Gender figures in math: Do female math teachers have an impact on female students' outcomes?

Wenhui Xu

1 Introduction

'Math isn't for girls.' All women have heard this statement at one time or another while growing up. While we can't claim the gender gap doesn't exist, this stereotype is not always true. There is practically no difference in math abilities between men and women. Studies have shown girls and boys had equal strength in math before entering kindergarten [Fryer and Levitt (2010)]. If nature does not explain the gap, nurture seems to fit the narrative better. For example, a study found that female and male students react differently to competitions [Lyons et al. (2022)], with females more likely to preform well in low-stakes competitions. People usually look at the rankings of competitions only, and their high-stake nature makes achievement favour boys. Another study suggested that families and schools in rural areas may not provide girls with the same educational resources and opportunities that are given to boys [Tsai et al. (2018)]. Other studies have pointed out that the gender gap in math is likely due to the internalisation of gender roles [Keller et al. (2021)].

Examining at a smaller scale, the environment shapes how one perceives themself. For example, if you remind a girl about her Asian identity before a math exam, she will do well. But if you emphasise the fact that she's a girl, the results tend to be less favourable [Steele and Ambady (2006)]. Growing up, girls subconsciously internalize their gender stereotypes from their environment, and sometimes even tailor themselves to better fit them. Women, in hindsight, are more aware of this and just play along to fit in [Lindner et al. (2022)]. To be clear: there is nothing wrong with protecting yourself. However silence, or so-called 'stigma consciousness,' often results in underrepresented figures in stereotyped fields; reinforcing male dominance. This is famously demonstrated in STEM fields, where women are underrepresented and stereotyped as preforming worse than men [Chan et al. (2021)]. This is known

as a negative feedback loop.

On average, female math teachers are less likely to gender bias their students or favour boys in the class, and there is a growing number of studies documenting the importance of female math teachers in the classroom. (e.g.Winters et al. (2013), Antecol et al. (2015))

Much of the literature on the subject discovered while writing this paper was published before the widespread use of machine learning. In these previous studies, associations were often concluded, but not causality. Understanding causality is important when designing policies in relevant fields. This study aims to fill in the blanks by answering two questions. First, what impact do female math teachers have on female students? That is, in what measurable outcomes and to what degree? Second, does exposing female students to female teachers help with closing the gender gap in these outcomes? This study is built on previous literature, where a large-scale RCT was conducted in rural China to study the effect of pay design on teachers' performances. I extend the previous research by exploring the effect of the gender gap in math, and the impact of gender-matching teachers on students for all subgroups in different teacher incentive pay design settings. The OLS model confirmed the gender gap in math, after accounting for individual-specific characteristics. On average, girls' mathematic achievements were lower. On the upside, girls grade improved if they had female math teachers, which is estimated to be significant and closing up the gender gap by one-third.

2 Data and Summary Statistics

2.1 Data Source

The data used in this study comes from a field experiment conducted in rural China. The original field experiment measured test scores in math, science, and reading to examine the impact of pay designs on teachers' performances. The sample comes from two provinces: Shaanxi (ranked 12th in GDP per capita among China's 1st level administrative regions) and Gansu (Ranked 27th). Among the 2 provinces, 216 schools were randomly selected from their nationally ranked "poverty counties", counties where annual earnings were below a certain threshold. All schools were publicly funded in rural China and had around 400 students. The data was collected through 4 surveys, conducted from 2012 to 2014, to explore different pay designs.

Student-level microdata was collected through two baseline surveys: one conducted in May 2012, and the other in September 2012. The surveys collected information such as age and gender, education

level and occupation of parents, number of siblings, and family assets. Another endline survey (May 2014) was also conducted, focusing on non-cognitive factors (self-concept, anxiety, and intrinsic and instrumental motivation scales) towards math, curricular coverage and difficulty of math curriculum, time spent on math and other subjects per week, perceptions of teaching practices, care shown by teacher, management of the classroom, and communication by the teacher; as well as the involvement of parents in schoolwork.

Teacher-level microdata was collected in one baseline survey, conducted at the start of the school year (Sep. 2013). This survey collected data on the teacher's gender, ethnicity, age, experience level, and credentials, as well as attitude towards performance pay alongside their current performance pay. The teacher's perceived returns on teaching effort for individual students within the class were also collected. At the end of sixth grade, a nearly identical survey was given to teachers.

During the 3 surveys issued to students, a 35-minute standardized math test was conducted. These tests were constructed by trained psychometricians in order to make sure a fair measurement was taken. The items on the math test were selected from the standardized math curricula for primary school students in Shaanxi and Gansu. The content validity was then verified by multiple experts, and then the tests were validated to ensure knowledge tested was evenly distributed (ie, no bottom or top coding). The test scores were normalized against a control group, with each test's scores being normalized separately, expressing the estimated effects in standard deviations. In this study, the entire original sample was used (N = 9072).

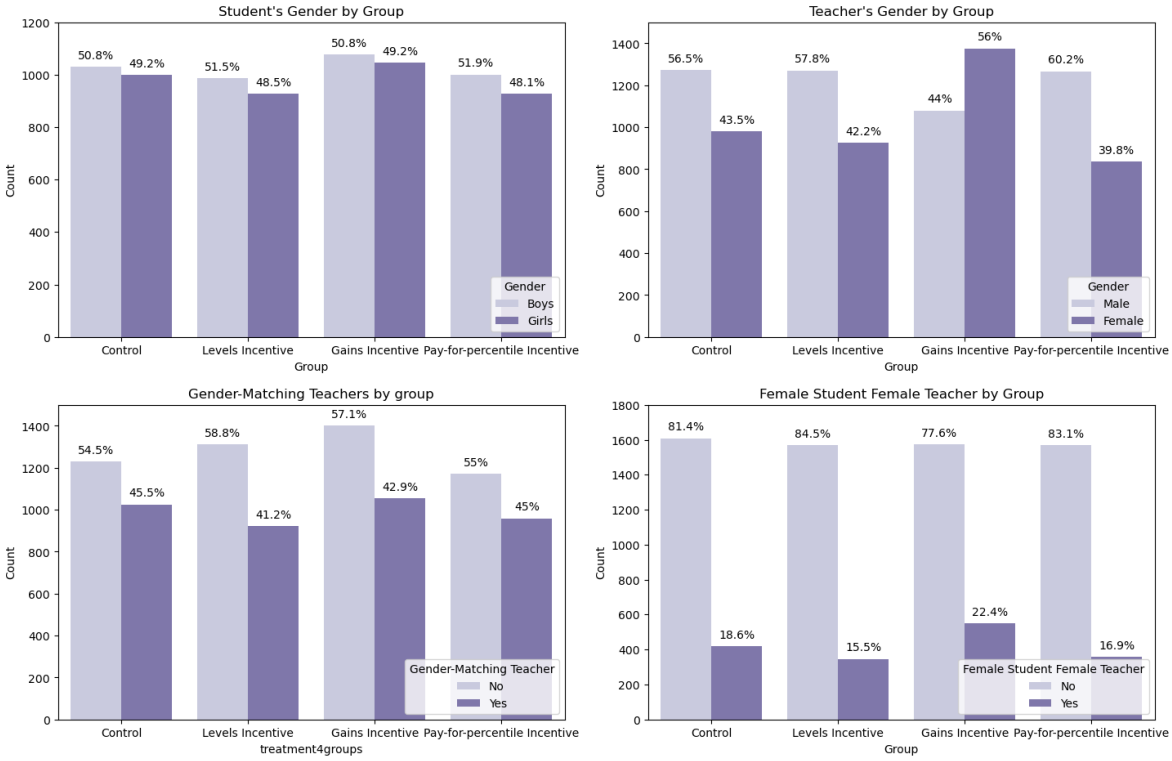
2.2 Key Variable and Summary Statistics

Table 1: Summary Statistics

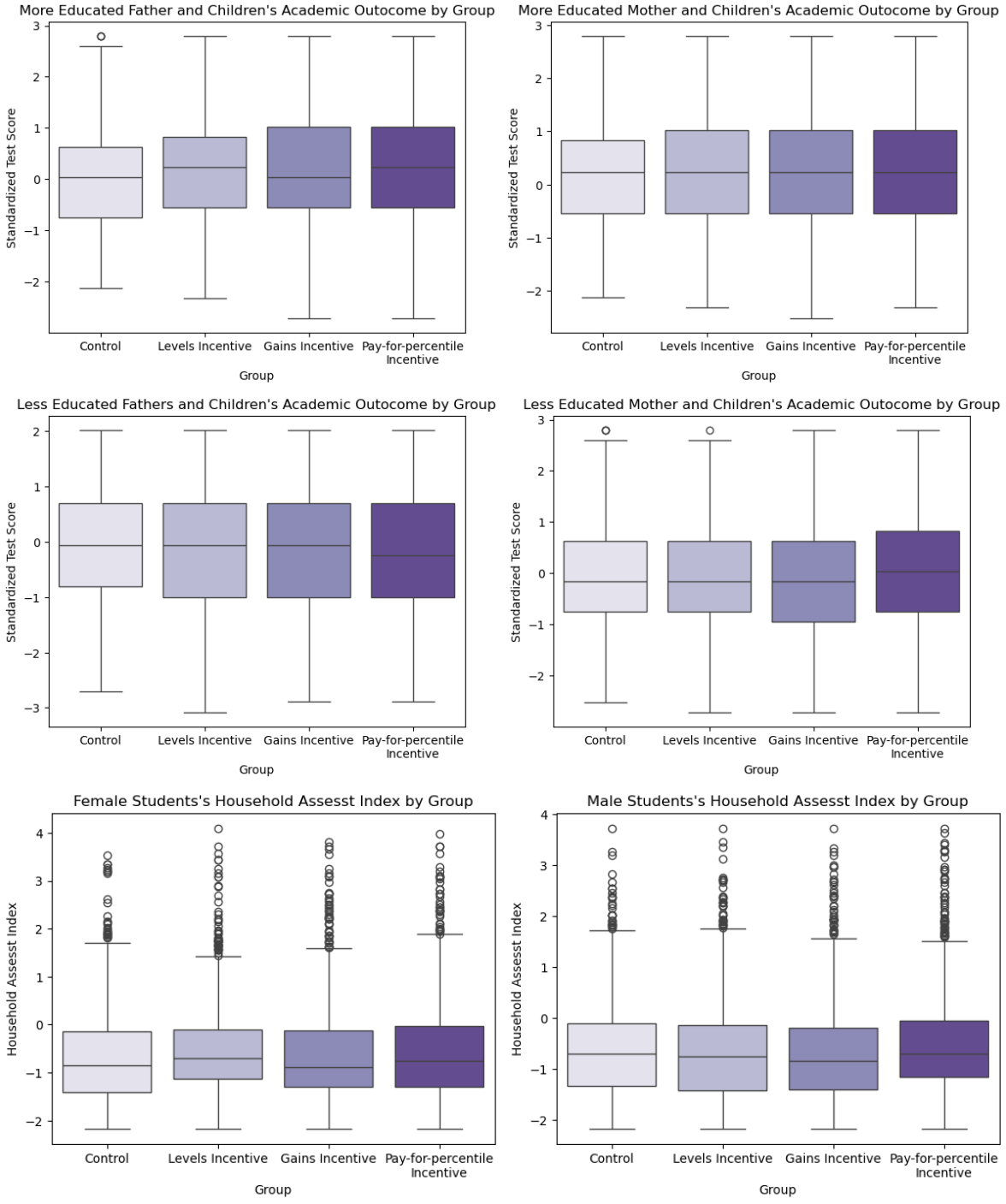
	count	mean	std	min	max
final score	8559.000000	0.069884	1.046860	-2.715281	2.800046
math baseline 1	7996.000000	-0.024245	1.013209	-3.270000	2.025307
math baseline 2	8136.000000	0.013555	1.026968	-3.209800	2.038579
student age	7992.000000	12.027277	1.033325	9.000000	19.000000
socioeconomic score	7996.000000	-0.601891	1.022967	-2.167915	4.097287
class size	9072.000000	44.054672	15.722632	7.000000	103.000000
teacher experience	9006.000000	12.339223	9.895316	0.083000	54.000000
teacher base pay	9006.000000	2935.605713	915.124634	0.000000	5220.000000

Table 1 displays a summary of statistics. Each student has three records of mathematic academic performance. I only have access to the standardized results, which are later on used in this study.

The standardized test score is the outcome variable obtained during the endline survey. The other two standardized exam scores were recorded at the beginning of the previous school year, and at the end of the school year, respectively.



There were approximately the same amount of female and male students in each design group, but with an imbalanced female:male teacher ratio. 60% of teachers were male in three out of the four groups. It was quite the opposite for the gain incentive design group which had 56% of the teachers being female. In each group, combining with the fact that less than half the student got a gender matching teacher, only an average of 18% of the female students had a female teacher .



Despite the entire sample being drawn from nationally designated poor counties in China, girls came from more disadvantaged families compared to their male peers. The variable household assets index was constructed by using polychronic principle components. This variable is an important measure of how much resources were available in a family for the child's education. Both boys and girls were more

likely to have a more educated father who at least finished junior high school (average 33.6% of the mothers and 53% of the fathers finished junior high school), however mother's education level matters more. In the best case, the father's education doesn't contribute towards students' grade as much as the mother's. By contrast, in the worst case less educated mothers are associated with lower median in each group.

3 Model Analysis and Implementation

3.1 Linear Regressions

Model Specification As the first step I created the following linear regression model to capture the gender gap in math, while capturing the impact of gender-matching teachers on students.

$$y_{ijc} = \alpha + Female_{ijc}\beta + T_{ijc}'^{t}\phi + X'_{ijc}\gamma_i + \tau'_c + \epsilon_{ijc}(2)$$

where, y_{ijc} is the *standardized math test score* for individual student i at school j in county c. $Female_{ijc}$ is a dummy variable if the individual i at school j in county c is female. $T_{ijc}'^{t}$ is a set of dummy variables for additional interests – the impact of gender matching teacher on students, where t = {Female Student Female Teacher, Male Student Male Teacher}. $X'_{ijc}$is a vector of control variables, including individual-specific data (two waves of baseline achievement scores and student age, gender, parent educational attainment, a household asset index, class size, teacher experience, and teacher base salary) *note, this also included the RTC assignments from the original studies, but as it is irrelevant to this paper would not be discussed and reported further[Loyalka et al. (2019)]. τ'_c is a set of variables account for county and school fixed effects.

Result The model shows strong evidence of a persistent gender gap, whether with or without gender-matching teachers. On average, girls' mathematic achievements are 0.142 standard deviations lower than boys' (Table 2 Column 1). The gap is more profound when gender-matching teachers are explicitly considered. Girls with teachers of the opposite gender (female student male teacher) performed 0.206 standard deviations less than boys with teachers of the opposite gender (male student female teacher). However, the evidence shows that the gender gap is reduced by 37% when girls have teachers of the same gender (female student female teacher), which is more than a third. Surprisingly, boys with the same gender teachers (male student male teacher) performed worse than those with the opposite gender teachers (male student female teacher). The coefficient of -0.052 is not significant (Column 2

Row 3) after clustering at the school level. I have attached a copy of the raw regression result, without clustering, in the appendix.

Table 2: Effect of Teacher-Student Gender Match on Math Test Scores

	Outcome Variable: standardized math test score	
	(1) Base	(2) Gender-Matching
Female	-0.142***	-0.206***
	(0.019)	(0.047)
Female student female teacher		0.078*
		(0.047)
Male student male teacher		-0.052
		(0.049)
Observations	7373	7373
R^2	0.476	0.477
Adjusted R^2	0.474	0.475
Residual Std. Error	0.756 (df = 7343)	0.756 (df = 7341)
F Statistic	137.696*** (df = 29; 7343)	130.007*** (df = 31; 7341)

Note: The standard errors are clustered at the school level *p<0.1; **p<0.05; ***p<0.01

3.2 Linear Regression with Regularization

In this subsection, linear regularization models Ridge and LASSO are used to consolidate the variable selection. The regression specification is the same as the linear regression, for simplicity I will list the objective functions instead.

3.2.1 Ridge Regression

Ridge Regression applies penalties to coefficients, resulting in a coefficient shrinkage. The shrinkage is proportional and makes the more important explanatory variables stand out.

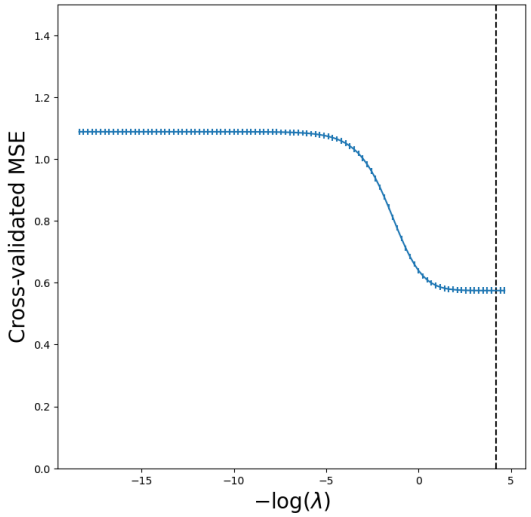
Objective Function: The shrinkage is achieved by optimizing the following objective function, where the penalty terms has a l_2 form.

$$\min_{b \in \mathbb{R}^p} \sum_i (y_i - b' X_i)^2 + \lambda \cdot \sum_{j=1}^{p} b_j^2 \hat{\psi}_j$$

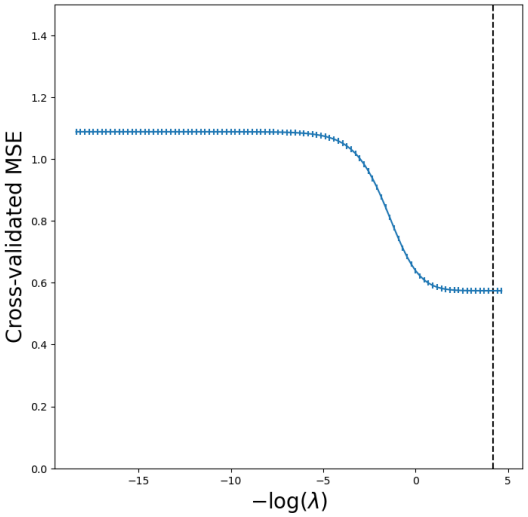
Result The optimal penalty value is determined by five-fold cross-validation, which minimizes the corresponding Mean-Square-Error. The vertical dashed lines in the figure 1 correspond to the best penalty value of 0.0152. Both models have the same penalty value, probably due to the fact that they

only differ by two extra variables. The penalty value 0.0152 is very small, suggesting the model can't shrink further without underfitting the data. Thus the optimal state of both models is to keep all variables.

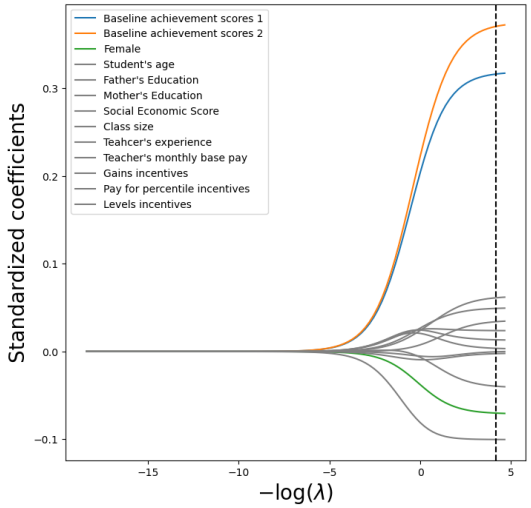
Figure 1: Ridge: regularization effect on model performance



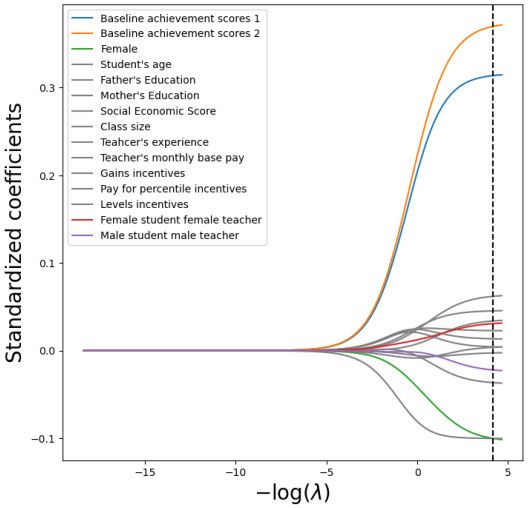
(a) Penalized ridge coefficient for the Base model



(b) Penalized ridge coefficient for the Gender-Matching model



(c) Five-folds cross-validation MSE for the Base model



(d) Five-folds cross-validation MSE for the Gender-Matching model

Note: School and county dummies are excluded from the above graphs for simplicity. The raw graphs are attached in the appendix

3.2.2 LASSO Regression

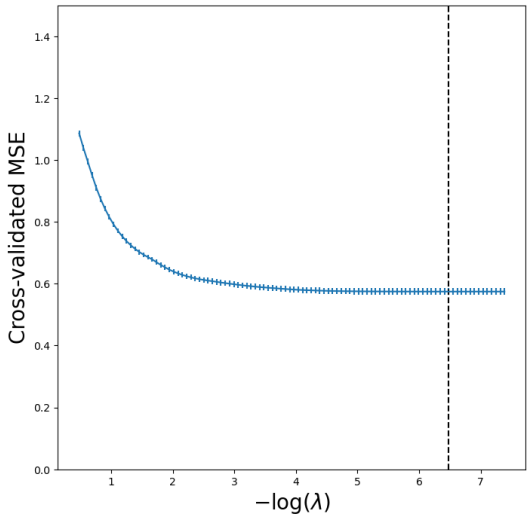
Unlike Ridge, LASSO Regression applies penalties to coefficients disproportionally, allowing coefficients to drop out.

Objective Function The shrinkage is achieved by optimizing the following objective function, where the penalty terms has a l_1 form.

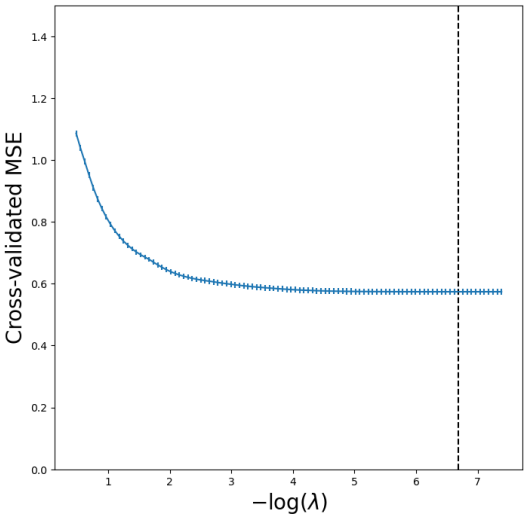
$$\min_{b \in \mathbb{R}^p} \sum_i (y_i - b' X_i)^2 + \lambda \cdot \sum_{j=1}^{p} |b_j| \hat{\psi}_j$$

Result The optimal penalty value is determined by five-fold cross-validation, which minimizes the corresponding Mean-Square-Error. The vertical dashed lines in the figure 2 correspond to the best penalty value of 0.0016. Similarly, this value suggests neither models needs a strong regularization to achieve better performance. LASSO suggests the same idea as Ridge for the base model, where all important variables (highted with colour) should be kept. However, for the Gender-Matching model, LASSO suggests otherwise. As the insignificance I find in the OLS, at the best penalty value of 0.0016, the coefficient on the dummy variable for male student male teacher almost drops out. This variable has limited power in explaining the students' outcome – standardized test score. However, I still need to keep the variable for economic interpretation in the future step.

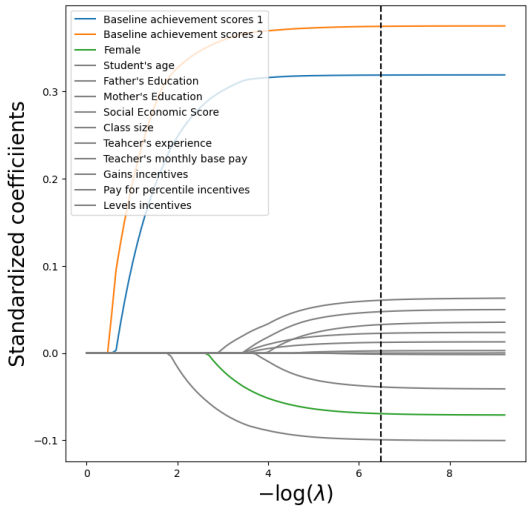
Figure 2: LASSO: regularization effect on model performance



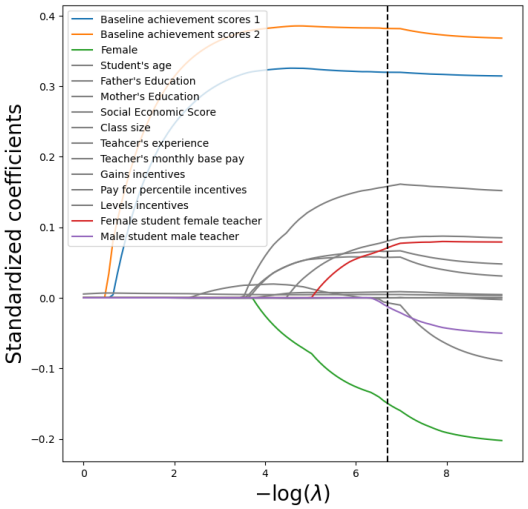
(a) Penalized ridge coefficient for the Base model



(b) Penalized ridge coefficient for the Gender-Matching model



(c) Five-folds cross-validation MSE for the Base model



(d) Five-folds cross-validation MSE for the Gender-Matching model

Note: School and county dummies are excluded from the above graphs for simplicity. The raw graphs are attached in the appendix

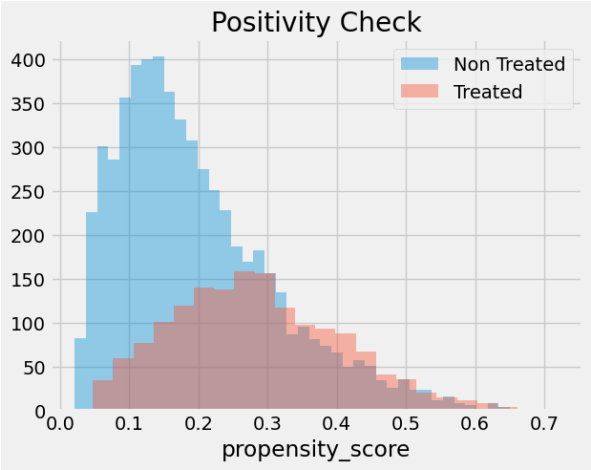


Figure 3

3.3 Inverse Propensity Score Weighting

Model Specification

Identification Strategy

The identification for the inverse propensity score weighting requires the treatment assignment is independent of the potential outcomes, conditional on observed confounder.

$$Y_1, Y_0 \perp T|X$$

The above assumption is usually not feasible in real life, thus an alternative condition involving propensity score is considered.

$$Y_1, Y_0 \perp \hat{T}(X)|X$$

Figure 3 shows the overlapping distribution of probability of both treated and non-treated. The positivity check has a moderate overlap, demonstrating the conditional independence assumption is satisfied. The ATE estimate is 0.071 for girls having female teacher, which is very close to the OLS estimate.

3.4 Ensemble Trees Method

3.4.1 Decission Tree

As shown in the figure 4, the tree's root node splits on baseline achievement scores. This suggests that the variable influences most predictions and that it is a critical determinant in the students' final achievement scores. This is not surprising, as past achievement scores are reflective of one's ability, which is also related to the individual's final outcome. Besides the academic factor, Social economic scores are a key splitting criterion on several branches, showing that the model can detect the disparities linked to socioeconomic factors. Teacher's experience, salary, and class size also appear, although they are further down the tree. Additionally, the fact that school ID appears as a splitting criterion shows a sign of school quality heterogeneity.

Looking at the extreme branches of the tree, high baseline achievement, socioeconomic status, and optimal learning conditions (smaller classes, well-paid teachers) lead to the highest prediction branch. This matches with research showing that prior achievement and access to resources significantly boost educational outcomes. Conversely, the opposite lead to the lowest prediction branch, reinforcing the importance of addressing systemic inequality to improve outcomes for disadvantaged students. Targeted interventions like smaller class sizes, improved teacher training, and better resource allocation could help students in the lowest prediction branch. A focus on improving baseline achievement, as well as addressing socioeconomic disparity, would yield the most significant impact.

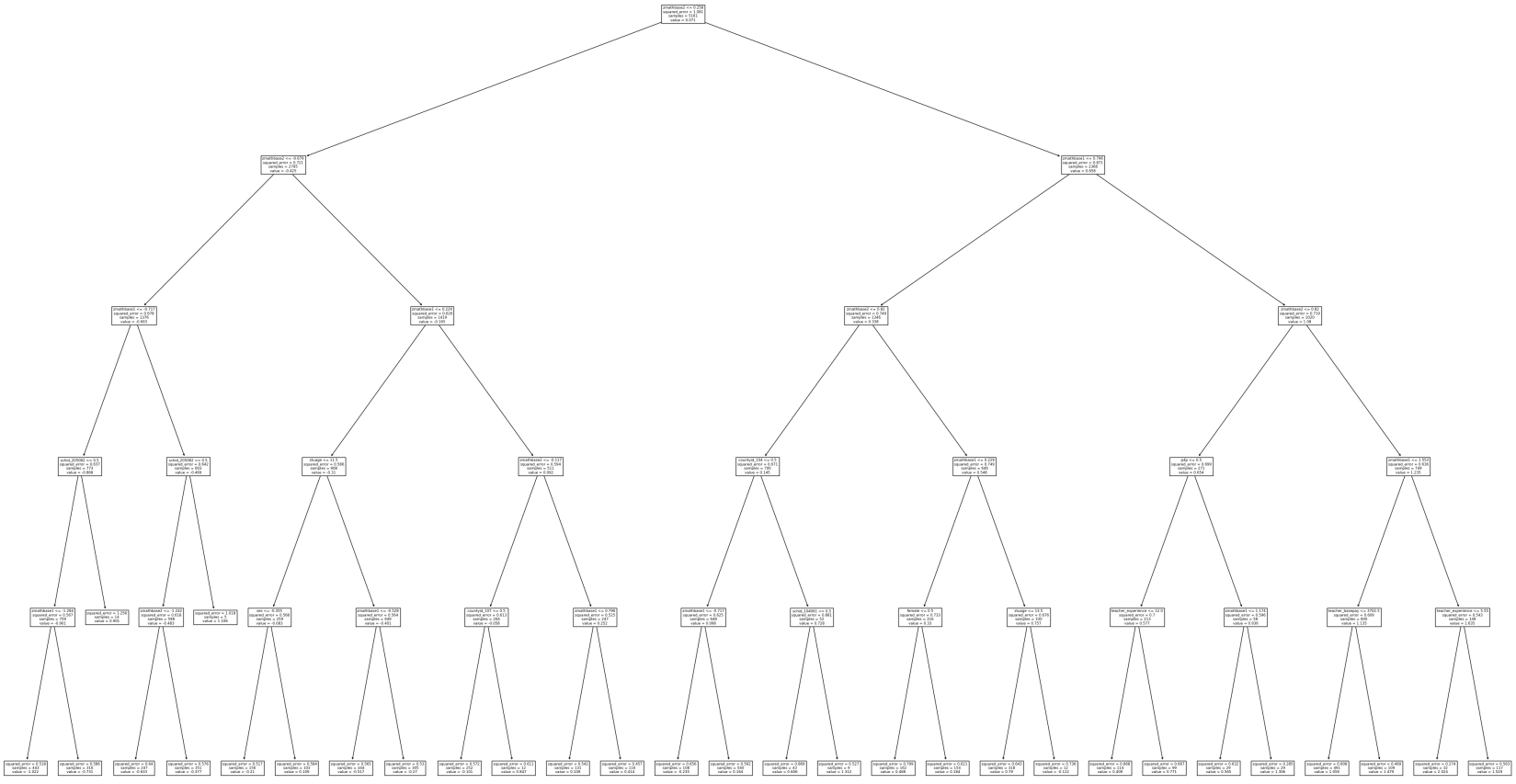


Figure 4: Regression Tree

The Feature Importance Index shows baseline achievement scores 1 and 2 to be the best predictors, suggesting a strong correlation with the target variable. Social economic score, teacher's base monthly salary, and class size are the next best predictors, though they have significantly lower Feature Importance Index results.

	Importance
Baseline achievement score 2	0.409373
Baseline achievement score 1	0.154122
Social economic score	0.055577
Teacher's base monthly salary	0.039246
Class size	0.037650
Teacher's experience	0.034000
Student age	0.028833
School ID 205082	0.014652
School ID 107081	0.007588
Female	0.007333

Table 3: Feature importance index (only top 10 included)

When it comes to model selection, the Boosting model has the lowest Mean Squared Error, meaning it has the most accurate predictions. Thus, in the following machine learning models, gradient boosting would be applied instead of the regression tree and the bagging estimators.

Mean Squared Error (MSE) Comparison

Table 4: Tuned MSE for Different Models

Model	**Tuned MSE**
Regression Tree	0.655
Bagging	0.565
Boosting	0.535

3.5 Direct Analytic Graphs

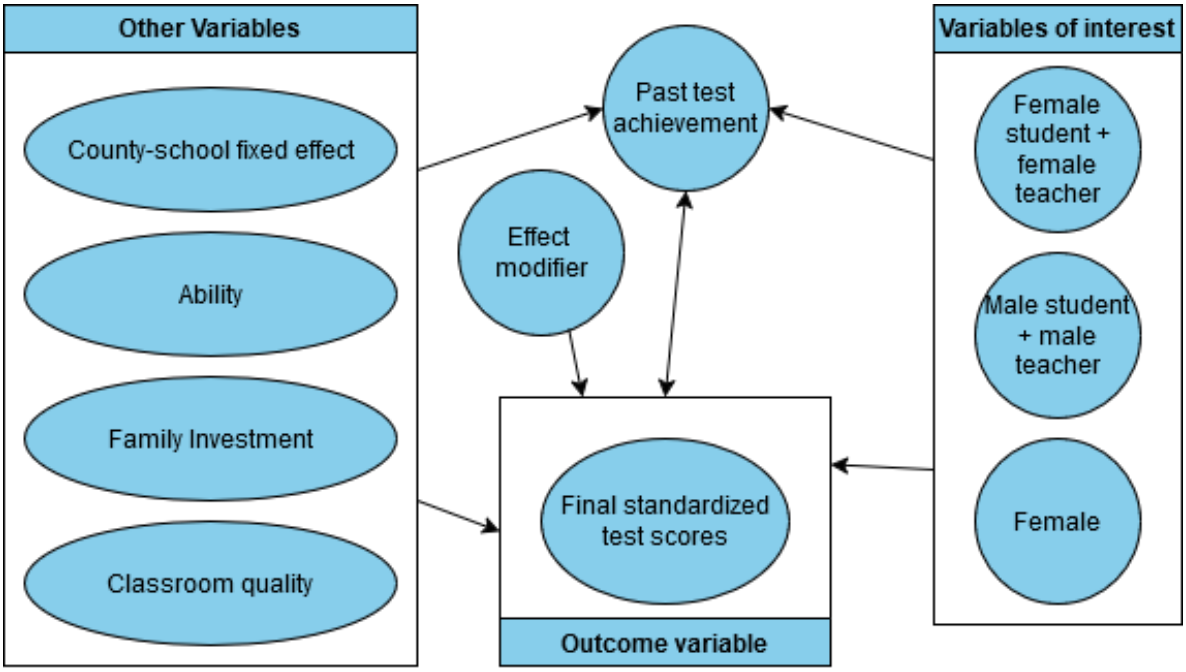
In this session, only one DAG graph is constructed, which captures both the gender gap and the gender-matching effects.

Identification Strategy

The important assumption for DAG is Unconfoundedness, or alternatively, there is no unobserved confounder affecting both the treatment and the outcome. I estimated three separate models accounting for three treatments: female, female-student-female-teacher, and male-student-male-teacher. As

shown in Figure 5, there exists a backdoor relationship between each treatment and the outcome. Thus, if relative confounder are properly controlled, a causal inference could be made under the backdoor framework. The dataset I used, could be sorted into 4 main areas of confoundedness – county-school fixed effect, individual's ability, family investment and classroom quality. County school fixed effect includes school and county id of the individual. Ability is measured through the individual's age. Family investment measures the potential support from one's family, this area includes variable such as Father and Mother's education level and family social economic status score. Classroom quality accounts for teacher's experience, teacher's monthly base salary and the classroom size. The effect-modifier is a randomized group assignment which took place where the data was collected, therefore must be included. Figure 5 is simplified for better understanding, the original DAG is attached in the appendix.

Figure 5: Simplified DAG



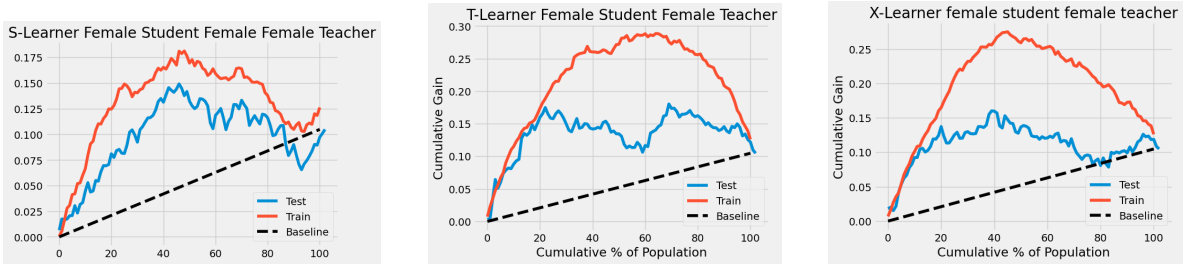
Result Under the DAG framework, the gender gap is smaller compared to the linear regression case. The causal impact of girls having female teachers almost eliminates the gap entirely.

	Estimate
Female	-0.159
Female student Female teacher	0.120
Male student Male teacher	-0.038

Table 5: DAG backdoor linear regression estimate

3.6 Meta Learners

The meta-learner estimates are implemented within the DAG framework. From the cumulative gain plot, as shown in figure 6, it is obvious that S-learner works the best given the data. Both T and X learners show a sign of overfitting. The cumulative gains not only shows the model performance but also hints on the treatment effect heterogeneity. S-Learner, for example, up to 50% of the population realize a huge gain from the treatment, in this case having a female math teacher as girls. The effect fades away, or realizing in a small quantity for the rest of the population. This heterogeneity would be further studied with the following machine learning models.



(a) S-Learner (b) T-Learner (c) X-Learner

Figure 6: Learner Cumulative Gains

There is no huge difference between different learner specification estimates. S-learner has relatively closer estimates to the DAG backdoor linear regression estimates. The gender gap is estimated to reduce the math outcomes by 0.158 standard deviations, while having a female teacher closing the gap by 0.123 standard deviations. The negative effect of boys having a same-gender is only -0.042, still much smaller compared to the OLS estimate of -0.052.

	S Learner	**T Learner**	**X-learner**
Female	-0.158	-0.159	-0.159
Female student Female teacher	0.123	0.124	0.124
Male student Male teacher	-0.042	-0.043	-0.043

Table 6: Comparison of S Learner, T Learner, and X-learner for different groups

3.7 Doubly Robust Methods

Identification Strategy

The double robust model estimates the ATE through the following equation.

$$\hat{\tau}_{\text{ATE}} = \frac{1}{N}\sum_i \left[T_i \frac{(Y_i - \mu_1(X_i))}{\hat{e}(X_i)} + \mu_1(X_i)\right] - \frac{1}{N}\sum_i \left[(1-T_i)\frac{(Y_i - \mu_0(X_i))}{1-\hat{e}(X_i)} + \mu_0(X_i)\right]$$

The biggest advantage of the double robust model is that it only needs to satisfy one of two assumptions – correctly specified propensity score or condition model μ_1, μ_0.

Result

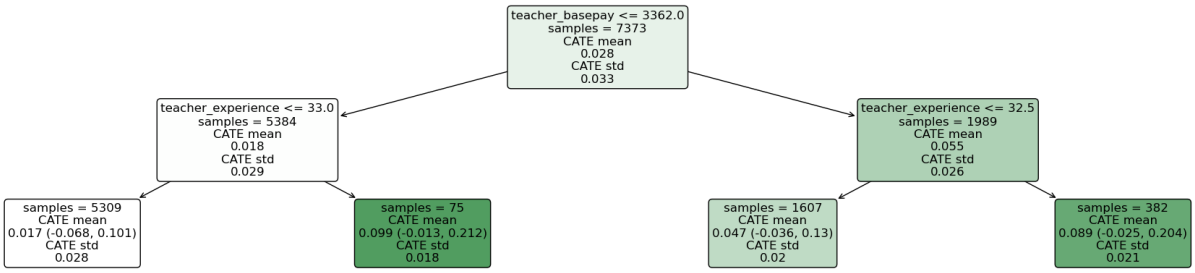
The estimates of DRL under the DAG are quite different, except the ATE of female- student-female-teacher, which remains at around 0.12. The estimated ATE almost doubled for both the gender gap and boys having male teachers.

	Estimate
Female	-0.237
Female student Female teacher	0.126
Male student Male teacher	-0.099

Table 7: DAG backdoor DRL estimate

The policy tree uses Conditional Average Treatment Effects (CATE) by splitting based on a feature's ability to explain heterogeneity. The first split in the policy tree occurs when the teacher's base salary is less than or equal to 3362, indicating it's the most significant. As well, subsequent splits are secondary, suggesting salary captures primary heterogeneity. All of this tells us that teacher's salary has the strongest predictive power for the treatment effect. In other words, the higher the teacher's salary, the higher the treatment effect. A teacher's salary often says something about the teacher's own abilities, and the resources the school has to offer. Salary is also something quantifiable and adjustable, so salary policies can optimize outcomes where the treatment effect is most beneficial.

Figure 7: Policy Tree: Identifying Key Predictors of Treatment Effect Heterogeneity



3.8 Causal Forest

Causal forest provides an insight into heterogeneous treatment effects as the tree are split based on the predictors' strength of in treatment heterogeneity, optimizing on causal estimands. It does feature selection automatically, without imposing a strict functional form.

Identification Strategy

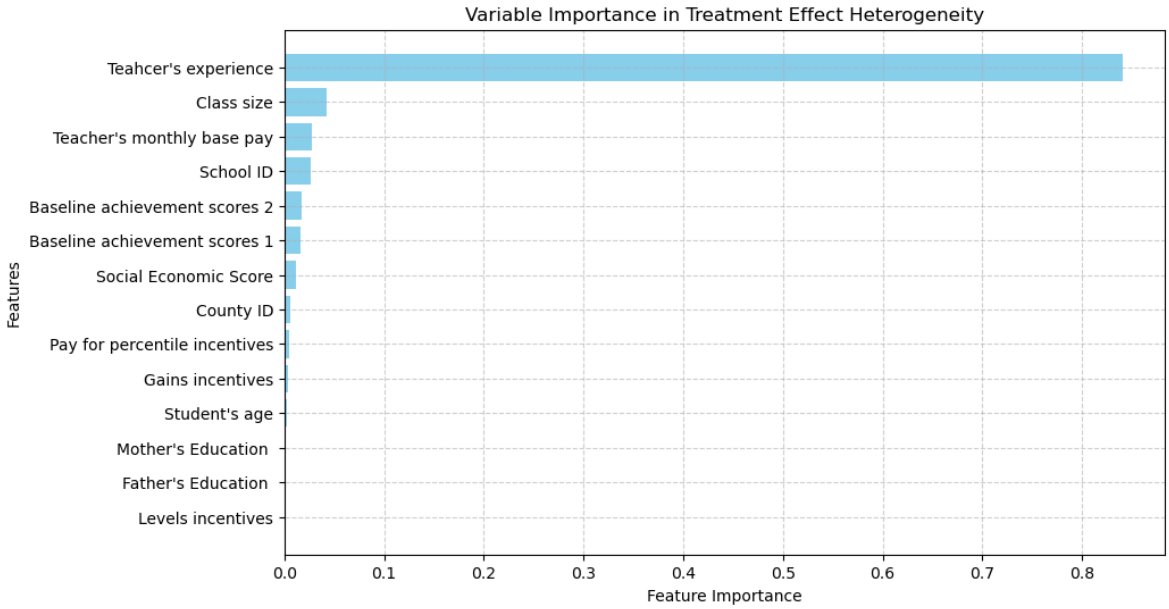
However, one of the key assumptions for this model is unconfoundedness, which fails here without considering the propensity score.

$$Y_1, Y_0 \perp T|X$$

Result

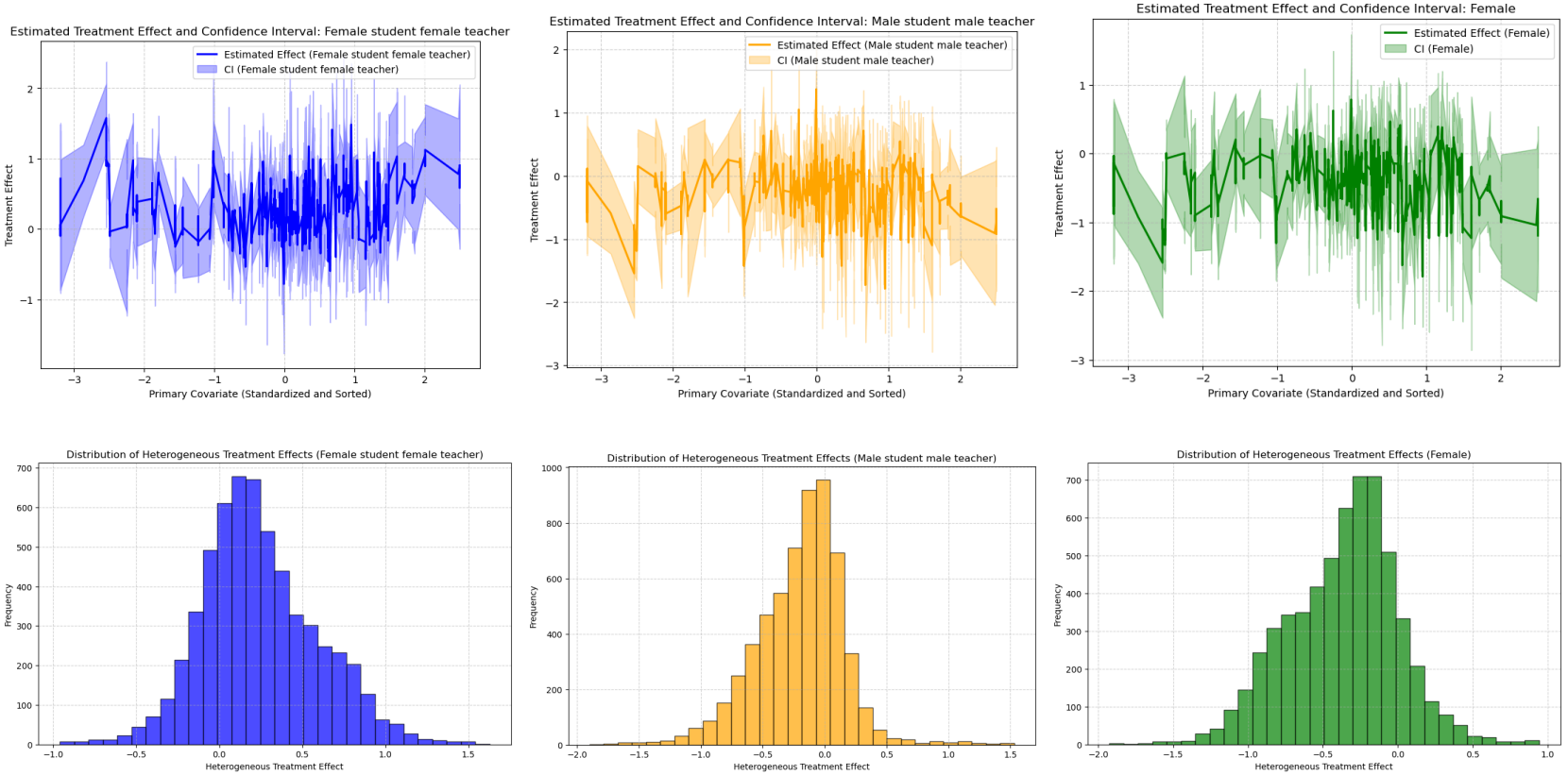
The causal forest variable importance says classroom-related factors like teacher's experience, class size and teacher's monthly base pay has the strongest effect in treatment effect heterogeneity amplification. This result aligns with the policy tree plot in the figure 7.

Figure 8: Causal Forest Variable Importance Index



I fitted two 3 models for each treatment. In figure 9, treatment effect and distribution for female-student-female-teacher , male-student-male-teacher, female are in blue, yellow and green respectively. The first row shows estimated treatment effects for individual predictors, with a 95% confidence interval band. The plots align with previous estimates, where the treatment effects for female and female-student-female-teacher are significant, while that of male-student-male-teacher is not. The bottom row histogram shows the spread of treatment effects. It suggests the amount of treatment effect varies significantly, which helps to further show the heterogeneous impacts. In particular, the treatment effect of having female teacher for female students are mostly positive, while that for male students having male teacher is approximately normal distributed at around zero. This is probably the reason why some point ATE estimates are negative but not significant.

Figure 9: Causal Forest Estimates and Distribution



3.9 Double Machine Learning

Given the fact that the treatment effect heterogeneity coming from teacher's characteristics, in this session I specialize in the subgroup treatment effect with the help of double machine learning model.

Identification Strategy

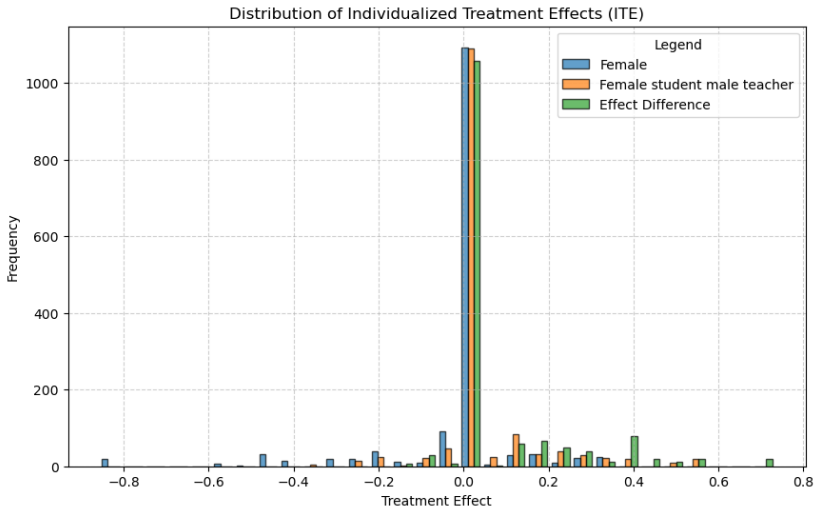
The double machine learning, as known as the debiased machine learning takes advantage of orthogonalization to eliminate potential biases. However, the key assumption of unconfoundedness is still required.

$$Y_1, Y_0 \perp T(X)|X$$

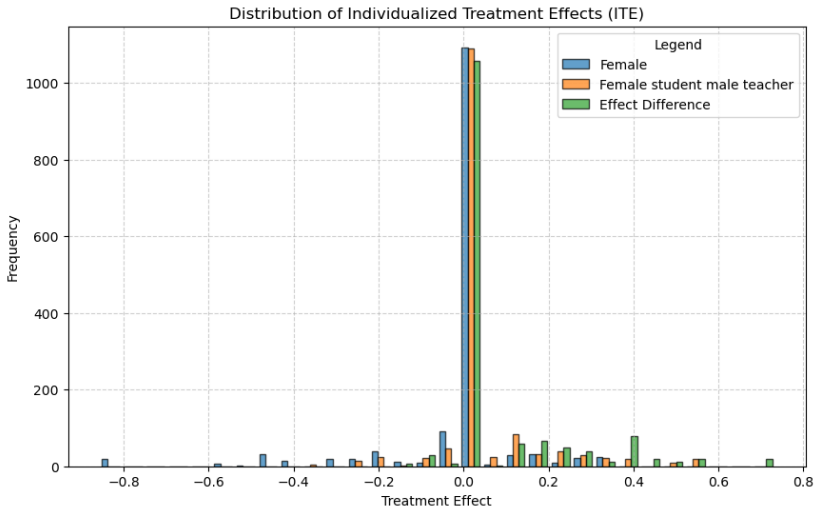
Since the model framework doesn't support propensity score and my treatment are not randomly assigned, ATEs and CATEs estimate would be biased in this case.

Result

Double machine learning model with python EconML library allows exploration of subgroup treatment effect. Specifically, I dig into the subgroup of teacher's base salary. In terms of Individual Treatment Effects, the linear double machine learning model has less variability, while the causal forest double machine learning captures a range of heterogeneity. This suggests that Causal Forest may be better when treatment effects are highly variable. However, in terms of the performance score, the linear model seems to slightly outperform Causal Forest in predictive accuracy on the testing data. This doesn't mean it is the better model for causal inference, though. Linear is best for interpretability and for a more structured approach. Nevertheless, when capturing heterogeneous treatment effects, Causal Forest is the better option. This result aligns with the fact that S-Learn fits the data the best. Overall, my data doesn't require a very complicate functional form assumption.



(a) LineaerDML Individual Treatment Effect



(b) Causalforest DML Individual Treatment Effect

Figure 10: Comparison between Linear and Causalforest DML Inidividual Treatment Effect

4 Discussion of Results

The results shed light on the influence of female math teachers on female students' performance. They offer meaningful insight into the research question and the broader theoretical context. Across all models, a pattern emerged: female students benefit significantly from having female math teachers. On average, they reduced the gender gap in math achievement by one-third, suggesting that gender-matched role models play a pivotal role in affecting educational outcomes. These results align with existing theories of stereotype internalization and the importance of representation in education. Not all results were straightforward, though. In some models, such as those checking gender-matched dynamics for male students, the findings were either statistically insignificant or indicated a small negative effect for boys with male teachers. This suggests that the benefits of gender-matching may not apply uniformly across genders. Models like Ridge and LASSO regression identified baseline math scores and socioeconomic factors as the most influential predictors of student outcomes, highlighting the importance of these broader contexts over singular factors like teacher-student gender dynamics. Advanced techniques like causal forest and doubly robust methods further illustrated the nuances in treatment effects. Teacher salary and experience emerged as key amplifiers of positive outcomes. These findings support the complex interaction of classroom characteristics, socioeconomic conditions, and gender dynamics in shaping educational achievement. Null or surprising results often pointed to the limitations of the models or the structure of the data. For example, the lack of randomized treatment assignments likely introduced bias into the double machine learning estimates, reducing their reliability in this context. These findings confirm earlier studies, such as those by Antecol et al. (2015) and Winters et al. (2013), which highlighted the positive impact of female teachers on female students. However, they challenge the assumption of uniform benefits across genders, painting a more complicated picture. This study also extends the literature by incorporating machine learning approaches, offering a deeper understanding of treatment heterogeneity and its implications. Practical and Policy Implications The policy implications of this study are both actionable and aligned with its findings. To close the gender gap in math and other STEM-related fields, prioritizing the recruitment and retention of female math teachers is critical. This is especially important in underprivileged areas, where systemic inequalities often worsen the challenges faced by female students. Increasing teacher salaries and reducing class sizes could further enhance these positive effects, as these factors were shown to significantly influence student outcomes. Targeted programs to support disadvantaged students, such as providing additional resources, parental engagement, and teacher training, are also

recommended. These interventions could address systemic barriers and improve outcomes for all students, particularly those from low socioeconomic backgrounds. Lastly, while this study shows the importance of gender dynamics in education, it also shows the need for nuanced approaches that consider the broader socioeconomic and institutional contexts affecting student achievement. This research underscores the complexity of the gender gap in education and the solutions required to address it. By linking these findings to broader economic discussions, such as the value of human capital and the economic potential of a more inclusive workforce, this study reinforces the importance of equitable educational policies in fostering long-term societal benefits.

5 Conclusion and Future Directions

This study demonstrates the significant role female math teachers play in improving female students' performance, reducing the gender gap in math achievement by nearly one-third. Using a variety of econometric and machine learning models, the analysis reveals that gender-matching between students and teachers is a critical factor in closing the gender gap. However, the effect is not uniformly beneficial across genders. These findings extend the existing literature by highlighting the nuanced interaction between gender dynamics, socioeconomic factors, and classroom characteristics. The results emphasize the importance of addressing systemic inequalities through policies that promote the recruitment and retention of female teachers, particularly in underserved areas, alongside efforts to improve teacher's salaries and classroom conditions. While this research confirms the importance of representation in education, it also underscores the need for more rigorous, randomized studies to further study the complex factors influencing student outcomes. By closing the gender gap in math, this work contributes to broader goals of equity in STEM fields and long-term economic inclusivity.

6 References

Antecol, H., O. Eren, and S. Ozbeklik (2015) "The effect of teacher gender on student achievement in primary school," *Journal of Labor Economics* 33(1), 63–89, ISSN 0734306X, 15375307

Chan, P. C. W., T. Handler, and M. Frenette (2021) "Gender differences in stem enrolment and graduation: What are the roles of academic performance and preparation?," *Statistics Canada Economic and Social Reports* ISSN 2563-8955

Fryer, R. G., and S. D. Levitt (2010) "An empirical analysis of the gender gap in mathematics," *American Economic Journal: Applied Economics* 2(2), 210–240, ISSN 19457782, 19457790

Keller, L., F. Preckel, J. Eccles, and M. Brunner (2021) "Top-performing math students in 82 countries: An integrative data analysis of gender differences in achievement, achievement profiles, and achievement motivation," *Journal of Educational Psychology* 114

Lindner, J., E. Makarova, D. Bernhard, and D. Brovelli (2022) "Toward gender equality in education—teachers' beliefs about gender and math," *Education Sciences* 12(6), ISSN 2227-7102

Loyalka, P., S. Sylvia, C. Liu, J. Chu, and Y. Shi (2019) "Pay by design: Teacher performance pay design and the distribution of student achievement," *Journal of Labor Economics* 37(3), 621–662

Lyons, E., A. Mesghina, and L. E. Richland (2022) "Complicated gender gaps in mathematics achievement: Elevated stakes during performance as one explanation," *Mind, Brain, and Education* 16(1), 36–47

Steele, J. R., and N. Ambady (2006) ""math is hard!" the effect of gender priming on women's attitudes," *Journal of Experimental Social Psychology* 42(4), 428–436, ISSN 0022-1031

Tsai, S.-L., M. L. Smith, and R. M. Hauser (2018) "Gender gaps in student academic achievement and inequality," *Research in the Sociology of Education*

Winters, M. A., R. C. Haight, T. T. Swaim, and K. A. Pickering (2013) "The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data," *Economics of Education Review* 34, 69–75, ISSN 0272-7757

7 Appendix

	Dependent variable: y	
	(1)	(2)
Female	-0.142***	-0.206***
	(0.018)	(0.027)
Female student female teacher		0.078***
		(0.027)
Male student male teacher		-0.052**
		(0.027)
Observations	7373	7373
R^2	0.476	0.477
Adjusted R^2	0.474	0.475
Residual Std. Error	0.756 (df = 7343)	0.756 (df = 7341)
F Statistic	230.266*** (df = 29; 7343)	216.034*** (df = 31; 7341)
Note: unclustered regression result		*p<0.1; **p<0.05; ***p<0.01

Figure 11: Original LASSO for Gender-Mathing Model

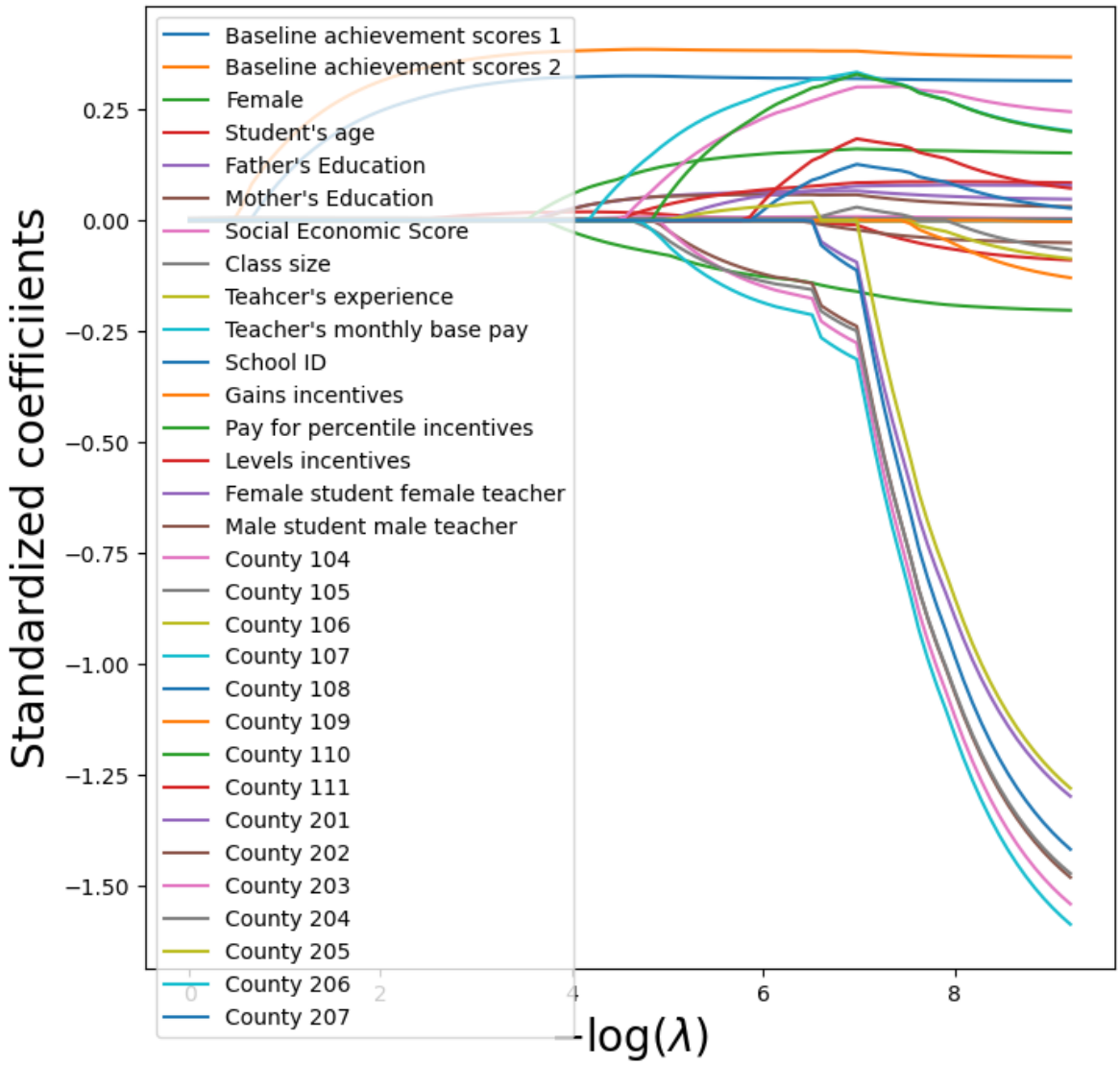


Figure 12: Original LASSO for Base Model

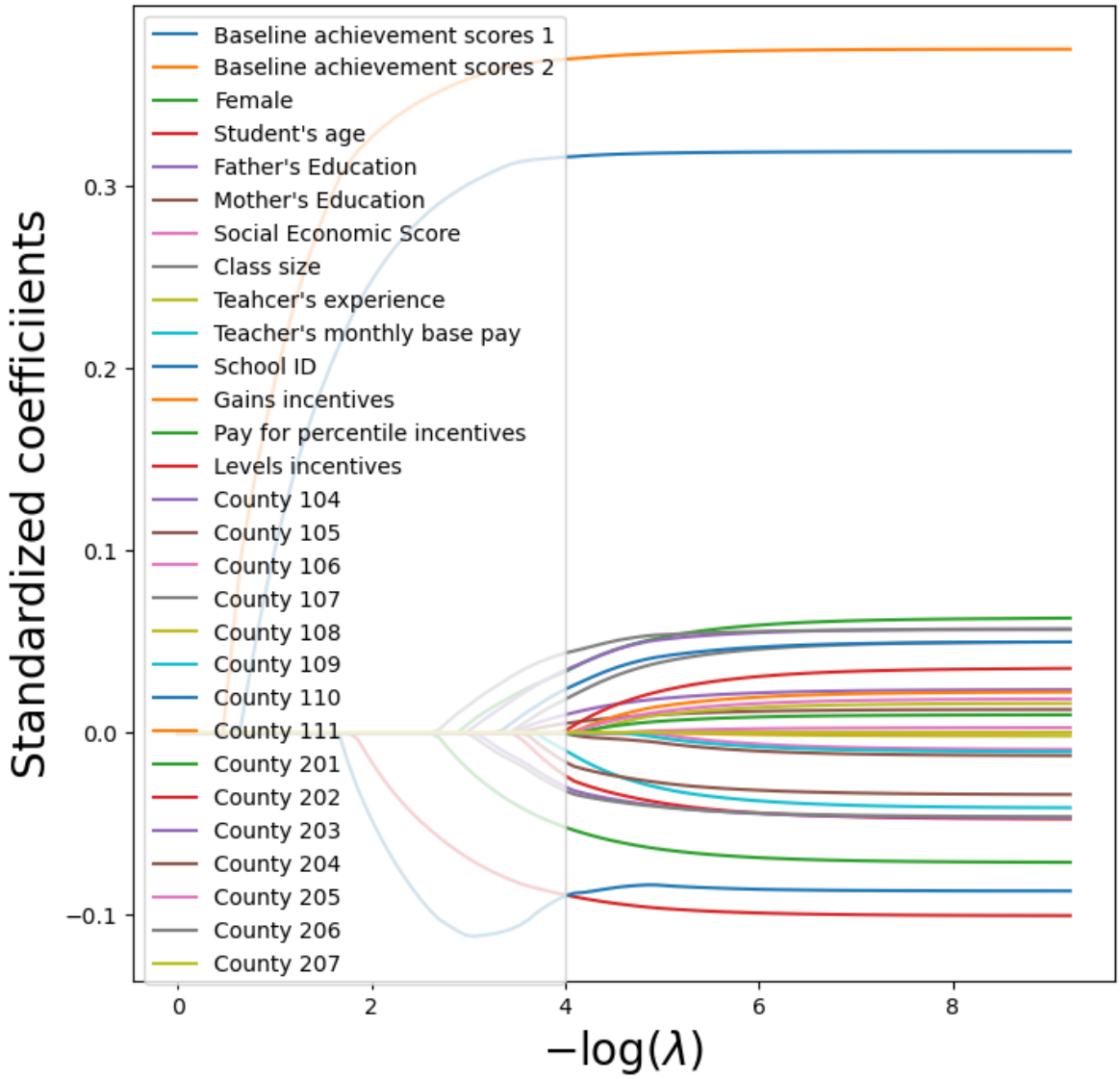


Figure 13: Original Ridge for Gender-Matching Model

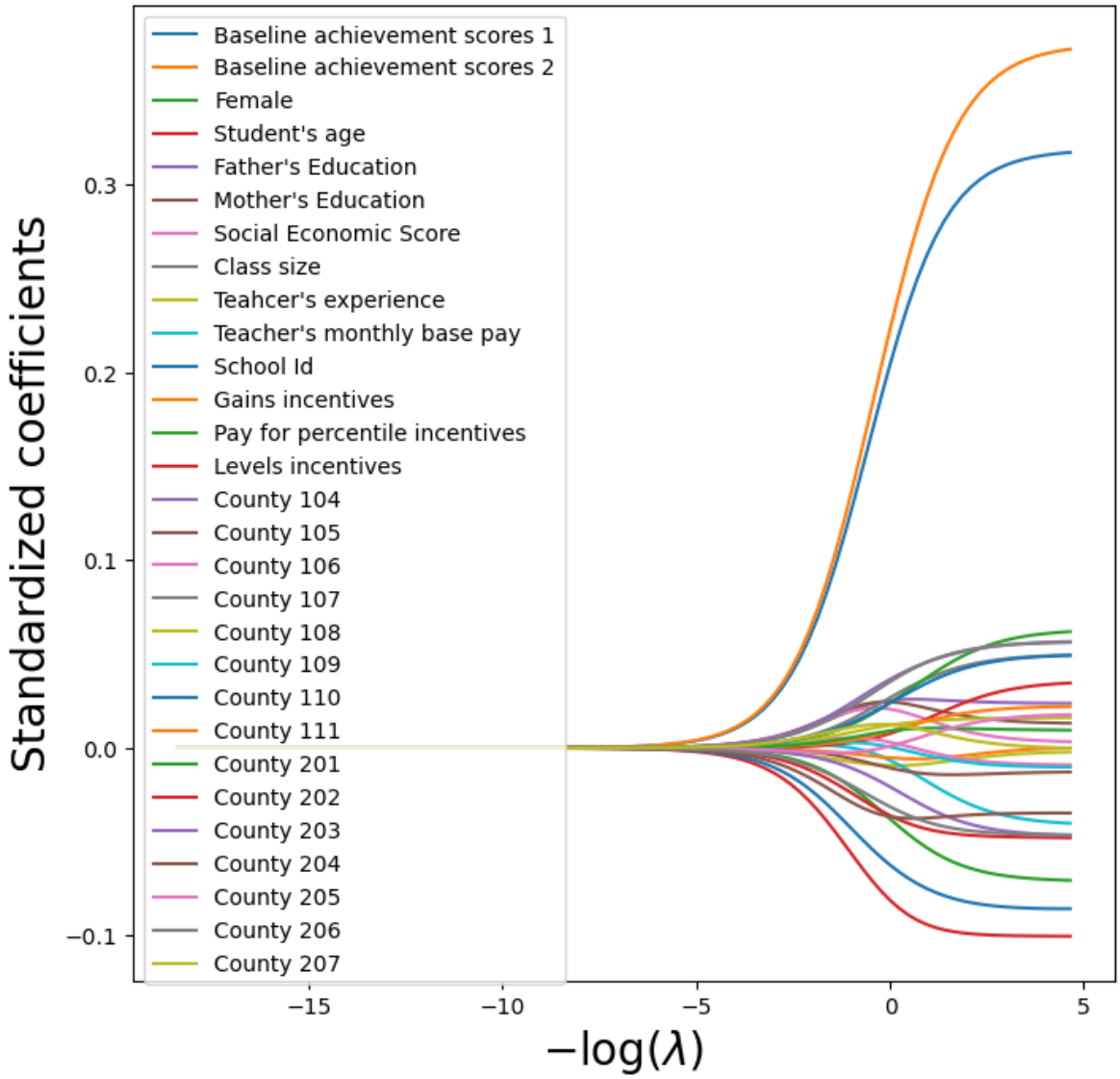
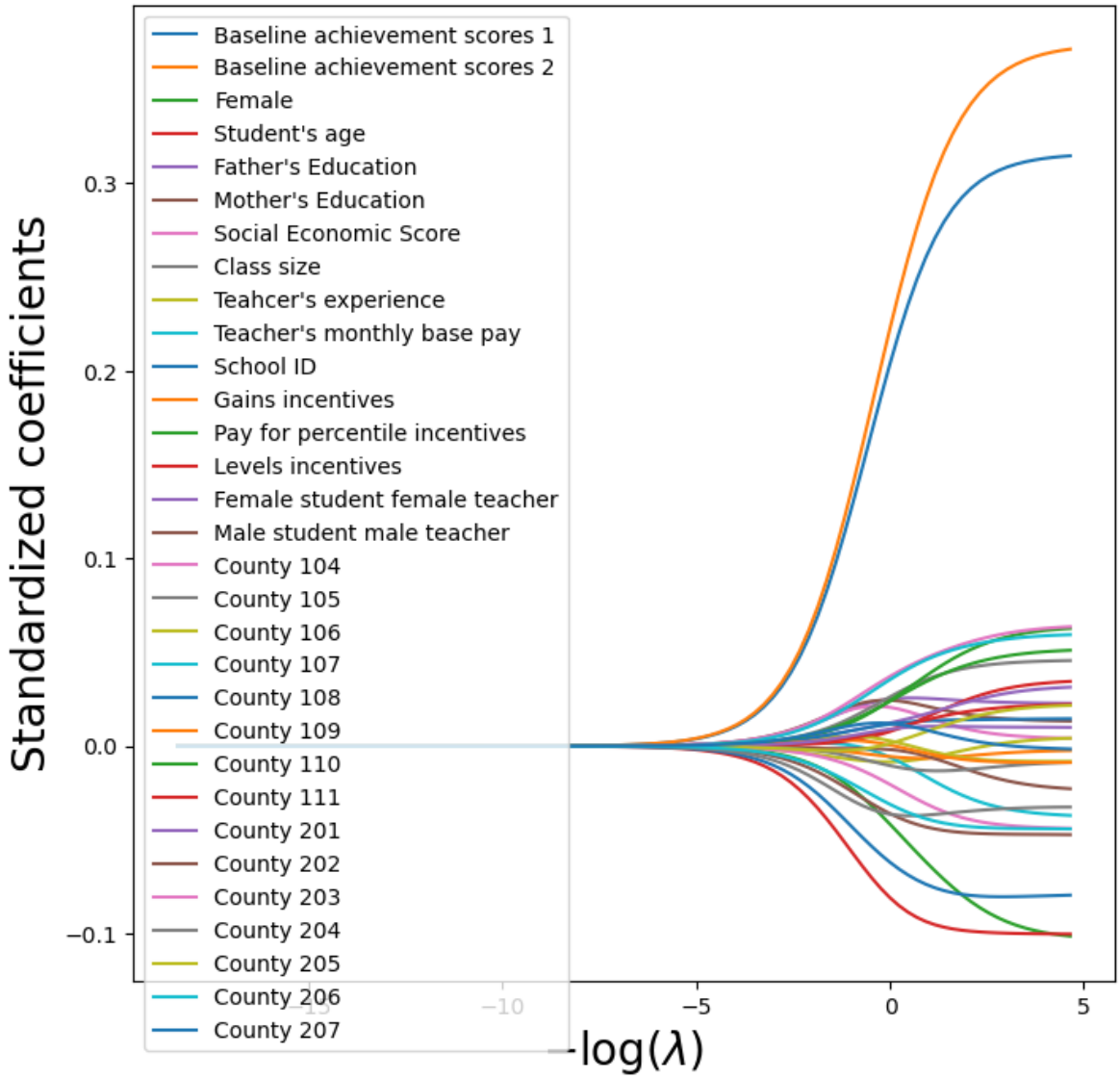


Figure 14: Original Ridge for Gender-Matching Model



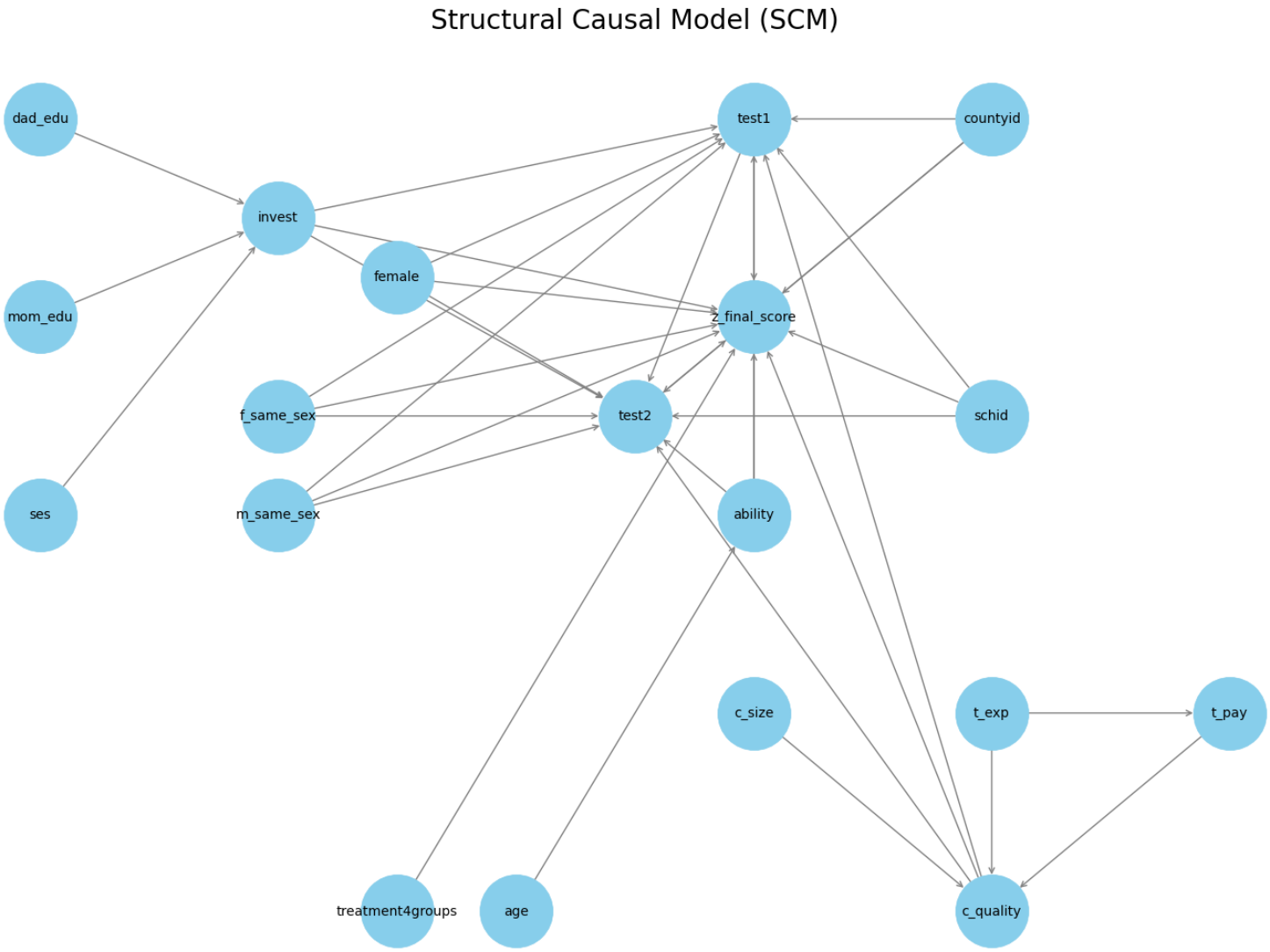


Figure 15: Regression Tree