Impact of Absent Father on Child's School Education

Aarshay Jain, Bindia Kalra, Keerti Agrawal
Data Science Institute, Columbia University
Broadway W 120th St, New York, NY 10027
{aj2713, bk2628, ka2601}@columbia.edu

Abstract

This study explores the impact of mother-father relationship, particularly biological father's involvement in upbringing of a child, to its educational outcome at the age of 15. We took the data from the Fragile Families and Child Wellbeing Study[1], which is a pioneer survey in this domain. The study focuses on children born to unwed parents. Around 5000 children were tracked from birth to the age of 15 and their parents were interviewed in 5 waves. The data has 12943 survey questions, 600 of which are summary features created by Fragile Family Challenge team. These have cryptic codings and it took us significant amount of time to understand them. We have created a data dictionary (11x6 table), which can be used to decipher these 600 features, and is one of the key contribution of this study. To validate our hypothesis, we engineered five factors, namely biological father's presence, biological father's involvement, mother's cohabitation with father, mother's cohabitation with any other partners and the number of non-marital relationships of the mother. We performed regression analysis and found that children with high or even mild involvement of their biological father's in their upbringing have higher GPAs on average. Also, presence of a fatherly figure, whether the biological father or some other partner of the mother, seems to positively correlate with GPA.

1 Introduction

The Fragile Families and Child Wellbeing Study is following a cohort of nearly 5000 children from 20 cities in the U.S. The data has been collected in 5 waves over 15 years between 1998 to 2015. Around three-forth of the interviewed parents were not married at the time of birth of the child. The study has important information on unmarried mothers but the major goal is to analyze the role of fathers in these families. The study tagged these families as "fragile" on account of these families being at greater risk of breaking up and having low household income as compared to the more traditional families.

The study was designed to help the researchers and policy makers to come up with solutions to help such "fragile" households. The study primarily address the following four questions:

1. What are the conditions and capabilities of unmarried parents, especially fathers?
2. What is the nature of the relationships between unmarried parents?
3. How do children born into these families fare?
4. How do policies and environmental conditions affect families and children?

We propose a hypothesis that the presence and level of involvement of father with the child at younger age plays a significant role in child's performance at school. We believe at younger ages, children are more impressionable and presence of a fatherly figure, which maybe the biological father or mother's other romantic partner, could effect their educational outcomes. The motivation behind this hypothesis has come from these three factors. First, as documented in Fragile Families survey[2], nearly a third of all children born in the United States today are born to unmarried parents with higher among poor and minority populations. The fragile families

[1] http://www.fragilefamilieschallenge.org

dataset serves an opportunity to come up with insights which could be used to formulate better policies for such disadvantage children. Second, already existing research around this field focuses on factors like effect of father's involvement on the behavior of the child (Shelley et al., 2016), effect of on/off relationship on father involvement with child (Turney and Halpern-Meekin, 2017), family instability and behavior changes (Fomby and Osborne, 2017). The effect of father's involvement on child's educational outcomes at school is not well studied. Third, fragile families dataset contains GPA at the age of 15 years which can be used as a measure of educational performance in school.

The paper is organized as follows: Section 2 (**Data**) contains the description on data and data collection process; Section 3 (**Feature Engineering**) describes the variable selection and variables construction, Section 4 (Exploration and Analysis) discusses, experiments and provide analysis on the features used for regression and various regression models tried out for the task and Section 5 (**Conclusion**); Section 6 (**Challenges**) discusses the shortcomings and lays the path for the future direction.

2 Data

We have used the Fragile Families and Child Wellbeing Dataset for our analysis. The data was collected on approximately 4700 births (3600 nonmarital, 1100 marital) in 75 hospitals in 20 cities across the United States over 15 years. Parents were interviewed five times which are classified as baseline, 1st year, 3rd year, 5th year and 9th year interview. The key difference between this dataset and other existing dataset in the same field is that they tracked and interviewed roughly 75% of the unwed fathers. Also, the missing data for the remaining fathers was completed based on the mother's interview[3].

This dataset contains 4242 observations and 12943 features. Around 600 variables are constructed and rest are raw. Variables represent interview answers related to the questions based on relationship, home environment, incarceration history, income level, employment status, race/religion etc. These observations were noted for five waves, child's birth, 1st year, 3rd year, 5th year and 9th year. There are following six outcome variables given: GPA, Grit, Material hardship, Eviction, Being laid off. The values for these have been only collected for year 15 only. Among these, we'll be studying GPA, the distribution of which seemed to be well spread as shown in Figure 1.

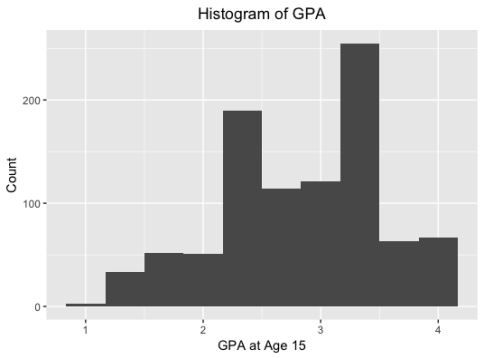


Figure 1: Distribution of GPA

3 Feature Engineering

In order to test our hypothesis, we began by understanding the data. We decided 5 factors, as discussed below, to test our hypothesis. A key step was to understand the questions asked in the survey and come up with proxy variables which can be usec to estimate our desired effects. These tasks are explained below in detail.

3.1 Data Dictionary

The fragile family data collector team has created a set of around 600 features from raw features which captures good amount of information about this survey. As our first task, we tried to decode these crypted variables by going through 25 documents that have questionnaire for each wave. We couldn't find any concrete document for such a cumbersome task. And therefore we have created a data dictionary of 66 cells, which contains all the information about these constructed features and can be used in future research. This will be one of our key constributions to the project, which will be open-sourced

[3] http://fragilefamilies.princeton.edu/sites/fragilefamilies/files/reichman_et_al_2001.pdf

via GitHub[2].

3.2 Data Cleaning

GPA being our outcome variable, we considered the samples only with non-missing GPA values, which accounted for 54% of the data. We also wanted to make sure that the mother is interviewed in all the waves (82%) and the data was filtered accordingly. Next is to remove the cases when the child is under the custody of the father (0.02%). We assumed that in the case of father custody, there is no point of looking at father's presence or involvement in child's life. Finally, we imputed few missing values in the data columns using the median.

3.3 Feature Selection

We took into consideration the following five features:

1. **Presence of father at home:** If the father saw the child in the last two years.
   We have categorized the data into None, Partial and Full presence based on the following features:
   - Using the m*c2 and m5b2(for the fifth wave) column we separated the fathers who had not seen the child in the last 2 years and added them in the None category.
   - We have checked the cm*relf on the remaining data from A, this feature identifies the married or cohabiting parents. All the fathers in this data were added to the Full category. (* is in 1,2,3,4,5)
   - The remaining fathers (who had seen the child in the last two years and are not living with the mother) were added to the Partial category.

   We observed that the behaviour was not constant throughout the years. Fathers have changed the categories in the interviews, e.g., Full to partial to None. To address this, we assigned weights to all the years. The underlying assumption in assigning the weights to the interview years is that the absence of a father would affect the child more in the later years (the child is 5 or 9) than in the early years (the child is 1 or 3). The distribution of weights have been shown in Figure 2

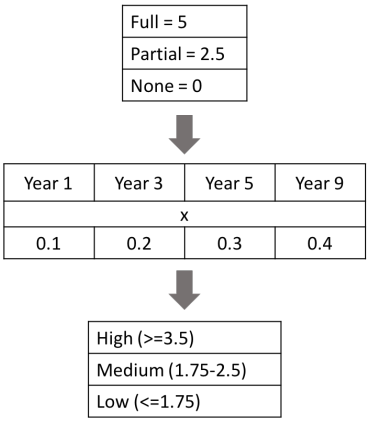


Figure 2: Weights Assigned To Each Wave and Category

2. **Father's involvement with the child:** We have categorized the data into Full, Partial and None based on the features m*c5 and m5b3 and saw how often the father spent 1+ hr per day with child.
   - The fathers who spend 1+ hour everyday with the child were added to the Full category.
   - The fathers who saw the child a few times a week or a few times a month were added to the Partial Category.
   - The fathers who saw the child one or two times per month or not at all were added to the No involvement category. The cases where the father didnt know about the child or where the father was deceased were also included in this category. The skipped cases where the father hadnt seen the child in the past month were also included in this category.

   We observed that the behaviour was not constant throughout the years. Fathers have changed the categories in the interviews, e.g., Low to High to Medium. The underlying assumption in assigning the weights to the interview years is that the absence of a father would affect the child more in the later years (the child

[2] https://github.com/aarshayj/FragileFamiliesChallenge

is 5 or 9) than in the early years (the child is 1 or 3). The distribution of weights have been shown in Figure 2.

3. **Cohabiting with biological father:** We checked the cm*relf to count the number of households where the mother was living with the biological father and we just sum the boolean values. For example, if the mother and the biological father were married or cohabiting during all four interviews, the sum of all the boolean values would be 4, if they were married or cohabiting for during 3 of the 4 interviewees, the number would be 3 and so on.

4. **Presence of a father figure at home (cohabiting with the biological father or any partner):** We checked cm*cohp and cm*marp to count the number of households where the mother living with a partner other than the biological father. We get our number by taking the sum of the count of the boolean values of all the interview years. For example, if the mother and the partner were married or cohabiting during all four interviews, the sum of all the boolean values would be 4. We then added the count from the result of the Cohabiting with biological father feature to get our final number. The number 4 in this case would mean that the mother was living with the biological father or a partner during all the four interviews.

5. **Number of partners (apart from father):** Counting number of partners apart from the babys father by checking the feature m4a13 and M5a101 to see the number of romantic relationships that the mother was involved in that lasted for at least one month since the last interview. The mothers who were married or were in a steady relationship were in the 0 category. The mothers who refused to answer were also in the 0 category. The number 3 here would mean that the mother had 3 relationships that lasted more than a month. This was only checked for wave three (the child was 5 years old) and wave four(the child was 9 years old) as the required question was asked only in these two interviews. The data here is for 6 years (between the 2nd wave and the fourth wave).

6. **Control Variables:** We have considered some control variables to account for other factors/confounders in our experiment:

   - **Parents education-** We checked the values of the features cm1edu and cf1edu to see the education level for both the parents (1- less than High school, 2 - HS or equiv, 3 - Coll; tech, 4 - coll; grad).
   - **CIDI-** We measured the cases where the mother was suffering from depression, alcohol or drug abuse and anxiety disorders. For example a value of 4 would mean the mother answered yes for 4 of the following features.
   - **Household Income-** We measured the income of the household by taking the log of the average of cm*hhinc for all five years, where cm*hhinc is the constructed variable for the household income for that particular wave.
   - **Cases of Spanking-** For this variable, we referred the childs interview at age 9 to find out if the child was being spanked by any or both of the parents. We checked the features k5b2d, k5b1d and k5b3d which is for if the child was spanked by the biological father, mother and social father respectively. The higher the value for this variable, the more are the number of times the child was hit or spanked.

4 Exploration and Analysis

Two types of exploratory analysis were performed to understand the correlations between the created variables and the outcome GPA. To begin with, univariate analysis was done to gauge the relation of each variable individually to GPA. We then found the effect of interactions between variables using ordinary least squares regression models.

4.1 Univariate Analysis

Figure 3 shows that 85% of the children had medium or high presence of father. But only high presence had higher GPA outcomes on average as compared to the low presence, i.e. medium or low presence have similar outcomes.

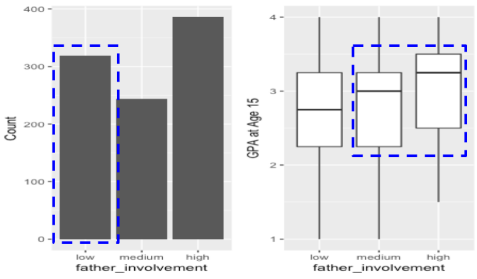


Figure 3: Father's Presence

In figure 4, we see that 33% of the children have low father involvement, though the presence numbers in figure 3 is higher. Also, children with even mildly involved fathers show a higher GPA on average. Thus, father's involvement is more important than mere presence.

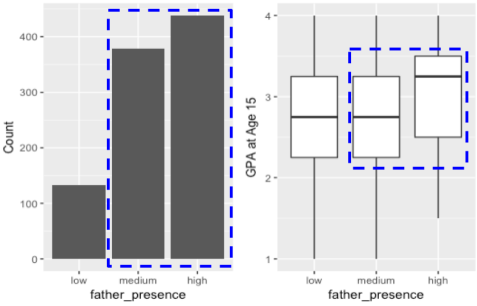


Figure 4: Father's Involvement

We also tested 3-way ANOVA for both father's presence and involvement and they showed significant p-values - 3.46e-12 for father's involvement and 3.79e-11 for father's presence.

Looking at the cohabitation with biological father or the any other partner have shown to be positive correlation around 0.2 with GPA. We can also see that about one-third of the mothers did not live with biological fathers in any of the wave.

Figure 6 shows that more than 50% of the mothres had non-marital relationships. As expected, we found a negative correlation of 0.2 with GPA.

4.2 Interaction Modeling with OLS

We took a step-wise approach to understand the impact of interactions between our features and GPA. We have used stargazer (Hlavac, 2013) package to compare the models and show results in tables.

As shown in figure 7, we started with the features

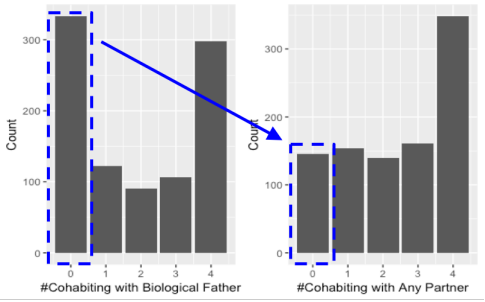


Figure 5: Cohabitation with a partner

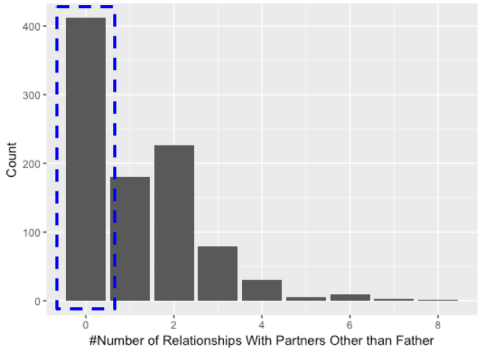


Figure 6: Non-marital partners of mother

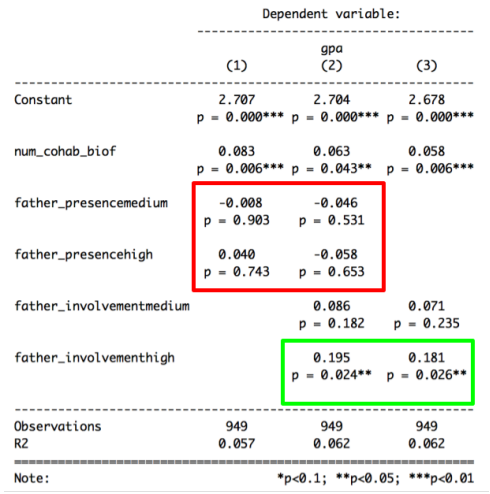
representing father's presence, i.e. father's presence itself and number of times mother was found cohabiting with father. These were individually significant but when combined, we found only one of them to work well so we'll keep it in the next model.

	Dependent variable:	
	gpa	
	(1)	(2)
Constant	2.714	2.707
	p = 0.000***	p = 0.000***
num_cohab_biof		0.083
		p = 0.006***
father_presencemedium	0.037	-0.008
	p = 0.576	p = 0.903
father_presencehigh	0.328	0.040
	p = 0.00000***	p = 0.743
Observations	949	949
R2	0.049	0.057
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 7: Linear Model Set 1

Next, father's involvement was added into the

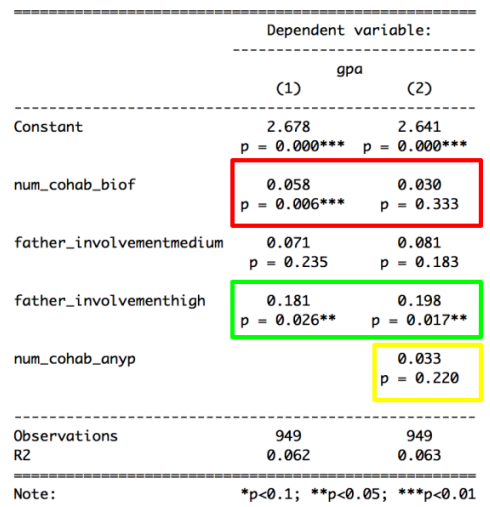
model and the results are shown in figure 8. Here we see that high involvement is more important than mere presence as the presence coefficients are no more significant. Even the medium involvement coefficient is not too bad.



	Dependent variable:		
	gpa		
	(1)	(2)	(3)
Constant	2.707	2.704	2.678
	p = 0.000***	p = 0.000***	p = 0.000***
num_cohab_biof	0.083	0.063	0.058
	p = 0.006***	p = 0.043**	p = 0.006***
father_presencemedium	-0.008	-0.046	
	p = 0.903	p = 0.531	
father_presencehigh	0.040	-0.058	
	p = 0.743	p = 0.653	
father_involvementmedium		0.086	0.071
		p = 0.182	p = 0.235
father_involvementhigh		0.195	0.181
		p = 0.024**	p = 0.026**
Observations	949	949	949
R2	0.057	0.062	0.062
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 8: Adding involvement into the model

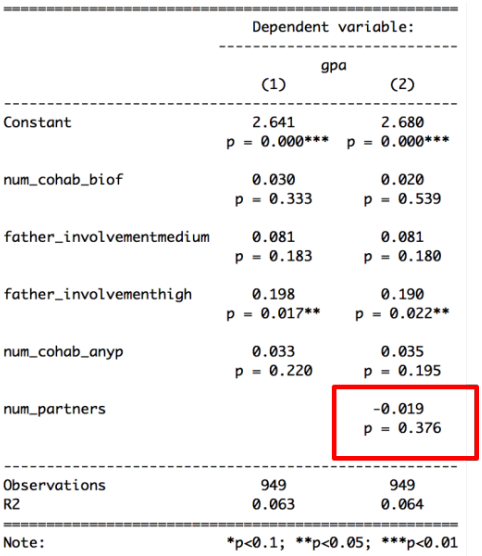
Moving on, cohabitation with any partner was added into the model as shown in figure 9. This is a proxy for presence of fatherly figure in the house. We don't see highly significant results, but the coefficient is on positive side.



	Dependent variable:		
	gpa		
	(1)	(2)	
Constant	2.678	2.641	
	p = 0.000***	p = 0.000***	
num_cohab_biof	0.058	0.030	
	p = 0.006***	p = 0.333	
father_involvementmedium	0.071	0.081	
	p = 0.235	p = 0.183	
father_involvementhigh	0.181	0.198	
	p = 0.026**	p = 0.017**	
num_cohab_anyp		0.033	
		p = 0.220	
Observations	949	949	
R2	0.062	0.063	
Note:	*p<0.1; **p<0.05; ***p<0.01		

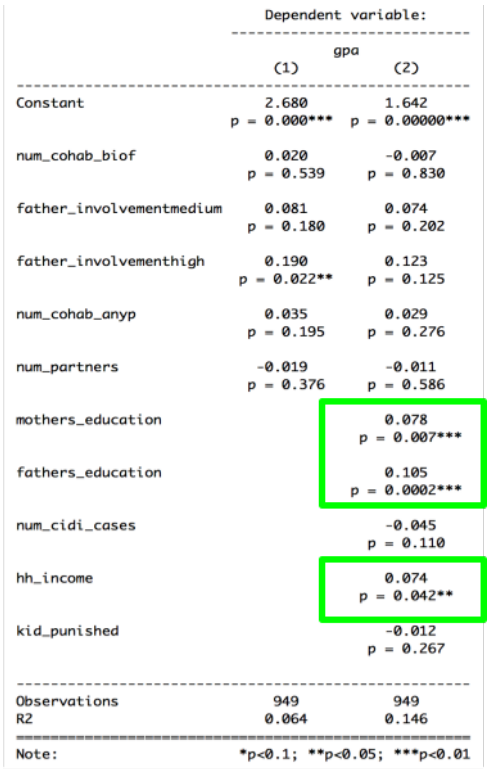
Figure 9: Adding cohabitation with father or other partner

Finally, we added the number of extra-marital partners of the mother into the model and found it to have a negative coefficient as expected. But the impact is not very significant.



	Dependent variable:	
	gpa	
	(1)	(2)
Constant	2.641	2.680
	p = 0.000***	p = 0.000***
num_cohab_biof	0.030	0.020
	p = 0.333	p = 0.539
father_involvementmedium	0.081	0.081
	p = 0.183	p = 0.180
father_involvementhigh	0.198	0.190
	p = 0.017**	p = 0.022**
num_cohab_anyp	0.033	0.035
	p = 0.220	p = 0.195
num_partners		-0.019
		p = 0.376
Observations	949	949
R2	0.063	0.064
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 10: Adding number of partners into the model



	Dependent variable:	
	gpa	
	(1)	(2)
Constant	2.680	1.642
	p = 0.000***	p = 0.00000***
num_cohab_biof	0.020	-0.007
	p = 0.539	p = 0.830
father_involvementmedium	0.081	0.074
	p = 0.180	p = 0.202
father_involvementhigh	0.190	0.123
	p = 0.022**	p = 0.125
num_cohab_anyp	0.035	0.029
	p = 0.195	p = 0.276
num_partners	-0.019	-0.011
	p = 0.376	p = 0.586
mothers_education		0.078
		p = 0.007***
fathers_education		0.105
		p = 0.0002***
num_cidi_cases		-0.045
		p = 0.110
hh_income		0.074
		p = 0.042**
kid_punished		-0.012
		p = 0.267
Observations	949	949
R2	0.064	0.146
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 11: Adding control variables

Finally, control variables were also added into the model as shown in figure 11. These tend to overpower the explanatory variables as these are more

predictive. But the overall interpretation is still the same.

5 Results

In conclusion, we found 3 factors to correlate with higher GPA outcomes. In order of priority, these are:

- Biological fathers high involvement (95% conf)
- Biological fathers mild involvement (80% conf)
- Presence of a fatherly figure (80% conf)

6 Challenges or Shortcomings

The key challenge we faced was in terms of understanding the data which took up 80% of our time. Some of the difficulties are:

- 12943 survey questions in the dataset
- The survey questions had a hierarchical structure, which was not very intuitive to figure out
- Some of the questions had response codes varying by wave, which was another pain
- The documentation was spread over 25 different files corresponding to different waves and respondents in each wave.

Also, the features that we created were not directly measurable. We also took some assumptions and selected proxy questions which could help us estimate the effect. For instance, father's involvement could depend on factors other than the frequency with which he spends more than an hour with the child.

A key shortcoming of the dataset is that we have information captured from birth to 9 years of age and we are using that to see impact on GPA at the age of 15. But that is how the study was planned.

7 Future Work

Including geographical information into the model is one of the key potential areas of improvement. But this information is present in a different contract dataset which is harder to gain access. We suspect the data to have some social desirability bias and response bias as well given that the questions asked are too personal. We saw a flavor of this in the outcomes. It would be interesting to investigate this further. Also, we have not made any causal claims here. Finding natural experiments in the data and determining the right instruments to perform causal inference is another area of research.

References

[Fomby and Osborne2017] Paula Fomby and Cynthia Osborne. 2017. Family instability, multipartner fertility, and behavior in middle childhood. *Journal of Marriage and Family*, 79(1):75–93.

[Hlavac2013] Marek Hlavac. 2013. stargazer: Latex code and ascii text for well-formatted regression and summary statistics tables. *URL: http://CRAN. R-project. org/package= stargazer*.

[Shelley et al.2016] Walter Shelley, Colleen Wynn, et al. 2016. Moving on down: The conflated impact of family instability and disadvantaged neighborhoods on cognitive, externalizing, and internalizing outcomes. Technical report.

[Turney and Halpern-Meekin2017] Kristin Turney and Sarah Halpern-Meekin. 2017. Parenting in on/off relationships: The link between relationship churning and father involvement. *Demography*, pages 1–26.