Executive Summary

For years now, Professor Teubner and I have spoken about me conducting a proper analysis of PD 12x and its effects on student persistence over time. This is that analysis. To properly carry this out, propensity score matching (PSM) was used to create a control group for the treatment group, the treatment being students enrolled in PD 12x. PSM is the gold standard of observational studies that allows us to most closely replicate the conditions we have in Randomized Controlled Trials (RCTs). Moreover, PSM is a *causal inference* analysis, which allows us to declare at its conclusion whether it is reasonable or not to suggest that PD 12x does not have a causal relationship with student persistence. After the control group was matched to the treatment group, logistic regression was employed to examine the effects of the predictors on the binary response variable of "Persisted"/"Not Persisted."

This analysis included the last five Fall and Spring semesters, examining students who persisted from Fall to Spring and Spring to Fall. When PD 12x is included in a multivariate analysis with other predictors like demographics, student type, age, credit hours attempted and earned, and GPA, **PD 12x is found to have no statistically significant effect on student persistence ($p = 0.564353$)**. Additionally, a Likelihood Ratio Test (LRT) found that the inclusion of whether a student took PD 12x or not in the predictive model had absolutely no effect on the model's ability to correctly identify students who would persist ($p = 0.8972$). Therefore, we can reasonably conclude that a student taking PD 12x *does not have a causal relationship with student persistence.*

If we were to require PD 12x (or its new code, "PD 13x") for all new students and provide it for free, it would conservatively amount to over $350,000 a year of credit free credit hours, not including the cost of salaries for the instructors.

This analysis is a great cautionary tale of why we should not rely on simple heuristics where you isolate students who took PD 12x and then just look at the proportion who persist to the next semester against the proportion of non-PD 12x students who persist. This simple univariate analysis will never be as robust as a multivariate statistical analysis that utilizes propensity score matching and logistic regression.

Feel free to reach out for clarifications and explanations of the analysis at

Understanding Propensity Score Matching

A challenge of assessing the effect of students enrolled in PD 12x on their persistence, retention, and completion is that we cannot make a Randomized Controlled Trial (RCT). For years, Cory and I have bandied about the challenges of comparing a cohort who took PD 12x with a cohort that did not. In observational research, and in causal inference analysis, there is a way that researchers have dealt with this problem for many decades. In 1983, Rosenbaum and Rubin pioneered Propensity Score Matching (PSM) as a way to try to maintain the rigor of RCTs but applied in the world of observational studies.

RCTs, in their most basic form, have a treatment group, a control group, and an outcome. Most often, we are looking for a binary outcome—Yes/No, Lived/Died, Passed/Failed, Persisted/Did Not Persist. In our case, we are wanting to see the if the treatment—did a student take PD 12x—has a *causal* effect on persistence, and if so, what is the strength of that association? To quantify the treatment effect, a few metrics are considered. One is the average treatment effect (ATE), which is the average effect, at the population level, of moving an entire population from untreated to treated. The other is the average treatment effect of the treated (ATT), which is the average effect of the treatment on those subjects who received the treatment. The goal of our research will be to ascertain the ATE, not the ATT.

In RCTs, the assignment of the treatment is randomized. Consequently, an *unbiased* estimate of the ATE can be directly computed from the study. Further, it allows the ATE to be defined in terms of the difference in means (if considering continuous outcomes) or a difference of proportions or absolute risk reduction (for binary outcomes). For binary outcomes, we can also measure relative risk and the odds ratio, usually done so through multiple logistic regression analysis.

In observational studies, the objective is still to identify cause-effect relationships even when it is not possible to use a controlled experiment. The primary difference here is that randomization of treatment and control group is not possible. There are usually enormous confounding variables related to outright comparison of treatment groups in observational studies to the larger population that is untreated. Much of the time, it is due to a form of selection bias called "volunteer bias." To try to account for as many confounding variables as possible, propensity score matching was devised.

The propensity score was defined by Rosenbaum and Rubin (1983) as the probability of treatment assignment given the baseline covariates. Typically, through logistic regression, the treatment and control groups are all given a propensity score, which is the probability a subject ends up in the treatment group. In propensity score matching, we match up the scores of each individual in the treatment group with an individual from the control group who has as similar a score as possible. By matching this way, the variables used to predict the outcome will be as similar as possible between the control group and the treatment group. This is what I have always told Cory we would need to do. You have to match, as best you can, a one-to-one comparison of subjects in both groups. The PSM allows us to systematize and mathematically accomplish this end.

In the case of this analysis, I conducted a one-to-one matching, also called "pair matching." Once this matching is complete, a causal inference analysis can be conducted to see what the effect of students taking PD 12x is on persistence. The analysis is carried out in R using a combination of packages, but the primary package for the PSM is the "matchit" package.

Exploratory Data Analysis

Before we continue with the PSM analysis, it is helpful to know what the data is, where it came from, what variables were kept, discarded, and why, as well as how any data imputation was done.

What data is included and where did it come from?

The data is retrieved from the official IR, 20th Day numbers, which are the numbers reported to the state. Here we are examining the last five Fall and Spring semesters—FA19, SP20, FA20-SP21, FA21-SP22, FA22-SP23, FA23-SP24. The initial dataframe has 60,370 students included, 28 predictors, and 2 response variables.

Feature Engineering

Feature engineering is the process of creating non-native data in order to better predict the response variable. There are two variables that were engineered here out of necessity: PD 12x/Not PD 12x and Persisted/Not Persisted.

To create a single column that has two classes, "PD 12x" and "Not PD 12x", I identified all the students who were in a PD 12x class in any of the semesters and labeled them as "PD 12x" while labeling the other students who were not in a PD 12x class as "Not PD 12x." I manually reviewed a hundred random samples to make sure that the code produced the correct labels. This label identifies the treatment group (PD 12x) and the *start of* the control group (Not PD 12x). The PSM will be done using this variable.

The second column that I had to create is the response variable on which the statistical model is trained, "Persisted"/"Not Persisted." Here, persistence is not being used as some clinical research uses it. Rather, in the instance of this analysis, *persistence* indicates the proportion of students from the previous semester that enrolled in the current semester, only examining Fall to Spring and Spring to Fall. Summer is excluded as it is never included in any clinical research about retention or persistence.

Formally, persistence is measured as follows:

$$\text{Persistence} = \frac{S_{t+1}}{S_t}$$

where S_t represents *all students* enrolled in the previous semester and S_{t+1} represents all the students from the previous semester enrolled in the current semester. For instance, if the current semester is Fall, then the persistence measures the proportion of students enrolled in Fall that were also enrolled in the Spring semester (i.e. the previous semester). For instance, if there are 6000 students enrolled in Spring 2020 and 3046 of those students reenrolled in the following Fall semester, the persistence would be:

$$\frac{\text{Spring Students Enrolled in Fall}}{\text{All Students Enrolled in Spring}} = \frac{3046}{6000} = 0.5076$$

which represents a persistence rate of 50.76% from Spring to Fall. Table 1 shows the percent of students who persisted from one semester to the next over the last five years.

Remember, there will not be a persistence value for Spring 2024 yet because that is represented by the proportion of students from Spring 2024 that are enrolled in Fall 2024; Fall 2024 enrollment is not yet complete.

Table 1: Persistence of Students

Semester	Not Persisted	Persisted
FA19	35.82%	64.18%
SP20	49.24%	50.76%
FA20	34.86%	65.14%
SP21	50.32%	49.68%
FA21	33.34%	66.66%
SP22	49.65%	50.35%
FA22	32.15%	67.85%
SP23	49.69%	50.31%
FA23	32.02%	67.98%

From Table 1, we see that with remarkable consistency, the proportion of students who persist from Spring to Fall is always around 50.00% (SP20 - 50.76%, SP21 - 49.68%, SP22 – 50.35%, SP23 – 50.31%). The percentage of students who tend to persist from Fall to Spring has changed from Fall 2019 to this last Fall, Fall 2023 (FA19 – 64.18%, FA20 – 65.14%, FA21 – 66.66%, FA22 – 67.85%, FA23 – 67.98%). It is important to note that these persistence rates are of *all students* in our student body. IR almost exclusively focuses on IPEDs data, which are a very limited cohort of the students at Butler. IPEDs data is almost entirely useless for understanding your students at an institutional level. It may be helpful for state and federal comparisons but is not helpful for understanding your students internally.

Combining Data

One of the necessities of PSM is to incorporate as many predictors as you can that you think might influence the outcome. Therefore, in addition to using the 20th day credit hour data from IR, I incorporated GPA and credit hour information from Argos. As we will cover below, there are many different ways in which institutions capture these two features, some overlap, others do not, nearly all of them, we will see, are statistically significant predictors of the outcome variable of persistence.

Variable and Variable Selection

Predictors

term (int): This is the six digit code used to identify the semester. It is necessary to include only in the propensity score matching part of the analysis so that we can match students from each term who took PD 12x and did not. It is not included in the final statistical analysis of student persistence.
age (int): This is the age of the student captured as an integer.
age_range (string): This is a discretized version of age. Ultimately it was not kept in the final model.

totcr (int): This is the number of credit hours in which a student is enrolled for a given semester.
stype (string): A single letter that denotes each student type.
resd_desc (string): A description of residency type the students is categorized as.
gender (string): The self-designated gender of each student.
ethn_desc (string): The description of the self-designated ethnicity of each student.
prevhrs (float): The number of credit hours a student has completed prior to entering the current semester.
acdstd (string): Records whether the student is in good academic standing at the end of the semester, on probation, or on academic suspension.
trmattmpt (float): Records the number of credit hours a student attempted in the current semester.
trmernd (float): Records the number of credit hours *earned* in the current semester. The proportion of earned/attempted has been found in the clinical literature to be a strong predictor of student persistence, retention, and completion.
trmgpa (float): GPA student earned for the semester.
instgpa (float): GPA of student for all courses taken at Butler.
instearn (float): All credit hours earned at Butler throughout the student's tenure.
ihrpass (float): All credit hours the student has passed at Butler.
ihratt (float): All credit hours student has attempted at Butler.
ihrgpa (float): GPA of the student for all Butler credit hours.
ogpa (float): GPA of student for all credit hours, including transfer hours.
ohrearn (float): All credit hours student has earned, including transfer hours.
ohrpass (float): All credit hours student has passed, including transfer hours.
ohratt (float): All credit hours student has attempted, including transfer hours.
ohrgpa (float): GPA for all credit hours, including transfer hours.

Response Variables

pdx (string): The binary variable that tracks whether a student was enrolled in a PD 12x course or not.
persistence (string): Binary variable that tracks whether a student persisted from one semester to another.

A keen observer would notice there are many potential predictors in this initial dataset that are remarkably similar. For whatever reason, whoever does the programming of Butler's Argos data enjoys having duplicate columns that essentially capture the same thing. Consequently, the first part of the analysis is to examine correlated predictors and eliminate multicollinearity. Figure 1 shows the strength of correlation between the numerical predictors.

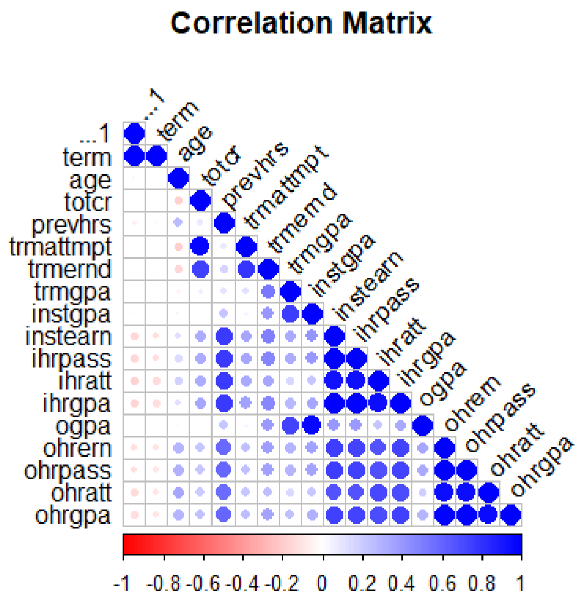

Figure 1: Correlation matrix of numerical predictors

Observations About Correlation

trmattmpt and *totcr* are 97% correlated and one is unnecessary. I will eliminate *totcr*. *ihratt* is 92% correlated with *instearn*, therefore I will eliminate *ihratt*.

ihratt is also 94% correlated with *ihrpass* so I will eliminate *ihrpass*.

ihrpass is 100% correlated with *instearn*. So I will drop the *ihrpass*.

ihrgpa is 99% correlated with *instearn*, *ihrpass*, and 95% correlated with *ihratt*. I will drop all but *ihrgpa*.

Ogpa is 97% correlated with *instgpa* so I will drop *ogpa*.

Overall GPA includes transfer credits, which not all students have. Therefore, it makes the most sence to drop that.

ohrpass and *ohearn* are 100% correlated, so I will drop the *ohrpass*. *Ohratt* is 94% correlated with *ohearn*. So I will drop *ohatt*.

Missing Values

The general rule of thumb is that if the missing values affect two percent or less, one could either drop them entirely or do some sort of data imputation. For values in which 5% or more are missing, one either needs to do data imputation or create a new class (for categorical variables) altogether. If a sufficient proportion of a predictor is missing, then it is sometimes prudent to leave it out altogether because data imputation would skew the data.

Butler's data typically does not have too many values that are missing given that so much of it is automatically generated. However, *gender* is missing 0.008% of the values. Given this miniscule number of values missing, the students were left out of the final analysis. The ethnic description variable is missing 4.35% of the values. Therefore, a new class was created for this variable simply called "Missing." The term attempted credit hours was missing 0.002% of all values, so those students with the missing values were removed from the final analysis.

Analysis of Violin Plots

Violin plots let us see both the distribution and the box plot of the variables in relation to the response variable. For this stage of the analysis, the response variable is the PD 12x/Not PD 12x. Figures 2, 3, and 4 (found in the Appendix) all show the violin plots of the numerical predictors as they relate to the PD 12x response.

Age: Not surprisingly, since it is new students who are supposed to take PD 12x, we see that the mean age of students who take PD 12x is 22.68, which is lower than those who do not take it at 23.57 years old.
Previous Credits: Similarly, we should not be surprised by the previous hours in relation to the PD 12x course. Students who are in the PD 12x course have completed fewer credits than those who did not take it. If a student has completed 21 or morecredits, they do not have to take the PD 12x course.
Attempted Credits: Attempted credits are in relation to the semester the student is in. They record the number of credits the student attempted to complete. It may seem surprising that students who took PD 12x attempted more credits, on average, than students who did not take it (11.36 crhrs vs 9.02 crhrs), but the reality is that students who take the course are typically in their first or second semester. Students in their first two semester, on average, take *more* credits than students who in later semesters. That is what we are seeing reflected here.
Earned Credits: These are the credits students actually earned in the semester (trmernd). This is going to be correlative with the attempted credits. Students in earlier semesters of their college careers attempt and earn more credits in the current semester than those who are deeper into their academic careers.
Term GPA: The term GPA *is* actually surprising. Students *who did not take* PD 12x, on average, had much higher GPAs than students who did take it. The violin plot shows there is a strong distribution above 3.0 GPA and a large spike at 4.0 GPA. Whereas the term GPA for students who took PD 12x has a comparatively slight distribution that peaks at 3.8 GPA.
Institutional GPA: Again, overall, we see a much higher institutional GPA, which is the GPA the student has earned over all their time at Butler, for students who *did not take* PD 12x versus those that did.
Institutional Earned CrHr: We see a larger spike at 60 credits for students' institutional earned credit hours for those who did not take PD 12x than those that did take it. The shape of the violin plot is roughly the same though, just more dramatic for those that did not take PD 12x than those that did.

Compare Categorical Variables With PD 12x Response

Analysis of χ^2 Test

The χ^2 Test is often used to determine if there is a significant association between two categorical variables. To test the association, the following hypothesis is offered:

H_0: The categorical variable and the binary response variable are independent and have no significant association.
H_a: The categorical variable and binary response variable are not independent and therefore have a significant association.

In its simplest form, the χ^2 Test is given by:

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where O represents the observed frequencies and E represents the expected frequencies. An α level of 0.05 will be used to test the hypothesis.

There is a statistically significant association between each of the categorical variables and the response variable *pdx*. All values approached zero and were p = 0.0001.

Propensity Score Matching

While we have covered what propensity score matching is, we have not covered its implementation. That is what this section is about. First, this process matches the students from the treatment groups and control groups with the nearest equal probability of being in the treatment group. This is a one-to-one match using the nearest neighbor without replacement. It was conducted using a logistic regression model in which each student in each group was given probability scores (propensity scores) and then those scores were matched. When the scores are matched, it means the variables of one person from the control group are as closely aligned to one person from the treatment group as possible.

We see from Table 2 that we started with 60,370 students from the five years examined. There are 4,162 students who took PD 12x and consequently make up the "treatment" group, and there are 56,208 students who did not take PD 12x who are used to match the 4,162 from the treatment group. This means 4,162 were matched with the treatment group and 52,046 were not matched. None were discarded.

Table 2: Total Values For PSM

	Control	Treated
All	56208	4162
Matched	4162	4162
Unmatched	52046	0
Discarded	0	0

When doing propensity score analysis, one should always validate the findings to make sure the treatment and control groups are behaving the way they are supposed to. Therefore, the discussion below peeks into the propensity score matching to validate its robustness.

Propensity Score Matching Validation

The first and easiest way to validate the propensity score matching is to examine the density plots of the treatment and control groups. If they perfectly overlap, then there is almost 100% matching between the treatment and control groups, meaning that each student from the treatment group is directly matched with a student from the control group in every variable considered.[1]

Recall, the propensity scores are a probability score given to each student relative to each semester and therefore are represented as a decimal falling between 0 and 1. "1" represents students who were in a PD 12x course in a given semester, and "0" represents students who were not enrolled in a PD 12x course in a given semester.

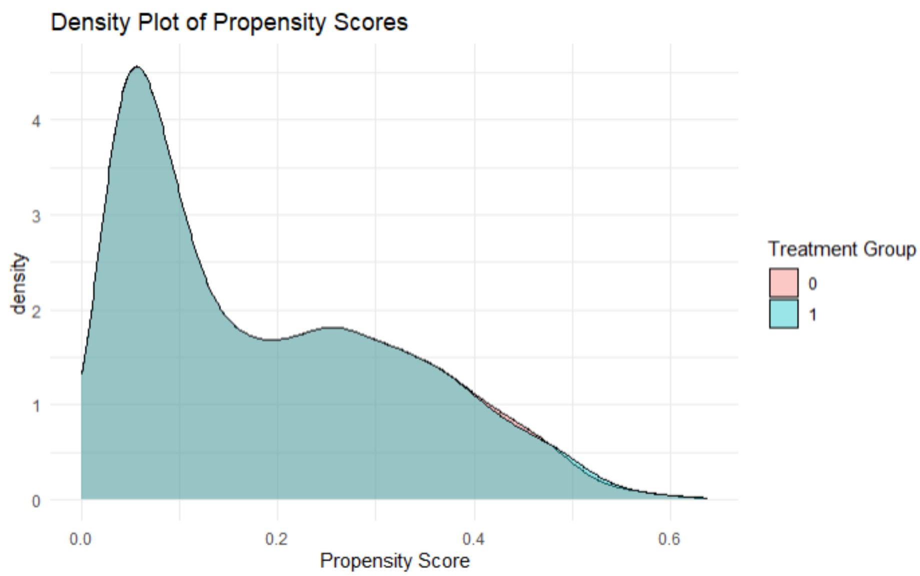

Figure 2: Matching of propensity scores in using density plot

This density plot shows remarkable overlap, so much so that you can hardly see any difference between the treatment and control groups, which means the one-to-one matching is almost identical in every way. This means each student in the treatment group is matched to a student from the control group with all the variables identically aligned—term, age, stype, resd_desc, gender, ethn_desc, prevhrs, acdstd, trmsttmpt, trmernd, trmgpa, instgpa, instearn, ohrern, and ohrgpa.

Figure 3 shows the histograms of the raw propensity scores versus the matched histograms. The visual inspection shows how radically different the propensity scores of the raw treatment and control groups were. The matched treatment and control groups are nearly identical, revealing the same as the density plot.

[1] It is common to share the tables of the data before the matching and after. These tables have the Mean Treated, Mean Control, Standardized Mean Difference, Variance Ratio (if there is one), the eCDF (empirical Cumulative Distribution Function) Mean, and the eCDF Max. These tables are in the Appendix (Table 3 and 4).

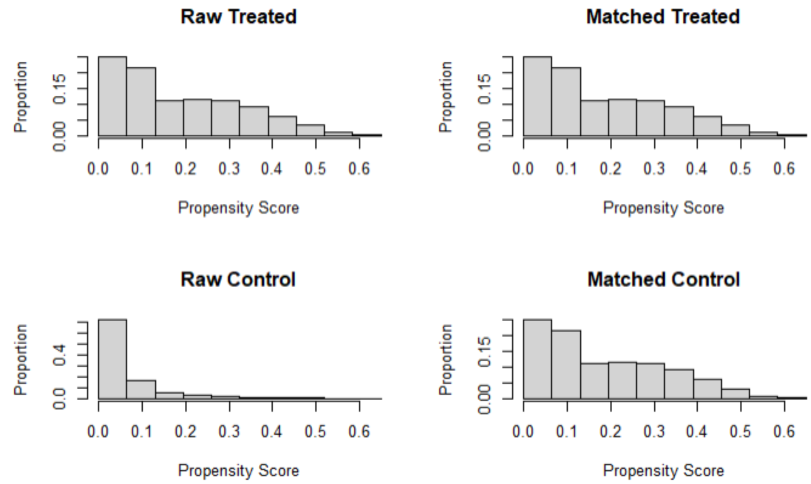

Figure 3: Histograms of effect of propensity score. Raw scores represent scores before matching. Matched scores show scores after matching.

Second, a common way to examine the robustness of the PSM is to calculate the difference between the propensity scores in each of the variables. Table 5 in the Appendix shows the difference between each variable and each class within the categorical variables. The general rule of thumb is any difference that is less than 0.10 is good. In this analysis, every single variable has an absolute difference of 0.02 or less, most having a difference of 0.00 or less, which is remarkable.

Third, we can visually inspect the original distance between the pairings and then the adjusted distance once the matches between control and treatment groups have occurred. Remember, the "adjustment" here is not due to manual manipulation of values but just represents the difference once there is a one-to-one match. Ideally, we want to see the blue dots of Figure 4 (the "adjusted" values) run almost precisely along 0.00. Figure 4 is a visual representation of Table 5.

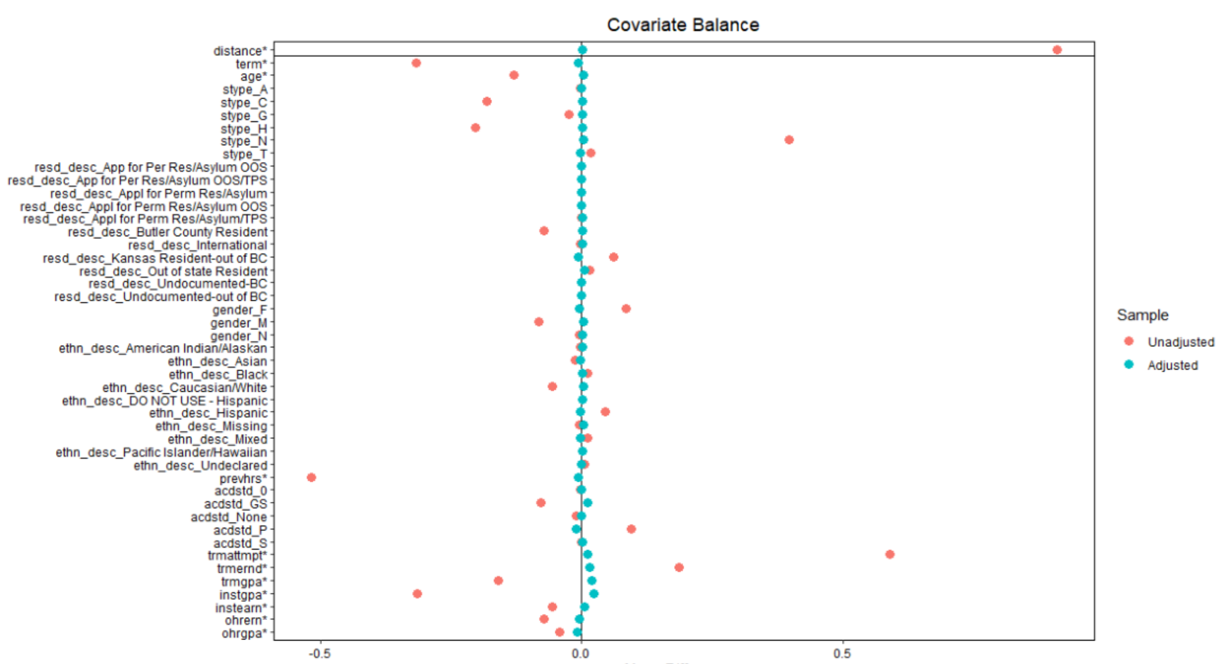

Figure 4: Visualization of corrected distance measures

Fourth, it is always prudent to conduct a sensitivity analysis to assess how sensitive the estimated treatment effects are to unmeasured confounding variables. This analysis showed a Γ value of 1, which indicates there are no unmeasured confounding variables. While this is an encouraging result, it should always be taken with caution. It does not require much thought to identify several potentially confounding variables that are uncaptured in this analysis.

All these analyses turned out as we would expect and hope. Now that the propensity score matching has been validated and we have both the treatment and control groups, we can turn our attention to the statistical analysis of the binary response variable of persistence.

Logistic Regression

A well-known form of regression analysis for binary classification is logistic regression. One of its core assumptions is that the data can be separated into roughly equal portions of both outcomes, 1 or 0, True or False, or in the case of this analysis, Persisted/Not Persisted. The model is predicated on the logistic function, also called the sigmoid function, $\sigma(z)$ and is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z is the input of the function and e is the base of the natural logarithm. z is the linear combination of the input features X and the model parameters β, which is given by the following equation:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n$$

The result of this calculation is then used to classify an observation into the positive or negative class via a simple heuristic of using a threshold of 0.5. If the predicted probability is greater than 0.5, then the prediction is rounded to 1, and if not, it is rounded to 0. To be clear, this threshold is not a fixed number and can be altered by the modeler if the cost of false positives or false negatives is of particular import.

To fit the parameters of β to the training dataset, this model *typically* uses a maximum likelihood estimator (MLE) to uncover the values of β that maximize the likelihood of observing the classification labels ('Persisted," "Not Persisted") given the features X and parameters β.

Set Up Dataframe for Logistic Regression

Split Data

To evaluate the logistic regression model, the data is randomly split into training and validation sets. The training set is made up of 80% of the random samples with 20% of the random samples held out for validation. After splitting the data, the numeric columns in the training and validation sets were scaled via standardization.

Apply Logistic Regression Model (LR)

The logistic regression model evaluates the predictors against the response variable of persistence ("Persisted" vs "Not Persisted"). In addition to the variables mentioned in the beginning exploratory data analysis, "pdx" now shifts to being a predictor as well.

LR Analysis

The LR has an 86.72% overall accuracy rate. Not too surprisingly, the *specificity*, which in this model is the measurement of students who are *most likely* to persist, is much higher than the *sensitivity*. Specificity measures the *True Negative Rate* (TNR) and sensitivity measures the *True Positive Rate* (TPR). In our case here, the positive class is those students who did not persist from one semester to another. The two measurements are given by the following equations:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{405}{405 + 97} = 0.8068$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} = \frac{1038}{1038 + 124} = 0.8933$$

It is not surprising that the specificity is better because nearly 70% of students in the matched groups persisted from one semester to another. Whatever the dominant group is, the model typically predicts with greater accuracy. The *TPR* was 0.8068, indicating that the model correctly identifies which students *will not persist* from one semester to another about 81% of the time. The *TNR* is 0.8933 and indicates that the model correctly predicts which students *will persist* from one semester to another 89% of the time. This is a model that has good explanatory power for the predictors. In a later section, we will examine what the predictors tell us, including whether students being enrolled in a PD 12x course has any statistically significant effect on a student's persistence from one semester to another.

Checking for Over Dispersion

Over-dispersion is used to assess the goodness of fit, particularly for models that assume a certain distribution of the error terms, such as with a binomial distribution in logistic regression. For logistic regression, the dispersion parameter is typically assumed to be one because the variabnce is a function of the mean. The formula for the dispersion parameter is:

$$\hat{\phi} = \frac{\text{Residual Deviance}}{\text{Degrees of Freedom}}$$

In this case, the residual deviance was 4802.60 and the degrees of freedom were 6626. Therefore:

$$\frac{4802.60}{6626} = 0.7248$$

A dispersion parameter ($\hat{\phi}$) of 0.7248 suggests a slight underdispersion in the model since the value is less than 1. Therefore, we can conclude the model is *not* overdispersed.

Model Significance

Just as the confusion matrix confirms, the model is statistically significant, meaning it is better than guessing, which is a low threshold. All it means when a binary classification model is statistically significant is that it is better than guessing. We have a 50/50 chance of guessing whether each student persists from one semester to another.

Deviance Residuals

Deviance residuals are a measure used to evaluate how well the logistic regression model fits the data. They help us quantify how well the model's predictions match the actual outcomes. The null and alternative hypothesis are:

H_0: The predictors do not improve the model fit.
H_a: The predictors do improve the model fit.

Here, an α value below 0.05 would cause us to reject the null hypothesis and conclude that the model is not a good fit. This is one of the few times that we are hoping for a value *above* 0.05 so that we fail to reject the null hypothesis and conclude that the model does fit the data well.

In our case, the deviance is 1 which means we fail to reject the null hypothesis and conclude that the model is a good fit.

Hosmer-Lemeshow Goodness of Fit Test

The Hosmer-Lemeshow GOF test also produced a p-value of 0.4586, indicating that we fail to reject the null hypothesis and conclude that the model adequately describes the relationship between the predictors and the outcome variable.

Assessing Multicollinearity

After calculating the Variance Inflation Factor to check for multicollinearity in the predictors, two predictors were identified—*ohern* and *ohrgpa*. An analysis of their correlation showed they were 98% correlated. An ANOVA test comparing the original model to a second model with these two removed found the model was not significantly improved by keeping them in the model. Furthermore, the AIC was not significantly diminished when removing them. Finally, the performance of the model remained identical; the accuracy, specificity, and sensitivity did not change. Consequently, these two predictors were removed from the model as they provide no additional explanatory power.

Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is a statistical test to compare the goodness of fit between two nested models. Nested models are models where one model is a special case of the other. The LRT evaluates whether the more complex model provides a significantly better fit to the data than the simpler model. In our case, the complex model is the one in which we leave the *pdx* variable in and the simpler model is the model without *pdx*. The goal is to determine if the model that has the *pdx* variable significantly improves the fit to the data over a model that excludes it altogether. If it improves it, then there is *some* reason to keep the predictor. Other analyses are also done to examine its impact on the response variable of persistence; this is just a starting point.

Model 1: Compare persistence against age, stype, resd_desc, gender, ethn_desc, prevhrs, pdx, acdstd, trmattmpt, trmernd, trmgpa, instgpa, and instearn.
Model 2: Compare persistence against age, stype, resd_desc, gender, ethn_desc, prevhrs, acdstd, trmattmpt, trmernd, trmgpa, instgpa, and instearn.

The final p-value of the LRT was 0.8972, indicating that there is no evidence to suggest that Model 1, which includes *pdx*, fits the data any better than Model 2, which excludes it. **In other words, the inclusion of whether a student was enrolled in a PD 12x course or not does not affect the model's ability to correctly identify which students persist or not. It has no statistically significant effect on persistence.**

Examining Logistic Regression Coefficients

Now that we have established that the model is statistically significant, that it is a good fit, and that it is not overdispersed, let us turn our attention to what the model tells us about each predictor in relation to the binary response variable of persistence versus non-persistence.

A key point to remember that will not be stated before every interpretation is that each conclusion below is written with the assumption of *ceteris paribus*, or "assuming all other variables are held constant," or, "all things being equal."

PD 12x

First, let us turn to the most burning question, the one that this whole analysis set out to examine. If a student took PD 12x in a given semester, does that result in the student having higher odds of persisting from one semester to another?

Recall, we have applied propensity score matching to align the control group with the treatment group. The treatment group being students who took PD 12x. This way, we have matched students with near identical demographic, credit hours, and GPA characteristics that both took and did not take PD 12x in a given semester.

When all of these other predictors are considered in tandem with whether a student took PD 12x or not, we find that the odds a student who took PD 12x will persist are only 4.20% higher than those who did not take it. Importantly, the p-value for this predictor is 0.564353, meaning it is not statistically significantly non-zero.

It has no statistical significance in predicting whether a student persists from one semester to another. Not only that, this is a *causal inference* analysis. **Therefore, we can reasonably suggest that PD 12x *does not have a causal* relationship with student persistence.**

This is a good lesson in why we do not do simple logistic regression–comparing just one predictor to the response–when we can include other predictors. The strength and significance of a predictor changes as other confounding variables are included in the analysis. **This multiple logistic regression clearly demonstrates that the PD 12x course has had no statistically significant effect on persistence over the last five years.** Therefore, forcing all new students to take this course and offering it for free would all be for naught. Not only that, ultimately the college would be giving away over $350,000 of credits, not including the faculty salary cost to teach the course. This would be a very bad investment. So, what predictors actually drive persistence if not enrollment in PD 12x?

Age

For each additional year that someone is alive, the odds they will persist from one semester to another, given they have completed PD 12x, raises by 1.68%.[2] What is interesting here is that this is counter to what is true of Butler's data when all students are considered and not just this limited cohort of students who took PD 12x matched with propensity score matching. Consequently, this tells us that the presence of PD 12x actually *inverts* the age predictor and its effect on the likelihood a student persists from one semester to another!

Normally, we see that as the age increases, the likelihood a student persists actually *decreases* rather than increases. Let's consider a few concrete examples. These probabilities are considering only if a student has taken PD 12x and their age. Probability is a different calculation mathematically than the log odds. This is the actual probability that a student will persist given they have taken PD 12x relative to their age.

For an 18 year old, the probability they will persist is 1.16%. Remember, this is a simplified regression equation; there are many more significant predictors than age. PD 12x, again, is not a significant predictor of persistence. Now a student who is 20 years old and has taken PD 12x has a probability of persisting to the next semester of 1.20%. If we carry this out to a 40 year old, the probability they will persist, give they have completed PD 12x is 1.66%. So again, as the age increases, the probability the student persists increases, but by vanishingly small margins.

This is interesting because the inverse of this relationship is that the younger a student is who takes PD 12x, the less effective it seems to be as the probability a student persists actually decreases as they get younger. It seems to have a *negative impact* on student persistence for younger students. My guess is there is more going on than this and it is a proxy for other issues about which we could only speculate without more information.

---

[2] These are the log odds, so it must be calculated with that in mind or your calculation does not come out correctly.

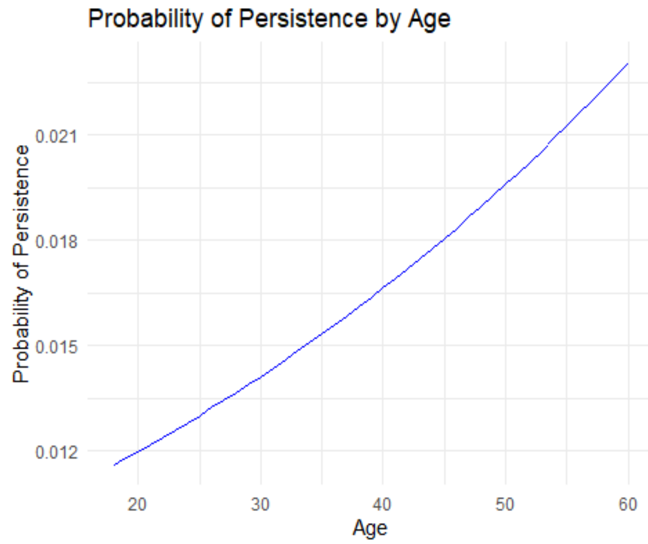

Figure 5: Probability of persistence when age and enrollment in PD 12x are considered

Student Type

The reference group for student type is continuing students. There were two student types that were statistically significant below the α-value of 0.05: Guest students (p = 0.0379) and New students (p = 0.0040). The odds a guest student persists from one semester to another are 3.38 times less than continuing students. The odds a new student persists is 0.3137 times higher than for continuing students. To translate this into an odds ratio (OR) as a percentage, one must use the natural log of the coefficients for each predictor.

$$\text{OR} = e^{coefficient}$$

If OR is greater than one, then you must subtract 1 and multiply by 100 to get the percentage $(OR - 1) * 100 = \text{percentage}$. If the OR is less than one, then you must take 1 minus the OR $(1 - OR) * 100 = \text{percentage}$.

So the odds a guest student persists is 96.58% less than that of a continuing student. The odds a new student persists is 3.99% more than a continuing student.

Residency Type and Ethnicity

When considered with these other predictors, not a single residency type or ethnicity was statistically significantly associated with the probability a student persists from one semester to another. This also demonstrates that when these demographic characteristics are considered along with other predictors, they lose their strength and significance compared to when they are considered in isolation. One should be weary of analysis that are univariate.

Previous Hours

The number of credit hours a student has completed going into the current semester is statistically significantly related to the odds of persistence (p = 0.0001). For each additional credit hour a student has completed going into the current semester, the odds he/she will persist are 0.1461 times less. This makes sense as students are more likely to graduate, transfer, or stop out the further they get into their academic career.

Academic Standing

The odds a student who has no academic standing at the end of the semester (denoted as simply '0' in Banner) will persist is 1.43 times lower than for a student who is in good academic standing. This predictor has a p-value of 0.0001. The predictor for students on academic probation has a p-value of 0.0001. The odds a student on probation persist from one semester to another are 0.8224 times less than students in good standing. The odds a student on academic suspension persist from one semester to another are 0.9508 times less than students in good standing ($p = 0.0001$).

Represented as a percentage, it is far more impressive. '0' students are 76.08% less likely to persist from one semester to another compared to students in good standing. Probation students are 56.06% less likely to persist from one semester to another than students in good standing. Academic Suspension students are 61.36% less likely to persist than students in good standing.

Credits and GPA

Credit hours attempted for the current term is statistically significant ($p = 0.0001$). Credit hours earned for the current term is statistically significant ($p = 0.0001$). End of Term GPA is statistically significant ($p = 0.0001$). Institutional GPA is statistically significant ($p = 0.0001$). Total institutionally earned credits is statistically significant ($p = 0.0001$). Overall earned credit hours is statistically significant ($p = 0.0001$). Overall GPA is statistically significant ($p = 0.0001$).

Without belaboring these other predictors, the undeniable reality sets in that the PD 12x course, when considered in conjunction with these other predictors, has no discernible impact on student persistence.

Concluding Remarks

This was a causal inference analysis to ascertain whether students taking the PD 12x course over the last five years persist at greater rates than students who did not take the course. The analysis shows clearly that taking the PD 12x course does not have a causal relationship with student persistence in any way. Consequently, we should neither offer to pay for this course for all new students nor require it for all new students, given its long track record of not having an impact on persistence.

Works Cited

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Appendix

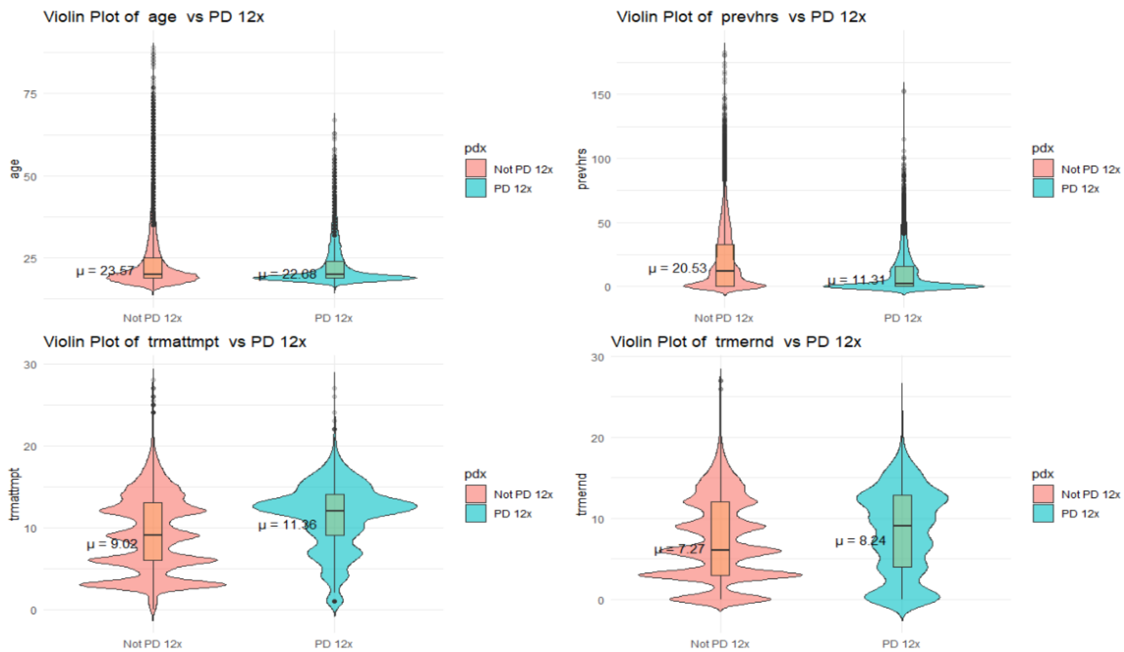

Figure 6: Violin plots of numerical predictors and PD 12x response

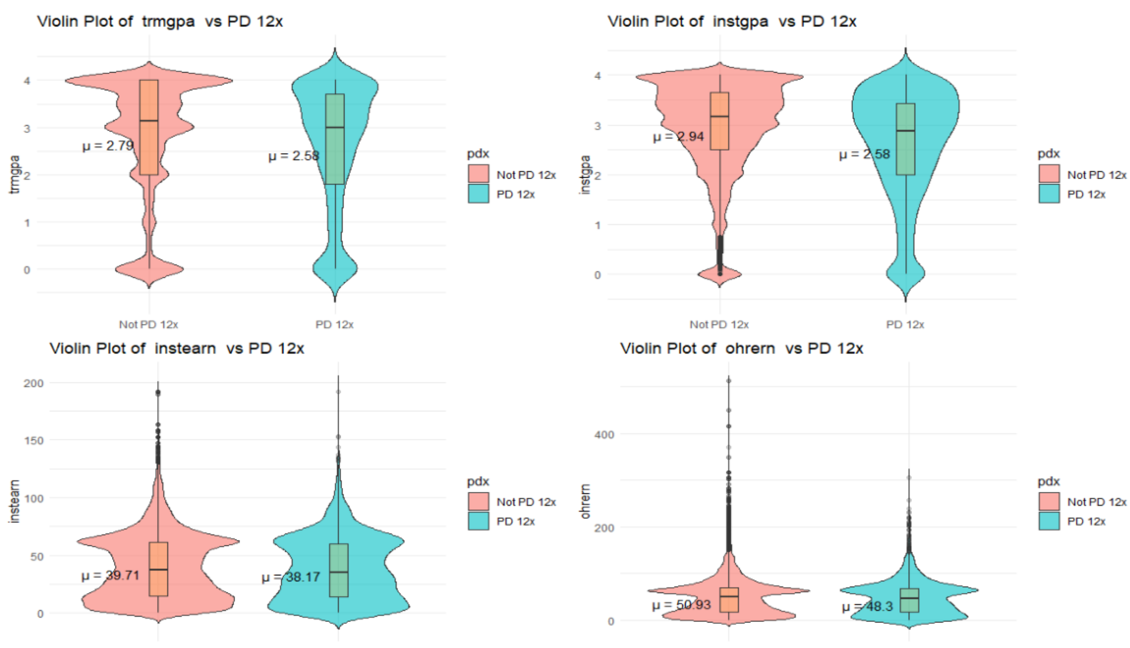

Figure 7: Violin plots of numerical predictors and PD 12x response

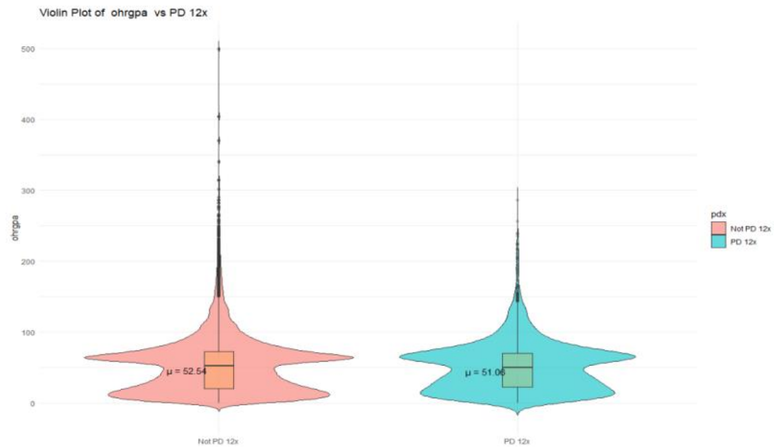

Figure 8: Violin plots of numerical predictors and PD 12x response

Table 3: Summary of Balance for All Data

Variable	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.1863	0.0603	0.9093	3.1404	0.3207	0.4689
term	202126.5401	202167.1298	-0.3164	0.9664	0.0978	0.1645
age	22.6826	23.5722	-0.1309	0.6851	0.0176	0.1116
stypeA	0	0.003	-0.0564	.	0.003	0.003
stypeC	0.3943	0.576	-0.3718	.	0.1817	0.1817
stypeG	0.001	0.0263	-0.8182	.	0.0254	0.0254
stypeH	0.0072	0.211	-2.4087	.	0.2038	0.2038
stypeN	0.5247	0.1287	0.7931	.	0.3961	0.3961
stypeT	0.0728	0.0551	0.0681	.	0.0177	0.0177
resd_descApp for Per Res/Asylum OOS	0	0.0001	-0.0087	.	0.0001	0.0001
resd_descApp for Per Res/Asylum OOS/TPS	0	0.0002	-0.0138	.	0.0002	0.0002
resd_descAppl for Perm Res/Asylum	0	0.0004	-0.0219	.	0.0004	0.0004
resd_descAppl for Perm Res/Asylum OOS	0	0.0001	-0.0076	.	0.0001	0.0001
resd_descAppl for Perm Res/Asylum/TPS	0.0005	0.0008	-0.0162	.	0.0004	0.0004
resd_descButler County Resident	0.1562	0.2278	-0.1973	.	0.0716	0.0716
resd_descInternational	0.0183	0.0205	-0.0168	.	0.0023	0.0023
resd_descKansas Resident-out of BC	0.7549	0.6935	0.1429	.	0.0614	0.0614
resd_descOut of state Resident	0.0656	0.0509	0.0594	.	0.0147	0.0147
resd_descUndocumented-BC	0	0.0003	-0.0175	.	0.0003	0.0003
resd_descUndocumented-out of BC	0.0046	0.0055	-0.0133	.	0.0009	0.0009
genderF	0.697	0.6114	0.1863	.	0.0856	0.0856
genderM	0.2989	0.3804	-0.1781	.	0.0815	0.0815
genderN	0.0041	0.0082	-0.064	.	0.0041	0.0041
ethn_descAmerican Indian/Alaskan	0.0067	0.0092	-0.0298	.	0.0024	0.0024
ethn_descAsian	0.0348	0.0473	-0.068	.	0.0125	0.0125
ethn_descBlack	0.0951	0.0831	0.0409	.	0.012	0.012
ethn_descCaucasian/White	0.5363	0.5924	-0.1125	.	0.0561	0.0561
ethn_descDO NOT USE - Hispanic	0.0007	0.0001	0.0242	.	0.0006	0.0006
ethn_descHispanic	0.1869	0.1411	0.1175	.	0.0458	0.0458
ethn_descMissing	0.0384	0.0438	-0.028	.	0.0054	0.0054
ethn_descMixed	0.0605	0.0485	0.0505	.	0.012	0.012
ethn_descPacific Islander/HawaiG	0.0022	0.0018	0.0083	.	0.0004	0.0004
ethn_descUndeclared	0.0382	0.0327	0.0287	.	0.0055	0.0055
prevhrs	11.3095	20.5332	-0.5173	0.6164	0.0645	0.2355
acdstd0	0.0211	0.0246	-0.0242	.	0.0035	0.0035
acdstdGS	0.778	0.8569	-0.1898	.	0.0789	0.0789
acdstdNone	0	0.0113	-0.1106	.	0.0113	0.0113
acdstdP	0.1634	0.0696	0.2537	.	0.0938	0.0938
acdstdS	0.0375	0.0377	-0.001	.	0.0002	0.0002
trmattmpt	11.3593	9.021	0.5888	0.7359	0.0939	0.2908
trmernd	8.242	7.2673	0.187	1.093	0.0475	0.1608
trmgpa	2.5778	2.7917	-0.1593	1.0411	0.0551	0.1127
instgpa	2.5835	2.9364	-0.3159	1.3031	0.0907	0.1412
instearn	38.1653	39.7147	-0.0566	0.9702	0.0108	0.0383
ohrern	48.3038	50.9319	-0.0729	0.8388	0.0115	0.0353
ohrgpa	51.061	52.5367	-0.042	0.8127	0.0111	0.031

Table 4: Summary of Balance for Matched Data

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.1863	0.186	0.0021	1.0089	0	0.0062	0.0023
term	202126.5401	202127.309	-0.006	1.0618	0.0106	0.0353	0.9875
age	22.6826	22.6651	0.0026	0.971	0.0039	0.0322	0.8013
stypeA	0	0	0	.	0	0	0
stypeC	0.3943	0.394	0.0005	.	0.0002	0.0002	0.3122
stypeG	0.001	0.0007	0.0078	.	0.0002	0.0002	0.0543
stypeH	0.0072	0.007	0.0028	.	0.0002	0.0002	0.0426
stypeN	0.5247	0.5223	0.0048	.	0.0024	0.0024	0.2435
stypeT	0.0728	0.0759	-0.012	.	0.0031	0.0031	0.4948
resd_descApp for Per Res/Asylum OOS	0	0	0	.	0	0	0
resd_descApp for Per Res/Asylum OOS/TPS	0	0	0	.	0	0	0
resd_descAppl for Perm Res/Asylum	0	0	0	.	0	0	0
resd_descAppl for Perm Res/Asylum OOS	0	0	0	.	0	0	0
resd_descAppl for Perm Res/Asylum/TPS	0.0005	0	0.0219	.	0.0005	0.0005	0.0219
resd_descButler County Resident	0.1562	0.155	0.0033	.	0.0012	0.0012	0.7261
resd_descInternational	0.0183	0.0178	0.0036	.	0.0005	0.0005	0.262
resd_descKansas Resident-out of BC	0.7549	0.7624	-0.0173	.	0.0074	0.0074	0.8519
resd_descOut of state Resident	0.0656	0.0598	0.0233	.	0.0058	0.0058	0.4755
resd_descUndocumented-BC	0	0	0	.	0	0	0
resd_descUndocumented-out of BC	0.0046	0.005	-0.0071	.	0.0005	0.0005	0.1426
genderF	0.697	0.7018	-0.0105	.	0.0048	0.0048	0.8815
genderM	0.2989	0.2953	0.0079	.	0.0036	0.0036	0.8749
genderN	0.0041	0.0029	0.0188	.	0.0012	0.0012	0.1017
ethn_descAmerican Indian/Alaskan	0.0067	0.0065	0.0029	.	0.0002	0.0002	0.1617
ethn_descAsian	0.0348	0.0377	-0.0157	.	0.0029	0.0029	0.3878
ethn_descBlack	0.0951	0.0942	0.0033	.	0.001	0.001	0.5929
ethn_descCaucasian/White	0.5363	0.5327	0.0072	.	0.0036	0.0036	0.9998
ethn_descDO NOT USE - Hispanic	0.0007	0	0.0269	.	0.0007	0.0007	0.0269
ethn_descHispanic	0.1869	0.1893	-0.0062	.	0.0024	0.0024	0.7802
ethn_descMissing	0.0384	0.0351	0.0175	.	0.0034	0.0034	0.3699
ethn_descMixed	0.0605	0.0627	-0.0091	.	0.0022	0.0022	0.4967
ethn_descPacific Islander/Hawaiian	0.0022	0.0019	0.0052	.	0.0002	0.0002	0.0879
ethn_descUndeclared	0.0382	0.0399	-0.0088	.	0.0017	0.0017	0.3999
prevhrs	11.3095	11.4231	-0.0064	1.0835	0.0051	0.0226	0.5874
acdstd0	0.0211	0.0216	-0.0033	.	0.0005	0.0005	0.2739
acdstdGS	0.778	0.7674	0.0254	.	0.0106	0.0106	0.8186
acdstdNone	0	0	0	.	0	0	0
acdstdP	0.1634	0.1737	-0.0279	.	0.0103	0.0103	0.696
acdstdS	0.0375	0.0372	0.0013	.	0.0002	0.0002	0.3858
trmattmpt	11.3593	11.3168	0.0107	0.9206	0.0154	0.0543	1.0106
trmernd	8.242	8.1664	0.0145	0.9407	0.0129	0.061	1.1264
trmgpa	2.5778	2.5535	0.0181	0.9503	0.0095	0.0274	1.0968
instgpa	2.5835	2.5568	0.0239	0.8983	0.0124	0.0252	1.0859
instearn	38.1653	38.0272	0.0051	1.0036	0.0077	0.0317	1.0441
ohrern	48.3038	48.4535	-0.0042	0.9259	0.0048	0.03	1.0224
ohrgpa	51.061	51.3577	-0.0084	0.9388	0.005	0.0259	1.0055

Table 5: Balance Measures

Variables	Type	Diff.Adj
distance	Distance	0.0021
term	Contin.	-0.006
age	Contin.	0.0026
stype_A	Binary	0
stype_C	Binary	0.0002
stype_G	Binary	0.0002
stype_H	Binary	0.0002
stype_N	Binary	0.0024
stype_T	Binary	-0.0031
resd_desc_App for Per Res/Asylum	Binary	0
resd_desc_App for Per Res/Asylum	Binary	0
resd_desc_Appl for Perm Res/Asyl	Binary	0
resd_desc_Appl for Perm Res/Asyl	Binary	0
resd_desc_Appl for Perm Res/Asyl	Binary	0.0005
resd_desc_Butler County Resident	Binary	0.0012
resd_desc_International	Binary	0.0005
resd_desc_Kansas Resident-out of	Binary	-0.0074
resd_desc_Out of state Resident	Binary	0.0058
resd_desc_Undocumented-BC	Binary	0
resd_desc_Undocumented-out of BC	Binary	-0.0005
gender_F	Binary	-0.0048
gender_M	Binary	0.0036
gender_N	Binary	0.0012
ethn_desc_American Indian/Alaska	Binary	0.0002
ethn_desc_Asian	Binary	-0.0029
ethn_desc_Black	Binary	0.001
ethn_desc_Caucasian/White	Binary	0.0036
ethn_desc_DO NOT USE - Hispanic	Binary	0.0007
ethn_desc_Hispanic	Binary	-0.0024
ethn_desc_Missing	Binary	0.0034
ethn_desc_Mixed	Binary	-0.0022
ethn_desc_Pacific Islander/Hawai	Binary	0.0002
ethn_desc_Undeclared	Binary	-0.0017
prevhrs	Contin.	-0.0064
acdstd_0	Binary	-0.0005
acdstd_GS	Binary	0.0106
acdstd_None	Binary	0
acdstd_P	Binary	-0.0103
acdstd_S	Binary	0.0002
trmattmpt	Contin.	0.0107
trmernd	Contin.	0.0145
trmgpa	Contin.	0.0181
instgpa	Contin.	0.0239
instearn	Contin.	0.0051
ohrern	Contin.	-0.0042
ohrgpa	Contin.	-0.0084