Friend Request Pending: A Comparative Assessment of Engineering- and Social Science–Inspired Approaches to Analyzing Complex Birth Cohort Survey Data

Claudia V. Roberts1

Abstract
The Fragile Families Challenge is a mass collaboration social science data challenge whose aim is to learn how various early childhood variables predict the long-term outcomes of children. The author describes a two-step approach to the Fragile Families Challenge. In step 1, a variety of fully automated approaches are used to predict child academic achievement. In total 124 models are fit, which involve most possible combinations of eight model types, two imputation strategies, two standardization approaches, and two automatic variable selection techniques using two different thresholds. Then, in step 2, an attempt is made to improve on the results from step 1 with manual variable selection on the basis of a detailed review of the codebooks. A total of 3,694 variables believed to be predictive of academic achievement, using a comprehensive review of student success literature to guide the decision-making process, were manually selected. The best models from step 1 were reestimated using the manually selected variables. The results show that manual variable selection improved the majority of the top 10 models in step 1 but did not improve the best of the top 10. The results indicate that variable selection inspired by social science methodologies can, in most cases, significantly improve models trained completely automatically.

Keywords
Fragile Families Challenge, survey research, social science data, predictive modeling, grade point average, fragile families, data analysis, machine learning, longitudinal survey, social science

The Fragile Families and Child Wellbeing Study (FFCWS) is a longitudinal, birth cohort study run by researchers at Princeton University and Columbia University (The Trustees of Princeton University 2018). The study follows a group of nearly 5,000 American children born between 1998 and 2000 and includes a large oversample of children born to unmarried parents (Reichman et al. 2001). The aim of the study is to characterize the relationships and conditions of unmarried parents and to study the cognitive development, mental and physical health, and social relationships of children born into such families.

The Fragile Families Challenge (FFC) is a mass collaboration social science data challenge designed to harness the predictive power of the FFCWS data set (Salganik et al. 2019). The FFC invites community members to use the data to build models that best predict six key outcomes: grade point average (GPA), grit, material hardship, eviction, job loss, and job training. In this article, we focus on predicting GPA only. It is our personal belief that a child’s GPA is very important, as it sets the tone for the rest of a child’s life and influences the range of opportunities afforded to the child (e.g., college acceptance, scholarships, admittance into competitive summer enrichment programs).

Out-of-the-box machine learning libraries such as scikit-learn and access to open data sets hosted on popular platforms such as Kaggle enable users from across the globe to create sophisticated predictive models with sometimes impressive predictive accuracy without ever needing to understand the underlying data (Kaggle 2017; Pedregosa et al. 2011). This is in stark contrast to traditional methods of
predictive modeling and data analysis undertaken by researchers in fields other than engineering, specifically the social sciences. In survey research, a popular measurement technique used in applied social research, the data are often very complex (Trochim 2006). They can span many years, in the case of longitudinal studies, and are susceptible to various sources of error: coverage error, sampling error, nonresponse error, and measurement error (Visser, Krosnick, and Lavrakas 2000). Thus, best practices in survey research call researchers to spend substantial time with the data—to “make friends with their data”—and to refrain from “throwing their data into a computer and trying to analyze it in minutes” (Wright 2003). Failure to do so could lead to spurious results and misleading conclusions, and researchers run the risk of misidentifying associations as statistically significant (Kelley et al. 2003).

McFarland, Lewis, and Goldberg (2016) argued that while sociologists are driven by theory and the desire to explain the patterns observed in the data, engineers are focused on creating algorithmic tools to increase the predictive accuracy of their models, without placing much importance on the explanation. But what if the only metric of success is predictive accuracy? To what extent would an engineer be rewarded for “befriending” the data? Using the FFC as our backdrop, we seek to answer whether engineers get better predictive results when they spend a little time learning the domain they are working in and, if so, how much better are these results.

In this article, we use the term variables to refer to survey questions in the codebook. The cookbook survey questions are our independent variables. We use the term outcome to refer to the dependent variable we are attempting to predict, which in this case is GPA. “Fitting” or “estimating a model” is the process by which we learn a mathematical relationship between a set of variables x and the dependent variable y. The term sample refers to a single observation or data point in the data set, which in this case is a child.

We divided the project into 2 steps. In step 1, we used a completely automatic approach that does not consider the data (the norm in data mining) to fit 124 models for GPA prediction. In step 2, we attempt to improve upon our results. We use a strategy that combines engineering-centric statistical analysis techniques with classical, more manual social science methodologies: we examine each variable in the codebook, manually selecting the ones believed to be predictive of academic achievement on the basis of a nonexpert reading of domain-specific research. Results indicate that in most cases, it pays for engineers to “make friends” with the FFCWS codebooks. We were able to improve the predictive accuracy of 6 of the 10 top step 1 models, of which 4 saw significant improvements. However, manual variable selection did not improve the predictive ability of the 2 most accurate models from step 1.

The article is structured as follows. First we describe the procedures used to create the initial set of 124 models. We then describe the process of creating the 15 manually curated variable sets. Next we present the results; we show that we were able to improve the predictive ability of almost all the models and demonstrate the effect of each variable subset on the models. We then look at the variables that most predict GPA as identified by the two most accurate models from this project. Finally, we end the discussion with closing remarks. Additional supporting materials can be found in the Appendix, available online.

### Step 1: Automatic Variable Selection

The goal of step 1 was to fit a model that could predict year 15 GPA as accurately as possible using a purely automated approach.

### Data Preprocessing

With 2,121 samples and 12,942 variables, the FFC data set is a high-dimensional data set. In settings in which the number of variables far exceeds the number of samples, overfitting becomes a problem, and the learned model loses its ability to generalize (Hua et al. 2005). Thus, it is important to preprocess the data to not only reduce the number of variables but also handle missing values and standardize the data.

We tried many different approaches to data preprocessing. We tried almost all combinations of four different decisions: two types of automatic variable selection ($F$ test and mutual information) using two thresholds (10 percent and 20 percent), two types of imputation strategies (median and mode), and two standardization approaches (no standardization and standardization). Detailed information of the preprocessing steps can be found in the Appendix.

### Model Selection

We used the following eight model types to fit a total of 124 models. This includes all possible combinations of eight different model types, two types of automatic variable selection ($F$ test and mutual information) using two thresholds (10 percent and 20 percent), two types of imputation strategies (median and mode), and two standardization approaches (no standardization and standardization):

1. Ordinary least squares linear regression
2. Least-angle regression (Efron et al. 2004)
3. Ridge regression (Tikhonov 1963)
4. Elastic net (Zou and Hastie 2005)
5. Orthogonal matching pursuit (Cai and Wang 2011)
6. Least absolute shrinkage and selection operator (LASSO) regression (Tibshirani 1994)
7. Least absolute shrinkage and selection operator (LASSO) regression (Tibshirani 1994)

In the case in which there were more variables than cases, scikit-learn finds the minimum $l_1$ norm solution via singular value decomposition (Pedregosa et al. 2011).
7. Decision tree regression (Quinlan 1986)
8. ε-support vector regression with linear kernel (Drucker et al. 1997)

The observant reader will notice that \(8 \times 2 \times 2 \times 2 \times 2 = 128\), but we fit only 124 models. We fit decision tree models using only some type of automatic variable selection. We did not fit these models using the full variable set, because decision trees are very susceptible to overfitting in high-dimensional settings such as this, in which the number of variables greatly outnumbers the number of samples (Pedregosa et al. 2011). This accounts for the missing 4 combinations. ²

Results
We used FFC holdout test set mean squared error (MSE) scores (FFC-HO-MSE) to evaluate the accuracy of the models. We chose the MSE metric because it is the metric used to rank and evaluate the predictive validity of the submissions made through the FFC web portal (Salganik et al. 2019).

Results from step 1 are summarized in Table 1.

| Model | Type | Imputation | Scaling | Feature Selection | Variable Set | FFC-HO-MSE |
|-------|------|------------|---------|-------------------|--------------|------------|
| 1     | LASSO| Median     | Standardize | None              | All          | 0.348      |
| 2     | EN   | Median     | Standardize | None              | All          | 0.35       |
| 3     | EN   | Mode       | Standardize | MI 20%            | All          | 0.381      |
| 4     | OMP  | Median     | Standardize | None              | All          | 0.389      |
| 5     | OMP  | Median     | None       | None              | All          | 0.389      |
| 6     | EN   | Median     | Standardize | MI 20%            | All          | 0.389      |
| 7     | DTR  | Median     | None       | F-Reg 10%         | All          | 0.404      |
| 8     | DTR  | Median     | None       | MI 20%            | All          | 0.412      |
| 9     | EN   | Mode       | Standardize | None              | All          | 0.474      |
| 10    | LASSO| Mode       | Standardize | MI 20%            | All          | 0.511      |

Note: Models are numerically labeled and ordered by increasing Fragile Families Challenge holdout test set mean squared error score (FFC-HO-MSE). The lower the mean squared error (MSE) the better. DTR = decision tree regression; EN = elastic net; LASSO = least absolute shrinkage and selection operator; MI = median imputation; OMP = orthogonal matching pursuit.

2The results for all 124 models can be found in the Appendix.

Results
We used FFC holdout test set mean squared error (MSE) scores (FFC-HO-MSE) to evaluate the accuracy of the models. We chose the MSE metric because it is the metric used to rank and evaluate the predictive validity of the submissions made through the FFC web portal (Salganik et al. 2019).

Results from step 1 are summarized in Table 1.

Step 2: Manual Variable Selection
The goal of step 2 was to improve the predictive accuracy of the models generated in step 1 by combining the previous automatic approaches with manual ones inspired by survey research best practices.

Manual Variable Selection
Our first step in this second phase of the project was to get friendly with the codebooks. We went through each of the 12,942 variables, manually selecting the ones believed to be predictive of future academic achievement. To inform the decision-making process, we turned to a comprehensive review of student success literature, “What Matters to Student Success,” a report commissioned for the National Postsecondary Education Cooperative in 2006 (Kuh et al. 2006). Specifically, we relied on the first section of the report, which discusses the effects of precollege experiences on student success, such as family and peer support, academic preparation, motivation to learn, socioeconomic status, and demographics. Although the report is targeted at student success in college, research has shown that high school grades are also highly correlated with socioeconomic factors such as family income and educational attainment (Zwick and Green 2007). From the National Postsecondary Education Cooperative report, we collated a list of 57 precollege factors that have been shown by social scientists to affect student success.³

Next, we manually examined each variable in the codebook and made judgement calls to determine whether it was directly related to any one of the 57 factors. It should be noted that we did not calculate intercoder reliability (Lombard, Snyder-Duch, and Bracken 2002). Calculating and reporting the intercoder reliability of this manual process is an area for future work. The aftermath of this process was a custom set of 3,694 variables.⁴

In an effort to identify the particular groups of variables most predictive of academic achievement, we created 14 additional, more granular subsets from the manually selected set of 3,694 variables. For example, we created a variable set that contained only wave 3 variables and a different subset that contained only wave 5 variables.

³The full list of 57 factors can be found in the Appendix.
⁴The complete list of 3,694 variable labels can be found in the Appendix.
Table 2. Descriptions of Each of the 16 Variable Sets Used in this Project.

| Data Set | Number of Features | Description |
|----------|-------------------|-------------|
| All      | 12,942            | All features from original data set |
| MF       | 3,694             | Manually selected features; subset of all |
| w1       | 138               | Wave 1 variables only; subset of MF |
| w2       | 613               | Wave 2 variables only; subset of MF |
| w3       | 659               | Wave 3 variables only; subset of MF |
| w4       | 755               | Wave 4 variables only; subset of MF |
| w5       | 1,458             | Wave 5 variables only; subset of MF |
| w1_5     | 1,595             | Wave 1 and wave 5 variables only; subset of MF |
| w1_t_kind| 482               | Wave 1, teacher, and kindergarten teacher variables; subset of MF |
| w1_5_t_kind| 1,670         | Wave 1, wave 5, teacher, and kindergarten teacher variables; subset of MF |
| c        | 423               | Constructed variables and variables containing the string “INT CHK”; subset of MF |
| child    | 1,628             | Variables containing the string “child”; subset of MF |
| t_kind   | 345               | Teacher and kindergarten teacher variables; subset of MF |
| k        | 103               | Child variables only; subset of MF |
| t_kind_k | 447               | Teacher, kindergarten teacher, and child variables; subset of MF |
| t_k      | 372               | Teacher and kid variables only; subset of MF |

Note: This count does not include the variable subsets created using the automated univariate feature selection routine in scikit-learn; it includes the number of variables in each before imputation and the shorthand label used to reference each of the variable sets. These 16 variable sets include the original set of 12,942 variables, our manually curated set of 3,694 variables, and 14 additional variable sets, each of which is a subset of the manually selected set of 3,694 variables (wave 3 only, wave 5 only, etc.).

We used a total of 16 variable sets in this project: (1) the original set of 12,942 variables, (2) our manually curated set of 3,694 variables, and (3) 14 additional variable sets, each of which is a subset of the manually selected set of the 3,694 variables (wave 3 only, wave 5 only, etc.). Table 2 summarizes each of these 16 variable sets and provides a shorthand label for each. We use these shorthand labels to reference the various variable sets for the remainder of this article.

Method

We reestimated the 10 most accurate models from step 1 on each of the 15 manually created variable subsets to produce a total of 150 models in this second step of the project. We used the same data-preprocessing procedures and imputation strategies used in step 1. As before, categorical variables were not identified and were not treated differently from the continuous ones. After data imputation, our manually curated variable set was reduced from 3,694 to 3,423 variables. The FFC submission pipeline remained the same.

Results

Manual variable selection indeed improved, and in some cases dramatically improved, the accuracy of the predictive models trained previously using purely automatic techniques. Table 3 shows that 8 of the 10 most accurate models were trained on the manually created variable sets. Figure 1 shows how substantially manual variable selection improved the FFC-HO-MSE values of the 3rd, 6th, 9th, and 10th most accurate models from step 1. After reestimating model 6 on the w5 variable set, the model rose to become the second most accurate model across both phases of the project, according to FFC-HO-MSE. The accuracies of models 4, 5, and 7 were also improved, but the change in FFC-HO-MSE was more tempered. The two most accurate models from step 1 saw no improvement.

Effect of Specific Variable Groups on Model Accuracy

A secondary goal was to understand how the various variable groups affected the predictive accuracy of the models trained in step 2 (e.g., do wave 5 variables yield better results than wave 3 variables?). Figure 2 is a 16 × 10 heatmap of FFC-HO-MSE scores from the 10 most accurate step 1 models trained on each of the 16 variable sets from the project, including the full set of 12,942 variables (labeled “All”). The lower the MSE value and the darker the color, the better. Variable sets w1, w2, and w3 appear to contain the weakest signal across almost all models, and variable set w5 appears to contain the strongest signal, closely followed by w1_5 and w1_5_t_kind. From w1 through w5 we see a gradual strengthening of color across several rows. This pattern and the previous observations suggest that later waves are more predictive of high school GPA than earlier waves. However, not all wave 5 data are created equally. Variable set k contains variables asked of only the child in wave 5, and t_k contains variables asked of the child and the child’s
Roberts

Looking across both columns, we can visually see how FFC-HO-MSE values improved across more than half of the models when input from the teacher was removed. We see a similar phenomenon when comparing the t_kind and t_kind_k columns. The majority of the models seem to improve with added input from the child. It appears that no matter how attentive a parent, teacher, or caretaker may (or may not) be, only the child really knows what he or she is feeling and experiencing on a day-to-day basis. And many of questions asked of the child in wave 5 attempt to tease out precisely this, questions such as “Frequency kids picked on you or said mean things to Table 3.

| Model | Type   | Imputation | Scaling | Univar. Feature Selection | Variable Set | FFC-HO-MSE |
|-------|--------|------------|---------|---------------------------|--------------|------------|
| 1     | LASSO  | Median     | Standardize | None                      | All          | 0.348      |
| 6     | EN     | Median     | Standardize | MI 20%                    | w5           | 0.349      |
| 2     | EN     | Median     | Standardize | None                      | All          | 0.35       |
| 1     | LASSO  | Median     | Standardize | None                      | w5           | 0.35       |
| 9     | EN     | Mode       | Standardize | None                      | w5           | 0.35       |
| 2     | EN     | Median     | Standardize | None                      | w5           | 0.351      |
| 6     | EN     | Median     | Standardize | MI 20%                    | wave1_5_t_kind | 0.351     |
| 3     | EN     | Mode       | Standardize | MI 20%                    | wave1_5_t_kind | 0.353     |
| 6     | EN     | Median     | Standardize | MI 20%                    | wave1_5      | 0.353      |
| 10    | LASSO  | Mode       | Standardize | MI 20%                    | wave1_5      | 0.353      |

Note: Models are listed in order of increasing Fragile Families Challenge holdout test set mean squared error score (FFC-HO-MSE) scores. The lower the mean squared error (MSE) the better. The column “Variable Set” contains the label name of the variable set used to train that particular model. Refer to Table 2 for a description of each variable set. EN = elastic net; LASSO = least absolute shrinkage and selection operator; MI = median imputation.

Figure 1. Effect of manual variable selection on the predictive ability of the 10 most accurate step 1 models. For the step 1 series, in which the full set of 12,942 variables was used to fit the models, the means squared error (MSE) value is plotted for each model. Recall that in step 2, we reestimated the top 10 step 1 models using each of the 15 manually created variable subsets (the full set of 3,694 manually curated variables plus 14 additional subsets taken from this set of 3,694 variables), giving us 15 MSE scores per model. Thus, for the step 2 series, for each model, we plot the holdout result on the basis of the result with the best leaderboard score.
Variables That Most Predict GPA

An important goal of the FFC is to gain insight into the specific variables that most predict the six outcomes of interest: GPA, grit, material hardship, eviction, layoff, and job training. The hope is that such insights may one day improve the lives of American children born into these “fragile families” (Salganik et al. 2019). Table 4 lists the variables that most predict year 15 GPA according to the two most accurate models from this project on the basis of FFC-HO-MSE scores. The most accurate model used LASSO, and the second most accurate model used elastic net. Coefficients are in parentheses, and variables are listed in order of decreasing absolute coefficient value. Because the data were standardized, we were able to compare variable coefficients according to their relative significance to the prediction task. That is, the higher the absolute value of the coefficient, the higher the level of importance of that particular variable in predicting the desired outcome, which in this case is GPA. In prediction tasks such as these, in which we are predicting a real-valued outcome in a setting with multiple independent variables, coefficients can be interpreted as the following: holding all other variables fixed, the predicted outcomes increase (if the sign of the coefficient is positive) or decrease (if the sign of the coefficient is negative) by a factor of $\beta_1$ units for every 1 unit increase in $x_1$, where $\beta_1$ is the coefficient associated with the variable $x_1$ (Princeton University Library 2007).

It is worth highlighting that although these two best models have almost equal predictive performance on the holdout data, 0.348 and 0.349, respectively, they exhibit very little overlap in the variables each model deems to be of most significance. We saw two different sets of variables returned by two
Table 4. Variables That Most Predict Grade Point Average within the Two Most Accurate Models from the Project.

| Rank | Most Accurate Model | Coefficient | Second Most Accurate Model | Coefficient |
|------|---------------------|-------------|-----------------------------|-------------|
| 1    | m13: What is the highest grade/years of school that BF have completed? | 0.046 | 1 | m5f26a: Number of charges you currently have pending | 0.064 |
| 2    | c5b1w: CHW. Child attends to your instructions | 0.042 | 2 | m5i7: In past 12 months you worked more than one regular job at the same time | 0.046 |
| 3    | hv5_ppvtp: PPVT percentile rank | 0.042 | 3 | m5e1: Highest grade of school that mother’s biological mother completed | -0.044 |
| 4    | flb20: Int chc: Are BM & BF living together? | -0.037 | 4 | k5h2: Frequency you wear a seatbelt when riding in a car | 0.040 |
| 5    | m11: What is the highest grade/years of school that you have completed? | 0.030 | 5 | m5g0: How satisfied you are with your life overall | 0.039 |
| 6    | hv5_wj10pr: Woodcock Johnson Test 10 percentile rank | 0.029 | 6 | m5d6: Highest grade of school current partner has completed | 0.034 |
| 7    | p5ml: M1. Number of families on block know well | -0.024 | 7 | k5a2d: Your mom misses events or activities that are important to you | -0.029 |
| 8    | cm3amrf: Constructed—Mother age when married father (years) | -0.022 | 8 | k5a3a: Your dad talks over important decisions with you | -0.027 |
| 9    | p5i13f: Gifted and talented program | -0.021 | 9 | k5a2a: Your mom talks over important decisions with you | -0.027 |
| 10   | (ffcc_famsurvey_b34_a) B34A. How many total hours do you usually work per week? Include regular overtime hours at (this job/all of your jobs). | -0.020 | 10 | m5i6b: Where I work it is difficult to deal with child care problems | 0.026 |

Note: With a Fragile Families Challenge holdout test set mean squared error score (FFC-HO-MSE) value of 0.348, the most accurate model used least absolute shrinkage and selection operator, median imputation, standardized variables, and no additional univariate feature selection and was trained on the full Fragile Families Challenge feature set (variable set labeled “All”). The second most accurate model, with an FFC-HO-MSE score of 0.349, used elastic net, median imputation, standardized variables, and univariate feature selection (20 percent) using the mutual information scoring function and was trained on the wave 5 (w5) manual feature subset. Because there is some disagreement about how to produce confidence intervals around estimates that come from regularized models, we chose not to include them. BF = birth father; BM = birth mother; PPVT = Peabody Picture Vocabulary Test.

different models of almost equal predictive accuracy, giving us two different pictures of which variables most predict year 15 GPA. In his analysis of the two cultures of statistical modeling, Breiman (2001) argued that in a situation in which “different models, all of them equally good . . . give different pictures of the relation between the predictor and response variables . . . the question of which one most accurately reflects the data is difficult to resolve.” As engineers, these difficulties are further compounded by a lack of domain knowledge in the social sciences. Thus, we leave intuitive explanation of these results for future work and collaborations with social scientists. Furthermore, further research is required to calculate confidence intervals for the coefficients listed in this section and to begin interpreting the magnitude of the values and features returned.

Conclusions

Using a two-step approach to the FFC, we were able to significantly improve the predictive accuracy of the majority of the models evaluated by using a combined approach of automatic and manual variable selection motivated by social science knowledge. We showed that such an approach, even though it is based on a nonexpert reading of domain-specific research, can improve the accuracy of models trained automatically. We demonstrated that if one is not careful choosing algorithms in such a data setting, it pays to take a look at the codebooks. But there is still room for improvement, as our strategy was unable to improve the accuracy of the two most accurate models from step 1; this is an area for future work.

Acknowledgments

We would like to thank Professor Matthew Salganik, Dr. Sara McLanahan, Professor Arvind Narayanan, and the Fragile Families Committee for helpful feedback and comments. We also want to thank Alex Kindel for providing additional support in the reanalysis of our data. The results in this paper were created with software written in R version 3.3.3 (R Core Team 2018) using the following packages: ggplot2 2.2.1 (Wickham 2009), broom 0.4.2 (Robinson and Hayes 2018), and caret 6.0-78 (Kohn 2017).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Computer Science Department at Princeton University. Funding for the FFCWS was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development through grants R01HD36916, R01HD39135, and R01HD40421 and by a consortium of private foundations, including the Robert Wood Johnson Foundation. Funding for the FFC was provided by the Russell Sage Foundation. This work was supported by the Computer Science Department at Anonymous University.
Supplemental Material

Supplemental material for this article is available with the manuscript on the Socius website.

References

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a rejoinder by the Author).” *Statistical Science* 16(3):199–231.

Cai, T. Tony, and Lie Wang. 2011. “Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise.” *IEEE Transactions on Information Theory* 57(7):4680–88.

Drucker, Harris, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. “Support Vector Regression Machines.” Pp. 155–61 in *Proceedings of the 9th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. “Least Angle Regression.” *Annals of Statistics* 32(2):407–99.

Max Kuhn. 2017. *caret: Classification and Regression Training*. 2017. R package version 6.0–78.

Hua, Jianping, Zixiang Xiong, James Lowey, Edward Suh, and Edward R. Dougherty. 2005. “Optimal Number of Features as a Function of Sample Size for Various Classification Rules.” *Bioinformatics* 21(8):1509–15.

Kaggle. 2017. “Home Page.” Retrieved October 9, 2017 (https://www.kaggle.com).

Kelley, Kate, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. “Good Practice in the Conduct and Reporting of Survey Research.” *International Journal for Quality in Health Care* 15(3):261–66.

Kuh, George D., Jillian Kinzie, Jennifer A. Buckley, Brian K. Bridges, and John C. Hayek. 2006. “What Matters to Student Success: A Review of the Literature.” National Postsecondary Education Cooperative. Retrieved December 14, 2018 (https://nces.ed.gov/nepc/pdf/kuh_team_report.pdf).

Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. “Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability.” *Human Communication Research* 28(4):587–604.

McFarland, Daniel A., Kevin Lewis, and Amir Goldberg. 2016. “Sociology in the Era of Big Data: The Ascent of Forensic Social Science.” *American Sociologist* 47(1):12–35.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. “scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12(October):2825–30.

Princeton University Library. 2007. “Interpreting Regression Output.” Retrieved March 26, 2018 (https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm).

Quinlan, J. R. 1986. “Induction of Decision Trees.” *Machine Learning* 1(1):81–106.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reichman, Nancy E., Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. 2001. “Fragile Families: Sample and Design.” *Children and Youth Services Review* 23(4):303–26.

Robinson, David, and Alex Hayes. 2018. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.1.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, and Sara McLanahan. 2019. “Introduction to the Special Collection on the Fragile Families Challenge.” *Socius* 5. doi:10.1177/2378023119871580.

Tibshirani, Robert. 1994. “Regression Shrinkage and Selection via the LASSO.” *Journal of the Royal Statistical Society, Series B* 58(1):267–88.

Tikhonov, A. N. 1963. “On the Solution of Incorrectly Formulated Problems and the Regularization Method.” *Soviet Mathematics Doklady* 151:501–504.

Trochim, William M. K. 2006. “Survey Research.” Retrieved October 9, 2017 (https://www.socialresearchmethods.net/kb/survey.php).

The Trustees of Princeton University. 2018. “About the Fragile Families and Child Wellbeing Study.” Retrieved December 14, 2018 (http://www.fragilefamilies.princeton.edu/about).

Visser, P. S., Jon A. Krosnick, and Paul J. Lavraka. 2000. “Survey Research.” Pp. 223–52 in *Handbook of Research Methods in Social and Personality Psychology*, edited by Harry T. Reis and Charles M. Judd. Cambridge, UK: Cambridge University Press.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wright, Daniel. 2003. “Making Friends with Your Data: Improving How Statistics Are Conducted and Reported.” *British Journal of Educational Psychology* 73(Pt. 1):123–36.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society, Series B* 67(Pt. 2):301–20.

Zwick, Rebecca, and Jennifer Greif Green. 2007. “New Perspectives on the Correlation of SAT Scores, High School Grades, and Socioeconomic Factors.” *Journal of Educational Measurement* 44(1):23–45.

Author Biography

**Claudia V. Roberts** is a second-year computer science master’s student at Princeton University. Claudia, a proud Dallas, Texas, native, GEM fellow, and Siebel scholar, earned a BS in computer science from Stanford University, after which she worked at Apple as a software engineer for four years. She presented her work at the Apple Worldwide Developers Conference on two occasions. While at Stanford, Claudia worked as an undergraduate researcher at the Transformative Learning Technologies Lab and interned at Cisco and Yahoo. Recently, she spent the last two summers at Adobe Research as an engineering intern and presented her work to the CEO and top-level executives. During her first year as a master’s student at Princeton, Claudia coauthored a technical paper on anomalous RSA keys in Tor relays and was awarded a faculty-nominated graduate teaching assistant award for her TA service.