Estimating Causal Returns to Compulsory Schooling: A Double Machine Learning Approach

Liam Henson

December 2025

1 Executive Summary

This paper examines the effect compulsory education has on the incomes of students in the United Kingdom. To understand the effect compulsory education has on earnings our analysis focuses on two compulsory schooling law reforms implemented in the U.K. in the mid 20th century. Compulsory schooling laws are educational policies which legally require students to remain in school up to a specific age, having the effect of keeping students who would otherwise choose to leave school in school longer. The two reforms analyzed increased the minimum age at which a student was legally allowed to leave school from 14 to 15 years old in 1947, and from 15 to 16 years old in 1972. Through the combination of U.K. General Household Survey data with Continuous Northern Ireland Household Survey data, we used the difference in the number of legally required years of schooling to estimate the causal effect additional compulsory education has on future real earnings. To understand the causal impact compulsory education has on earnings, this paper utilizes double machine learning. Double machine learning is a modern econometric approach which applies machine learning methods to understand causal relationships, while flexibly controlling for complex structures in the data. To examine regional differences in the effects of compulsory education we separately analyze how each reform affected students in Great Britain and Northern Ireland. We find that the increased compulsory education induced by these policy changes had a pronounced effect on earnings and differed greatly across each reform and region. We find that for the reform which raised the minimum school leaving age from 14 to 15 years old in Great Britain, remaining in school for an additional year corresponded to a 30-35% increase in real earnings. Similarly, for students in Northern Ireland the additional educational attainment induced by the reform results in a 20-30% increase in real earnings. In contrast, the 1972 reform which raised the minimum school leaving age from 15 to 16 years old produces smaller estimates of a 10-17% increase in real earnings in Great Britain. However, are unable to draw conclusions regarding the 1972 reform in Northern Ireland as our estimates are not statistically significant. Our findings highlight the benefits of compulsory education and the importance of policies designed at keeping students who would otherwise have chosen to leave school in school longer.

2 Introduction

Education is a significant factor in shaping one's human capital, and is a key determinant of labour market outcomes, social mobility, and overall economic well being. It is important to understand the causal impact education has on these outcomes both in the context of economic research and educational policy analysis. Despite a large body of economic research, the causal link between one's level of educational attainment and their earnings remains a highly debated topic, as the true causal impact education has on earnings is obscured by endogeneity concerns. This paper further examines returns to education using exogenous variation in the amount of education attained through the introduction of two significant compulsory schooling law reforms in the United Kingdom.

The first compulsory schooling law reform raised the minimum age an individual was legally allowed to leave school from 14 to 15 years old. This reform was implemented in Great Britian in 1947 and in Northern Ireland in 1957. Subsequently, the second reform implemented in Great Britain and Northern Ireland in 1972 further raised the minimum school leaving age from 15 to 16 years old. These policy changes had the effect of forcing some students, those who would have otherwise chosen to leave school, to remain in school for an additional year, so called compliers. The same policy change however, would have had no effect on students who would have otherwise chosen to remain in school. By exploiting the variation in the number of years an individual was legally required to remain in school, the focus of this paper is to estimate the causal effect additional compulsory education has on the future real earnings of compliers.

To accomplish this we apply Double Machine Learning (DML) in an interactive instrument variable (IV) model to estimate the Local Average Treatment Effect (LATE) for the causal effect additional compulsory educational attainment has on real earnings. Unlike common instrument variable approaches traditionally used in the returns to education literature which imposes restrictive linear modeling assumptions, double machine learning methods are highly flexible and can be used to identify causality in the context of treatment effects while flexibly allowing for arbitrary non-linear relationships in the data. While a large portion of the wider returns to education literature generally has the goal of estimating the marginal returns to education for the population at large, that is the return to an additional year of education on average. We instead specifically set out to examine the causal return to the additional compulsory education induced by each respective policy reform for the subpopulation of individuals whose educational attainment was directly impacted by changes to the minimum school leaving age. By doing so, we seek to analyze how compulsory education affects the economic outcomes of those most likely to leave school early, providing evidence as to the long run value of compulsory education rather than simply the value of education for the general population at large.

Although most compulsory schooling law policies were implemented decades ago examining how increased education benefits marginal students highlights the importance of policy actions specifically targeting marginal students and offers useful social and economic insights as to the value of keeping such students in school for a longer time period. Moreover by applying double machine learning to such a foundational economic question, this paper contributes to the literature

by allowing for a more modern and highly flexible estimation process for the causal effects education has on earnings.

3 Literature Review

A central focus of the empirical educational economics literature examining causal returns to education is the issue of endogeneity. This problem arises as both individuals' income and educational choices are jointly determined by a variety of factors, such as their inherent ability, family characteristics, and one's perceived marginal costs and benefits of education. As a result this issue can complicate assessing the causal link between education and income and makes it difficult to isolate the effect education has on earnings.

To address this problem, Instrument Variable (IV) estimation is a common practice in seeking to model the causal returns to education as traditional ordinary least squares (OLS) is unable to overcome endogeneity issues. Common instruments used in the literature include include quarter of birth Angrist and Krueger (1991), changes in compulsory schooling laws Oreopoulos (2006a) and proximity to colleges Card et al. (1995). Different instruments produce varying estimates of educational returns and there is no clear consensus as to which instrument is most appropriate Dickson and Harmon (2011). Furthermore, a common concern with IV estimation is that estimates often exceed OLS estimates when in theory OLS estimates should be biased upward due to omitted variable bias of neglecting ability Card (1999). In general, IV estimates are only able to capture local effects, that is IV estimates only capture the average returns to education for the subpopulation that alters their behavior in response to the chosen instrument, rather than the overall population Oreopoulos (2006a). Analyzing returns to education for a subset of the population is however often of interest, particularly in addressing returns to education in the context of policy changes Card (2001).

Compulsory schooling laws are an example of a prominent policy based instrument frequently used in the returns to education literature. Since compulsory schooling laws only truly impact individuals who would have otherwise dropped out of school rather than the population as a whole, using compulsory schooling laws as an instrument offers an interesting example of not only a relevant IV but also in assessing the local returns to education for possibly a more vulnerable and policy relevant segment of the population. Oreopoulos (2006a) compared average returns to education in U.K. and Ireland with Canada and the United States using compulsory schooling laws reforms implemented in each country throughout the 20th century as instrument variables. After controlling for year of birth and regional fixed effects, returns to education were found to be between 10 to 14% for each country analyzed, despite the fact that a significantly higher proportion of individuals were affected by compulsory schooling laws in the U.K. and Ireland as compared to Canada and the United States. As a result, Oreopoulos (2006a) argued that returns to education are large regardless of the proportion of the population that they affect, and that the local returns to education in the U.K. and Ireland would have approximated, in this case, the average returns

to education since such a high proportion of individuals were affected by compulsory schooling reforms.

In contrast, other researchers have used compulsory schooling laws as instruments and have reached different conclusions. Stephens and Yang (2014) explored returns to education using compulsory schooling laws as an instrument in the United States using census data from 1960 to 1980. Stephens and Yang (2014) argued that only including year of birth and regional fixed effects is insufficient to understand returns to education, as they fail to capture how factors such as school quality change disproportionally in some states over others during the same time period. After accounting for the interaction between state and year of birth fixed effects, Stephens and Yang (2014) found that returns to education estimates no longer become statistically significant.

While traditional IV approaches remain the standard model used throughout the returns to education literature, recent econometric research has led to the development of modern approaches which incorporate IV estimation procedures with modern machine learning models into so called double machine learning. Chernozhukov et al. (2018) outlined the theoretical framework of double machine learning in the context of partial linear models and demonstrated how the double machine learning estimation process can be used to flexibly estimate average and local average treatment effects. The advent of double machine learning allows for robust causal inference in the presence of high dimensional data with potentially highly complex non-linear relationships without imposing any parametric assumptions on the underlying data generating process.

Applying the double machine learning framework developed by Chernozhukov et al. (2018), this paper contributes to the literature by estimating the causal effect increased compulsory education has on earnings, without imposing restrictive modeling assumptions inherent to the standard IV modeling procedure. In doing so, we aim to provide more robust and accurate parameter estimates for the causal effect compulsory education has on economic outcomes.

4 Data

The data used for our analysis is a subset of the replication data from Oreopoulos (2006b). The data set is repeated cross-sectional micro-data consisting of a combination of 15 U.K. General Household Surveys that span from 1983 to 1998 along with 14 Northern Ireland Continuous Household Surveys that span from 1985 to 1998. The data set contains demographic information, labour market outcomes, along with schooling history. Schooling history is given by the age an individual left full time education. Northern Ireland. The raw unfiltered data set contains 416,324 observations. The data was filtered and relevant variables were constructed in the same manner as Oreopoulos (2006a). The data was restricted to individuals less than 65 years of age at the time of the given survey to avoid including retirees in our sample. We also dropped all observations that had missing schooling or earnings information, and removed observations of individuals who left school before age 10 or after age 18. This was done to remove both outliers along with individuals who continued with post secondary education, which were not relevant for our analysis. To identify whether an individual

was exposed to a given compulsory schooling law reform, we define the indicator variables:

$$Law_{14} = 1\{(YearAt14 \geq 1947 \text{ and } Nireland = 0) \text{ or } (Yearat14 \geq 1957 \text{ and } Nireland = 1)\}$$

$$Law_{15} = 1\{Yearat15 \geq 1972\}$$

Where Law_{14} defines if an individual was exposed to the reform which raised the minimum school leaving age from 14 to 15 years old and $Law15$ defines whether an individual was exposed to the reform which raised the minimum school leaving age from 15 to 16 years old. To construct indicator variables for reform exposure we define $YearAt14$ and $YearAt15$ which denote the year at which a survey respondent was 14 or 15 years old and $Nireland$ is an indicator variable defining whether the individual was a respondent from Northern Ireland or not (Great Britain). We also define indicator variables to capture whether an individual remained in school beyond the minimum school leaving age for each of the two reforms:

$$School_{14} = 1\{\text{age left school} > 14\}, School_{15} = 1\{\text{age left school} > 15\}$$

For individuals whose behavior was altered by a given reform, that is students who would have otherwise left school had the reform not been in place, $School_{14}$ and $School_{15}$ identify policy induced increases in educational attainment.

In an effort to examine potential heterogeneity in returns in each region/reform and to avoid unobserved time and regional trends influencing our results, the data was then segmented into four mutually exclusive samples surrounding each educational reform in each region. Each region/reform sample was constructed by first dividing the data set by region (Great Britain and Northern Ireland) and then considering a four-year window surrounding each reform in each region. For the 1947 Great Britain reform, we restricted the data to consider only observations of individuals who were aged 14 from 1943 to 1951. For the 1957 Northern Ireland reform, we consider observations of individuals aged 14 between 1953 and 1961. For the 1972 reform, we consider observations of individuals who were aged 15 between 1968 and 1976 in either Great Britain or Northern Ireland.

Below we present descriptive statistics for relevant variables in each reform/region sample. Tables 1,2 depict the Great Britain and Northern Ireland samples which encompass and reflect the policy reform which raised the minimum school leaving age from 14 to 15 years old. Before the reform came into effect, roughly only a third of students remained in school past age 14 in Great Britain and a half of students remained in school past age 14 in Northern Ireland. These values reflect significantly high dropout rates in both regions. Following the implementation of the reform, the proportion of students who stayed in school past age 14 increased drastically to roughly 90% in each region. In addition, the average age at which students left full time education increased by roughly 0.69 years in Great Britain and 0.53 years in Northern Ireland. These changes reflect that the policy reform had a significant impact on increasing the number of students who stayed in school beyond age 14, highlighting the reform's effectiveness at increasing educational attainment among students who otherwise would have left school at age 14.

Table 1: Descriptive Statistics: Great Britain 1947 Sample 14 to 15 reform (N=12421)

Variable	Pre Reform Non Exposed Cohorts	Post Reform Exposed Cohorts
Log Real Annual Earnings	8.66 (1.02)	8.79 (0.97)
Age Left Education	14.64 (1.11)	15.33 (0.9)
Age at Time of Survey	57.37 (3.05)	54.54 (4.33)
Proportion Stayed Past 14	0.34 (0.47)	0.9 (0.3)

Note: Values are in presented in mean (sd). Real annual earnings reflect 1998 U.K. pounds

Table 2: Descriptive Statistics: Northern Ireland Sample 1957 14 to 15 reform (N=3207)

Variable	Pre Reform Non Exposed Cohorts	Post Reform Exposed Cohorts
Log Real Annual Earnings	8.7 (0.89)	8.76 (0.9)
Age Left Education	15.09 (1.31)	15.62 (1.09)
Age at Time of Survey	50.11 (4.16)	45.8 (4.47)
Proportion Stayed Past 14	0.54 (0.5)	0.89 (0.31)

Note: Values are in presented in mean (sd). Real annual earnings reflect 1998 U.K. pounds

Relative to the earlier reform samples, the 1972 samples exhibited much lower pre-reform dropout rates. Tables 3,4 similarly report descriptive statistics for the 1972 samples in Great Britain and Northern Ireland. Here we can see the majority of students already remained in school past age 15 in each region before the reform came into effect; 62% of students remained in school beyond age 15 in Great Britain and 71% stayed past age 15 in Northern Ireland. The subsequent policy change which raised the minimum school leaving age from 15 to 16 years old again resulted in a large jump in the proportion of students who stayed in school beyond 15, which increased to over 90% in both regions. The average age at which students left full time education increased by 0.39 years in Great Britain and 0.34 years in Northern Ireland. The smaller gains in educational attainment observed in the 1972 samples reflect that by 1972 a much smaller proportion of students were initially planning on leaving school. This implies that the 1972 reform truly affected a much narrower range of marginal students.

Table 3: Descriptive Statistics: Great Britain 1972 Sample 15 to 16 reform (N=30724)

Variable	Pre Reform Non Exposed Cohorts	Post Reform Exposed Cohorts
Log Real Annual Earnings	9.05 (1.11)	9.06 (1.1)
Age Left Education	15.94 (1.02)	16.33 (0.84)
Age at Time of Survey	39.78 (7.27)	35.45 (7.23)
Proportion Stayed Past 15	0.61 (0.49)	0.92 (0.27)

Note: Values are in presented in mean (sd). Real annual earnings reflect 1998 U.K. pounds

Table 4: Descriptive Statistics: Northern Ireland 1972 Sample 15 to 16 reform (N=4029)

Variable	Pre Reform Non Exposed Cohorts	Post Reform Exposed Cohorts
Log Earnings	8.9 (0.79)	8.88 (0.73)
Age Left Education	16.16 (1.05)	16.5 (0.84)
Age at Time of Survey	36.56 (4.7)	32.05 (4.66)
Proportion Stayed Past 15	0.7 (0.46)	0.95 (0.23)

Note: Values are in presented in mean (sd). Real annual earnings reflect 1998 U.K. pounds

Notably, in each sample we do not observe a substantial change in average earnings before and after each reform was implemented. However, examining changes in overall population average earnings obscures any potential heterogeneity in earnings that could result from individuals' decisions to remain in school. That is, examining average population earnings before and after each reform was implemented are subject to the same endogeneity concerns as examining returns to education for the whole population. Average population earnings includes information regarding both individuals' who would have chosen to remain in school had the respective reform not been in place, along with individuals' who had their schooling choices altered as a result of the reform. As a result, the observed small change in average earnings depicted in each samples' summary statistics does not reflect the true effect each educational reform had on earnings of marginal individuals.

Figures 1,2 plot the proportion of students which stayed in school beyond age 14 and age 15 over time in each respective region. Figure 1, depicts a substantial discontinuous jump in the proportion of students who remained in school beyond age 14 due to the 1947 reform in Great Britain. This highlights just how high dropout rates were in Great Britain surrounding the year 1947 and shows that the majority of the population consisted of marginal individuals whose schooling choices were directly influenced by the policy change. In contrast, the 1972 Great Britain policy change also produced a large jump in the proportion of students who stayed in school beyond age 15. However, unlike the 1947 reform, we observe that the number of students who stayed in school past age 15 was increasing over time, which suggests that the reform raised educational attainment for a smaller segment of the overall population.

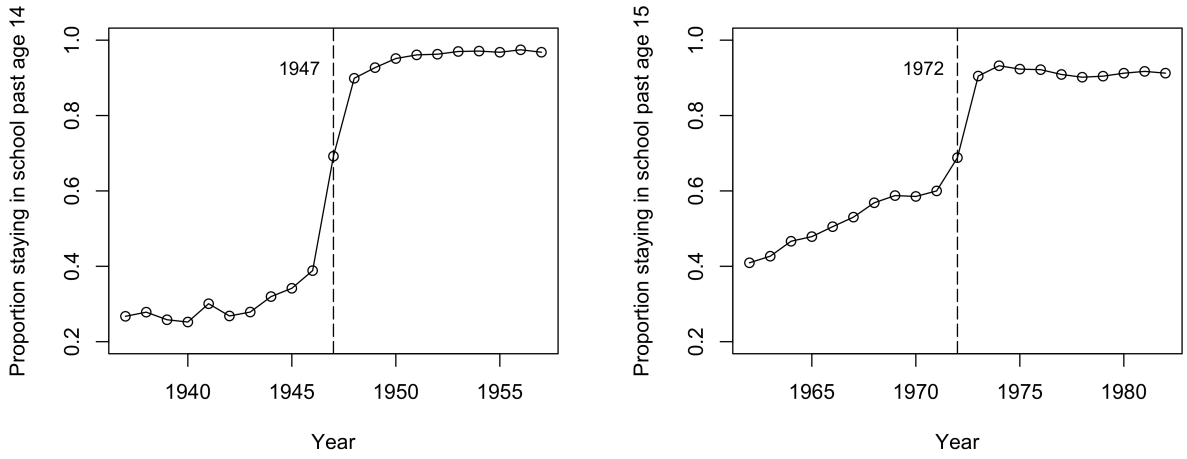

Figure 1: 1947 and 1972 reform compliance (Great Britain). Right: Proportion of students who stayed in school past age 14. Left: Proportion of students who stayed in school past age 15

For both the 1957 and 1972 Northern Ireland samples, Figure 2 shows the proportion of students who remained in school past the given legal cutoff was also rising over time before each reform was implemented. As a result of the 1957 reform, we again see a sizable jump in the proportion of students who remained in school beyond age 14. However, the increase was less pronounced relative to the equivalent policy change in Great Britain. As for the 1972 sample, we do observe a clear jump in the number of students who stayed in school beyond age 15 as a result of the policy change. This suggests that very few students' behavior was ultimately impacted by the 1972 Northern Ireland policy change.

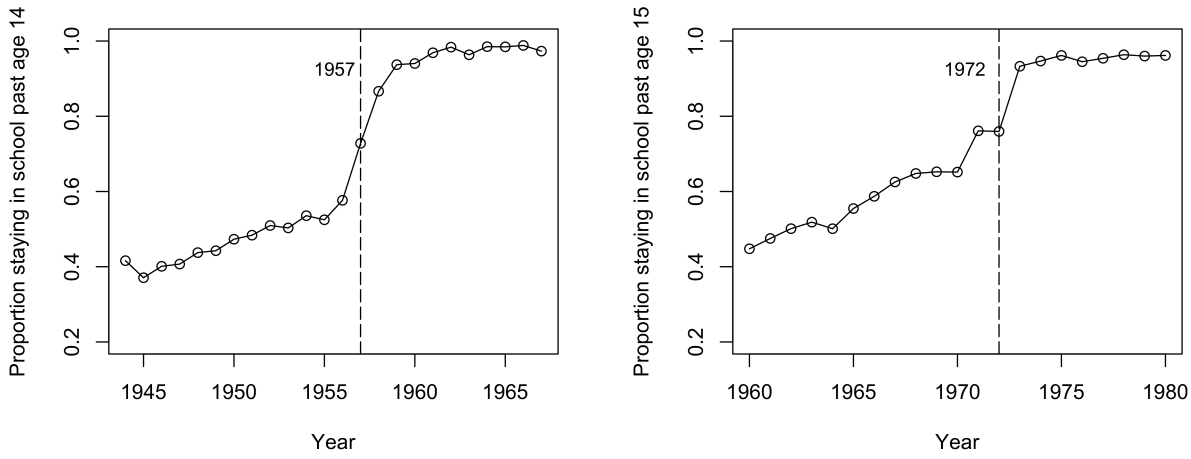

Figure 2: 1957 and 1972 reform compliance (Northern Ireland). Right: Proportion of students who stayed in school past age 14. Left: Proportion of students who stayed in school past age 15.

5 Model

In this section we outline the methodological framework used in our analysis. The focus of our analysis is to estimate the causal treatment effect additional compulsory education has on real earnings, and to analyze heterogeneity in returns across each region/reform sample. We begin by defining the causal parameter of interest, then outline the specification of the interactive IV model used in our analysis along with how it identifies the causal parameter of interest. Finally, we outline the empirical estimation process taken, which forms the basis of our results.

The casual parameter of interest we seek to estimate is the local average treatment effect (LATE) of the respective policy change. In the context of our problem, LATE is the appropriate parameter of interest as it specifically identifies the returns to additional compulsory education induced by the given policy reform on compliers, individuals whose schooling increased strictly because the law compelled them to stay in school past a given legal threshold. We estimate LATE using an interactive IV model, a double machine learning model which specifically targets LATE estimation.

To analyze heterogeneity in returns, we fit separate interactive IV models to estimate LATE in each region/reform sample which were constructed in the previous section. In each region/reform sample the variables used in our analysis are defined in the same way. The outcome variable Y denotes log real annual wages. The treatment variables corresponding to each educational reform are defined to be indicators for staying in school beyond the given minimum school leaving age $School_{14}$, $School_{15}$. Additionally, our instrument variables define exposure to each policy change, Law_{14} for the 1947/1957 reform and Law_{15} for the 1972 reform. The covariates included in each model are controls for the year a given student was 14 or 15 based on the specific reform analyzed, the year in which the survey was conducted, age at the time of survey, and respondent sex.

Unlike traditional linear IV modeling approaches, an interactive IV model does not require an explicit functional form in any stage of the estimation process and in the context of our problem can be defined as:

$$Y_i = g_0(School_i, X_i) + U \quad E(U_i|X_i, Law_i) = 0$$

$$School_i = r_0(Law_i, X_i) + V \quad E(V_i|X_i, Law_i) = 0$$

$$Law_i = m_0(X_i) + \epsilon \quad E(U_i|\epsilon) = 0$$

Here g_0, m_0, r_0 are arbitrary unknown and potentially non-linear functions estimated through machine learning algorithms referred to as nuisance functions, and X denotes a vector of our covariates. Since both the treatment and instrument variables are binary, $r_0(Law_i, X_i)$ describes the conditional probability of remaining in school beyond the minimum school leaving age given law exposure and the covariates. $m_0(X_i)$ describes the conditional probability of law exposure given the covariates. The double machine learning algorithm used to fit the model works by first estimating g_0, m_0, r_0 through a given machine learning model. The algorithm then partials out the effect of

the covariates on the given treatment and instrument variable before then using the remaining variation in the residuals as an instrument to estimate the casual parameter of interest. By doing so, the process of double machine learning removes the regularization bias that traditional naive machine learning approaches impart on the causal parameter of interest. Defining our model in this way, LATE, the parameter of interest is then given by:

$$\hat{\theta} = \frac{E[g_0(1, X) - g_0(0, X)]}{E[r_0(1, X) - r_0(0, X)]}$$

This approach allows us to apply modern flexible machine learning models to estimate the causal treatment effect that additional compulsory education has on real earning while flexibly allowing for potentially nonlinear relationships between the outcome variable, covariates, and treatment at each stage of the estimation process.

We now outline the foundation of our estimation procedure. For a baseline specification, we begin our analysis by reporting weak instrument tests along with two stage least square (2SLS) estimates for the causal parameter of interest for each region/reform sample. We then provide estimates for the interactive IV models for each region/reform sample. In an effort to achieve robust parameter estimates, we fit each interactive IV model in each region/reform sample using three separate machine learning models in the DML estimation process. We considerer two highly flexible models, random forests and extreme gradient boosting (XGBoost), along with a more restrictive regularized linear model elastic net. For each interactive IV model estimated, all nuisance functions hyperparameter are tuned using 5-fold cross-validation. In addition, to prevent overfitting bias in the DML estimation process we report estimates using both 5 and 10-fold cross-fitting.

6 Results

Depicted in Table 5, we report the standard 2SLS estimates for the causal parameter of interest in each region/reform sample. In each sample, we can easily reject the null hypothesis for a weak instrument as each model results in a highly significant F-statistics. The F-statistics are much smaller in both Northern Ireland samples relative to the Great Britain samples, reflecting both a weaker first stage relationship and smaller sample sizes in both Northern Ireland samples. However the F-Statistics in the two Northern Ireland samples are still well above the threshold to reject the weak instrument hypothesis.

Table 5: 2SLS estimates for the effect of staying in school past the minimum school leaving age on log real earnings

Sample	Weak Instrument Test F-Stat	Coefficient	95% Confidence Interval
GB 1947 Sample	624.065***	0.227*	[0.051,0.403]
NI 1957 Sample	31.32***	0.105	[-0.52,0.729]
GB 1972 Sample	634.86***	-0.124	[-0.309,0.062]
NI 1972 Sample	41.952***	-0.332	[-0.899,0.235]

Note: Covariates included are age, year at time of reform, sex, year of survey

From the 1947 Great Britain sample, the estimated coefficient for the effect of additional compulsory education attained beyond age 14 is positive and significant. Although the estimate has a relatively wide confidence interval, the point estimate of 0.227 gives the interpretation that students compelled to stay in school beyond age 14 have roughly 23% higher real earnings. The subsequent three samples corresponding to the 1957 reform in Northern Ireland along with the 1972 reform in both Great Britain and Northern Ireland yield parameter estimates which are highly variable and not statistically significant. While the 2SLS model provides a standard benchmark to estimate the returns to additional compulsory education among compliers, it relies on restrictive linear modeling assumption which may not capture the true underlying relationships within our data. As such we next present interactive IV results for each region/reform sample.

Table 6: Interactive IV LATE estimates for the effect of staying in school past age 14 on log real earnings: Great Britain, 1947 reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.344***	[0.31,0.378]
	XGBoost	0.329***	[0.295,0.363]
	Elastic Net	0.317***	[0.283,0.352]
K = 10	Random Forest	0.347***	[0.314,0.381]
	XGBoost	0.299***	[0.264,0.334]
	Elastic Net	0.331***	[0.296,0.366]

Note: Covariates included are age, year at time of reform, sex, year of survey

Table 7: Interactive IV LATE estimates for the effect of staying in school past age 14 on log real earnings: Northern Ireland, 1957 reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.279***	[0.181,0.377]
	XGBoost	0.241***	[0.15,0.331]
	Elastic Net	0.263***	[0.177,0.35]
K = 10	Random Forest	0.293***	[0.19,0.396]
	XGBoost	0.202***	[0.11,0.295]
	Elastic Net	0.216***	[0.131,0.301]

Note: Covariates included are age, year at time of reform, sex, year of survey

The results depicted in Tables 6, 7 report the interactive IV model LATE estimates for the 1947 Great Britain and 1957 Northern Ireland samples. Both tables show the causal returns to additional compulsory education attained by raising the minimum school leaving age from 14 to 15 years old. Unlike the 2SLS estimates, the causal parameters of interest are positive and significant in both regions. From Table 6 we can see the returns in Great Britain estimated through the interactive IV model are larger than the 2SLS estimates. Additionally, the estimates are robust across the three machine learning models used in the DML estimation process and across the number of folds used when cross-fitting, with point estimates ranging from 0.3 to 0.35. Examining the equivalent policy reform for the 1957 Northern Ireland sample, LATE estimates are again positive and significant ranging from 0.2 to 0.29. We further note that for the 1957 Northern Ireland reform LATE estimates are also robust to the choice of base learner and number of cross-fitting folds.

Table 8: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Great Britain, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.176***	[0.134,0.219]
	XGBoost	0.111***	[0.054,0.168]
	Elastic Net	0.174***	[0.136,0.212]
K = 10	Random Forest	0.179***	[0.136,0.222]
	XGBoost	0.105***	[0.063,0.148]
	Elastic Net	0.23***	[0.192,0.268]

Note: Covariates included are age, year at time of reform, sex, year of survey

Table 8 depicts LATE estimates for the causal returns to additional compulsory education attained through increasing the minimum school leaving age from 15 to 16 years old in Great Britain. Each estimate is again positive and statistically significant suggesting that additional compulsory education attained past age 15 had a positive effect on real wages. Comparing the 1947 Great Britain sample with the 1972 Great Britain sample, we can see the parameter estimates for the 1972 sample are sizably lower than those from 1947. This suggests that for Great Britain the causal return to the additional compulsory education achieved through raising the minimum

school leaving age from 14 to 15 years old had a significantly larger effect on real incomes than the additional compulsory education achieved by raising the minimum school leaving age from 15 to 16 years old. Furthermore, the parameter estimates for the 1972 Great Britain sample are much closer in size to the linear IV returns to education parameter estimates reported in the wider returns to education literature, see Oreopoulos (2006a); Card (2001).

Table 9: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Northern Ireland, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.252***	[0.153,0.352]
	XGBoost	0.274***	[0.178,0.37]
	Elastic Net	0.087	[-0.001,0.175]
K = 10	Random Forest	0.257***	[0.154,0.36]
	XGBoost	0.225***	[0.147,0.303]
	Elastic Net	0.113*	[0.025,0.2]

Note: Covariates included are age, year at time of reform, sex, year of survey

We are however much less confident in our results for the interactive IV models fit on the 1972 Northern Ireland sample. The LATE estimates presented in Table 9 are much less robust to the choice of machine learning model relative to the pervious samples analyzed. Examining the coefficients across the same machine learning models but different folds we can see the estimated parameters are similar across different folds but highly variable across different machine learning models. Random forests and XGboost produces very large estimates of roughly 0.25, while elastic net results in estimates of roughly half the size and in the case of the model fit with 5-fold cross-fitting are not statistically significant. As such we are less confident in the robustness of the results for the 1972 Northern Ireland sample and we are not able to draw any conclusions regarding the size of the causal parameter of interest. The variability in our estimates for the 1972 Northern Ireland sample is likely due to the weaker first stage relationship between being affected by the reform and the probability of a given student remaining in school beyond the minimum school leaving age. This contrasts with the other three region/reform samples which all saw much larger discontinuous jumps in the number of students who stayed in school past the given minimum school leaving age after the respective reform was implemented. Furthermore, the Northern Ireland samples are much smaller in size relative to the Great Britain samples, which could be a reason as to why we observe much less robust parameter estimates in the 1972 Northern Ireland sample.

Apart from the 1972 Northern Ireland sample, we find that mandating marginal students to stay in school an additional year results in additional educational attainment, which produces a substantial effect on their real wages. This highlights just how high the welfare benefits of additional compulsory education are for students who would otherwise choose not to remain in school. When we compare the estimates from the 1947 Great Britain sample with the 1972 Great Britain sample, we find that returns to additional compulsory education decrease as the minimum school

leaving age increases. Furthermore, the large positive significant parameter estimates identified by the interactive IV models suggest that there may exist complex non-linear relationships between the outcome, covariates and treatment variable which the restrictive linear modeling assumptions inherent in standard 2SLS estimates fail to capture. The robustness of our estimates suggest that our interactive IV estimates are most accurate for lager samples where the implementation of the respective policy change results in a strong first stage jump in the probability a student remains in school beyond the given minimum school leaving age. However, for smaller samples which do not show a substantial jump in the probability of a student remaining in school past the given minimum school leaving age, our interactive IV estimates produce much weaker less significant results.

7 Limitations and Future Work

Our results come with several limitations. First in examining heterogeneity in returns to compulsory education in each region/reform sample, we are limited by how each sample is constructed. Defining each region/reform sample by only considering a four year window surrounding each reform results in large variation in the respective size of each sample. Although we justify this approach by attempting to limit the effect of unobserved time and regional trends on our results, defining our samples in this way results in the Great Britain samples being much larger than the comparable Northern Ireland samples. The Great Britain samples have 12,421 and 30,724 observations for the respective 1947 and 1972 reforms, whereas the Northern Ireland samples only contain 3,207 and 4,029 observations for the 1957 and 1972 samples. Although we observe positive and significant parameter estimates for the 1957 reform, the inconclusive results from the 1972 Northern Ireland sample suggest that these results should be interpreted with caution. This highlights the importance of using sufficiently large data sets to produce accurate parameter estimates when applying double machine learning methods. To demonstrate robustness in the method by which each sample is constructed we also report results for samples constructed using three and five-year windows surrounding each reform which are presented in the Appendix. Notably constructing our samples using different windows does not significantly influence our results.

We are also constrained by the covariates included in each model. The data set used in the analysis does not contain a set of rich covariates. As such we are clearly omitting other relevant factors such as family background, school quality and other labour market and demographic variables which could influence both earnings and educational attainment. It is therefore possible that the large interactive IV model parameter estimates may be biased upwards due to omitted variable bias present in our analysis. This issue is especially important when considering the 1947 Great Britain sample. The 1947 reform came into effect shortly after the end of the second world war which represented a time of rapid and substantial economic and labour market disruption. As a result, the large parameter estimates observed in the 1947 Great Britain sample may simply reflect the effects of unobserved post war economic transformation rather than strictly increased returns to education. Since double machine learning methods can identify causal relationships in

the context of high dimensional data, future work to explore the analysis presented here with a much larger set of relevant covariates could offer more nuanced and robust results.

We are also limited by the generalizabillity of our findings. Since we examine only two U.K. specific compulsory schooling law reforms, which have clearly defined discrete legal cutoffs, we cannot reliably say how returns to additional compulsory education extends to other contexts where the implementation and enforcement of compulsory schooling laws may differ. For instance, similar reforms implemented in different regions, at different times, or with different legal cutoffs for the required number of years that students are compelled to stay in school may produce different results. Additionally, other regions may differ in factors such as school quality, baseline educational attainment, and labour market characteristics, which could drastically influence the our findings. It would be interesting for future work to apply our interactive IV framework to other compulsory schooling law reforms implemented in different regions or during different time periods.

8 Conclusion

In this paper we examined causal returns to additional compulsory education in the United Kingdom through the application of double machine learning in an interactive IV model. This was accomplished by exploiting exogenous variation in the number of years of compulsory schooling completed through two policy reforms implemented in Great Britain and Northern Ireland which raised the minimum school leaving age from 14 to 15 years old beginning in 1947, and subsequently from 15 to 16 years old in 1972. This strategy allows us to estimate the Local Average Treatment Effect (LATE) additional compulsory education has on real earnings.

Our results show that returns to additional compulsory education are positive and substantial. Examining heterogeneity in returns across the two different reforms in both Great Britain and Northern Ireland, we find that the additional compulsory education attained by increasing the minimum school leaving age from 14 to 15 years old in Great Britain increases real earnings by roughly 30 to 35% and in Northern Ireland by between 20 to 30%. These findings are much larger than previous linear IV returns to education estimates reported in the literature. In contrast estimates for the 1972 reform, which increased the minimum school leaving age from 15 to 16 years old, result in much smaller estimates of returns to additional compulsory education. For Great Britain, returns are roughly 10 to 17%, which are more in line with the linear IV returns to education estimates reported in the literature. We are however unable to draw any conclusions regarding the 1972 reform in Northern Ireland.

These findings emphasize two key points. Firstly, increased compulsory education has a pronounced effect on real earnings among marginal students who would otherwise have chosen not to continue with their education. The causal parameter estimates obtained underscore the effectiveness and importance of compulsory schooling law policies with the goal of keeping students in school longer. Secondly, we find that applying double machine learning methods to analyze returns to education results in parameter estimates that in some cases exceed those from traditional IV

estimates. This suggests that linear IV approaches may fail to capture the true underlying relationship education has on earnings. Our results however caution that sufficiently large samples are necessary to achieve robust parameter estimates through the implementation of double machine learning models. This is reflected in the results from our much smaller Northern Ireland samples, which produce noisier parameter estimates, and in the case of the 1972 reform, do not offer any conclusive results. Overall, this paper emphasizes how applying modern econometric techniques can produce new and interesting results regarding causal returns to education, suggesting that the returns to education debate may still be far from over.

9 Appendix

9.1 Results with each sample constructed using a three-year window surrounding each reform:

Table 10: 2SLS estimates for the effect of staying in school past the minimum school leaving age on log real earnings

Sample	Weak Instrument Test F-Stat	Coefficient	95% Confidence Interval
GB 1947 Sample	324.098***	0.194	[-0.039,0.427]
NI 1957 Sample	14.758***	-0.037	[-0.926,0.852]
GB 1972 Sample	418.325***	-0.204	[-0.432,0.024]
NI 1972 Sample	22.253***	-0.637	[-1.466,0.192]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 11: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Great Britain, 1947 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.336***	[0.297,0.375]
	XGBoost	0.275***	[0.236,0.313]
	Elastic Net	0.301***	[0.257,0.344]
K = 10	Random Forest	0.345***	[0.306,0.384]
	XGBoost	0.288***	[0.249,0.326]
	Elastic Net	0.306***	[0.262,0.35]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 12: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Northern Ireland, 1957 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.325***	[0.205,0.444]
	XGBoost	0.261***	[0.145,0.376]
	Elastic Net	0.28***	[0.18,0.38]
K = 10	Random Forest	0.298***	[0.177,0.42]
	XGBoost	0.245***	[0.131,0.36]
	Elastic Net	0.291***	[0.19,0.393]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 13: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Great Britain, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.177***	[0.128,0.226]
	XGBoost	0.083***	[0.037,0.129]
	Elastic Net	0.076***	[0.032,0.12]
K = 10	Random Forest	0.157***	[0.107,0.208]
	XGBoost	0.085***	[0.036,0.134]
	Elastic Net	0.143***	[0.099,0.186]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 14: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Northern Ireland, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.245***	[0.127,0.362]
	XGBoost	0.354***	[0.239,0.469]
	Elastic Net	0.112*	[0.008,0.216]
K = 10	Random Forest	0.249***	[0.133,0.366]
	XGBoost	0.339***	[0.224,0.454]
	Elastic Net	0.103	[-0.001,0.208]

Note:

Covariates included are age, year at time of reform, sex, year of survey

9.2 Results with each sample constructed using a five-year window surrounding each reform:

Table 15: 2SLS estimates for the effect of staying in school past the minimum school leaving age on log real earnings

Sample	Weak Instrument Test F-Stat	Coefficient	95% Confidence Interval
GB 1947 Sample	1055.446***	0.186*	[0.044,0.329]
NI 1957 Sample	56.096***	-0.021	[-0.506,0.464]
GB 1972 Sample	814.362***	-0.132	[-0.294,0.03]
NI 1972 Sample	57.829***	-0.266	[-0.743,0.211]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 16: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Great Britain, 1947 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.343***	[0.315,0.372]
	XGBoost	0.301***	[0.273,0.33]
	Elastic Net	0.334***	[0.306,0.362]
K = 10	Random Forest	0.34***	[0.311,0.369]
	XGBoost	0.321***	[0.291,0.351]
	Elastic Net	0.358***	[0.329,0.387]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 17: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Northern Ireland, 1957 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.317***	[0.23,0.404]
	XGBoost	0.336***	[0.241,0.43]
	Elastic Net	0.246***	[0.169,0.323]
K = 10	Random Forest	0.329***	[0.24,0.418]
	XGBoost	0.311***	[0.23,0.391]
	Elastic Net	0.223***	[0.145,0.302]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 18: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Great Britain, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.216***	[0.179,0.254]
	XGBoost	0.097***	[0.059,0.134]
	Elastic Net	0.084***	[0.05,0.119]
K = 10	Random Forest	0.221***	[0.183,0.259]
	XGBoost	0.105***	[0.068,0.143]
	Elastic Net	0.166***	[0.132,0.2]

Note:

Covariates included are age, year at time of reform, sex, year of survey

Table 19: Interactive IV LATE estimates for the effect of staying in school past age 15 on log real earnings: Northern Ireland, 1972 Reform

Cross-fitting Folds	Learner	Coefficient	95% Confidence Interval
K = 5	Random Forest	0.272***	[0.18,0.363]
	XGBoost	0.26***	[0.176,0.344]
	Elastic Net	0.13***	[0.053,0.207]
K = 10	Random Forest	0.265***	[0.178,0.353]
	XGBoost	0.252***	[0.163,0.341]
	Elastic Net	0.146***	[0.067,0.224]

Note:

Covariates included are age, year at time of reform, sex, year of survey

References

Angrist, J. D. and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings?*. *The Quarterly Journal of Economics 106*(4), 979–1014.

Card, D. (1999). The Causal Effect of Education on Earnings. In *Handbook of Labor Economics*, Volume 3, pp. 1801–1863. Elsevier.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160. Num Pages: 34 Place: Evanston, United Kingdom Publisher: Blackwell Publishing Ltd.

Card, D., L. N. Christofides, E. K. Grant, and R. Swidinsky (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In *Aspects of labour market behaviour: Essays in honour of John Vanderkamp*, pp. 201–222. Toronto; Buffalo and London: University of Toronto Press.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Dickson, M. and C. Harmon (2011). Economic returns to education: What We Know, What We Don't Know, and Where We Are Going—Some brief pointers. *Economics of Education Review 30*(6), 1118–1122.

Oreopoulos, P. (2006a). Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter. *The American Economic Review 96*(1), 152–175.

Oreopoulos, P. (2006b). Replication data for: Estimating average and local average treatment effects of education when compulsory schooling laws really matter.

Stephens, M. and D.-Y. Yang (2014). Compulsory Education and the Benefits of Schooling. *American Economic Review 104*(6), 1777–1792.

Liam Henson ECO1400

2025-12-01

```
rm(list=ls())
set.seed(198)
library(mlr3tuning)
```

```
## Loading required package: mlr3
```

```
## Loading required package: paradox
```

```
library(haven)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ivreg)
library(DoubleML)
library(mlr3)
library(mlr3learners)
library(paradox)
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
```

```
##
## The following object is masked from 'package:dplyr':
##
##     add_rownames
##
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:huxtable':
##
##     add_footnote
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(tinytex)
library(knitr)
```

```
#reads, selects relevant variables and filterers data
data = read_dta("combined-general-household-survey.dta")
filtered_data = data %>%
  select("earn",
         "agelfted",
         "yobirth",
         "age",
         "datyear",
         "sex",
         "nireland") %>%
  filter(!is.na(agelfted) & !is.na(earn) & age <= 64 & agelfted >= 10 & agelfted <= 18) %>%
  mutate(
    yearat14 = yobirth + 14,
    yearat15 = yobirth + 15,
    learn = log(earn)
  )
```

```r
# Defines treatment and instrument variables
filtered_data = filtered_data %>%
  mutate(
    Law14 = ifelse((nireland ==0 & yearat14 >= 47) | (nireland ==1 & yearat14 >= 57),1,0),
    Law15 = ifelse(yearat15>= 73,1,0),
    School14 = as.integer(agelfted >= 15),
    School15 = as.integer(agelfted >= 16)
  )

write.csv(filtered_data,"filtered_data.csv")
filtered_data = read.csv("filtered_data.csv")

# Specifies window used to construct samples
# (band = 4 produces main results in the paper, band = 3, band = 5 produces results given in appendix)
band = 4

# Constructs samples for each reform/region
GB_47_sample = filtered_data %>%
  filter(yearat14 >= 47 - band & yearat14 <= 47 + band & nireland ==0)

NI_57_sample = filtered_data %>%
  filter(yearat14 >= 57 - band & yearat14 <= 57 + band  & nireland ==1)

GB_72_sample = filtered_data %>%
  filter(yearat15 >= 72 - band, yearat15 <= 72 + band & nireland ==0)

NI_72_sample = filtered_data %>%
  filter(yearat15 >= 72 - band, yearat15 <= 72 + band & nireland == 1)
```

```r
# Create a function to create tables to summary statistics for each region/reform sample
summary_stats = function(data, vars, var_names,Law,caption) {

  #Creates two separate data frames for pre and post law expose
  pre_treat_data = data[data[[Law]]==0,vars]
  post_treat_data = data[data[[Law]]==1,vars]

  # Computes mean and sd for each data frame rounding to 2 decimal places
  pre_teat_mean = round(apply(pre_treat_data, 2, mean),2)
  pre_teat_sd = round(apply(pre_treat_data, 2, sd),2)
  post_teat_mean = round(apply(post_treat_data, 2, mean),2)
  post_treat_sd = round(apply(post_treat_data, 2, sd),2)

  # Creates a table of estimated means and sds for each variable in each a sample
  stats_data = data.frame(
    Variable = var_names,
    Non_exposed = paste0(pre_teat_mean, " (",pre_teat_sd,")"),
    Exposed = paste0(post_teat_mean," (",post_treat_sd,")")
  )

  # Output as Latex table
  stats_table = kable(
    stats_data,
    col.names = c(
```

```
      "Variable",
      "Pre Reform Non Exposed Cohorts",
      "Post Reform Exposed Cohorts"
    ), format = "latex",
    booktabs = TRUE,
    align = "lcc",
    caption = caption
  )
  stats_table = footnote(stats_table,
                         "Values are in presented in mean (sd). Real annual earnings reflect 1998 U.K. p
  return(cat(stats_table))
}

#Defines variables and table titles
vars15 = c("learn","agelfted","age","School14")
labels15 = c("Log Real Annual Earnings","Age Left Education",
             "Age at Time of Survey","Proportion Stayed Past 14")

vars16 = c("learn","agelfted","age","School15")
labels16 = c("Log Real Annual Earnings","Age Left Education",
             "Age at Time of Survey","Proportion Stayed Past 15")


# Creates tables 1..4 in the report
captionGB1947 = paste0("Descriptive Statistics: Great Britain 1947 Sample 14 to 15 reform",
                       " (N=", nrow(GB_47_sample),")")
summary_stats(GB_47_sample, vars15, labels15,"Law14",captionGB1947)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-4}Descriptive Statistics: Great Britain 1947 Sample 14 to 15 refor
## \centering
## \begin{tabular}[t]{lcc}
## \toprule
## Variable & Pre Reform Non Exposed Cohorts & Post Reform Exposed Cohorts\\
## \midrule
## Log Real Annual Earnings & 8.66 (1.02) & 8.79 (0.97)\\
## Age Left Education & 14.64 (1.11) & 15.33 (0.9)\\
## Age at Time of Survey & 57.37 (3.05) & 54.54 (4.33)\\
## Proportion Stayed Past 14 & 0.34 (0.47) & 0.9 (0.3)\\
## \bottomrule
## \multicolumn{3}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{3}{l}{\rule{0pt}{1em}Values are in presented in mean (sd). Real annual earnings reflect
## \end{tabular}
## \end{table}
```

```
captionNI1957 = paste0("Descriptive Statistics: Northern Ireland 1957 Sample 14 to 15 reform",
                       " (N=", nrow(NI_57_sample),")")
summary_stats(NI_57_sample, vars15, labels15,"Law14",captionNI1957)
```

```
## \begin{table}
##
```

```
## \caption{\label{tab:unnamed-chunk-4}Descriptive Statistics: Northern Ireland 1957 Sample 14 to 15 re
## \centering
## \begin{tabular}[t]{lcc}
## \toprule
## Variable & Pre Reform Non Exposed Cohorts & Post Reform Exposed Cohorts\\
## \midrule
## Log Real Annual Earnings & 8.7 (0.89) & 8.76 (0.9)\\
## Age Left Education & 15.09 (1.31) & 15.62 (1.09)\\
## Age at Time of Survey & 50.11 (4.16) & 45.8 (4.47)\\
## Proportion Stayed Past 14 & 0.54 (0.5) & 0.89 (0.31)\\
## \bottomrule
## \multicolumn{3}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{3}{l}{\rule{0pt}{1em}Values are in presented in mean (sd). Real annual earnings reflect
## \end{tabular}
## \end{table}
```

```
captionGB1973 = paste0("Descriptive Statistics: Great Britain 1972 Sample 15 to 16 reform",
                       " (N=", nrow(GB_72_sample),")")
summary_stats(GB_72_sample, vars16, labels16,"Law15",captionGB1973)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-4}Descriptive Statistics: Great Britain 1972 Sample 15 to 16 refor
## \centering
## \begin{tabular}[t]{lcc}
## \toprule
## Variable & Pre Reform Non Exposed Cohorts & Post Reform Exposed Cohorts\\
## \midrule
## Log Real Annual Earnings & 9.05 (1.11) & 9.06 (1.1)\\
## Age Left Education & 15.94 (1.02) & 16.33 (0.84)\\
## Age at Time of Survey & 39.78 (7.27) & 35.45 (7.23)\\
## Proportion Stayed Past 15 & 0.61 (0.49) & 0.92 (0.27)\\
## \bottomrule
## \multicolumn{3}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{3}{l}{\rule{0pt}{1em}Values are in presented in mean (sd). Real annual earnings reflect
## \end{tabular}
## \end{table}
```

```
captionNI173 = paste0("Descriptive Statistics: Northern Ireland 1972 Sample 15 to 16 reform",
                      " (N=", nrow(NI_72_sample),")")
summary_stats(NI_72_sample, vars16, labels16,"Law15",captionNI173)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-4}Descriptive Statistics: Northern Ireland 1972 Sample 15 to 16 re
## \centering
## \begin{tabular}[t]{lcc}
## \toprule
## Variable & Pre Reform Non Exposed Cohorts & Post Reform Exposed Cohorts\\
## \midrule
## Log Real Annual Earnings & 8.9 (0.79) & 8.88 (0.73)\\
## Age Left Education & 16.16 (1.05) & 16.5 (0.84)\\
## Age at Time of Survey & 36.56 (4.7) & 32.05 (4.66)\\
```

```
## Proportion Stayed Past 15 & 0.7 (0.46) & 0.95 (0.23)\\
## \bottomrule
## \multicolumn{3}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{3}{l}{\rule{0pt}{1em}Values are in presented in mean (sd). Real annual earnings reflect
## \end{tabular}
## \end{table}
```

```
# A function to plot proportion of students which stay in school past given legal threshold over time
year_plot = function(data,year_var,school_var,min_year,max_year,ylab,reform_year) {

  # Filters which years we want to include in the plot based on min_year and max_year
  data = data[data[[year_var]]>=min_year & data[[year_var]]<=max_year,]

  # Defines years plotted as a factor for easier indexing as defining as a factor automatically orders
  data[[year_var]] = factor(data[[year_var]])
  years = levels(data[[year_var]])

  # Creates empty vector of proportions for each year
  avg_left_school = rep(0, length(years))

  # Creates vector of proportions for each year
  for (i in 1:length(years)) {
    avg_left_school[i] = mean(data[data[[year_var]]==years[i],school_var])
  }


  plot(as.numeric(years) + 1900, avg_left_school,xlab = "Year",ylab = ylab,type = "o",ylim = c(0.2,1))

  #Adds vertical ling to the plot to specify the year the reform was implemented
  abline(v=reform_year,lty = 5)

  #Adds year od reform next to vertical line on the plot
  text(reform_year-3, 0.92, labels = reform_year, pos = 4, offset = 0.3, cex = 0.9)
}
```

```
#Filters data to have just Great Britain observations
filtered_data_GB = filtered_data %>% filter(nireland ==0)

# Creates Figure 1 in the report
png("GB_Figure1.png", width = 10, height = 4,units = "in", res = 500)
par(mfrow = c(1,2), mar = c(5, 5, 2, 2))
ylab = expression("Proportion staying in school past age 14")
year_plot(filtered_data_GB,"yearat14","School14",37,57,ylab,1947)


ylab = expression("Proportion staying in school past age 15")
year_plot(filtered_data_GB,"yearat15","School15",62,82,ylab,1972)
dev.off()
```

```
## pdf
##   2
```

```r
#Filters data to have just Great Britain observations
filtered_data_GB = filtered_data %>% filter(nireland ==1)

# Creates Figure 2 in report
png("NI_Figure2.png", width = 10, height = 4,units = "in", res = 500)
par(mfrow = c(1,2), mar = c(5, 5, 2, 2))
ylab = expression("Proportion staying in school past age 14")
year_plot(filtered_data_GB,"yearat14","School14",44,67,ylab,1957)


ylab = expression("Proportion staying in school past age 15")
year_plot(filtered_data_GB,"yearat15","School15",60,80,ylab,1972)
dev.off()
```

```
## pdf
##   2
```

```r
#Defines the learners to be used in the analysis

#Defines random forest
ml_gRf = lrn("regr.ranger")
ml_mRf = lrn("classif.ranger", predict_type = "prob")
ml_rRf = lrn("classif.ranger", predict_type = "prob")

#Defines XGboost
ml_gXG = lrn("regr.xgboost")
ml_mXG = lrn("classif.xgboost", predict_type = "prob")
ml_rXG = lrn("classif.xgboost", predict_type = "prob")

#Defines elastic net
ml_gGlm = lrn("regr.cv_glmnet")
ml_mGlm = lrn("classif.cv_glmnet", predict_type = "prob")
ml_rGlm = lrn("classif.cv_glmnet", predict_type = "prob")
```

```r
# Define tuning parameters for Random Forest, XGBoost and Elastic Net
Rf_tuning_params = ps(
  num.trees = p_int(lower = 150, upper = 650),
  max.depth = p_int(lower = 3, upper = 9),
  min.node.size = p_int(lower = 55, upper = 220)
)

Xgb_tuning_params = ps(
  nrounds = p_int(lower = 60, upper = 250),
  max_depth = p_int(lower = 3, upper = 9),
  min_child_weight = p_dbl(lower = 1, upper = 8),
  gamma = p_dbl(lower = 0, upper = 4.),
  eta = p_dbl(lower = 0.01, upper = 0.4)
)

Net_tuning_param = ps(alpha = p_dbl(lower = 0.01, upper = 0.99))
```

```r
# Fits IIVM models for Great Britain 1947 reform 5-fold cross-fitting
x_15 = c("yearat14", "age", 'datyear', 'sex')
IIVM_14_GB_data = DoubleMLData$new(
  data = GB_47_sample,
  y_col = "learn",
  d_cols = "School14",
  z_cols = "Law14",
  x_cols = x_15
)

#Fits Random Forest model
IIVM_14_GB_RF_5_fold = DoubleMLIIVM$new(
  IIVM_14_GB_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_RF_5_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                           "ml_m" = Rf_tuning_params,
                                           "ml_r" = Rf_tuning_params))
IIVM_14_GB_RF_5_fold$fit()
IIVM_14_GB_RF_5_fold$coef


#Fits XGBoost model
IIVM_14_GB_XG_5_fold = DoubleMLIIVM$new(
  IIVM_14_GB_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_XG_5_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                           "ml_m" = Xgb_tuning_params,
                                           "ml_r" = Xgb_tuning_params))
IIVM_14_GB_XG_5_fold$fit()
IIVM_14_GB_XG_5_fold$coef

#Fits Elastic Net model
IIVM_14_GB_Glm_5_fold = DoubleMLIIVM$new(
```

```
  IIVM_14_GB_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_Glm_5_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                            "ml_m" = Net_tuning_param,
                                            "ml_r" = Net_tuning_param))
IIVM_14_GB_Glm_5_fold$fit()
IIVM_14_GB_Glm_5_fold$coef



# Fits IIVM model for Great Britain 1947 reform 10-fold cross-fitting

#Fits Random Forest model
IIVM_14_GB_RF_10_fold = DoubleMLIIVM$new(
  IIVM_14_GB_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_RF_10_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                            "ml_m" = Rf_tuning_params,
                                            "ml_r" = Rf_tuning_params))
IIVM_14_GB_RF_10_fold$fit()
IIVM_14_GB_RF_10_fold$coef


#Fits XGBoost  model
IIVM_14_GB_XG_10_fold = DoubleMLIIVM$new(
  IIVM_14_GB_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
```

```
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_XG_10_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                            "ml_m" = Xgb_tuning_params,
                                            "ml_r" = Xgb_tuning_params))
IIVM_14_GB_XG_10_fold$fit()
IIVM_14_GB_XG_10_fold$coef

#Fits Elastic Net model
IIVM_14_GB_Glm_10_fold = DoubleMLIIVM$new(
  IIVM_14_GB_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_GB_Glm_10_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                            "ml_m" = Net_tuning_param,
                                            "ml_r" = Net_tuning_param))
IIVM_14_GB_Glm_10_fold$fit()
IIVM_14_GB_Glm_10_fold$coef
```

```
# Fits IIVM model for Northern Ireland 1957 reform 5-fold cross-fitting
x_15 = c("yearat14", "age", 'datyear', 'sex')
IIVM_14_NI_data = DoubleMLData$new(
  data  = NI_57_sample,
  y_col = "learn",
  d_cols = "School14",
  z_cols = "Law14",
  x_cols = x_15
)

#Fits Random Forest model
IIVM_14_NI_RF_5_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)
```

```
IIVM_14_NI_RF_5_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                           "ml_m" = Rf_tuning_params,
                                           "ml_r" = Rf_tuning_params))
IIVM_14_NI_RF_5_fold$fit()
IIVM_14_NI_RF_5_fold$coef


#Fits XGBoost  model
IIVM_14_NI_XG_5_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_NI_XG_5_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                           "ml_m" = Xgb_tuning_params,
                                           "ml_r" = Xgb_tuning_params))
IIVM_14_NI_XG_5_fold$fit()
IIVM_14_NI_XG_5_fold$coef

#Fits Elastic Net model
IIVM_14_NI_Glm_5_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_NI_Glm_5_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                           "ml_m" = Net_tuning_param,
                                           "ml_r" = Net_tuning_param))
IIVM_14_NI_Glm_5_fold$fit()
IIVM_14_NI_Glm_5_fold$coef




# Fits IIVM model for Northern Ireland 1957 reform 10-fold cross-fitting

#Fits Random Forest model
```

```
IIVM_14_NI_RF_10_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_NI_RF_10_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                            "ml_m" = Rf_tuning_params,
                                            "ml_r" = Rf_tuning_params))
IIVM_14_NI_RF_10_fold$fit()
IIVM_14_NI_RF_10_fold$coef


#Fits XGBoost  model
IIVM_14_NI_XG_10_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_14_NI_XG_10_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                            "ml_m" = Xgb_tuning_params,
                                            "ml_r" = Xgb_tuning_params))
IIVM_14_NI_XG_10_fold$fit()
IIVM_14_NI_XG_10_fold$coef

#Fits Elastic Net model
IIVM_14_NI_Glm_10_fold = DoubleMLIIVM$new(
  IIVM_14_NI_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)
```

```
IIVM_14_NI_Glm_10_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                             "ml_m" = Net_tuning_param,
                                             "ml_r" = Net_tuning_param))
IIVM_14_NI_Glm_10_fold$fit()
IIVM_14_NI_Glm_10_fold$coef
```

```
# Fits IIVM model for Great Britain 1973 reform 10-fold cross-fitting
x_16 = c("yearat15", "age", 'datyear', 'sex')
IIVM_15_GB_data = DoubleMLData$new(
  data  = GB_72_sample,
  y_col = "learn",
  d_cols = "School15",
  z_cols = "Law15",
  x_cols = x_16
)

#Fits Random Forest model
IIVM_15_GB_RF_5_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_RF_5_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                           "ml_m" = Rf_tuning_params,
                                           "ml_r" = Rf_tuning_params))
IIVM_15_GB_RF_5_fold$fit()
IIVM_15_GB_RF_5_fold$coef


#Fits XGBoost  model
IIVM_15_GB_XG_5_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_XG_5_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                           "ml_m" = Xgb_tuning_params,
```

```
                                      "ml_r" = Xgb_tuning_params))
IIVM_15_GB_XG_5_fold$fit()
IIVM_15_GB_XG_5_fold$coef

#Fits Elastic Net model
IIVM_15_GB_Glm_5_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_Glm_5_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                      "ml_m" = Net_tuning_param,
                                      "ml_r" = Net_tuning_param))
IIVM_15_GB_Glm_5_fold$fit()
IIVM_15_GB_Glm_5_fold$coef




# Fits IIVM model for Great Britain 1973 reform 10-fold cross-fitting
#Fits Random Forest model
IIVM_15_GB_RF_10_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_RF_10_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                      "ml_m" = Rf_tuning_params,
                                      "ml_r" = Rf_tuning_params))
IIVM_15_GB_RF_10_fold$fit()
IIVM_15_GB_RF_10_fold$coef


#Fits XGBoost  model
IIVM_15_GB_XG_10_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gXG,
```

```
  ml_mXG,
  ml_rXG,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_XG_10_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                            "ml_m" = Xgb_tuning_params,
                                            "ml_r" = Xgb_tuning_params))
IIVM_15_GB_XG_10_fold$fit()
IIVM_15_GB_XG_10_fold$coef

#Fits Elastic Net model
IIVM_15_GB_Glm_10_fold = DoubleMLIIVM$new(
  IIVM_15_GB_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_GB_Glm_10_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                            "ml_m" = Net_tuning_param,
                                            "ml_r" = Net_tuning_param))
IIVM_15_GB_Glm_10_fold$fit()
IIVM_15_GB_Glm_10_fold$coef
```

```
# Fits IIVM model for Northern Ireland 1973 reform 5-fold cross-fitting
x_16 = c("yearat15", "age", 'datyear', 'sex')
IIVM_15_NI_data = DoubleMLData$new(
  data  = NI_72_sample,
  y_col = "learn",
  d_cols = "School15",
  z_cols = "Law15",
  x_cols = x_16
)

#Fits Random Forest model
IIVM_15_NI_RF_5_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 5,
```

```
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_RF_5_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                           "ml_m" = Rf_tuning_params,
                                           "ml_r" = Rf_tuning_params))
IIVM_15_NI_RF_5_fold$fit()
IIVM_15_NI_RF_5_fold$coef


#Fits XGBoost  model
IIVM_15_NI_XG_5_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 5,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 10,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_XG_5_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                           "ml_m" = Xgb_tuning_params,
                                           "ml_r" = Xgb_tuning_params))
IIVM_15_NI_XG_5_fold$fit()
IIVM_15_NI_XG_5_fold$coef

#Fits Elastic Net model
IIVM_15_NI_Glm_5_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_Glm_5_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                           "ml_m" = Net_tuning_param,
                                           "ml_r" = Net_tuning_param))
IIVM_15_NI_Glm_5_fold$fit()
IIVM_15_NI_Glm_5_fold$coef
```

```r
# Fits IIVM model for Northern Ireland 1973 reform 10-fold cross-fitting

#Fits Random Forest model
IIVM_15_NI_RF_10_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gRf,
  ml_mRf,
  ml_rRf,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_RF_10_fold$tune(param_set = list("ml_g" = Rf_tuning_params,
                                            "ml_m" = Rf_tuning_params,
                                            "ml_r" = Rf_tuning_params))
IIVM_15_NI_RF_10_fold$fit()
IIVM_15_NI_RF_10_fold$coef


#Fits XGBoost  model
IIVM_15_NI_XG_10_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gXG,
  ml_mXG,
  ml_rXG,
  n_folds = 10,
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_XG_10_fold$tune(param_set = list("ml_g" = Xgb_tuning_params,
                                            "ml_m" = Xgb_tuning_params,
                                            "ml_r" = Xgb_tuning_params))
IIVM_15_NI_XG_10_fold$fit()
IIVM_15_NI_XG_10_fold$coef

#Fits Elastic Net model
IIVM_15_NI_Glm_10_fold = DoubleMLIIVM$new(
  IIVM_15_NI_data,
  ml_gGlm,
  ml_mGlm,
  ml_rGlm,
  n_folds = 10,
```

```
  apply_cross_fitting = TRUE,
  score = "LATE",
  n_rep = 5,
  trimming_threshold = 0.05,
  subgroups = list(always_takers = TRUE, never_takers = TRUE)
)

IIVM_15_NI_Glm_10_fold$tune(param_set = list("ml_g" = Net_tuning_param,
                                             "ml_m" = Net_tuning_param,
                                             "ml_r" = Net_tuning_param))
IIVM_15_NI_Glm_10_fold$fit()
IIVM_15_NI_Glm_10_fold$coef
```

```
# Defines IV models for each region/reform sample
IvGB47 = ivreg(learn ~ School14 + age + sex + yearat14 + datyear
               | Law14 + age + sex + yearat14 + datyear,data = GB_47_sample)

IvNI57 = ivreg(learn ~ School14 + age + sex + yearat14 + datyear
               | Law14 + age + sex + yearat14 + datyear,data = NI_57_sample)

IvGB72 = ivreg(learn ~ School15 + age + sex + yearat15 + datyear
               | Law15 + age + sex + yearat15 + datyear,data = GB_72_sample)

IvNI72 = ivreg(learn ~ School15 + age + sex + yearat15 + datyear
               | Law15 + age + sex + yearat15 + datyear,data = NI_72_sample)
```

```
# A function to create table of results (coef, CI) for each IV model
get_params = function(model) {
  #Creates empty list of model parameters
  params = list()

  #gets p-value for variable of interest
  pval = summary(model)$coefficients[2,4]
  pval_F = as.numeric(summary(model, diagnostics = TRUE)$diagnostics["Weak instruments", "p-value"])
  #Adds parameters to the list and rounds to 3 decimal places
  params[[1]] = round(as.numeric(model$coefficients[2]), 3)
  params[[2]] = round(as.numeric(confint(model)[2, 1]), 3)
  params[[3]] = round(as.numeric(confint(model)[2, 2]), 3)
  params[[4]] = case_when(pval<0.001~"***",
                          pval<0.01~"**",
                          pval<0.05~"*",
                          .default = "")
  params[[5]] = round(as.numeric(summary(model,
                                 diagnostics = TRUE)$diagnostics["Weak instruments", "statistic"]), 3
  params[[6]] = case_when(pval_F<0.001~"***",
                          pval_F<0.01~"**",
                          pval_F<0.05~"*",
                          .default = "")
  return(params)
}

# Creates a list of each IV model
models = list(IvGB47, IvNI57, IvGB72, IvNI72)
```

```r
# Creates a matrix of model parameters
params = sapply(models, get_params)

#Creates data frame of, learners, estimated coefs, estimated CIs, and weak instrument F-stats
results = data.frame(
  Sample = c("GB 1947 Sample", "NI 1957 Sample", "GB 1972 Sample", "NI 1972 Sample"),
  Fstat = paste0(as.numeric(params[5, ]),params[6,]),
  Coef = paste0(as.numeric(params[1, ]),params[4,]),
  CI = paste0("[", params[2, ], ",", params[3, ], "]")
)

#Outputs results in a latex table (Creates table 5 in the report)
Iv_table = kable(
  results,
  col.names = c(
    "Sample",
    "Weak Instrument Test F-Stat",
    "Coefficient",
    "95 Confidence Interval"

  ),
  format = 'latex',
  booktabs = TRUE,
  align = "lccc",
  caption = "2SLS estimates for the effect of staying in school
  past the minimum school leaving age on log real earnings"
)

Iv_table = footnote(
  Iv_table,
  "Covariates included are age, year at time of reform, sex, year of survey"
)

cat(Iv_table)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-15}2SLS estimates for the effect of staying in school
##   past the minimum school leaving age on log real earnings}
## \centering
## \begin{tabular}[t]{lccc}
## \toprule
## Sample & Weak Instrument Test F-Stat & Coefficient & 95 Confidence Interval\\
## \midrule
## GB 1947 Sample & 624.065*** & 0.227* & {}[0.051,0.403]\\
## NI 1957 Sample & 31.32*** & 0.105 & {}[-0.52,0.729]\\
## GB 1972 Sample & 634.86*** & -0.124 & {}[-0.309,0.062]\\
## NI 1972 Sample & 41.952*** & -0.332 & {}[-0.899,0.235]\\
## \bottomrule
## \multicolumn{4}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{4}{l}{\rule{0pt}{1em}Covariates included are age, year at time of reform, sex, year of
## \end{tabular}
## \end{table}
```

```
# A function to create tables of results coef, CI for IIVM model
dml_results = function(models, caption) {

  # A function that will be used to extract coefs and CIs from each given model
  get_params = function(model) {
    #Creates empty list of model parameters
    params = list()

    #Adds parameters to the list and rounds to 3 decimal places
    params[[1]] = round(as.numeric(model$coef), 3)
    params[[2]] = round(as.numeric(model$confint()[1]), 3)
    params[[3]] = round(as.numeric(model$confint()[2]), 3)
    params[[4]] = case_when(as.numeric(model$pval)<0.001~"***",
                            as.numeric(model$pval)<0.01~"**",
                          as.numeric(model$pval)<0.05~"*",,
                          .default = "")
    return(params)
  }

  # Creates a matrix of model parameters
  params = sapply(models, get_params)

  #Creates data frame of folds, learners, estimated coefs, estimated CIs
  results = data.frame(
    K = c("K = 5", "", "", "K = 10", "", ""),
    Learner = rep(c("Random Forest", "XGBoost", "Elastic Net"), 2),
    Coef = paste0(as.numeric(params[1, ]),params[4, ]),
    CI = paste0("[", params[2, ], ",", params[3, ], "]")
  )

  #Outputs results in a latex table
  dml_results = kable(
    results,
    col.names = c("Cross-fitting Folds", "Learner", "Coefficient", "95 Confidence Interval"),
    format = 'latex',
    booktabs = TRUE,
    align = "lcc",
    caption = caption
  )
  dml_results = footnote(dml_results,"Covariates included are age,
                         year at time of reform, sex, year of survey")
  return(cat(dml_results))
}
```

```
#Outputs IIVM model results in latex tables for each region/reform combination
#Creating tables 6..9 in the report
dml_results(
  list(
    IIVM_14_GB_RF_5_fold,
    IIVM_14_GB_XG_5_fold,
    IIVM_14_GB_Glm_5_fold,
    IIVM_14_GB_RF_10_fold,
    IIVM_14_GB_XG_10_fold,
```

```
    IIVM_14_GB_Glm_10_fold
  ),
  "Interactive IV LATE estimates for the effect of staying
  in school past age 15 on log real earnings: Great Britain, 1947 Reform"
)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-17}Interactive IV LATE estimates for the effect of staying
##   in school past age 15 on log real earnings: Great Britain, 1947 Reform}
## \centering
## \begin{tabular}[t]{lccl}
## \toprule
## Cross-fitting Folds & Learner & Coefficient & 95 Confidence Interval\\
## \midrule
## K = 5 & Random Forest & 0.344*** & {}[0.31,0.378]\\
##  & XGBoost & 0.329*** & {}[0.295,0.363]\\
##  & Elastic Net & 0.317*** & {}[0.283,0.352]\\
## K = 10 & Random Forest & 0.347*** & {}[0.314,0.381]\\
##  & XGBoost & 0.299*** & {}[0.264,0.334]\\
## \addlinespace
##  & Elastic Net & 0.331*** & {}[0.296,0.366]\\
## \bottomrule
## \multicolumn{4}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{4}{l}{\rule{0pt}{1em}makecell[l]{Covariates included are age, \
## \end{tabular}
## \end{table}
```

```
dml_results(
  list(
    IIVM_14_NI_RF_5_fold,
    IIVM_14_NI_XG_5_fold,
    IIVM_14_NI_Glm_5_fold,
    IIVM_14_NI_RF_10_fold,
    IIVM_14_NI_XG_10_fold,
    IIVM_14_NI_Glm_10_fold
  ),
  "Interactive IV LATE estimates for the effect of staying
  in school past age 15 on log real earnings: Northern Ireland, 1957 Reform"
)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-17}Interactive IV LATE estimates for the effect of staying
##   in school past age 15 on log real earnings: Northern Ireland, 1957 Reform}
## \centering
## \begin{tabular}[t]{lccl}
## \toprule
## Cross-fitting Folds & Learner & Coefficient & 95 Confidence Interval\\
## \midrule
## K = 5 & Random Forest & 0.279*** & {}[0.181,0.377]\\
##  & XGBoost & 0.241*** & {}[0.15,0.331]\\
##  & Elastic Net & 0.263*** & {}[0.177,0.35]\\
```

```
## K = 10 & Random Forest & 0.293*** & {}[0.19,0.396]\\
##  & XGBoost & 0.202*** & {}[0.11,0.295]\\
## \addlinespace
##  & Elastic Net & 0.216*** & {}[0.131,0.301]\\
## \bottomrule
## \multicolumn{4}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{4}{l}{\rule{0pt}{1em}makecell[l]{Covariates included are age, \
## \end{tabular}
## \end{table}
```

```
dml_results(
  list(
    IIVM_15_GB_RF_5_fold,
    IIVM_15_GB_XG_5_fold,
    IIVM_15_GB_Glm_5_fold,
    IIVM_15_GB_RF_10_fold,
    IIVM_15_GB_XG_10_fold,
    IIVM_15_GB_Glm_10_fold
  ),
  "Interactive IV LATE estimates for the effect of staying
  in school past age 15 on log real earnings: Great Britain, 1972 Reform"
)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-17}Interactive IV LATE estimates for the effect of staying
##   in school past age 15 on log real earnings: Great Britain, 1972 Reform}
## \centering
## \begin{tabular}[t]{lccl}
## \toprule
## Cross-fitting Folds & Learner & Coefficient & 95 Confidence Interval\\
## \midrule
## K = 5 & Random Forest & 0.176*** & {}[0.134,0.219]\\
##  & XGBoost & 0.111*** & {}[0.054,0.168]\\
##  & Elastic Net & 0.174*** & {}[0.136,0.212]\\
## K = 10 & Random Forest & 0.179*** & {}[0.136,0.222]\\
##  & XGBoost & 0.105*** & {}[0.063,0.148]\\
## \addlinespace
##  & Elastic Net & 0.23*** & {}[0.192,0.268]\\
## \bottomrule
## \multicolumn{4}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{4}{l}{\rule{0pt}{1em}makecell[l]{Covariates included are age, \
## \end{tabular}
## \end{table}
```

```
dml_results(
  list(
    IIVM_15_NI_RF_5_fold,
    IIVM_15_NI_XG_5_fold,
    IIVM_15_NI_Glm_5_fold,
    IIVM_15_NI_RF_10_fold,
    IIVM_15_NI_XG_10_fold,
    IIVM_15_NI_Glm_10_fold
```

```
  ),
  "Interactive IV LATE estimates for the effect of staying
  in school past age 15 on log real earnings: Northern Ireland, 1972 Reform"
)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-17}Interactive IV LATE estimates for the effect of staying
##   in school past age 15 on log real earnings: Northern Ireland, 1972 Reform}
## \centering
## \begin{tabular}[t]{lccl}
## \toprule
## Cross-fitting Folds & Learner & Coefficient & 95 Confidence Interval\\
## \midrule
## K = 5 & Random Forest & 0.252*** & {}[0.153,0.352]\\
##  & XGBoost & 0.274*** & {}[0.178,0.37]\\
##  & Elastic Net & 0.087 & {}[-0.001,0.175]\\
## K = 10 & Random Forest & 0.257*** & {}[0.154,0.36]\\
##  & XGBoost & 0.225*** & {}[0.147,0.303]\\
## \addlinespace
##  & Elastic Net & 0.113* & {}[0.025,0.2]\\
## \bottomrule
## \multicolumn{4}{l}{\rule{0pt}{1em}\textit{Note: }}\\
## \multicolumn{4}{l}{\rule{0pt}{1em}makecell[l]{Covariates included are age, \
## \end{tabular}
## \end{table}
```