Research Article

Marco Giese*

Predicting Higher Education Grades using Strategies Correcting for Panel Attrition

<https://doi.org/10.1515/edu-2020-0123>

Received Jun 12, 2020; accepted Aug 28, 2020

Abstract: This study aims to forecast the final grade of the first higher education degree which can be of considerable interest for higher education institutions to implement early warning systems, students themselves, or potential employers. The analysis is based on the National Education Panel Study (NEPS), a large German dataset covering many aspects of students' (educational) life. Since panel attrition concerns 35% of participants the Heckman correction and the inverse probability weight (IPW) estimator are used to reduce the estimation bias. A distinction is made between two scenarios, excluding dropout students and including them with a grade of 5.0. Some predictors reveal significant parameter estimates in the first but not in the second scenario, or vice versa, which means that dropout and study performance is not driven by the same variables. To get an early prediction of grades only variables of a pre-university episode were included in the first step. Afterward, variables of the early study phase are included. For the IPW estimator, the R^2 improves from 0.202 to 0.593 (dropouts included) when adding the additional variables. The best predictors are the grades at secondary school, grades in the first exams, and the type of institution.

Keywords: grade prediction, higher education, students' performance, dropout, Tweedie glm

1 Introduction

In the last two decades, numerous studies investigated students' dropout from higher education. But there are only a few studies that forecast the final grade of students. The prediction of grades at an early time of study, long before graduation, can be helpful to implement an early warning system for students at risk (Beck and Davidson,

2001) which assists universities to help students in a more targeted way, for example through special tutorials. Especially the public sector is interested in students with good grades (Velasco et al., 2012) so this can also help universities in recruiting the best student assistants.

The database used in this study is the fifth starting cohort of the National Education Panel Study (NEPS), which is a broad German panel dataset containing almost 18,000 freshmen students of winter term 2010/11 and covering various aspects of students' academic and personal life (Blossfeld et al., 2011)¹. The grades of the German higher education system are in the range from 1.0 (the best possible grade) to 5.0 (failure), where 4.0 is the worst grade which is just enough to pass.

The study distinguishes between two different scenarios. In the first scenario dropout students are included in the predictions, which are students who finally leave the higher education system without a degree. In the second scenario, only students who earned a first higher education degree are of interest. Comparing the estimated regression coefficients in both scenarios the most interesting aspect is the question which coefficient estimates change dramatically. This would mean that the dropout students have a huge influence on the parameter estimate and the particular variable influence dropout decision and higher education performance in different ways.

Variables of two different points of students' academic careers were used for grade prediction. Firstly, only variables of the pre-university phase were used, which are, for example, demographic variables, migration, and information about secondary schooling and possible vocational training. The advantage of this approach is that we have prediction results at a very early stage just before the start

*Corresponding Author: Marco Giese: Chair of Statistics, University of Duisburg-Essen, 45117 Essen, Germany; Email: Marco.Giese@wiwinf.uni-due.de; Tel.: +49 201 18-33748

¹ This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:12.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg in cooperation with a nationwide network.

of the first semester. The disadvantage is that predictions are less accurate. This problem is mitigated in the second regression step where variables from the early study phase are also included, e.g. grades in the first exams, study satisfaction, working status etc. This improves the prediction accuracy at this stage.

From a statistical perspective, this approach leads to two major challenges. The first is caused by panel attrition which is a very common problem when analyzing survey data (Behr et al., 2005). To avoid misunderstandings, students who finally leave tertiary education without a degree, are designated as (study) dropouts, and students who finally leave the panel (panel attrition) are labeled panel leavers or attriters. It stands to reason that the probability to leave the panel depends on academic performance and satisfaction and further variables covering information about the interview process. Since this would lead to biased results in the ordinary least squares (OLS) estimation the Heckman estimation and the inverse probability weight estimator (IPW) shall reduce the bias (Little and Rubin, 2019).

The second major problem occurs in the scenario where dropout students are included in the study. This leads to a mixture of a distribution which is continuous in the range of 1.0 to 4.0 (the graduates) and discrete with a value of 5.0 for the dropouts. The problem can be solved by a general linear model (glm) with the Tweedie distribution as an exponential family which is a novel approach in the education context. The Tweedie glm was, for example, utilized in modeling the zero-catch problem in the fishing industry (Shono, 2008) where the distribution is also a mixture of a continuous and a discrete (in case no fish are caught) distribution.

This study is structured as follows. The second section gives a short overview of previous literature in the field of educational data mining with a focus on the prediction of study performance of higher education students. Furthermore, some aspects of panel attrition are discussed in this section, whereby the methodological aspects are discussed in section 4. Some more information about the dataset and a short discussion about the missing values is given in section 3. The results, including a comparison of different approaches on how to deal with the panel attrition problem, are presented in section 5. Section 6 discusses the results in the higher education context and concludes.

2 Related work

Dropout prediction

A large number of studies in the research field of educational data mining investigate students' dropout of the tertiary education system as performance indicator using various data mining techniques. Behr et al. (2020a) give a comprehensive literature review regarding dropout of higher education. Widely used methods for this binary classification problem are, among others, artificial neural networks (Rios et al., 2013; Jadrić et al., 2010), decision trees and/or random forests (Superby et al., 2006; Aulck et al., 2016; Baradwaj and Pal, 2011), logistic regression (Knowles, 2015) and support vector machines (Mayra and Mauricio, 2018). Although students' dropout and students' grades are strongly correlated, poor study performance is not the only reason to leave university without a degree. Blüthmann et al. (2012) find four different clusters for the dropout students. Only in the cluster "overwhelmed" is the main reason for dropping out the poor study performance. The students in this cluster mainly suffering from a lack of interest in the study field also reveal a poor study performance. There are mainly poor study performance reasons for leaving university. Nevertheless, the grade point average (GPA) is the strongest predictor for study dropout (Stinebrickner and Stinebrickner, 2014).

Grade prediction

The number of studies predicting the final grade at tertiary education is much smaller. The problem of just analyzing students' dropout is, that no distinction is made between excellent students graduating with honors and graduates earning the degree with poor grades that are just enough to pass the specific exam. A regression analysis with the final grade as dependent variable can be seen as a generalization of the binary dropout prediction since all graduates have grades from 1 to 4 and all dropouts get the final grade of 5. Strecht et al. (2015) compares different algorithms for the two problems 1) dropout/graduate (using classification methods) and 2) final grade (using regression methods) with a focus on model performance. The performance for the regression algorithms in terms of the root mean squared error (RMSE) was approximately equal to the classification analysis. Beck and Davidson (2001) find that the academic efficacy and apathy are the most relevant determinants to predict students' final GPA. Sherman (1979) predicts the mathematics performance at high school using linear regression, where the mathematics grade in the previous years is the best predictor.

Differences to other studies

This study stands out from the few studies trying to predict higher education grades because it follows two new approaches in this research field. On the one hand, it compares methods Heckman correction and IPW estimation) to correct the distortion of panel attrition. Some other studies in social sciences make use of these methods, e.g. Behr (2006), but, to the best of my knowledge, not in the field of higher education grade prediction. Many studies further ignore missing values in the data that are often a result of the attrition problem. Asendorpf et al. (2014) investigate 35 articles in the International Journal of Behavioral Development in the years 2012 and 2013 and find that 20% of the studies completely ignore the problem of missing data and further 26% use inadequate methods. On the other hand, the Tweedie distribution is used to achieve better modeling of the scores. Other studies use this approach to model the monthly rainfall (Hasan and Dunn, 2011), or the zero-catch problem in the fishing industry (Shono, 2008). As far as I know, this approach is innovative in the field of higher education research. This allows for improved comparison of the two regression models that include and exclude dropouts. Determinants that are mainly significant due to the inclusion of dropouts can be detected by this approach. The central aspect of this article is still on the new finding of relevant results in the research field of higher education. But since it needs advanced statistical methods for the reasons described above (which is probably the reason why there are so few studies trying to predict university grades) the methodology cannot be neglected and should also motivate other researchers to go beyond the statistical standard methods.

3 Survey dataset

3.1 The National Education Panel Study

The National Education Panel Study (NEPS) is a comprehensive German survey data set. This study uses the starting cohort 5 covering 17,910 freshman students of the winter term 2010/2011 enrolled at German higher education institutions and more than 3,000 variables of various aspects in students' life (Blossfeld et al., 2011). In December 2019 the fifth cohort comprised twelve waves. An overview of the waves, the term of the survey, the number of participants, temporal- and final dropouts is given in Table 1.

The dependent variable is the final grade of the first higher education degree, which is, in general, the Bachelor's degree, where 1.0 is the best possible grade and 4.0 the worst possible grade for graduates. Since one scenario also includes dropout students with a final grade of 5.0 it is essential to define dropout. As Tinto (1975) states, the definition can have a huge impact on the study results. Spady (1970) declares there are two general different dropout definitions. The first definition regards dropout from a micro perspective, meaning from a universities or faculties viewpoint. The second definition is from a macro perspective, i.e. dropouts are defined as students who never receive a degree from any higher education institution. Here, the second definition is used since the focus is not on a single faculty or university, but on the entire German higher education system. Furthermore, the data is well suited to use this definition which is generally not possible with administrative data. This dropout definition considers students who changed the institution or the study

Table 1: Participants and panel dropouts in the current scientific use file (SUF) (LifBi, 2017; Zinn, 2019, and own calculations). CATI: Computer assisted telephone interview, CAWI: Computer assisted web interview.

wave	instrument	partic. survey	temp. attrition	final attrition	survey term
1st	CATI	17,910	0	0	winter 10/11
2nd	CAWI	12,273	5,591	46	autumn 11
3rd	CATI	13,113	4,560	237	spring 12
4th	CAWI	11,202	6,424	284	autumn 12
5th	CATI	12,694	3,444	600	spring/summer 13
6th	CAWI	10,183	7,039	688	autumn 13
7th	CATI	9,547	7,161	1,139	summer 14
8th	CAWI	8,629	6,024	3,257	autumn 14
9th	CATI	10,096	4,321	3,493	spring/summer 15
10th	CATI	9,090	4,192	4,628	spring/summer 16
11th	CAWI	7,020	5,042	5,848	autumn 16
12th	CAWI	8,551	3,041	6,318	spring-autumn 17

program as graduates. In this case, the final grade of the subject where the student obtained the first degree was used. All relevant variables up to wave 12 were used to construct the status of a student (dropout, graduate, still studying, or the status is not available). The status variable is truncated on the right side after wave 12 which means that it is missing for students who are still studying after wave 12 (6 years/12 semesters after they start studying) and do not have a higher education degree (Fox, 2015). According to Heublein et al. (2008) 75% of the study programs in Germany have a standard period of six semesters and 25% seven or eight semesters. Only 3% of Bachelor dropouts leave their study program after the 10th semester (Heublein et al., 2017). The median study duration of Bachelor students in Germany was 7.6 semesters in 2018, including study interruptions (DESTATIS, 2019), so it can be expected that the number of students who are still studying after wave 12 without obtaining their first degree is small.

The explanatory variables used in this study were selected based on a prior descriptive analysis of the NEPS data (Behr et al., *ress*). The most relevant variables with sufficient data quality (not more than 50% missing values) were used in this study. These are variables that become relevant already before study, e.g. demographic variables (e.g. migration, age, gender), secondary education (e.g. final school grade, type of school), parental background, variables describing the phase immediately before study, e.g. is the student studying his subject of choice, what do parents and friends think about the study choice or was there an alternative to study. Finally, variables that are of importance during the study program, e.g. study satisfaction, study commitment, academic integration, off-study work, or financial situation were used to predict the final degree grade. The Tables A1, A2 and A3 in the appendix reveal a more detailed list of variables in the two episodes pre-university and early study phase.

Another educational panel study in Germany is the “Studienberechtigtenpanel” published by the German Centre for Higher Education and Science Research (DZHW) approximately every three years. This panel has, compared to the NEPS, a much smaller number of variables, contains only two waves and reveals a much larger num-

ber of panel leavers in wave 2 (Birkelbach et al., 2019). One of the largest survey datasets for educational research covering 15-year old students in OECD countries in 2018 is the Program for International Student Assessment (PISA) (Sellar and Lingard, 2014). In contrast to the NEPS PISA is a cross-sectional dataset whereas the NEPS has a panel structure.

3.2 Problems due to non-response and initial selection bias

From a target population of 31,082 freshmen students 13,172 students did not respond, which leads to 17,910 participants in the first wave (Zinn et al., 2017). These students, who did not respond, were not asked for participation in further waves and no information about them is available in the data. Design weights were introduced to overcome the initial bias due to nonresponse and different selection probabilities, e.g. women, students without migration background and students born in 1990 or later are overrepresented in the initial sample (LifBi, 2017).

The more severe problem of the data regards students who finally leave the panel before graduation or dropout. The extent of final and temporary panel leavers in each wave is displayed in Table 1. The contingency Table 2 reveals the students' status (graduation, dropout, continue studying, or the status is not available) and the panel attrition (whether a student finally left the panel up to wave 12). Even if the true frequencies for the dropouts are near the expected frequencies under the assumption of independence, this does not hold for the other three groups. The χ^2 -test of independence (Hartung et al., 2009) rejects the null with a p-value near zero. The group of students who are still studying and finally left the panel is comparably large in the data. This is mainly caused by students who finally left the panel before graduation or dropout but after wave two (since these students have no available status). Furthermore, one can see that the graduation rate in the sample ($9815/10,657 = 0.921$, if just dropouts and graduates are used) is above the graduation rate of 85.3% that

Table 2: Contingency table with students' status and panel attrition

attrition	status	dropout	graduate	still studying	status not available	Σ
no final attrition		545	8,832	2,215	0	11,592
final attrition		297	983	3,739	1299	6,318
Σ		842	9,815	5,954	1,299	17,910

Schnepf (2014) found for Germany using a similar dropout definition.

Since only observations, where the final grade is available, are useful for later regression models the final sample contains 8,727 observations in the situation where university dropouts are included in the study. This disregards all students with unavailable status, who are still studying and all graduates who did not state their final grade. It stands to reason that the probability to leave the panel also depends on the final grade. The point biserial correlation (Bortz and Schuster, 2010) between the grade and the binary attrition variable in the sample containing dropouts and graduates is 0.280 which indicates that the probability to leave the panel rises as grades worsen. But this is mainly caused by dropout students. Excluding the dropouts leads to a point biserial correlation of -0.004 which suggests that grades and panel attrition are uncorrelated. Therefore, in the later analyses both samples are regarded, graduates and dropouts ($n = 8,727$) and only graduates ($n = 7,884$). In the latter sample, the sample size is smaller and dropouts are simply ignored but the attrition bias might be smaller.

4 Methodological approach

This section describes the statistical methods needed for empirical analysis. As already described in the introduction there are two major statistical challenges in this study. The first problem covers panel attrition and the resulting missing values in the data. Most, but not all missing values are caused by panel attrition. Section 4.1 describes the three major types of missing values and the types of missing values in the NEPS. The following subsection 4.2 explains imputation strategies that fill in the missing values. The subsections 4.3 and 4.4 explain the two methods that reduce (or - if all assumptions met - eliminate) the bias in the parameter estimates which is caused by attrition. The following subsection 4.5 introduces the Tweedie distribution which is needed to handle the second major problem of a zero-inflated continuous distribution in the scenario where dropouts are included. Lastly, measures to evaluate the performances of the different models are introduced in 4.6. In order not to interrupt the flow of reading by too many formulas, methodological details are included in the appendix.

To compare the different strategies the ordinary least squares estimation (OLS) is used as a benchmark (see Appendix).

4.1 Types of missing data

In general, a distinction is made between three types of missing data (Little and Rubin, 2019; Fox, 2015).

(1) Data is missing completely at random (MCAR) if missing data appears randomly independent of the missing variables or any other study variables. MCAR rarely occurs in real data.

(2) Data is missing at random (MAR) if the missing mechanism is not completely random and depends on the observed data. But, conditioned on the observed data, the missing mechanism is independent of the missing data. Statistical methods to prove for MAR do not exist. For example, if students were asked for their actual grades at university, there will be more missing values for freshmen students because they did not write any exams. If the missing mechanism does not depend on the grade itself, the data is MAR.

(3) If the missing mechanism depends on the missing variable itself, the data is not missing at random (NMAR). To continue the example above, if the willingness to disclose the actual university grade depends on the grade itself, e.g. students with bad grades may be less willing to disclose their grades, then the data is NMAR. In this situation, the missing mechanism is nonignorable, since ignoring it would lead to biased results.

NEPS

The NEPS distinguishes between three broad classes of missing data (LifBi, 2017): a) Item nonresponse, e.g. refused answers or the participant does not know the answer. b) Not applicable, e.g. the variable was not included in a specific survey wave or the variable was filtered (e.g. men were not asked for pregnancy). c) Edition missings, i.e. for some (very special and for this analysis not relevant) variables a remote access is needed, otherwise, the variable is not available. Furthermore, category d) of missing values can be introduced which includes temporal and final panel leavers who are not contained in the NEPS data.

Type c) of missing values is not relevant for this analysis. The 46 cati and 35 cawi variables are deleted. The missing type b) is also of minor interest in this study, since just the survey waves where the specific variable was included were used. Type a) and especially type d) are more problematic because it stands to reason that these kind of missings are nonignorable, even if there is no statistical method to test that hypothesis without making special assumptions (Little and Rubin, 2019). Whereas missing type a) occurs rarely, i.e. in only 0.71% of all non-missing cati variables, type d) emerges frequently from wave 2 as dis-

played in Table 1. It will not be possible to completely eliminate the bias from the estimation since the assumptions of the following sections are very strict and might not be entirely fulfilled. Nevertheless, the bias should be reduced as far as it is possible.

Köhler et al. (2015) investigate the response behavior in competence tests in the NEPS starting cohorts 3, 4 and 6 and come to the result that the response probability is strongly related to the competence of a person but also other person-specific attributes are relevant. This indicates the importance of an adequate estimation of the response probabilities that are needed for the inverse probability weighting in section 4.4.

4.2 Imputation methods

Imputation methods complete the missing entries in a dataset and make the application of statistical standard methods possible (Fox, 2015). There are two general imputation strategies: 1) Single imputation where all missing values are completed once. The most simple imputation methods are mean or median imputation which usually reduce the variance of the imputed variable dramatically (Little and Rubin, 2019). 2) To mitigate the problem of single imputation, multiple imputation completes all missing entries D times, where each imputed value is drawn from the predictive distribution of the missing value given the observed values. This technique can reflect the uncertainty of the missing data on the costs of additional complexity and computation time (Fox, 2015).

This study uses predictive mean matching (PMM) as imputation technique, introduced by Rubin (1986), which has less stringent assumptions compared to some parametric imputation methods. It has the advantage that real values are sampled being from the same sample space as the original variable which makes it applicable to metric as well as ordinal or categorical variables. The basic idea of PMM is to find possible matching candidates for the missing values in the set of observed values by minimizing the distance of predictive regression values on the variable that is imputed. From these candidates, one value is randomly drawn. In contrast to other imputation techniques, the PMM uses linear regression not for directly imputing missing values, but for matching missing cases with the most similar observed cases (Van Buuren, 2018).

The step of PMM where random values are drawn makes it possible to repeat this step D times to generate different datasets which is known as multiple imputation (Van Buuren, 2018). Averaging the results leads to the combined estimate. Note that the variance of multiple imputa-

tion has a within and between component (Little and Rubin, 2019).

Whereas in the previous decade $D = 5$ imputations were the usual, Asendorpf et al. (2014) suggest using at least $D = 20$ imputations since the computation power increased rapidly.

Only the explanatory variables were imputed in this study. It follows that just the complete cases of the dependent variable are included in the OLS-model.

4.3 Heckman correction

To overcome the problem of self-selection, i.e. that the students with available final study grades are not representative for the whole population, Heckman (1976) suggested a two-step approach. In the first step the dichotomous response variable

$$R_i = \begin{cases} 1, & \text{if } Y_i \text{ is observed} \\ 0, & \text{else} \end{cases} \quad (1)$$

is defined. Via probit regression (Bishop, 2006) the probability of an observed grade given a set of variables $\mathbf{Z}$ is calculated: $P(R = 1|\mathbf{Z}) = \Phi(\mathbf{Z}\gamma)$. Here, γ is a regression parameter estimated by the probit model that is used to estimate the inverse mills ratio $\hat{\lambda}$ (see the Appendix for details) and Φ is the probability function of the Gaussian distribution. In the second step, the estimates $\hat{\lambda}$ are used as additional regression parameter to estimate the final grade (Fox, 2015). The matrix $\mathbf{Z}$ in the probit regression contains variables used to estimate the response (1 for respondents and 0 for non-respondents), γ is the parameter vector optimized by the model and δ_i is a normal distributed error term. The matrix $\mathbf{Z}$ can contain variables that are also in the design matrix $\mathbf{X}$ but there it also contains additional variables describing the response behavior but have no influence on the target variable y , e.g. information about the interviewer and the number of contact attempts. Table A3 gives an overview of these interview specific variables used in the study.

4.4 Weighting methods

Weighting is one of the most widely used methods when panel attrition induces a bias in the estimates in the common OLS model (Vandecasteele and Debels, 2007). The two steps of the inverse probability weighted estimator, described by Robins et al. (1995), are very simple and intuitive.

In the first step one estimates the response probabilities R_i using the variables in $\mathbf{Z}$, defined above for the Heckman correction, for example via logistic regression.

These are denoted with $\hat{\pi}_i = \hat{P}(R_i = 1 | \mathbf{Z}_i)$, $i = 1, \dots, n$ and the $n \times n$ matrix $\hat{\Pi} = \text{diag}(\hat{\pi})$ contains the vector $\hat{\pi}$ on the diagonal and all other entries are zero.

In the second step a weighted OLS estimation with $\mathbf{X}$ as exploratory and Y as dependent variable is conducted, where the inverse weights from first step are used to put more weight on participants with a large non-response probability:

$$\hat{\beta}_{IPW} = [\mathbf{X}' \hat{\Pi}^{-1} \mathbf{X}]^{-1} \mathbf{X}' \hat{\Pi}^{-1} Y. \quad (2)$$

The inverse probability weighted (IPW) estimator results in a consistent estimation of β if the response probabilities are known Robins et al. (1995). Therefore, it is essential to get unbiased estimates of the response probabilities in the first step.

4.5 Tweedie distribution

The final grade from 1 to 4 in the German higher education system is rounded down on one decimal. Indeed, the grade distribution would be continuous in the interval [1,4] if it were not rounded to one decimal place. A problem occurs if dropout students are included in the model with a 5.0 which leads to a semicontinuous distribution. Mixture models (Van Buuren, 2018) can handle such distributions where a discrete and a continuous part occurs. These are often used in zero-inflated models. Standard applications are the modeling of daily rainfall (many days without rain) (Hasan and Dunn, 2011) or for insurance companies the loss amount of individual policyholders in a certain period (Jørgensen and Paes De Souza, 1994). To transform the grade variable to a zero-inflated model the transformation

$$\tilde{y}_i = \sqrt{5} - \sqrt{y_i} \quad (3)$$

is used, where y_i is the original grade variable of the i -th student and $\tilde{y}_i$ the transformed grade. The square root is used because it best eliminates the skewness of the data.

The Tweedie distribution, introduced by Tweedie (1984), can overcome the problem of zero-inflation. It is a generalization of some other distributions, including the Gaussian distribution, but here the Poisson-Gamma distribution is of interest to model the zero-inflated data (Shono, 2008). If the random variable K is discrete Poisson-distributed and $Z_1, \dots, Z_K$ are independent, identical random variables following a Gamma distribution, a Poisson-

Gamma distributed variable Y can be written as

$$Y = \sum_{k=1}^K Z_k. \quad (4)$$

This leads to the zero-inflation in cases where $K = 0$. This distribution is used as exponential family in a generalized linear models (glm) (Fahrmeir and Tutz, 2013) to model the final grades including dropouts.

4.6 Model comparison

To evaluate the model performance on new, unseen observations the dataset is divided into training data to fit the model and test data (50% of the complete dataset in each group) (Hastie et al., 2009). The training and test sets are different samples for all of the $D = 20$ imputed datasets and the results were aggregated as explained in section 4.2. As evaluation measures, I used the $R^2 \in [0, 1]$, which quantifies the variance explained by the model, and the mean squared error (MSE) (Aggarwal, 2015), which measures the squared error between the predicted values $\hat{y}_i$ and the observed values

$$\text{MSE} = \sum_{i=1}^n (\hat{y}_i - y_i)^2 / n. \quad (5)$$

A good model should have a large R^2 and a small MSE, whereby the latter strongly depends on the variance of the dependent variable.

Furthermore, the parameter estimates are compared especially to the OLS model where the parameters are expected to be biased. Other studies, such as Behr (2006), conduct a bias analysis but this is only possible under strong assumptions which do not apply to the NEPS data.

Note that (1.) model performance and (2.) parameter estimation are two completely different topics. The IPW estimator and the Heckman correction mainly correct for the bias in the parameter estimates but they may also improve the model performance. In the two-step approach of the Heckman model, we have an additional explanatory variable (the inverse Mills-ratio) that might also improve the model performance. Since underrepresented students in the training data are also underrepresented in the test data it can be expected that the model performance gap between the OLS model and the two model correcting for attrition increases in favor of the latter two models if they were applied in real situations where no group of students is over- or underrepresented.

To eliminate implausible continual values larger than 4.0, all predicted grades larger than 4.0 were set to 5.0. This

also brings us closer to the true mixture distribution which is discrete for dropouts (grade 5.0) and continuous otherwise.

5 Empirical results

This section presents the empirical results to answer the two major research questions 1) How do parameter estimates differ when dropouts are included? 2) How far does the model improve when additional variables from a later point in time are added to the model?². To answer the second major research question this section is divided into two main parts. In section 5.1 just pre-university variables are used where the number of missing values in the explanatory variables is small and therefore the data is less sensitive to imputation. Adding additional variables of the early study phase, what is done in section 5.2, improves the model performance in terms of MSE and R^2 due to the additional information in the data. However, the added variables are not only from wave 1 but mainly from waves 2 and 3 where the data contains more missing values caused by panel attrition. This makes the model more sensitive for a potential bias caused by missing data and the prediction is only possible at a later time in study.

To find answers on the first major research question, four models were compared in the scenario including dropouts and three models when dropouts are excluded. In the latter case, the Tweedie glm model is missing since the problem a zero-inflated mixture distribution only applies for the scenario with dropouts included. One further important aspect of this article is an adequate handling of panel attrition and missing values in survey data and therefore two models (IPW and Heckman) are compared. While the Tweedie glm should mainly improve the model performance, the IPW and Heckman model should reduce the bias in the parameter estimates. The IPW estimator and the Heckman correction were also embedded in the Tweedie model. The OLS model serves as a benchmark model.

5.1 Pre-university variables

Here only pre-university variables are used as explanatory variables which means variables up to the end of secondary education or vocational training that have nothing

to do with higher education or the study decision process. An overview of the variables is given in the Tables A1, A2 and A3 in the appendix. The number of missing values in each variable (% NA) is calculated on basis of students ($n = 8,727$) where the degree grade is available including study dropouts, except for the degree grade itself where the percentage of missing values is calculated based on all 17,910 participants of wave 1.

Table 3 reveals the out-of-sample performance results of the four models in both scenarios. Note that the grades were retransformed to the original form for better interpretation, which is equally applicable to Table 5. The transformation in equation 3 was only used for better modeling properties. Since a general linear model with a Gaussian exponential family is nothing else as the usual OLS regression, the Tweedie model reveals the same results as the OLS model when dropouts are excluded.

The glm with the Tweedie distribution slightly outperforms the OLS benchmark model in terms of R^2 and MSE in the situation with dropouts. The best models regarding the model performance are the Heckman model and the IPW estimator which slightly outperform the two other models. The Heckman correction includes estimates of the inverse Mills ratio $\hat{\lambda}$ as an additional explanatory variable in the second step which has a significant influence as demonstrated in Table 4. The relatively small amount of variance explained by the models is caused by the fact that just pre-university variables were used but the dropout process and study performance are also affected by many study related variables as stated in section 5.2.

Figure 1 reveals the kernel density estimation of the true grades and the kernel density of the out-of-sample predictions of the four models. The distribution of the models is concentrated near the median of the true distribution. In the left panel, one can see that only the Heckman estimator accomplishes to predict a notable amount of university dropouts. It is just too early to get reasonable predictions of study performance. The variance of the predicted grades is much smaller when dropouts are excluded.

Table 4 shows the parameter estimates of the four models. The parameter estimates were averaged over the 20 imputations as in equation A3 the standard deviation was calculated following equation A4. All observations have been used for the regression since it is not necessary to hold out test data as in the performance analysis.

Note that in Table 4 as well as in Table 6 and 7 in the next section the transformed degree grade of equation 3 is used since back-transformation is not possible here. This means that a positive sign of the coefficient estimates mean better study grades if the value of the regressor is increasing.

² For all calculations the statistical software R version 3.6.1 was used (R Core Team, 2019)

The most important variables (measured by the lowest p -value) of the pre-university episode to predict the final degree grade are the overall grade at secondary school (better school grades generally lead to a better university grade), the school type (students who attended a general Gymnasium perform better at higher education), the gender (females perform better), the year of birth, the number of repeated classes in their school career (more repeated classes lead to worse university grades) and the final points in the school subject German (the better the results in German the better the university grades). The large importance of German grades at school is mainly caused by the fact that students who have good school grades in German tend to choose study fields like linguistics and cultural sciences more frequently. Students in these “soft” study fields have on average better grades than students of science, technology, engineering, mathematics (STEM)

fields (Heublein et al., 2017), which can also be found in this data. For the same reasons students who had Mathematics as advanced course at school in some models perform significantly worse. These students are more frequently enrolled in “hard” study fields like Engineering or Mathematics.

Comparing the various modeling strategies minor differences in the parameter estimates can be found. The IPW estimates for some variables slightly differ from the other estimation strategies. The Heckman correction does generally not change the estimation results dramatically. Even though, the coefficient of the inverse Mills ratio is significant. Its negative coefficient indicates that students with worse grades are more prone for panel attrition as it was expected.

Table 3: R^2 and MSE of the three different methods using only pre-university variables (standard errors over the 20 imputations in parenthesis)

Measure	Include dropouts	OLS	Tweedie	IPW	Heckman
MSE	yes	0.857 (0.022)	0.855 (0.021)	0.852 (0.022)	0.851 (0.021)
	no	0.197 (0.003)	0.197 (0.003)	0.199 (0.003)	0.194 (0.003)
R^2	yes	0.192 (0.012)	0.204 (0.016)	0.202 (0.021)	0.205 (0.023)
	no	0.169 (0.009)	0.169 (0.009)	0.198 (0.011)	0.183 (0.011)

Figure 1: Kernel density estimation of true grades and the out-of-sample predictions of the four models including dropout students (left panel) and excluding them (right panel) using only data of the pre-university phase.

5.2 Early study phase

In this section, 54 additional variables describing the early study phase were added to the pre-university variables to a total number of 80 explanatory variables modeling the first higher education degree grade. This includes the selected study field and type of the higher education institution, the average grade of the first higher education exams, early study satisfaction, academic and social integration, financial aspects, off-study work, study commitment, and the big five personality traits.

Table 5 illustrates the performance results of the four models regarding MSE and R^2 . Additional information on the first semesters at university improves the regression performances of all models by far especially in terms of the R^2 . Regarding the MSE the IPW estimator performs slightly worse than the other models but this model explains most of the variance. In the situation with zero-inflated distribution, where dropouts are included, the benchmark OLS model underperforms dramatically. In the situation without dropouts, all models reveal similar results since the usual OLS estimation is less problematic.

Figure 2 visualizes the out-of-sample predictions of all models by its kernel density estimation. In the left panel, where dropouts are included, one can see that the OLS estimator has massive problems to model the tails of the true distribution. Nevertheless, the other models also have problems in the prediction of dropouts with a grade of 5.0. If the interest is mainly on classification (dropout or gradu-

Table 5: R^2 and MSE of the three different methods using pre-university and early study phase variables (standard errors over the 20 imputations in parenthesis)

Measure	Include dropouts	OLS	Tweedie	IPW	Heckman
MSE	yes	0.705 (0.013)	0.703 (0.021)	0.731 (0.028)	0.693 (0.013)
	no	0.120 (0.003)	0.120 (0.003)	0.124 (0.003)	0.120 (0.003)
R^2	yes	0.415 (0.016)	0.550 (0.029)	0.593 (0.054)	0.550 (0.031)
	no	0.522 (0.022)	0.522 (0.022)	0.545 (0.025)	0.523 (0.022)

ate) and the specific grade estimation is not relevant, classification models are preferable to regression models. But even classification models tend to underestimate the proportion of the minority class (here the dropout students) by far. Behr et al. (2020b) used random forests to classify dropouts and graduates, whereby they adjusted the probability threshold to generate a larger number of classified dropouts.

The right panel of Figure 2 highlights that the IPW estimator is slightly left-shifted. The Heckman- and the OLS estimator are very similar in this situation.

Table 6 and 7 present the parameter estimates. The regression contains 90 explanatory variables since two character variables (region of origin and study field) were con-

Figure 2: Kernel density estimation of true grades and the three models including dropout students (left panel) and excluding them (right panel) using variables of pre-university and early study phase.

Variable	including dropouts						no dropouts							
	OLS		Tweedie glm		Tweedie IPW		Heckman		OLS		IPW		Heckman	
	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.
socint_instructors	-0.010	0.007	-0.017**	0.006	-0.032***	0.006	-0.026***	0.006	-0.002	0.004	0.001	0.004	-0.003	0.004
socint_students	0.005	0.004	0.014**	0.004	0.014**	0.004	0.011*	0.004	-0.007**	0.002	-0.007**	0.002	-0.007**	0.002
commit_necessary	-0.005	0.003	-0.010***	0.003	-0.011***	0.003	-0.011***	0.003	-0.001	0.002	-0.003*	0.002	-0.001	0.002
commit_enjoy	0.006	0.004	0.013***	0.004	0.008	0.004	0.015***	0.004	0.004	0.002	0.002	0.002	0.004	0.002
commit_identification	0.003	0.004	0.003	0.004	0.006	0.004	0.007	0.004	-0.005*	0.002	-0.004	0.002	-0.005*	0.002
job_semester	-0.001***	0.000	-0.001***	0.000	-0.001***	0.000	-0.002***	0.000	0.000	0.000	0.000	0.000	0.000	0.000
costs_direct	0.008*	0.003	0.011***	0.003	0.019***	0.003	0.010***	0.003	0.001	0.002	0.000	0.002	0.001	0.002
financialaid_bafog	-0.013*	0.005	-0.020***	0.006	-0.018**	0.006	-0.020***	0.006	-0.003	0.003	0.002	0.003	-0.002	0.003
change_field	0.015	0.011	0.027*	0.013	0.068***	0.014	0.021	0.013	0.009	0.006	0.005	0.006	0.009	0.006
satisf_enjoy	0.006*	0.003	0.006*	0.003	0.001	0.003	0.007**	0.003	0.004*	0.001	0.002	0.001	0.003*	0.001
satisf_whole	0.012***	0.002	0.019***	0.002	0.009***	0.002	0.016***	0.002	0.000	0.001	0.001	0.001	0.000	0.001
satisf_interesting	-0.005*	0.002	-0.01***	0.002	-0.005*	0.002	-0.011***	0.002	0.000	0.001	0.000	0.001	0.000	0.001
satisf_tired	0.001	0.001	0.001	0.001	0.005***	0.001	0.003	0.001	-0.001	0.001	-0.002**	0.001	-0.001	0.001
insttype	-0.054***	0.007	-0.065***	0.008	-0.058***	0.008	-0.032***	0.008	-0.035***	0.004	-0.031***	0.004	-0.030***	0.004
big5_extraversion	0.002	0.003	0.003	0.004	0.011**	0.004	0.004	0.003	-0.004*	0.002	-0.003	0.002	-0.004*	0.002
big5_agreeable	-0.002	0.005	-0.005	0.005	0.001	0.005	-0.006*	0.005	-0.004	0.003	-0.007**	0.003	-0.004	0.003
big5_conscientious	0.003	0.004	0.002	0.004	0.006	0.004	0.003	0.004	0.006**	0.002	0.005*	0.002	0.006**	0.002
big5_neuroticism	0.020***	0.004	0.031***	0.004	0.028***	0.004	0.029***	0.004	0.006**	0.002	0.007***	0.002	0.006**	0.002
arts	0.063**	0.022	0.071*	0.028	0.089**	0.028	0.075**	0.027	0.056***	0.012	0.076***	0.012	0.055***	0.012
mathematics and natural sciences	0.001	0.017	0.016	0.018	0.024	0.018	0.015	0.018	0.022*	0.009	0.029**	0.009	0.021*	0.009
linguistics and cultural sciences	0.013	0.017	0.015	0.018	0.012	0.018	0.021	0.018	0.021*	0.009	0.028**	0.009	0.021*	0.009
inv. Mills ratio	.	.	.	.	.	.	-0.175***	0.010	.	.	.	.	-0.337***	0.008

verted to dummies. To reduce the number of coefficients displayed in the two tables, only coefficients were the estimate is significant to the 5% level for at least two of the seven models or where it is significant for at least one model to the 0.1% level. The other variables listed in the appendix but not shown in one of the Tables 6 or 7 were used for the regression but do not have a significant parameter estimate for more than one model.

The coefficient stronger varying between the models including dropouts. Furthermore, coefficient estimates from variables of the pre-university episode slightly differ from the estimates in Table 4 caused by the inclusion of other correlated variables which are also significant. For example, the points in the school subject German were highly significant in Table 4 but after adding the subject groups this effect decreases in the situation without dropouts and disappears in the situation including dropouts.

The most important new variables from the early study phase are the grade point average after the first exams at university, the own performance evaluation of the students, the type of institution and neuroticism (students with more confidence generally perform better). For most of these variables it is obvious why they are important and these findings are already widely discussed in the literature. The grade point average is even included in the output variable if only by a small percentage of the final grade. Students at universities of applied sciences have better grades, which might be the result of a lower requirement level. The variable indicating study restrictions is significant in all models but this is mainly caused by the large correlation with the secondary school grade.

There are much less significant coefficients in the situation where dropouts are excluded. Students who state higher values of conscientiousness have significantly better grades only in the models where dropouts are excluded.

Students in arts, linguistics and cultural sciences get significantly better grades. This also applies to mathematics and natural sciences but only if dropouts are excluded because there are larger dropout rates that were also found by Heublein et al. (2012) caused by many exams with large failure rates at the beginning of the study program.

6 Discussion and conclusion

This analysis aims to estimate the final degree grade of the first higher education degree of German freshman students who first enrolled in the winter term 2010/11. The data used for the study comes from the National Education

Panel Study and contains in total 17,910 students and more than 3,000 variables.

Two different scenarios were analyzed in the study: 1) including study dropouts with a grade of 5.0 and 2) excluding study dropout in the regression models.

Furthermore, two sets of variables were used. The first set of variables only contains pre-university variables, which became relevant after the secondary education degree even before the final study decision process. In the second step variables of the early study phase were added to the models to investigate the model improvement in the second step when additional information is available.

A glm with the Tweedie distribution as exponential family is used to model the zero-inflation of the data if dropouts are included in the model. The predictive performance improves markedly when adding the additional variables of the early study phase from 0.204 to 0.550 in terms of R^2 when dropout students are included. The benefit of the pre-university model is that predictions at a very early point directly after secondary school graduation are possible at the expense of model performance. Behr et al. (2020b) found similar results in a binary dropout-graduate classification analysis.

When dropouts are included in the regression model the results regarding influencing variables are predominantly in line with the previous dropout literature presented at the beginning of section 2. Interestingly, significant coefficients in the scenario including dropouts are not significant if dropouts are excluded from the model. These are mainly parameters that were found to be significant in the dropout literature if the reason for dropping out is not performance-related. Consequential, these variables mainly influence the dropout process but have only a minor influence on the final grade of graduates.

In some situations even the sign of the estimated coefficient changes. This applies, for example, for the age of a student, where a rising age has a negative influence in the dropout literature (Sarclotti and Müller, 2011), but has a positive influence on the grade if the student does not drop out. Müller and Schneider (2013), Lassibille and Gómez (2009) and Montmarquette et al. (2001) also found a larger dropout probability for older students. The possible reasons are higher opportunity costs for older students who already have experience in the labor market and the increasing financial (and social, if they have children) pressure if they already have a family. However, if older students graduate they can profit from their higher life experience.

Other variables that are good predictors for dropout but not for performance are study alternative, enjoying the degree program, study satisfaction, direct study costs, or

the working hours during the semester. Many students are forced to work during the semester to raise the costs of their studies and already have a study alternative (their job). If they do not enjoy their degree program and are not satisfied they may leave the higher education institution despite good performance. Stinebrickner and Stinebrickner (2014) have an economic explanation for the dropout phenomenon. Students want to maximize their lifetime utility and if opportunity costs become too high or they expect only a minor increase of their salary with a higher education degree they frequently tend to leave the system without a degree.

The most important variables from the pre-university episode are the final grade at secondary school, the number of repeated classes in students' school career, the school type, and the age. In the first semesters at the higher education institution, especially the average grades of the first exams become relevant. Whether the student is studying at the institution of choice, has an alternative (e.g. vocational training) to the degree program, whether there are study restrictions, and the type of institution, university or university of applied sciences, are important determinants already before the start of the first semester. During the early study phase, also the study satisfaction, the match of study workload and curriculum plan, and a weakly developed neuroticism have a positive influence on the final degree grade.

The limitations of the study are mainly data-driven. As in most survey datasets in panel design, the NEPS data also suffers under panel attrition, which leads to an overrepresentation of well-performing graduates. The Heckman correction and the inverse probability weight estimation should correct for this problem to get (ideally) unbiased parameter estimates. Strongly related to (temporary) panel attrition is the problem of missing values in the explanatory variables. The MAR assumption that is made for imputation is presumably not fulfilled for all missing values. This can also influence parameter estimates especially if the relative number of missing values is large which is especially the case for some early study variables (see Tables A1, A2 and A3).

The models presented in this study can help higher education institutions to implement early warning systems for students at risk. In contrast to other early warning systems that are only based on dropout prediction, e.g. (Knowles, 2015), this system can also send a warning to the students if they fail to meet their performance targets (e.g. a specific grade they want to reach). Students themselves can get extra motivation if they get early feedback from their institution.

A more detailed dataset, for example combining survey data with administrative data, covering more detailed information about the credit points earned and grades in single exams, would lead to further improvement of the models.

While this study predicts the final grade of the first higher education study program, which is generally a Bachelor program, a further research question would be to predict the grades of a Master program using information from the Bachelor courses. Since dropout rates in Master programs are lower (Heublein et al., 2017) different results can be expected and the estimated parameters of models including and excluding dropout students will be less markedly different. When doing this with the NEPS data the right censoring problem arises since there is a considerable number of students actually still studying in the Master's program. Therefore also administrative data can be used where the problem of panel attrition does not exist but many "soft" variables as satisfaction are not available.

References

- Aggarwal, C. C. (2015). *Data Mining: The Textbook*, volume 1. Springer Science & Business Media.
- Asendorpf, J. B., van de Schoot, R., Denissen, J. J. J., and Hutteman, R. (2014). Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation. *International Journal of Behavioral Development*, 38(5):453–460.
- Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.
- Baradwaj, B. K. and Pal, S. (2011). Mining educational data to analyze students' performance. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 2(6):63–69.
- Beck, H. P. and Davidson, W. D. (2001). Establishing an early warning system: Predicting low grades in college students from survey of academic orientations scores. *Research in Higher Education*, 42(6):709–723.
- Behr, A. (2006). Comparing estimation strategies for income equations in the presence of panel attrition. *Jahrbücher für Nationalökonomie und Statistik*, 226(4):361–384.
- Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and determinants of panel attrition in the european community household panel. *European Sociological Review*, 21(5):489–512.
- Behr, A., Giese, M., Kamdjou, H. D. T., and Theune, K. (2020a). Dropping out of university: a literature review. *Review of Education*.
- Behr, A., Giese, M., Tegum, H. D., and Theune, K. (in press). Dropping out from higher education in germany - an empirical evaluation of determinants for bachelor students. *Open Education Studies*.
- Behr, A., Giese, M., Theune, K., et al. (2020b). Early prediction of university dropouts—a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, 1(ahead-of-print).
- Birkelbach, R., Vietgen, S., and Wallis, M. (2019). *DZHW-Studienberechtigtenpanel 2012*. Deutsches Zentrum für Hochschul-

- und Wissenschaftsforschung (DZHW), Hannover, Germany.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (2011). Education as a Lifelong Process—The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft [Special Issue]*, 14.
- Blüthmann, I., Lepa, S., and Thiel, F. (2012). Überfordert, enttäuscht, verwöhlt oder strategisch? eine typologie vorzeitig exmatrikulierter bachelorstudierender. *Zeitschrift für Pädagogik*, 58(1):89–108.
- Bortz, J. and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien*. Springer-Lehrbuch. Springer Berlin Heidelberg.
- DESTATIS (2019). Bildung und kultur - prüfungen an hochschulen. [https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420187004.pdf?__blob=publicationFile\(08.01.2020\)](https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420187004.pdf?__blob=publicationFile(08.01.2020)).
- Fahrmeir, L. and Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.
- Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.
- Garciaarena, U. and Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89:52–65.
- Hartung, J., Elpelt, B., and Klösgener, K. (2009). *Statistik: Lehr- und Handbuch der angewandten Statistik ; [mit zahlreichen durchgerechneten Beispielen]*. Oldenbourg.
- Hasan, M. M. and Dunn, P. K. (2011). Two tweedie distributions that are near-optimal for modelling monthly rainfall in australia. *International Journal of Climatology*, 31(9):1389–1397.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. pages 475–492.
- Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J., and Woisch, A. (2017). Zwischen Studieneurwartungen und Studienwirklichkeit. *Forum Hochschule* 1/2017.
- Heublein, U., Schmelzer, R., Sommer, D., and Wank, J. (2008). Die entwicklung der studienabbruch- und schwundquotenquoten an den deutschen universitäten und fachhochschulen. *HIS-Projektbericht. Hannover*.
- Heublein, U., Schmelzer, R., Sommer, D., and Wank, J. (2012). Die entwicklung der schwund- und studienabbruchquoten an den deutschen hochschulen. In *HIS: Forum Hochschule*, volume 3, page 2012.
- Jadrić, M., Garača, Ž., and Čukušić, M. (2010). Student dropout analysis with application of data mining methods. *Management: Journal of Contemporary Management Issues*, 15(1):31–46.
- Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.
- Kleinke, K., Reinecke, J., Salfrán, D., and Spiess, M. (2020). *Applied Multiple Imputation: Advantages, Pitfalls, New Developments and Applications in R*. Springer Nature.
- Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. *JEDM-Journal of Educational Data Mining*, 7(3):18–67.
- Köhler, C., Pohl, S., and Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling*, 57(4):499.
- Lassibille, G. and Gómez, M. L. N. (2009). Tracking students' progress through the spanish university school sector. *Higher Education*, 58(6):821–839.
- LifBi (2017). Startkohorte 5: Studierende (SC5) - Studienübersicht Wellen 1 bis 9. Technical report, Leibniz Institut für Bildungsverläufe e.V.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Mayra, A. and Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in ecuador. In *Global Engineering Education Conference (EDUCON), 2018 IEEE*, pages 1238–1242. IEEE.
- Montmarquette, C., Mahseredjian, S., and Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of Education Review*, 20(5):475–484.
- Müller, S. and Schneider, T. (2013). Educational pathways and dropout from higher education in germany. *Longitudinal and Life Course Studies*, 4(3):218–241.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rios, G. et al. (2013). Predicting early students with high risk to drop out of university using a neural network-based approach. In *ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology*.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.
- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.
- Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 8(Jul):1623–1657.
- Saracchetti, A. and Müller, S. (2011). Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien. *Zeitschrift für Bildungsforschung*, 1(3):235–248.
- Schnepf, S. V. (2014). Do tertiary dropout students really not succeed in european labour markets? *IZA Discussion Paper No. 8015*.
- Sellar, S. and Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. *British Educational Research Journal*, 40(6):917–936.
- Sherman, J. (1979). Predicting mathematics performance in high school girls and boys. *Journal of Educational Psychology*, 71(2):242.
- Shono, H. (2008). Application of the tweedie distribution to zero-catch data in cpue analysis. *Fisheries Research*, 93(1-2):154–162.
- Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85.
- Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3):601–644.
- Strech, P., Cruz, L., Soares, C., Mendes-Moreira, J., and Abreu, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. *International Educational Data Mining Society*.

- Superby, J.-F., Vandamme, J., and Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining*, volume 32, page 234.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125.
- Tweedie, M. C. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.
- van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.
- Vandecasteele, L. and Debels, A. (2007). Attrition in panel data: the effectiveness of weighting. *European Sociological Review*, 23(1):81–97.
- Velasco, M. S. et al. (2012). More than just good grades: candidates' perceptions about the skills and attributes employers seek in new graduates. *Journal of Business Economics and Management*, 13(3):499–517.
- Vink, G., Frank, L. E., Pannekoek, J., and Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90.
- Zinn, S. (2019). Samples, weights, and nonresponse: Neps starting cohort 5 – first-year students - from higher education to the labor market (wave 12). Technical report, Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.
- Zinn, S., Steinhauer, H. W., and Aßmann, C. (2017). Samples, Weights, and Nonresponse: the Student Sample of the National Educational Panel Study (Wave 1 to 8) (NEPS Survey Paper No. 18). Technical report, Leibniz Institut für Bildungsverläufe e.V.

Appendix

OLS regression

In the OLS regression one is interested to get an estimate of the parameter vector β in the following equation

$$Y = \mathbf{X}\beta + \varepsilon \quad (\text{A1})$$

where $\mathbf{X} \sim (n \times k)$ denotes the design matrix containing the explanatory variables with n observations and k variables and $Y \sim (n \times 1)$ is the dependent variable containing the final grades for the first higher education degree and $\varepsilon \sim (n \times 1)$ is an error term (Hastie et al., 2009). The best estimate $\hat{\beta}$ is calculated as follows

$$\hat{\beta} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'Y. \quad (\text{A2})$$

Details of the imputation methods

Predictive mean matching (PMM)

Let $x_i, i = 1, \dots, k$ be one specific variable of the design matrix $\mathbf{X}$ containing missing values that should be imputed and x_{-i} be the other variables of $\mathbf{X}$ without x_i . Vink et al. (2014) suggest to split the PMM algorithm into the following steps:

1. Produce a linear regression, where the observed values x_i^{obs} with the regressor variables x_{-i}^{obs} are used to estimate a coefficient vector $\hat{\beta}$.
2. Randomly draw a coefficient vector β^* from the distribution of $\hat{\beta}$. The vector β is multinomial normal distributed with mean $\hat{\beta}$ and the empirical covariance estimation $Cov(\hat{\beta}) = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}[(x_{-i}^{obs})'(x_{-i}^{obs})]$ with regression residuals $\hat{\varepsilon}$.
3. Generate predictions with the coefficient vector β^* for observed (x_i^{obs}) as well as for missing values (x_i^{mis}).
4. For each estimated missing value $x_{j,i}^{mis}, j = 1, \dots, m_i$ (where m_i indicates the number of missing values in variable i) calculate the distances to the observed values $\Delta_j = |x_{j,i}^{mis} - x_i^{obs}|$.
5. Take one random value of the d closest distances of Δ_j and insert the corresponding x_i^{obs} as imputed value for $x_{j,i}^{mis}$. Van Buuren (2018) suggests a value of $d \in \{5, 6, \dots, 10\}$ if the dataset is not too small (less than 100 observations). Since the number of observations is relative large $d = 10$ is a good choice, even though the difference is negligible.
6. Follow the previous steps for all variables with missing values and repeat this to generate D complete datasets.

Multiple imputation

In multiple imputation D different datasets without any missing values are generated. After imputation, statistical standard methods can be applied to the complete data sets. Let $\hat{\theta}_d$ denote a point estimate and V_d the variance of a single imputed dataset $d, d = 1, \dots, D$. The calculation of the combined estimate is done by simply averaging

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d. \quad (\text{A3})$$

The variance of A3 has a within-component $\bar{W}_D = \frac{1}{D} \sum_{d=1}^D V_d$ averaging the D single imputation variances, and a between-imputation component $B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2$. The total variance of the combined estimate $\bar{\theta}_D$ is

$$V_D^{total} = \bar{W}_D + \frac{D+1}{D} B_D. \quad (\text{A4})$$

MICE-algorithm

As one can see in the first step of the PMM-algorithm, the method is constructed for univariate imputation, which means that only the imputed variable has missing entries. Otherwise, the linear regression would not be possible since linear regression cannot handle missing data. Multivariate imputation by chained equation (MICE), described by Van Buuren (2018), uses PMM as base model for imputation. Any other univariate imputation model can also be used. MICE is an iterative algorithm where random samples from the observed values are used in the first iteration to make use of PMM which is applied to every variable with missing values. The missing values are updated in every iteration of the MICE-algorithm. Usually, five iterations are sufficient, which is also the default in the R-package MICE (van Buuren and Groothuis-Oudshoorn, 2011).

The best imputation method is a widely discussed topic in the literature. Saar-Tsechansky and Provost (2007) and Garcarena and Santana (2017) find different methods and strategies of dealing with missing data leading to the best out-of-sample performance results depending on the analyzed dataset. Several imputation strategies, like random forest imputation, hot deck imputation and pmm have been compared. The best out-of-sample R^2 was found for the pmm-method. Trivial methods like mean or median imputation can just be used for single imputation and have the disadvantage that the variance in the imputed variables is severely underestimated with consequences on confidence intervals and statistical tests (Kleinke et al., 2020).

Heckman correction

The mathematical details of the Heckman correction model are illustrated in Fox (2015). The first step is a regression on the latent response variable ξ which illustrates the observed values of Y :

$$\xi_i = \mathbf{X}_i\beta + \varepsilon_i. \quad (\text{A5})$$

In a second step the variable ϑ describes whether ξ is observed or not:

$$\vartheta_i = \mathbf{Z}_i\gamma + \delta_i. \quad (\text{A6})$$

We just observe the variable

$$Y_i = \begin{cases} \text{not available, if } \vartheta_i \leq 0 \\ \xi_i, \text{ if } \vartheta_i > 0. \end{cases} \quad (\text{A7})$$

The errors are assumed to be bivariate normal with mean zero, variances $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$, $\text{Var}(\delta) = 1$ and correlation $\text{Cor}(\varepsilon, \delta) = \rho_{\varepsilon\delta}$. The expected value of Y_i given that Y_i is observed is

$$\begin{aligned} E(Y_i | \vartheta_i > 0) &= \mathbf{X}_i\beta + E(\varepsilon_i | \vartheta_i > 0) \\ &= \mathbf{X}_i\beta + \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\mathbf{Z}_i\gamma), \end{aligned} \quad (\text{A8})$$

where $\lambda_i \equiv m(-\mathbf{Z}_i\gamma) = \phi(-\mathbf{Z}_i\gamma) / [1 - \Phi(-\mathbf{Z}_i\gamma)] = \phi(\mathbf{Z}_i\gamma) / [\Phi(\mathbf{Z}_i\gamma)]$ is the inverse Mills ratio. In a regression model where Y is only regressed on the design matrix $\mathbf{X}$ the additional effect λ is omitted which leads to biased estimates if the coefficient $\beta_\lambda \neq 0$ in the Heckman correction model in the following equation

$$(Y_i | \vartheta_i > 0) = \mathbf{X}_i\beta + \beta_\lambda \lambda_i + v_i. \quad (\text{A9})$$

Via probit regression we get estimates $\hat{\gamma}$ as defined in section 4.3. These are used to get estimates $\hat{\lambda}_i = \phi(\mathbf{Z}_i\hat{\gamma}) / [\Phi(\mathbf{Z}_i\hat{\gamma})]$. In the second step model A9 can be estimated via OLS where the estimates $\hat{\lambda}$ are used instead of the true λ .

The main criticism of the model is that the estimation is inconsistent if the assumption of joint normality of the error terms (ε, δ) is not fulfilled.

Tweedie distribution

Since the assumption of Gaussian distributed error terms in the OLS model is strongly violated for the data in the scenario where dropouts are included and the distribution of the modeled variable is not completely continuous generalized linear models (glm) are briefly introduced. More detailed information is in Fahrmeir and Tutz (2013). These models have three main components:

1. The exponential family, which is the Gaussian distribution in the standard OLS model.
2. The linear prediction $\eta = \mathbf{X}\beta$ which is also well known from the OLS model.
3. The link function g so that $\eta = g(\mu)$, where $\mu = E(\mathbf{Y}|\mathbf{X})$. The link function can also be nonlinear in glm's. In the OLS model g is just the identity function.

The family of Tweedie distributions are special cases of exponential dispersion models, which are a generalization of the exponential family. Therefore, the Tweedie distribution can be used in glm's (Shono, 2008). The density function of a Tweedie distributed random variable $\mathbf{Y}$ can be written as

$$f_{\mu, \sigma^2, p}(y) = a_{\sigma^2, p}(y) \exp\left[-\frac{1}{2\sigma^2} d_{\mu, p}(y)\right], \quad (\text{A10})$$

where μ is the location parameter with $E(\mathbf{Y}) = \mu$, σ^2 is the dispersion parameter and p is the power parameter with $\text{Var}(\mathbf{Y}) = \sigma^2 \mu^p$. For a power parameter p of 1, 2 and 3 one gets the Poisson, Gamma and the inverse Normal distribution, respectively. In this situation, with a zero-inflated continuous distribution one should select $p \in (1, 2)$ which leads to a compound Poisson-Gamma distribution. The parameter p is tuned via a 5-fold cross-validation in the set of values $p \in \{0, 1, 1.1, 1.2, \dots, 1.9, 2, 3\}$. The Poisson-Gamma distribution combines a discrete Poisson distributed random variable $K \sim \text{pois}(\frac{\mu^{2-p}}{(2-p)\sigma^2})$ and iid random variables $Z_1, \dots, Z_K \sim \Gamma(\frac{2-p}{p-1}, \frac{\mu^{1-p}}{(p-1)\sigma^2})$ to the mixture distribution of equation 4.

The model comparison between the Tweedie distribution model and other models can be difficult and is usually limited to the comparison of the predictive results, as explained in section 4.6. Another disadvantage of the Tweedie model is that quasi-likelihood estimation is used to transform the Tweedie distribution to an exponential family. This impedes the calculation of widely used information criterions such as the Akaike information criterion or the Bayesian information criterion.

The tuned power parameter of the Tweedie distribution is $p = 1.6$ in the scenario with dropouts, which is a Poisson-Gamma distribution, and $p = 0$ when dropouts are excluded from the prediction, which is the Gaussian distribution.

Table A1: Description of pre-university variables.

Name	Description	Values and Notes	Wave	% NA
genstat	Generation status in Germany	1.0 to 4.0 (no immigration background)	1	0.000
rep_class	Number of repeated school classes	0 to 4	1	0.000
familylife	Living together with parents up to the age of 14	1 = with both parents, 0 else	1	0.023
school_type	Type of secondary school attended	1 = upper secondary education, 0 = other	spell	2.189
qualif_max	School-leaving qualification obtained	2 = general university entrance qualification, 1 = university of applied science qualification, 0 = other	spell	0.046
grade_school	Overall grade on final school certificate	1.0 (excellent) to 5.0 (poor)	1	1.765
math_points	Points in Mathematics in the last school year	0 (poor) to 15 (excellent)	spell	30.675
ger_points	Points in German in the last school year	0 (poor) to 15 (excellent)	spell	30.159
exam_german	German as examination subject for graduation	0 = no, 1 = yes	spell	6.337
exam_adv_german	German as first examination subject	0 = no, 1 = yes	spell	5.901
exam_maths	Mathematics as examination subject for graduation	0 = no, 1 = yes	spell	6.234
exam_adv_maths	Mathematics as first examination subject	0 = no, 1 = yes	spell	5.855
birthyear	Birth year of the target person	1950 to 1994	1	0.000
gender	Gender of the target person	0 = female, 1 = male	1	0.000
mother_qualif	CASMIN: mother's highest general school leaving qualification	0 to 8 (highest)	1	0.630
father_qualif	CASMIN: father's highest general school leaving qualification	0 to 8 (highest)	1	2.292
mother_job	ISEI-08: mother's occupation	11.74 to 88.96	1	26.366
father_job	ISEI-08: father's occupation	11.56 to 88.96	1	21.061
voctrain	Successfully completed vocational trainings before study	0 = no, 1 = yes	spell	0.000
fail_prestudy	Number of dropouts from other trainings before study	0 = no, 1 = yes	spell	0.000
mother_alive	Mother alive	0 = no, 1 = yes	1	0.046
fath_alive	Father alive	0 = no, 1 = yes	1	1.111
bilingual	Grew up bilingual	0 = no, 1 = yes	1	0.000
lang_ger	Speak German fluently	0 = no, 1 = yes	1	0.000
household	Number of household members	0 to 20	1	0.057
country	Region of origin	Germany, Western Europe, Eastern Europe, Afrika, Asia, America, Australia	1	0.011
Dependent variable				
degree_grade	Final degree grade in the first higher education degree	1 (excellent) to 4/5 (poor/dropout)	all waves	51.273

Table A2: Description of study related variables (first part).

Name	Description	Values and Notes	Wave	% NA
gpa_cur	Current grade point average	1.0 (excellent) to 4.0 (poor)	2	24.155
fieldofchoice	Studying the field of first choice	0 = no, 1 = yes	1	18.345
instinfochoice	Studying at the institute of higher education of choice	0 = no, 1 = yes	1	14.151
study_alternative	Would you have started something else instead of a degree	0 = no, 1 = yes	1	14.335
study_judge_parent	Parents' opinion concerning study/field	1 to 5 (strong support)	1	14.060
study_judge_friend	Friends' opinion concerning study/field	1 to 5 (strong support)	1	14.048
study_restrict	Study with admission restrictions or selection procedure?	0 = no, 1 = yes	1	12.513
partic_people	Participation in university events aimed getting to know people	0 = no, 1 = yes	1	12.937
partic_organ	Participation in university events on study organization	0 = no, 1 = yes	1	13.682
partic_facil	Participation in university events with respect to central facilities	0 = no, 1 = yes	1	14.461
partic_acadskills	Participation in university events on academic skills	0 = no, 1 = yes	1	14.610
partic_course	Participation in university events on bridging courses	0 = no, 1 = yes	1	13.601
preparation	General preparation for study	0 (poor) to 4 (perfect)	2	22.917
skills_prep	Necessary knowledge acquired for subject field	1 (poor) to 4 (perfect)	2	23.261
workload_match	Matching study workload to curriculum plan?	1 (much less) to 5 (much more)	2	26.492
performance_eval	Satisfaction with academic performances till yet	1 (low) to 4 (high)	2	25.438
selfconcept	Academic self-concept (talent, learning new material etc.)	1 (low) to 7 (high)	2	26.263
study_informed	Informed about study	1 (poor) to 5 (good)	1	0.286
socint_instructors	Social integration: acceptance by instructors	1 (poor) to 4 (good)	2	25.450
socint_students	Social integration: contacts with students etc.	1 (poor) to 4 (good)	2	25.576
commit_necessary	Commitment to degree course: do no more than necessary	1 (does not apply) to 5 (applies completely)	2	25.656
commit_enjoy	Enjoyment of degree program	1 (does not apply) to 5 (applies completely)	2	25.839
commit_demands	High demands on self	1 (does not apply) to 5 (applies completely)	2	25.931
commit_energy	Invest a lot of energy for academic success	1 (does not apply) to 5 (applies completely)	2	25.828
commit_identification	Identification with degree program	1 (does not apply) to 5 (applies completely)	2	26.080
helplessness	Study-specific helplessness	1 (does not apply) to 5 (applies completely)	2	26.309
job_semester	Time spent in a week for employment during semester time	0 to 60 hours	2	24.682
study_semester	Time spent in study-oriented activities during semester	0 to 142 hours	2	24.648
job_break	Time spent in a week for employment during semester break	0 to 99 hours	2	24.728
study_break	Time spent in study-oriented activities during semester break	0 to 99 hours	2	24.739
costs_direct	Difficulty to pay direct costs of higher education	1 (difficult) to 5 (easy)	1	0.321
costs_opportunity	limitation of the possibilities to earn own money	1 (not at all) to 5 (a lot)	1	0.401
financialaid_bafoeg	Currently receive student financial aid (BAfoeG)	0 = no, 1 = yes	2	46.884
funding	Total amount of financial resources per month	0 to 10899 Euro	2	26.871

Table A3: Description of study related variables (second part) and interviewer variables.

Name	Description	Values and Notes	Wave	% NA
change_field	Have changed at least once the subject field	0 = no, 1 = yes	spell	0.000
change_uni	Have changed at least once the institution	0 = no, 1 = yes	spell	0.000
change_degree	Have changed at least once the type of degree (e.g. Bachelor instead of state examination)	0 = no, 1 = yes	spell	0.000
satisf Enjoy	Study satisfaction: enjoy degree course	0 (no) to 10 (absolutely)	3	16.501
satisf_conditions	Wishing better study conditions	0 (no) to 10 (absolutely)	3	16.523
satisf_match	Degree course and other obligations are hard to match	0 (no) to 10 (absolutely)	3	16.512
satisf_whole	On the whole, satisfied with actual studies	0 (no) to 10 (absolutely)	3	16.501
satisf_frustrating	External circumstances are frustrating	0 (no) to 10 (absolutely)	3	16.741
satisf_kill	Degree course kills me	0 (no) to 10 (absolutely)	3	16.523
satisf_interesting	Degree course is interesting	0 (no) to 10 (absolutely)	3	16.501
satisf_concerns	Concerns of students are not taken into account	0 (no) to 10 (absolutely)	3	16.695
satisf_tired	Often tired due to degree course	0 (no) to 10 (absolutely)	3	16.501
insttype	Type of institution	1 = university of applied sciences, 2 = university	spell	0.023
nontrad	Non traditional student	0 = no, 1 = yes	2	1.134
big5_extraversion	Big five personality traits: Extraversion	1 (outgoing) to 5 (reserved)	3	13.074
big5_agreeable	Agreeableness	1 (challenging) to 5 (friendly)	3	13.029
big5_conscientious	Conscientiousness	1 (efficient) to 5 (easy going)	3	13.017
big5_neuroticism	Neuroticism	1 (nervous) to 5 (confident)	3	13.017
big5_openness	Openness to experience	1 (curious) to 5 (cautious)	3	13.017
field	First field of study	engineering, mathematics and natural sciences law, economics and social sciences, linguistics and cultural sciences, arts, medicine, other minor subject groups	1	0.321
Interview specific variables to calculate the attrition probability				
nr_contact	Number of contact tries	0 to 129	all waves	0.000
minutes	Minutes of interview	13 to 136	all waves	0.000
int_gender	Gender of the interviewer	0 = female, 1 = male	all waves	0.000
int_age	Age of the interviewer	4 levels from 1 (young) to 4 (old)	all waves	0.000
int_school	Highest school degree of the interviewer	1 (no degree) to 4 (Gymnasium)	all waves	0.000
int_problems	Problems during the interview	1 = no problems, 0 = problems	all waves	0.000