA data contribute to the explanatory power of our four final models with the objective of narrowing down characteristics that increase site success. This will allow the PhysTEC team and staff from those sites to better target key programmatic aspects related to site success.
PhysTec 2021 In-Depth Analyses: Phase I

Remy Dou

4/15/2021

PhysTec 2021 In-Depth Analyses: Phase I

Importing Dataframe: PhaseI

Dataframe Cleaning in R

```r
library(lmPerm)

PhaseI = PhaseI[c(1:43), c(-138:-246)]
PhaseI = PhaseI[, c(-12, -17, -91)]
```

Data Cleaning and Descriptives in R: Output Variables

Avg. Grads During+1 (BW), Avg. Grads Post (BX), Change in Avg Post-During+1 (BZ)

```r
library(psych)

names(PhaseI)[84] <- "AvgDuring"
names(PhaseI)[85] <- "AvgPost"
names(PhaseI)[87] <- "ChngDuringPost"
names(PhaseI)[77] <- "ChngPostPre"
names(PhaseI)[75] <- "ChngDuringPre"

PhaseI$AvgPost = as.numeric(PhaseI$AvgPost)
PhaseI$ChngDuringPost = as.numeric(PhaseI$ChngDuringPost)

PhaseI$DidChngPostPre[PhaseI$ChngPostPre > 0.49] = 1
PhaseI$DidChngPostPre[PhaseI$ChngPostPre < 0.5] = 0
PhaseI$DidChngPostPre = as.factor(PhaseI$DidChngPostPre)

OutDescrp = PhaseI[, c(84, 85, 87)]

describe(OutDescrp)
multi.hist(OutDescrp)

describe(PhaseI$AvgDptSizePost)
describe(PhaseI$ChngDptSizePreDuring)
```
Data Cleaning and Descriptives in R: Input Variables

```r
names(PhaseI)[5] <- "HDG"
names(PhaseI)[9] <- "FundedYrs"
names(PhaseI)[72] <- "AvgGradsPre"
names(PhaseI)[126] <- "AvgDptSizePre"
names(PhaseI)[127] <- "AvgDptSizePost"
names(PhaseI)[130] <- "ChngDptSizePreDuring"
```

```r
PhaseI$HDG = as.character(PhaseI$HDG)
PhaseI$HDG[31] = NA
PhaseI$HDG = as.factor(PhaseI$HDG)
print(PhaseI$HDG)
```

```r
PhaseI$AvgDptSizePre = as.numeric(PhaseI$AvgDptSizePre)
PhaseI$ChngDptSizePreDuring = as.numeric(PhaseI$ChngDptSizePreDuring)
PhaseI$Site.Code = as.factor(PhaseI$Site.Code)
PhaseI$CompDum = as.factor(PhaseI$CompDum)
PhaseI$TargDum = as.factor(PhaseI$TargDum)
PhaseI$RecDum = as.factor(PhaseI$RecDum)
```

```r
PhaseI$AvgDptSizePost = as.numeric(PhaseI$AvgDptSizePost)
PhaseI$ChngDptSizePreDuring = PhaseI$AvgDptSizePost - PhaseI$AvgDptSizePre
PhaseI$ChngDptSizePreDuring = PhaseI$AvgDptSizePost - PhaseI$AvgDptSizePre
PhaseI$DidDptSizePreDuring = PhaseI$ChngDptSizePreDuring
PhaseI$DidDptSizePreDuring[PhaseI$DidDptSizePreDuring < 0.5] = 0
PhaseI$DidDptSizePreDuring[PhaseI$DidDptSizePreDuring > 0.49] = 1
```

```r
PhaseI$NmbrYrsSnceFndng <- substr(PhaseI$FundedYrs, 0, 4)
PhaseI$NmbrYrsSnceFndng = as.numeric(PhaseI$NmbrYrsSnceFndng)
PhaseI$NmbrYrsSnceFndng = 2020 - PhaseI$NmbrYrsSnceFndng
```

```r
str(PhaseI$HDG)
str(PhaseI$AvgDptSizePre)
str(PhaseI$Site.Code)
str(PhaseI$ChngDptSizePreDuring)
str(PhaseI$DidDptSizePreDuring)
str(PhaseI$NmbrYrsSnceFndng)
```

```r
describe(PhaseI$AvgDptSizePre)
```

Running Phase I Models

Naming Convention:
Outcome: Average Graduation Rate During Funding +1

t.test(PhaseI$AvgGradsPre ~ PhaseI$RecDum) ### (Recruiting Sites With Lower Mean)
t.test(PhaseI$AvgGradsPre ~ PhaseI$CompDum)
t.test(PhaseI$AvgGradsPre ~ PhaseI$TargDum)
t.test(PhaseI$AvgOptSizePre ~ PhaseI$RecDum) # Recruiting sites with lower department sizes
t.test(PhaseI$AvgOptSizePre ~ PhaseI$CompDum)
t.test(PhaseI$AvgOptSizePre ~ PhaseI$TargDum)
t.test(PhaseI$ChngOptSizePreDuring ~ PhaseI$RecDum, na.omit = TRUE) ## Smaller change in recruiting

t.test(PhaseI$ChngOptSizePreDuring ~ PhaseI$CompDum, na.omit = TRUE) # Bigger change in comprehensive

t.test(PhaseI$ChngOptSizePreDuring ~ PhaseI$TargDum, na.omit = TRUE)
t.test(PhaseI$NmbrYrsSinceFndng ~ PhaseI$RecDum) ### Fewer years since funding
t.test(PhaseI$NmbrYrsSinceFndng ~ PhaseI$CompDum) ## Greater years since funding
t.test(PhaseI$NmbrYrsSinceFndng ~ PhaseI$TargDum)

Model 1.0 = lm(AvgDuring ~ AvgOptSizePre, data = PhaseI)
Model 1.1 = lm(AvgDuring ~ AvgOptSizePre + AvgGradsPre, data = PhaseI) ### (only AvgGradsPre)

Model 2.0 = lm(AvgDuring ~ HDG, data = PhaseI)
Model 2.1 = lm(AvgDuring ~ HDG + AvgGradsPre, data = PhaseI) ### (only AvgGradsPre)

Model 3.0 = lm(AvgDuring ~ Site.Code, data = PhaseI) # (Estimate yes; model almost)
Model 3.1 = lm(AvgDuring ~ Site.Code + AvgGradsPre, data = PhaseI) ## (only AvgGradsPre)
pModel 3.0 = lm(AvgDuring ~ Site.Code, data = PhaseI) # (Estimate yes; model almost)
pModel 3.1 = lm(AvgDuring ~ Site.Code + AvgGradsPre, data = PhaseI) ## (only AvgGradsPre)

Model 4.0 = lm(AvgDuring ~ CompDum, data = PhaseI) #
Model 4.1 = lm(AvgDuring ~ CompDum + AvgGradsPre, data = PhaseI) ## (only AvgGradsPre)
t.test(PhaseI$AvgOptSizePre ~ PhaseI$CompDum) # (Comprehensive Sites with Higher Mean)

Model 5.0 = lm(AvgDuring ~ RecDum, data = PhaseI) #
Model 5.1 = lm(AvgDuring ~ RecDum + AvgGradsPre, data = PhaseI) ## (only AvgGradsPre)
t.test(PhaseI$AvgOptSizePre ~ PhaseI$RecDum) ### (Recruiting Sites With Lower Mean)

Model 6.0 = lm(AvgDuring ~ TargDum, data = PhaseI)
Model 6.1 = lm(AvgDuring ~ TargDum + AvgGradsPre, data = PhaseI) ## (only AvgGradsPre)

Phases$ChngOptSizePreDuring = unlist(PhaseI$ChngOptSizePreDuring)
Phases$ChngOptSizePreDuring = as.numeric(PhaseI$ChngOptSizePreDuring)
Model 7.0 = lm(AvgDuring ~ PhaseI$ChngOptSizePreDuring, data = PhaseI)
Model 7.1 = lm(AvgDuring ~ PhaseI$ChngOptSizePreDuring + AvgGradsPre, data = PhaseI) ### (only AvgGradsPre)

Model 8.0 = lm(AvgDuring ~ PhaseI$DidOptSizePreDuring, data = PhaseI)
Model 8.1 = lm(AvgDuring ~ PhaseI$DidOptSizePreDuring + AvgGradsPre, data = PhaseI) ### (only AvgGradsPre)
Outcome: Average Graduation Rate Post Funding

\texttt{library(lmPerm)}

Model2.1.0 = \texttt{lm(AvgPost ~ AvgDptSizePre, data = PhaseI)}
Model2.1.1 = \texttt{lm(AvgPost ~ AvgDptSizePre + AvgGradsPre, data = PhaseI)}  # (only AvgGradsPre; model no)

Model2.2.0 = \texttt{lm(AvgPost ~ HDG, data = PhaseI)}  # (almost PhD higher; model no)
Model2.2.1 = \texttt{lm(AvgPost ~ HDG + AvgGradsPre, data = PhaseI)}  # (only AvgGradsPre)

Outcome: Change in Average Graduation Rate Post Minus During (Note: This would ideally be small)

\texttt{library(lmPerm)}

Model3.1.0 = \texttt{lm(ChngDuringPost ~ AvgDptSizePre, data = PhaseI)}
Model3.1.1 = \texttt{lm(ChngDuringPost ~ AvgDptSizePre + AvgGradsPre, data = PhaseI)}  # (only AvgGradsPre)
 Outcome: Change in Average Teacher Graduation Rate Post Minus Pre

```
library(lmPerm)

Model3.2.0 = lm(ChngDuringPost ~ HDG, data = PhaseI)
Model3.2.1 = lm(ChngDuringPost ~ HDG + AvgGradsPre, data = PhaseI)  ### (AvgGradsPre; HDGMast; no model)
pModel3.2.1 = lmp(ChngDuringPost ~ HDG + AvgGradsPre, data = PhaseI)  # (NOT CONFIRMED)

Model3.3.0 = lm(ChngDuringPost ~ Site.Code, data = PhaseI)
Model3.3.1 = lm(ChngDuringPost ~ Site.Code + AvgGradsPre, data = PhaseI)  # (only AvgGradsPre; no model)

Model3.4.0 = lm(ChngDuringPost ~ CompDum, data = PhaseI)
Model3.4.1 = lm(ChngDuringPost ~ CompDum + AvgGradsPre, data = PhaseI)  # (only AvgGradsPre)

Model3.5.0 = lm(ChngDuringPost ~ RecDum, data = PhaseI)
Model3.5.1 = lm(ChngDuringPost ~ RecDum + AvgGradsPre, data = PhaseI)  # (only AvgGradsPre; no model)

Model3.6.0 = lm(ChngDuringPost ~ TargDum, data = PhaseI)
Model3.6.1 = lm(ChngDuringPost ~ TargDum + AvgGradsPre, data = PhaseI)

Model3.7.0 = lm(ChngDuringPost ~ PhaseI$ChngDptSizePreDuring, data = PhaseI)
Model3.7.1 = lm(ChngDuringPost ~ PhaseI$ChngDptSizePreDuring + AvgGradsPre, data = PhaseI)

Model3.8.0 = lm(ChngDuringPost ~ PhaseI$DidDptSizePreDuring, data = PhaseI)
Model3.8.1 = lm(ChngDuringPost ~ PhaseI$DidDptSizePreDuring + AvgGradsPre, data = PhaseI)

Model3.9.0 = lm(ChngDuringPost ~ PhaseI$NmbrYrsSnceFndng, data = PhaseI)
Model3.9.1 = lm(ChngDuringPost ~ PhaseI$NmbrYrsSnceFndng + AvgGradsPre, data = PhaseI)
```

```
Model4.1.0 = lm(ChngPostPre ~ AvgDptSizePre, data = PhaseI)
Model4.1.1 = lm(ChngPostPre ~ AvgDptSizePre + AvgGradsPre, data = PhaseI)  ### AvgGradsPre

Model4.2.0 = lm(ChngPostPre ~ HDG, data = PhaseI)
Model4.2.1 = lm(ChngPostPre ~ HDG + AvgGradsPre, data = PhaseI)  ### AvgGradsPre

Model4.3.0 = lm(ChngPostPre ~ Site.Code, data = PhaseI)
Model4.3.1 = lm(ChngPostPre ~ Site.Code + AvgGradsPre, data = PhaseI)  ### AvgGradsPre #Code3
pModel4.3.1 = lmp(ChngPostPre ~ Site.Code + AvgGradsPre, data = PhaseI)  ### AvgGradsPre #Code3

Model4.4.0 = lm(ChngPostPre ~ CompDum, data = PhaseI)
Model4.4.1 = lm(ChngPostPre ~ CompDum + AvgGradsPre, data = PhaseI)  # CompDum ###AvgGradsPre
pModel4.4.1 = lmp(ChngPostPre ~ CompDum + AvgGradsPre, data = PhaseI)  # CompDum ###AvgGradsPre

Model4.5.0 = lm(ChngPostPre ~ RecDum, data = PhaseI)
Model4.5.1 = lm(ChngPostPre ~ RecDum + AvgGradsPre, data = PhaseI)  ### AvgGradsPre

Model4.6.0 = lm(ChngPostPre ~ TargDum, data = PhaseI)
Model4.6.1 = lm(ChngPostPre ~ TargDum + AvgGradsPre, data = PhaseI)  ### AvgGradsPre

Model4.7.0 = lm(ChngPostPre ~ PhaseI$ChngDptSizePreDuring, data = PhaseI)
Model4.7.1 = lm(ChngPostPre ~ PhaseI$ChngDptSizePreDuring + AvgGradsPre, data = PhaseI)  ### AvgGradsPre
```
Model 4.8.0 = \text{lm}(\text{ChngPostPre} \sim \text{PhaseI}$\text{DidDptSizePreDuring}$, \text{data} = \text{PhaseI})

Model 4.8.1 = \text{lm}(\text{ChngPostPre} \sim \text{PhaseI}$\text{DidDptSizePreDuring} + \text{AvgGradsPre}$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 4.9.0 = \text{lm}(\text{ChngPostPre} \sim \text{PhaseI}$\text{NmbrYrsSnceFndng}$, \text{data} = \text{PhaseI})

Model 4.9.1 = \text{lm}(\text{ChngPostPre} \sim \text{PhaseI}$\text{NmbrYrsSnceFndng} + \text{AvgGradsPre}$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

\text{library(lmPerm)}

Model 5.1.0 = \text{lm}(\text{ChngDuringPre} \sim \text{AvgDptSizePre}, \text{data} = \text{PhaseI})

Model 5.1.1 = \text{lm}(\text{ChngDuringPre} \sim \text{AvgDptSizePre} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.2.0 = \text{lm}(\text{ChngDuringPre} \sim \text{HDG}, \text{data} = \text{PhaseI})

Model 5.2.1 = \text{lm}(\text{ChngDuringPre} \sim \text{HDG} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.3.0 = \text{lm}(\text{ChngDuringPre} \sim \text{Site.Code}, \text{data} = \text{PhaseI})

Model 5.3.1 = \text{lm}(\text{ChngDuringPre} \sim \text{Site.Code} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

pModel 5.3.1 = \text{lm}(\text{ChngDuringPre} \sim \text{Site.Code} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.4.0 = \text{lm}(\text{ChngDuringPre} \sim \text{CompDum}, \text{data} = \text{PhaseI})

Model 5.4.1 = \text{lm}(\text{ChngDuringPre} \sim \text{CompDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{CompDum} \# \text{AvgGradsPre}

pModel 5.4.1 = \text{lm}(\text{ChngDuringPre} \sim \text{CompDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.5.0 = \text{lm}(\text{ChngDuringPre} \sim \text{RecDum}, \text{data} = \text{PhaseI})

Model 5.5.1 = \text{lm}(\text{ChngDuringPre} \sim \text{RecDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.6.0 = \text{lm}(\text{ChngDuringPre} \sim \text{TargDum}, \text{data} = \text{PhaseI})

Model 5.6.1 = \text{lm}(\text{ChngDuringPre} \sim \text{TargDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.7.0 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{ChngDptSizePreDuring}$, \text{data} = \text{PhaseI})

Model 5.7.1 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{ChngDptSizePreDuring} + \text{AvgGradsPre}$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.8.0 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{DidDptSizePreDuring}$, \text{data} = \text{PhaseI})

Model 5.8.1 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{DidDptSizePreDuring} + \text{AvgGradsPre}$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 5.9.0 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{NmbrYrsSnceFndng}$, \text{data} = \text{PhaseI})

Model 5.9.1 = \text{lm}(\text{ChngDuringPre} \sim \text{PhaseI}$\text{NmbrYrsSnceFndng} + \text{AvgGradsPre}$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

\text{library(lmPerm)}

Model 6.1.0 = \text{glm}(\text{DidChngPostPre} \sim \text{AvgDptSizePre}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI})

Model 6.1.1 = \text{glm}(\text{DidChngPostPre} \sim \text{AvgDptSizePre} + \text{AvgGradsPre}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 6.2.0 = \text{glm}(\text{DidChngPostPre} \sim \text{HDG}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \#

Model 6.2.1 = \text{glm}(\text{DidChngPostPre} \sim \text{HDG} + \text{AvgGradsPre}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 6.3.0 = \text{glm}(\text{DidChngPostPre} \sim \text{Site.Code}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \#

Model 6.3.1 = \text{glm}(\text{DidChngPostPre} \sim \text{Site.Code} + \text{AvgGradsPre}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \# \text{AvgGradsPre}

Model 6.4.0 = \text{glm}(\text{DidChngPostPre} \sim \text{CompDum}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \#

Model 6.4.1 = \text{glm}(\text{DidChngPostPre} \sim \text{CompDum} + \text{AvgGradsPre}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI}) \# \text{CompDum}

Model 6.5.0 = \text{glm}(\text{DidChngPostPre} \sim \text{RecDum}, \text{family} = "\text{binomial}"$, \text{data} = \text{PhaseI})
Based on the above, I suggest we consider the following primary control variables: RecDum, CompDum, AvgGradsPre and with or without HDG (Tested Below)

### Comparing Phase I Models

#### Model 1a
Model 1a = \texttt{lm(AvgDuring ~ RecDum + CompDum + AvgGradsPre, data = PhaseI)}  
\texttt{## (only AvgGradsPre)}

#### Model 1b
Model 1b = \texttt{lm(AvgDuring ~ RecDum + CompDum + AvgGradsPre + HDG, data = PhaseI)}  
\texttt{### (only AvgGradsPre)}

#### Model 2a
Model 2a = \texttt{lm(AvgPost ~ RecDum + CompDum + AvgGradsPre, data = PhaseI)}  
\texttt{## (only CompDum)}

#### Model 2b
Model 2b = \texttt{lm(AvgPost ~ RecDum + CompDum + AvgGradsPre + HDG, data = PhaseI)}  
\texttt{### (only AvgGradsPre)}

#### Model 3a
Model 3a = \texttt{lm(ChngDuringPost ~ RecDum + CompDum + AvgGradsPre, data = PhaseI)}  
\texttt{## (only AvgGradsPre; not model)}

#### Model 3b
Model 3b = \texttt{lm(ChngDuringPost ~ RecDum + CompDum + AvgGradsPre + HDG, data = PhaseI)}  
\texttt{### (only AvgGradsPre)}

#### Model 4a
Model 4a = \texttt{lm(ChngPostPre ~ RecDum + CompDum + AvgGradsPre, data = PhaseI)}  
\texttt{## (only AvgGradsPre)}

#### Model 4b
Model 4b = \texttt{lm(ChngPostPre ~ RecDum + CompDum + AvgGradsPre + HDG, data = PhaseI)}  
\texttt{### (only AvgGradsPre)}

#### Model 5a
Model 5a = \texttt{glm(ChngPostPre ~ RecDum + CompDum + AvgGradsPre, family = "binomial", data = PhaseI)}  
\texttt{## AvgGradsPre # CompDum}

#### Model 5b
Model 5b = \texttt{glm(ChngPostPre ~ RecDum + CompDum + AvgGradsPre + HDG, family = "binomial", data = PhaseI)}  
\texttt{### (only AvgGradsPre)}

#### Model 6a
Model 6a = \texttt{lm(ChngDuringPre ~ RecDum + CompDum + AvgGradsPre, data = PhaseI)}  
\texttt{## AvgGradsPre}

#### Model 6b
Model 6b = \texttt{lm(ChngDuringPre ~ RecDum + CompDum + AvgGradsPre + HDG, data = PhaseI)}  
\texttt{# AvgGradsPre}

Final suggested control variables based on R-squared comparisons: RecDum, CompDum, AvgGradsPre For the outcome variable, so far best proxies are AvgDuring, AvgPost, & DidChngPostPre
Phase I: Exploration of Potentially Viable Input and Output Variables

Input Recoding

\[
\text{PhaseI}\$\text{PhDum} = \text{PhaseI}\$\text{HDG} \\
\text{PhaseI}\$\text{PhDum} = \text{as.character(PhaseI}\$\text{PhDum)} \\
\text{PhaseI}\$\text{PhDum}[\text{PhaseI}\$\text{PhDum} == "PhD"] = 1 \\
\text{PhaseI}\$\text{PhDum}[\text{PhaseI}\$\text{PhDum} == "Mast"] = 0 \\
\text{PhaseI}\$\text{PhDum}[\text{PhaseI}\$\text{PhDum} == "Bach"] = 0 \\
\text{PhaseI}\$\text{PhDum} = \text{as.factor(PhaseI}\$\text{PhDum)} \\
\text{print(PhaseI}\$\text{PhDum)} \\
\text{PhaseI}\$\text{MastDum} = \text{PhaseI}\$\text{HDG} \\
\text{PhaseI}\$\text{MastDum} = \text{as.character(PhaseI}\$\text{MastDum)} \\
\text{PhaseI}\$\text{MastDum}[\text{PhaseI}\$\text{MastDum} == "PhD"] = 0 \\
\text{PhaseI}\$\text{MastDum}[\text{PhaseI}\$\text{MastDum} == "Mast"] = 1 \\
\text{PhaseI}\$\text{MastDum}[\text{PhaseI}\$\text{MastDum} == "Bach"] = 0 \\
\text{PhaseI}\$\text{MastDum} = \text{as.factor(PhaseI}\$\text{MastDum)} \\
\text{print(PhaseI}\$\text{MastDum)} \\
\text{PhaseI}\$\text{BachDum} = \text{PhaseI}\$\text{HDG} \\
\text{PhaseI}\$\text{BachDum} = \text{as.character(PhaseI}\$\text{BachDum)} \\
\text{PhaseI}\$\text{BachDum}[\text{PhaseI}\$\text{BachDum} == "PhD"] = 0 \\
\text{PhaseI}\$\text{BachDum}[\text{PhaseI}\$\text{BachDum} == "Mast"] = 0 \\
\text{PhaseI}\$\text{BachDum}[\text{PhaseI}\$\text{BachDum} == "Bach"] = 1 \\
\text{PhaseI}\$\text{BachDum} = \text{as.factor(PhaseI}\$\text{BachDum)} \\
\text{print(PhaseI}\$\text{BachDum)}
\]

Output Recoding

\[
\text{PhaseI}\$\text{F3} = \text{as.numeric(PhaseI}\$\text{F3)} \\
\text{PhaseI}\$\text{AvgGradF1F3} = (\text{PhaseI}\$\text{F3} + \text{PhaseI}\$\text{F2} + \text{PhaseI}\$\text{F3})/3 \\
\text{PhaseI}\$\text{PF1} = \text{as.numeric(PhaseI}\$\text{PF1)} \\
\text{PhaseI}\$\text{PF2} = \text{as.numeric(PhaseI}\$\text{PF2)} \\
\text{PhaseI}\$\text{AvgGradPF1PF2} = (\text{PhaseI}\$\text{PF1} + \text{PhaseI}\$\text{PF2})/2
\]

Comparing Additional Phase I Models

\[
X = 1 = \text{AvgDuring} \quad X = 2 = \text{AvgPost} \quad X = 7 = \text{AvgGradF1F3} \quad X = 8 = \text{AvgGradPF1PF2} \quad Y = 4 = \text{CompDum} \\
Y = 5 = \text{RecDum} \quad Y = 7 = \text{PhDum} \quad Y = 8 = \text{MastDum} \quad Y = 9 = \text{BachDum} \quad Z = 1 = \text{AvgGradsPre}
\]

Model7.4.0 = \text{lm}\left(\text{AvgGradF1F3} ~ \text{CompDum}, \text{data} = \text{PhaseI}\right) \\
Model7.4.1 = \text{lm}\left(\text{AvgGradF1F3} ~ \text{CompDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}\right) \text{### (only AvgGradsPre but really}

Model7.5.0 = \text{lm}\left(\text{AvgGradF1F3} ~ \text{RecDum}, \text{data} = \text{PhaseI}\right) \\
Model7.5.1 = \text{lm}\left(\text{AvgGradF1F3} ~ \text{RecDum} + \text{AvgGradsPre}, \text{data} = \text{PhaseI}\right) \text{### (only AvgGradsPre but really}

Model7.7.0 = \text{lm}\left(\text{AvgGradF1F3} ~ \text{PhDum}, \text{data} = \text{PhaseI}\right)
library(lmPerm)

Model1.7.0 = lm(AvgDuring ~ PhDum, data = PhaseI)  ### (only AvgGradsPre)
Model1.7.1 = lm(AvgDuring ~ PhDum + AvgGradsPre, data = PhaseI)  ### (only AvgGradsPre)
Model1.8.0 = lm(AvgDuring ~ MastDum, data = PhaseI)
Model1.8.1 = lm(AvgDuring ~ MastDum + AvgGradsPre, data = PhaseI)  ### (only AvgGradsPre)
Model1.9.0 = lm(AvgDuring ~ BachDum, data = PhaseI)
Model1.9.1 = lm(AvgDuring ~ BachDum + AvgGradsPre, data = PhaseI)  ### (only AvgGradsPre)

Model2.7.0 = lm(AvgPost ~ PhDum, data = PhaseI)
Model2.7.1 = lm(AvgPost ~ PhDum + AvgGradsPre, data = PhaseI)  ### (only AvgGradsPre)
Model2.8.0 = lm(AvgPost ~ MastDum, data = PhaseI)
Model2.8.1 = lm(AvgPost ~ MastDum + AvgGradsPre, data = PhaseI)  ### (only AvgGradsPre)
Model2.9.0 = lm(AvgPost ~ BachDum, data = PhaseI)
Model2.9.1 = lm(AvgPost ~ BachDum + AvgGradsPre, data = PhaseI)  #

Based on the above, I will add “BachDum” as an additional control variable and AvgGradF1F3 as an additional output variable.
Phase I: Comparing Final Models

Dealing with missingness (simple)

Output Variables (Dependent): A = AvgDuring B = AvgPost C = AvgGradF1F3 D = DidChngPostPre
Input/Control Variables (Independent): AvgGradsPre, CompDum, RecDum, BachDum

library(Amelia)
library(psych)
ImpPhaseI = PhaseI[, c(13, 15, 72, 84, 85, 137, 138, 142)]
str(ImpPhaseI)
ImpPhaseI = amelia(ImpPhaseI, m = 1, noms = c("CompDum", "RecDum", "BachDum", "DidChngPostPre"))
ImpPhaseI.out = ImpPhaseI$imputations$imp1
write.csv(ImpPhaseI.out, file = "ImputedPhaseIData.csv")

multi.hist(PhaseI$ChngDuringPost)

summary(ImpPhaseI)
missmap(ImpPhaseI)

Comparing Final Models

library(lmPerm)
library(lm.beta)

ModelA1 = lm(AvgDuring ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out) ### (AvgGradsPre)
ModelA2 = lm(AvgDuring ~ AvgGradsPre + CompDum + BachDum + RecDum, data = ImpPhaseI.out) ### (AvgGradsPre)
anova(ModelA1, ModelA2) # No differences. R-squared difference is negligible.

ModelB1 = lm(AvgPost ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out) ## (CompDum)
ModelB2 = lm(AvgPost ~ AvgGradsPre + CompDum + BachDum + RecDum, data = ImpPhaseI.out) # (Barely CompDum)
anova(ModelB1, ModelB2) # No differences. R-squared difference is 2%

ModelC1 = lm(AvgGradF1F3 ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out) ### (AvgGradsPre)
ModelC2 = lm(AvgGradF1F3 ~ AvgGradsPre + CompDum + BachDum + RecDum, data = ImpPhaseI.out) ### (AvgGradsPre)
anova(ModelC1, ModelC2) # No differences. R-squared difference is negligible.

ModelD1 = glm(DidChngPostPre ~ AvgGradsPre + CompDum + BachDum, family = "binomial", data = ImpPhaseI.out) ## (AvgGradsPre) # CompDum
ModelD2 = glm(DidChngPostPre ~ AvgGradsPre + CompDum + BachDum + RecDum, family = "binomial", data = ImpPhaseI.out) ## (AvgGradsPre)
# No real differences based on AIC.

Based on the above we would use the simpler model with the following independent variables: AvgGradsPre, CompDum, BachDum
Testing Final Base Model

```r
library(lmPerm)
library(lm.beta)
library(reghelper)
library(glmperm)

ModelA1 = lm(AvgDuring ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (AvgGradsPre)
pModelA1 = lm(AvgDuring ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (AvgGradsPre)
lm.beta(ModelA1)  ### (B = 0.54)

ModelB1 = lm(AvgPost ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (CompDum)
pModelB1 = lm(AvgPost ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (CompDum)
lm.beta(ModelB1)  ### (B = 0.54)

ModelC1 = lm(AvgGradF1F3 ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (AvgGradsPre)
pModelC1 = lm(AvgGradF1F3 ~ AvgGradsPre + CompDum + BachDum, data = ImpPhaseI.out)  ### (AvgGradsPre)
lm.beta(ModelC1)  ### (B = 0.63)

ModelD1 = glm(DidChngPostPre ~ AvgGradsPre + CompDum + BachDum, family = "binomial",
               data = ImpPhaseI.out)  ### (AvgGradsPre) # CompDum
exp(cbind(OR = coef(ModelD1), confint(ModelD1)))
# Probability = 98% CompDum Probability = 14% AvgGradsPre
```

These inputs matter differently based on what outcomes we’re looking at...the average rate of graduates during funding + 1 period, as well as during the first three years of funding, is better (and only) predicted by the average rate of graduates prior to funding...but the average rate of graduates post funding (lagged) is better (and only) predicted by whether or not a site is a Comprehensive Site, supporting the plausibility that these sites have longer lasting impact. This is further supported by the logistic regression revealing a 98% probably that if a site is a Comprehensive Site it will exhibit a positive change in teacher graduation rates from pre to post. Note also the large Betas and R-squared values in the final report.