

**An Econometric Analysis on
Factors that Affect Teenage Pregnancy Based on 1979 National
Longitudinal Survey of Youth**

By: Athip Tantivorawong
Advised By Prof. Stephen Spear

Introduction

Teenage pregnancy in the United States presents complex problems that affect the lives of the teenage parents themselves, their children, as well as the government. The problem such as low education attainment rate for adolescent parents is an important example. Without gaining sufficient education, the adolescent parents end up working in less prestigious jobs and earning lower salary than their peers.¹ Moreover, this problem incurs cost to the government and society as well since these adolescent teens usually turn to the government for support through social welfare benefits.² This is an important issue as well since the government has to spend billions of dollars per year on these adolescents instead of using that amount of money to spend on goals such as improving the education or health care system. In order for the policymakers to effectively solve the problem of teenage pregnancy, the factors that affect this problem must be determined first. Once they are identified, the policymakers can come up with appropriate policies that would specifically influence these factors in a way that reduces the rate of teenage pregnancy.

The Effects of Teenage Pregnancy

Low education attainment

One of the problems is the fact that adolescent child bearers do not complete as many years of schooling as their classmates.³ For the female cohort that gave birth when they were less than 17 year-old, they only received an educational score of 3, meaning

¹ Grogger J and Bronars SG (1993)

² Hotz VJ, McElroy SW and Sanders SG (1997)

³ Card JC and Wise LL (1978)

that they have only completed grade 11 in high school, 5 years after they were out of high school. For the same cohort 11 years out of high school, their educational score only improved by 0.7, which means that they do not even have an equivalent of high school education. As for the less than 17 year-old male cohort, they receive slightly more educational score than the female counterpart. Nevertheless, they still fail to attain an equivalent of high school education 11 years after they are out of high school. The scores for both male and female cohorts that decide to bear children between the ages of 25-29 are significantly higher than the less than 17 year-old cohorts. Both cohorts receive the score of around 7, which is an educational level of a college graduate, when they are 5 and 11 years out of high school.

There is a linear relationship between the parents' level of education at the periods of 5 and 11 years out of high school and the age in which the parents decide to have their first child. The main reason is that the parents are forced to drop out of high school when the first child is born. Instead of being able to go to school to receive education, they need to spend a lot of time caring for their new born baby.

Low Family Income

There are also occupational differences associated with an early childbearing. After high school, most of the male parents usually enter labor force right away. However, as for the female parents, they have a low rate of entry into labor force. But after 11 years, a high number of them enter labor force, even more so than the females who bear their first child in their mid 20's. This is due to the fact that right after females gave birth, they tend to stay home in order to care for the child and take care of

household works. But when the child has grown up, they no longer need to stay home.⁴ However, even if the parents decided to enter labor force, they usually end up getting less prestige jobs such as such as domestic servants or nurses' aides. This is because occupational achievement correlates with the time spent in school, or human capital investment. That means that it is really difficult for these parents to catch up with their classmates, who delay their first childbearing and obtain college education, in terms of amount of salary made.⁵

Higher number of children

Teenage parents tend to have higher number of children than their classmates at 5 and 11 years out of high school.⁶ As for the less than 17 year-old female cohort, their average number of children 5 years after high school, or when they are about 23 year-old, is 2.4, where as that of the 20-24 year-old cohort 11 years after high school, or when they are about 29 year-old, is 2.2. These two numbers reveal the fact that the rate of childbearing is almost the same for both cohorts within the first six years after they have their first child. However, the less than 17 year-old cohort have a harder time meeting their goal of having an average of 2.4 children 11 years after school since their actual number is 3.1. On the other hand, the preferred number of children 11 years after high school for the 20-24 year old cohort is at 2.6, which is close to their actual number of 2.2.

Exceeding the preferred number of children for the less than 17 year-old cohort can present a problem. Since, as adolescent parents, it is more difficult for them to find prestigious jobs that pay well, they would have a harder time budgeting their income for

⁴ Ibid

⁵ Hoffman SD, Foster EM and Furstenberg FF Jr. (1993)

⁶ Card JC and Wise LL (1978)

their families. They might have to force some of their children to go out and work so that they would have enough money to buy them food and other basic necessities.

Poor birth outcome

Early child bearing increases the risk of many complications in the infants. *Olausson et al.* showed that the risk of infant mortality is inversely related to the age of the mother; that is the earlier the mothers give birth to the children, the more likely it is for them to have a high mortality rate.⁷ The researchers looked obtained the data from medical birth register which contained the information on child births, demographic data, maternal medical complications during pregnancy, and infant complications after birth. The researchers defined neonatal death as death that happened before 27 days after birth and postneonatal death as death that happened between 28 days and one year after birth. They found a correlation between the risk of neonatal death and the age that the mothers gave birth. For the 13-15 year-old cohort of mothers, the rate of neonatal mortality was 14.5 per 1,000 while the rates for 16-17, 18-19, and 20-24 year-old were 7.6, 5.5, and 4.6 per 1,000, respectively. This showed that the neonatal mortality rate for 13-15 year-old cohort was three time that of 20-24 year-old cohort. The trend for postneonatal death was also similar to neonatal death with the highest rate for 13-15 year-old cohort, except that the differences between cohorts were less apparent.⁸ The researchers proposed that biological factor played a big role in causing poorer mortality rate among the younger cohorts. Since teenagers were still having growth spurt, their bodies required more nutrients than adults. This greater nutrient requirement may have diverted the nutrients away from the womb. Without getting adequate nutrients, these children were less like to

⁷ Olausson PO, Cnattingius S, Haglund B (1999)

⁸ Ibid

successfully survive past the neonatal and postneotal stages. Therefore, poor clinical outcome among children is also directly related to early child bearing.

High cost to society

An article published in 1985 by Martha R. Burt attempts to estimate the cost of teenage childbearing to the government.⁹ The three main public goods that the author uses to measure the cost are Aid to Families with Dependent Children (AFDC), food stamps, and Medicaid. In total, the teenage mothers received about \$17 billion in aid while the total amount available is \$31.40 billion. Teenage mothers alone absorb nearly 53 percent of total public expenditures. The researchers also performed a cross-sectional analysis of the annual AFDC and food stamp benefits in billions of dollars received for women who had a teen birth and had no teen births. According to their results, the amount of aid received by mothers who had a teen birth is significantly higher than those with no teen birth in all age groups. Between ages 17 and 20, the graph for mothers who had teen birth exponentially increases due to the fact that these mothers tend to drop out of high school and college after they gave birth to their child. So between these ages, the mothers rely heavily on social benefits as their source of income. However, after their child has grown up, the mothers enter labor force again at the age of around 26-27, which is consistent with the earlier finding. This is shown by a decline in amount of aid received after the age of 26.

Trends

⁹ Burt MR (1986)

The rate of teenage pregnancy has been on the rise from 1980, peaked in 1990 and dropped down, as shown in figure 1. For the 15-17 year-old cohort, it was estimated that there are about 77 pregnancies for every 1,000 children in year 1980. This number rose to 80.3 in the year 1990. However, the pregnancy rate experienced a sharp decline from year 1990 to 1997 in which it dropped down to 63.7, representing a 21 percent decline. The trend for 18-19 year-old cohort also represents a similar trend in which it peaked at 183.4 per 1,000 women in the early 1990's and dropped down to 141.7 in 1997.¹⁰

General Factors Causing Teenage Pregnancy

In the article written by *Kandel et. al.*, the researchers examined the trend for drug and alcohol usage among teenagers. In this study, the researchers did a survey on a sample of New York public high school students regarding their behavior of drug usage.¹¹ The study breaks down the term “drugs” into four main categories, which are alcohol, marijuana, cigarettes, and other illicit substances. They then plot the proportion of drug usage versus age in which the teens use the drugs, as shown in figure 2. The plot for both alcohol and marijuana have the same trend in which the teens start using these two substances at around the age of 15, peak at the age of 20 and then taper off after that. The trend for cigarettes is slightly different in a way that it keeps on increasing after the age of 20. The graph suggests that once the teens start using substances, they tend to keep on using it more and more, suggesting that the initial usage causes the rise in subsequent usage. Also in a study conducted by *Rosenbaum et. al.*, the researchers were able to determine that drug and alcohol usage in childhood strongly correlates with the

¹⁰ Ventura SJ, Mosher WD, Curtin SC, and Abma JC (2001)

¹¹ Kandel DB and Logan JA (1984)

likelihood of having sexual precocity.¹² When other factors are controlled, early involvement with alcohol, cigarettes, marijuana and other illicit substances significantly increase the risk of having an intercourse before the age of 16, the earlier the involvement, the higher the risk. Not only does using substances during childhood promotes subsequent usage as the child gets older, it also increases the likelihood that the child will engage in sexual activity. This means that an early involvement with one risky activity promotes an involvement with another risky activity.

In a study conducted by *Young et al.*, the researchers tried to determine the various factors that contribute to teenage pregnancy.¹³ With the data from National Education Longitudinal Study, researchers performed logistic regressions on the participant's education background and aspiration, parents' education, family income, and other key socioeconomic status indicators. They found that girls who are confident in getting at least a high school diploma are associated with being pregnant later on in their lives. This is in contrast with the group of girls who do not feel confident in graduating from high school due to various reasons; this group of girls would in fact report being pregnant before they graduate. Also, girls whose parents report having low education are usually pregnant before girls whose parents are highly educated. Moreover, the researchers also found that teenage girls who come from low income families are at a higher risk of getting pregnant than those who come from rich families. From this study, one can conclude that girls whose families are of lower socioeconomic status have higher chance of getting pregnant.

Data

¹² Emily R and Kandel DB (1990)

¹³ Young T, Turnet J, Denny G et al (2004)

The data used in the analysis was taken from the 1979 National Longitudinal Survey of Youth conducted by the Bureau of Labor Statistics, or BLS. It was a survey of 12,686 individuals with ages between 14 and 21 in 1979. These individuals were asked with the same set of interview questions every year from 1979 to 1994 and every two years from 1996 until present. The interview questions were geared towards labor force experiences, education, income, other socioeconomic status indicators, and many other information pertain to each individual's behaviors and characteristics. The samples were also heterogeneous in that they vary in terms of race, ethnicity, and family background.

For the thesis, since the emphasis is on teenage pregnancy, the only type of data extracted from the database is for female responses. Also, the range of the ages of children employed in the data analysis is between 14 to 16 year-old when they were interviewed in 1979. To capture the factors and their effects on teenage pregnancy, the data from the year 1982 were extracted from the database, giving the age range between 17 to 19 year-old, capping the upper bound of normal teenage years. Cross-sectional analysis was performed on this sample because of the unavailability of panel-data due to privacy issue. Specifically, if panel data were made available to the public, the public would be able to deduce who these interviewees were through either logical reasoning or using computer programming, thus exposing their anonymity. So to keep their identity confidential, such data was not disclosed by BLS.

Descriptions of variables

There are a total of seven main variables used in the analysis are covered in the survey, which are pregnancy, education level, contraceptive use, alcohol abuse, family income, job, and race.

Pregnancy is covered by the question “have you ever been pregnant?” If the respondents have, then they were requested to put down number 1. If they have not, then they were requested to put down number 0. This means that the dependent variable is a binary dummy variable, suggesting that specific regression model needs to be used to run the analysis. Moreover, if the respondents put down anything below 0, then that means they either refused to answer the question or skipped it all. Such answers can be coded as missing data.

Education level is a commonly used indicator of an individual’s socioeconomic status and is captured by the question “number of years in school.” Such question includes such answers from first grade, which is coded as 1, to 4th year in college, which is coded as 16. Given the scope of the age of children who took part in the interview, having either 3rd or 4th year college level education is very rare since the oldest age is only 19 year-old. Also, any responses that are below 0 can be considered as missing data.

Since the level of education varies with the age of respondents, an adjusted education (adjedu) variable is created to take into account this difference; that is to take into account of the fact that it is natural for an older respondent to have higher level of education.

Contraceptive use is covered by the question “do you usually use contraceptive when you have sex?” The responses are binary, number 1 as yes and 0 as no. In this case, binary dummy variable has to be employed in the regression models to account for such responses. Responses below 0 are either refusing to answer, do not know, or invalid answers and they can be treated as missing variables in the analysis.

Childhood alcohol abuse is measured by the question “how many cans of beer do you consume per week?” Even though there are many other types of alcoholic beverages available, the most easily accessible one is beer since they are available at all most all liquor stores and even some supermarkets in certain states. Hence, looking at the number of cans of beer consumed gives a relatively accurate picture of childhood alcohol abuse and its correlation with teenage pregnancy. Respondents can put down any number of cans of beer that they think they consume per week on average. Any responses below 0 would be regarded as missing data.

Family income is also a socioeconomic status indicator, along with education, and is captured by the question “total net family income per year.” Family income basically measures the financial resources that people in the household, particularly the children, gain an access to. There are studies showing that family income is related to the well-being of one’s socioeconomic status and health.¹⁴ Hence, it is important to control for this factor in any social science or public health analysis. Respondents were asked to indicate an estimate of their net family income per year in their responses. Any responses below 0 would be treated as missing data.

Job is measured by the following question in the survey: “how many hours per week do you spend on working at your current job?” The respondents were then asked to indicate the approximate number of hours worked per week on their answer sheet. Any answer below 0 means the respondents refused to answer, did not the answer, or skipped the question all together. Such answers can be treated as missing variables in the data analysis. The average number of hours reported were 17.29, with a median of 17.

¹⁴ Sturm R and Gresenz CR (2002)

There can also be differences between the number of hours worked for those who do not go to college after finishing high school and those who go to college after finishing high school. Hence, a binary college variable, after adjusting for age differences by using the adjusted education variable, is created to take into account of this fact.

Race is an important variable that needs to be controlled for in such analysis. Many studies suggest a different level of health status among individual of different races so this factor must be included in the regression models as well.¹⁵ The question titled “racial/ethnicity” captures this effect. Respondents were asked to put down the number 1 if they are Hispanic, 2 for Black, and 3 for White. When used in the regression, such factor needs to be broken down into two newly created binary variables, namely Hispanic and Black. These Hispanic and Black variables are then included in the model just like normal binary variables. Thus, a Hispanic person would have a 1 for the Hispanic variable and 0 for black variable. A Black person would have a 1 for the black variable and a 0 for Hispanic variable. A white person would just have both variables as 0 because that is the baseline data that the model uses to compare the racial difference with. A race variable capturing the racial difference might also be used as well but this would just capture the difference between a group of minority as a whole and white, leaving out the differences among minority group itself.

Missing Data

Due to some missing data, which can almost always happen in social science research, some data adjustments must be made to accommodate such issue. Table 1 contains the raw data as extracted from the NLSY database. As mentioned earlier, this

¹⁵ Williams DR, Mourey DL, and Warren RC (1994)

data is for female children whose age were between 17 to 19 year-old by the time of their interviews in 1982. This raw data contains 1979 observations for each variable in total, meaning a total of 1979 responses were recorded for female children whose age were between 17 to 19 year-old. However, this raw data cannot be used in the regression models right away since it still contains missing responses.

There are many ways that one can adjust for missing data but two most common ways are omitting the observations missing one or more data and substituting in the mean of the reported data for the missing data.¹⁶ Table 2 shows the adjusted values for missing data for each variable. In this case, the dependent binary variable, which is the pregnancy variable, was missing 373 observations. Since it is a dependent variable, the most effective way of dealing with missing data is to omit them. This gives a total of 1606 observations, which is still number of observation.

The education variable was missing a total of 572 observations. To account for missing data, the average of the available number of years of education was calculated and substituted in for the missing data. This procedure was to ensure that not too many observations would be omitted due to missing data and shrinking the sample size. One characteristic about this procedure is that it does not add any new information. The overall mean of the variable is still the same for both before and after substitution. However, this procedure should still be used because the total observation would still be preserved. Also in this study, this procedure will be performed for all missing continuous variables for consistency of the model. That is none of such variable will be dropped from the model.

¹⁶ Allison, P.D. (2001)

The contraceptive variable was missing 1536 observations. Since it is a binary dummy variable, with the number 1 as using and 0 as not using, it is wise to just drop the missing observations for now, giving a total of 443 observations. The issue with the shrinking of sample size when this variable is run in regression models will be explored later on. If the inclusion of this variable results in a substantial drop in the number of observations, it will be dropped out of the model.

The beer variable stands for the number of cans of beer consumed per week on average. This variable was missing 1377 observations. Since it is a continuous variable, one can substitute in the mean of the available data, which is 3.36 cans. This procedure helps preserve the number of observations for this variable.

The income variable is for the family income of the respondent. It was missing 895 observations. These observations were substituted with an income average of 23580.9. It is interesting to note that this average is very close to 1982 average of 24309.0, suggesting that the sample data is a representation of the national data.¹⁷

The job variable measures the number of hours worked per week on average. This variable was missing 786 observations in total. Again, since it can be considered as a continuous variable, the average number of hours, which is 26.64, can be substituted in to preserve the number of total observations.

Analysis

Initial Analysis

Pairwise Correlation

In doing an analysis for social science research, it is always wise to check for the degree of correlation for between each independent variable. To perform such correlation, Pearson correlation coefficient must be calculated. This correlation coefficient is a strong tool in performing such analysis because it can be used independently of the unit of measurement of each variable.¹⁸ The coefficient value ranges from -1 to 1, with -1 as a perfect negative correlation and +1 as a perfect positive correlation. Note that correlation is different from causal effect. For example, the causal effect would be A causes B but correlation just means A and B are associated with each other, whether it be A causes B or B causes A. The correlation analysis is important because the correlation among independent variables can result in the biasness of the coefficient estimation in the regression models.

Table 3 shows the results from Pearson pairwise correlation analysis with the associated p-value. All of the independent variables are positively correlated with dependent variable, except for hispanic variable which shows a negative correlation. When looking at the p-values, it turns out that only the only significant correlation is between pregnancy and job variable. This is consistent with the finding from regression analysis since the only significant independent variable in the fitted models is also the job variable.

Education variable is shown to be positively correlated with beer, income, and job variables while negatively correlated with hispanic and black variables. The positive correlation, even though it is not significant from looking at p-value, between education and beer variables is surprising because one would expect an individual with higher education level to be less addicted to drinking. The positive correlation between

¹⁸ Moradi, A (2007)

education and family income variables is not as surprising because one would expect a family with high income to be able to successfully support the children's education. This evidence is supported by a significant p-value. Also, with higher level of education, the children can utilize more of their knowledge in their occupation, resulting in the higher number of hours worked per week. The negative correlation for both hispanic and black minority groups suggest that minorities attain lower level of education.

From table 3, none of the beer, income, job, hispanic, and black variables have significant correlation among each other due to high p-values. This shows that most of the independent variables are not associated with one another except for income and education. This result provides a support that the regression models with these variables are relatively unbiased, and that the biasness due to correlation between independent variable comes from the correlation between education and income variables.

Also, the two extra variables, which are adjusted education and college variable, must be included in the correlation test. It is not surprising to see that adjusted education as well as college, which is a function of adjusted education, variables are significantly correlated with income variable. This is mainly due to the fact that education variable itself and income are significantly correlated, as mentioned above. Of course, it is also expected that both adjusted education and college variables are significantly correlated with the given education variable since both variables use the given education variable as a baseline for adjustment. Furthermore, college and adjusted education variables are correlated since college variable is a function of adjusted education variable. None of these two variables are significantly correlated with any other variables. This suggests

that the bias in the regression models is not caused by the introduction of these two variables.

Two-Sample T Test

In performing statistical analysis, it is crucial to determine whether the difference in the mean of the two populations is significant or not. If such difference is statistically significant, then the data used to run the regression models can give biased results. To see whether such difference is significant, two-sample t test must be used. In this case, the mean of the regression variables between the group of people with missing and non-missing pregnancy data, and those who reported and did not report ever being pregnant should be compared.

Table 4 shows the comparison of the mean of the group of people with missing and non-missing pregnancy data as well as their p-value after two-sample t test was performed. This analysis is important because if the difference is significant, then the data from the non-missing group is not a good representation of the data at national level; there is homogeneity in the data among the group of people who responded to the pregnancy question in the survey. Upon a close examination, the differences that are significant are for the contraceptive, income, and black variables. People who responded to the pregnancy question tend to report using contraceptive and have a higher average income than those who did not respond. The majority of the people who did not respond to the question also belong to the black minority group. Table 5 shows the comparison of the mean of the variables among those who reported ever being pregnant and non-pregnant. There are slight differences among all the variables but none of them are

significant. This means that the data used in the regression analysis is heterogeneous enough and there is no bias in terms of the data being congregated around certain values.

Table 6 shows the comparison between the group of people who responded to the question regarding the use of contraceptive. This comparison is important because there are a lot of people who did not respond to this question so any distinct differences in the data between those who did and did not respond must be noted. From the table, one can infer from the p-value that the differences in education level and family income are significantly different between these two groups. The difference between the level of education is at 2.2% significant level while the difference between family income is at 1,6% significant level. Therefore, the inclusion of this variable for the non-missing data cases can cause biasness to a certain degree, especially for the education and income variables.

Regression Analysis

Several regressions were used to fit the adjusted variables for missing data. All regression models were run on Intercooled Stata Version 9.2 statistical package, available on the computers at Johns Hopkins University Computer Lab.

Sample Regression

In beginning the regression analysis, sample ordinary least square regression was used to obtain a rough estimate of the effect of independent variable on dependent variable. The following equation was employed in the first sample model:

$$\text{Pregnancy} = B_0 + B_1 * \text{edu} + B_2 * \text{contraceptive} + B_3 * \text{beer} + B_4 * \text{income} + B_5 * \text{job} + B_6 * \text{hispanic} + B_7 * \text{black} + \text{residual}$$

The variables were selected based on the various variables used in the mentioned studies above with job variable added in to test the hypothesis that the number of hours worked per week should help reduce the probability of getting pregnant. The education, contraceptive, beer, and income are used to as controls for analysis done by prior studies. It was found that education, contraceptive, and family income are negatively correlated while the rate of alcohol abuse is positively correlated with teenage pregnancy. The race variables are often used in many public health researches as control for any racial differences.

Since the pregnancy variable is a binary variable, the model can be considered as a linear probability model. The coefficients of each variable used in this model give the change in the probability of the respondent being pregnant when the associated variable marginally increases. Table 7 gives the coefficients of each variable with their p-value; the lower the p-value, the higher the significance level. From examining the results of the first sample model, the income variable is the only one that is significant with p-value of 0.012. This means that the variable is significant at 1.2%. Though it seems like this model can be used to fit the data, the number of observations suggest otherwise. The original dataset contained 1979 observations in total, but when the first sample model was run, the number of observations was 133, or only about 6.7% of the original dataset. Such a stark reduction in the number of observations can cause the sample data to deviate from national data, introducing biasness into the model. A potential variable that causes such a drop is contraceptive, which is missing 1536 observations.

In the second sample model, contraceptive was dropped to see if the number of observations would increase. Remarkably, the number of observations rise to 1606, an increase of 1473 observations. Since 1606 observations is close to 1979 observations of the raw data, dropping contraceptive variable would help make the adjusted sample more representative of the national data. When the contraceptive variable is dropped, the p-value of income variable decreases to -0.262, making the variable no longer significant. However, the job variable becomes significant with a p-value of 0.026. This means that the job variable becomes significant at 2.6%. The fact that the job variable becomes significant after the exclusion of this variable might be due to the fact that the differences of some of the variables are statistically significant between the group of people with missing and non-missing contraceptive data. As mentioned above, the differences are significant for education and income variables. So the inclusion of this variable restricts the model to the data reported by those with non-missing contraceptive data, which is skewed around certain values and can be bias. However, it should be noted that the dropping of this variable may also introduce biasness to the model in another way in that it is commonly known contraceptive use is usually negatively correlated with pregnancy. Still, such drop must be performed to increase the number of observations.

Linear Probability Model

As stated earlier, dropping contraceptive variable is necessary for the number of observations to be large enough to represent the national data. In the linear probability model analysis, three different variations were used to fit the data. The results are shown in table 8.

The first variation of linear probability model includes the following variables: education, beer, income, job, hispanic, black, and education². The contraceptive variable is left out of the model to preserve the high number of observations as stated previously. The mean of the data is also substituted in for all missing data that are continuous variables to preserve the number of observations as well as the consistency in data substitution. The squared term of education variable is introduced into the model to capture nonlinearity effect of level of education on the change in probability of an individual being pregnant, that is to capture the increasing or decreasing marginal effect. That is if the education term is positive and if the education² term is also positive, that means that the effect is increasing at an increasing rate. However, if the education term is positive while the education² term is negative, that means that the effect of education on pregnancy is increasing at a decreasing rate. This is also true on the opposite. If the education term is negative but the education² term is positive, then the effect is decreasing at an increasing rate and vice versa. As for the hypothesis of this variable, since the knowledge obtained from higher level of education can be accumulated over time, it can be inferred that more knowledge at one point promotes even further knowledge at subsequent points. Therefore, the education² variable should be positive. With the mentioned education variable already hypothesized to be negative, that means that higher education should reduce the probability of getting pregnant at an increasing rate. After the quadratic term of education variable is introduced, only the job variable becomes significant with a p-value of 0.027.

The second variation of linear probability model includes the following variables: education, beer, income, job, hispanic, black, and the interaction term between income

and beer. The interaction term is included to capture the fact that a beer consumer having a certain level of income might have a different level of likelihood of getting pregnant when comparing to another beer drinker having different level of income; thus one should expect this term to be associated with an increase in probability. In this variation, the coefficient of education variable reduces to 0.001 from the first variation, though the variable remains insignificant from the p-value. Also, the job variable is still significant even though the p-value increases slightly 0.029 while the coefficient still remains fixed at 0.001.

The third variation of linear probability model includes the followings: education, beer, income, job, hispanic, black, education², and the interaction term between income and beer. In this variation, the education level coefficient goes back up to 0.022 while the variable itself still remains insignificant. The job coefficient remains rigid at 0.001, but the p-value goes up to 0.03. The fixture of the job coefficient at 0.001 among all three variations of the model suggests a strong evidence for a positive correlation between the number of hours worked per week and the change in the probability of a teenager getting pregnant. Also, when comparing the p-value of the job variable, the first variation of the model has the lowest p-value. This suggests the fact that the model containing variables from the first variation makes the job variable the most significant among all three variations.

In order to account for heteroskedasticity, or the violation of the constant variance of the error term, heteroskedasticity-robust procedure must be employed.¹⁹ This procedure will correct for heteroskedasticity, whether or not the error term violates constant variance assumption and always work for large sample size. By performing

¹⁹ Wooldridge, J.M. (2005)

heteroskedasticity-robust procedure, heteroskedasticity-robust standard and t statistic would be obtained as well as the adjusted p-value of each variable. In doing the analysis, heteroskedasticity-robust procedure was performed on the three variations of the mentioned linear probability model. The results are shown in table 9.

The first variation of the robust linear probability model includes education, beer, income, job, hispanic, black, and education² variables. In this variation, the job variable is still the only significant variable. When compared, with its non-robust counterpart, the p-value decreases from 0.027 to 0.019, making the job variable slightly more significant when heteroskedasticity is adjusted for. The job coefficient still remains at 0.001, whether heteroskedasticity was adjusted for or not.

The second variation of the robust linear probability model includes education, beer, income, job, hispanic, black, and the interaction term between income and beer. Again, the job coefficient remains at 0.001 but the p-value increases slightly to 0.022. The beer variable coefficient increases slightly to 0.006 but its p-value still suggests that this variable is insignificant. As a matter of fact, all the coefficients of both robust and non-robust models are the same, but the p-value changes, however, not to the extent of making any other variables significant.

The third variation contains all the variables except for contraceptive, which is dropped to increase the number of observations. In this composite third variation, the only significant variable is the job variable, with a coefficient of 0.001 and p-value of 0.021. There is a drop of 0.009 in job variable p-value when compared with the third variation of non-robust model. The coefficients of other variables are still the same as that of the third variation of non-robust model as well.

There are some drawbacks in using the linear probability model to fit binary dependent variable. The main problem is the coefficient only shows a linear relationship between independent and dependent variables. However, the probability does not always increase linearly for a given change in independent variable. Therefore, another model must be used to fit the data.

Probit Model

In order to capture the nonlinear form of the independent variable coefficient, probit model can be employed. The general form of this model is given by:

$$P(Y=1) = \Phi(B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots)$$

The dependent variable in this case is still the binary pregnancy variable, Φ is the cumulative normal distribution function, and B's are the coefficients associated with each independent variable. The S-shape curve of this equation means that the nonlinearity form of probability is accounted for. The data can easily be fitted using this model on Stata. The results are reported in table 10.

Three variations of probit model were run to assess the significance of each variable. The first variation contains the following variables: education, beer, income, job, hispanic, black, and education². Heteroskedasticity-robust procedure was also performed on this variation to account for heteroskedasticity, with results in table 11. Since interpreting the coefficients of probit model is complicated and can turn out to be impractical, dprobit command as well as the robust procedure were performed on Stata was used to see the effect of marginal change in each variable on the probability, as

reported in table 12 and 13 respectively. Dprobit allows one to interpret the coefficients as the effect of marginal change for each variable on the probability of an individual getting pregnant. The robust version of this model is also performed to account for any heteroskedasticity.

Job variable is significant in the first variation for both probit and dprobit as well as their heteroskedasticity adjusted robust version. The p-values are 0.027, 0.014, 0.027 and 0.014 for probit, robust probit, dprobit, and robust dprobit respectively. To look at the marginal effect of the job variable, its coefficient from dprobit must be looked at. Both the non-robust and robust dprobit gives a coefficient of 0.001. This means that a marginal increase in the job variable, or an increase in the number of hours worked per week by one hour, is associated with an increase in the probability of a teenager getting pregnant by 0.001, or an increase the chance of getting pregnant by 0.1%.

The second variation of the probit model contains the following variables: education, beer, income, job, hispanic, black, and the interaction term between income and beer. The p-values are 0.028, 0.017, 0.028, 0.017 for probit, robust probit, dprobit, and robust dprobit versions. The job coefficient for both non-robust and robust dprobit is 0.001. None of the other coefficients for other variables are significant.

The education² variable is added to the second variation of the probit model to obtain the third variation. The p-values for all four versions of the third variation are 0.029, 0.015, 0.029, and 0.015. The p-values shows a slight decrease when robust procedure was performed on the model, suggesting that the variable becomes more significant once heteroskedasticity is accounted for. The job variable coefficient still remains fixed at 0.001. After performing probit and dprobit models as well as their

robust version, one can conclude that job variable is an important predictor for the pregnancy probability because of its statistical significance in all the models.

Final Regression

Robust probit model is used to run the final regression. This regression contains the following variables: $adjedu$, $adjedu^2$, $beer$, $income$, $income^2$, job , job^2 , $hispanic$, $black$, and $college$. Adjusted education variable is used to account for the fact that differences in level of education might be due to differences in age. The squared term of this variable is also introduced to capture any nonlinearity effect in the effect of adjusted education level on the probability of getting pregnant. It can then be hypothesized that since the squared term of unadjusted education is positive, the squared term of adjusted education variable should be positive as well for the same reason already mentioned. The squared term of the family income variable can be hypothesized to be positively associated with the probability of getting pregnant; that is the higher the family income, the probability of getting pregnancy decreases at an increasing rate. Also the squared term of the job variable can be said to be positively associated with the probability of getting pregnant; that it affects the probability at an increasing rate. Lastly, the binary college variable is used to differentiate between the group of respondents who did and did not go onto getting college level education, after adjusting for age. This difference must be controlled for in the regression model since an individual may end up working more hours if that individual do not go on to college after finishing high school. Table 14 shows the results of the final robust probit model. A dprobit model with the same

variables is also used to fit the data to obtain the marginal effect of each variable, with results shown in table 14. As before, even when age and the fact if an individual is in college are accounted for, job variable is the only one that is statistically significant enough at 4.7% level. The coefficient for the job variable also goes up to 0.003 from the coefficient of previous probit and dprobit models of 0.001. The inclusion of more squared terms as well as the controlling for age and college education differences makes the effect of job variable on the probability more pronounced.

Discussion

From looking the results from all the variations of the models, the following probit model best captures the effect of factors associated with teenage pregnancy:

$$\Pr(\text{pregnancy}=1) = \Phi(B_0 + B_1*\text{adjedu} + B_2*\text{adjedu}^2 + B_3*\text{beer} + B_4*\text{income} + B_5*\text{income}^2 + B_6*\text{job} + B_7*\text{job}^2 + B_8*\text{hispanic} + B_9*\text{black} + B_{10}*\text{college})$$

where Φ represents the cumulative normal distribution function. Probit model should be employed instead of linear probability model because the probability of being pregnant does not always increase linearly with the marginal increase of each associated variable. The adjusted education², income², and job² terms are added to the model while the interaction term between income and beer is dropped. With the p-value of 0.047, the job variable is said to be significant at 4.7% level. To assess for this variable's marginal effect on pregnancy, dprobit function on Stata was used on the model, giving a coefficient of 0.003. This means that for an increase in one hour of average number of

hours worked per week, the probability of a teenager getting pregnant increases by 0.003, or 0.3%.

From Pearson pairwise correlation results, the only variable that is significantly correlated with pregnancy is the job variable. With a correlation coefficient of 0.0525, these independent and dependent variables are positively correlated with each other, which is consistent with the regression results. However, the correlation test also shows that education and income variables are also significantly positively correlated with each other. The variables used to adjust for age and college education differences are not significantly correlated with any other variables, except for the ones that are used as baselines for adjustments. This means that the introduction of these variables do not cause bias on the regression models. From looking at two-sample t test, there is a significant difference in the level of contraceptive use and family income between the groups of people with missing and non-missing pregnancy data. There is also a racial difference in that black individuals tend to leave out such data. However, the differences among those who reported being pregnant and non-pregnant are insignificant. It can be concluded the data used for contraceptive, income, and black variables do not give a good representation of the national data and can cause a biased regression results because there are significant differences between the groups of people with missing and non-missing pregnancy data.

The regression of finalized model also yields different results from the hypothesis. Table 15 compares the hypothesis with the regression results, it can be easily seen that all of the variables in the finalized regression model are insignificant except for the job variable. The actual effect of the job variable is different from the hypothesis. One of

the explanations is that this variable acts on the dependent variable through unobserved factors. Teenagers who work more also earn more total salary. This figure is usually not factored into family income since teenagers tend to keep the amount of money that they earn to themselves; no other household members have access to this fund. With more financial leverage, the teens can engage in more risky activities. From one of the studies mentioned above, since one risky behavior causes another, these teenagers are more prone to getting pregnant. Since the model does not capture all the possible risky behavior among teens, it may seem as if higher number of hours worked is associated with an increase probability in getting pregnant.

This model is, however, not without flaws. The job variable might be affected by the fact that the higher number of hours worked per week is caused by pregnancy, not because pregnancy is a causal effect of the job variable. There is a possibility that teenagers who become pregnant work more hours to support for their forthcoming child, should they expect to have one. In social science and especially public health research, there are many related factors that are the causal effect of each other. For example, the health and socioeconomic status of an individual are often thought to be the causal effect of each other, that is low socioeconomic status causes poor health and vice versa. The job and pregnancy variables can also have such similar effect. Therefore, to only measure the causal effect of job variable on teenage pregnancy, the longitudinal data for the number of hours worked should be adjusted for both before and after pregnancy.

Conclusion

Teenage pregnancy and childbearing is one of the most important public health problems. Early childbearing has been shown to be associated with the low maternal

educational attainment, low family income, and poor health of the children due to the mother's biological unpreparedness. Therefore, it is important the factors that are associated with the increase in likelihood of getting pregnant in teenagers. In doing such an analysis, probit regression model is used to capture the effects of key socioeconomic status indicators on binary dependent pregnancy variable. From the regression result, it turns out that an increase in the number of hours worked is correlated with an increase in the probability of getting pregnant in teenagers. One explanation of this relationship is given by the fact that the higher number of hours worked, the more financial resources these teens have to engage in risky behaviors, such as substance abuse, gambling, etc. Thus, it can be inferred that the effect of working more hours affects the chance of getting pregnant through these risky behaviors. However, one still cannot consider this explanation to be totally valid due to the flaws in the model. Particularly, the data does not specifically contain information regarding whether the individuals become pregnant before or after they start working more hours. They might also be forced by their parents to work more hours, whether they are pregnant or not. Therefore, even though the model suggests that the job variable is statistically significant, one can only infer that there is a correlation between the number of hours worked per week and the probability of getting pregnant. More research should be devoted to elucidate the actual causal effect of this factor on pregnancy.

Figure 1: The rate of pregnancy by age from 1980 to 1997²⁰

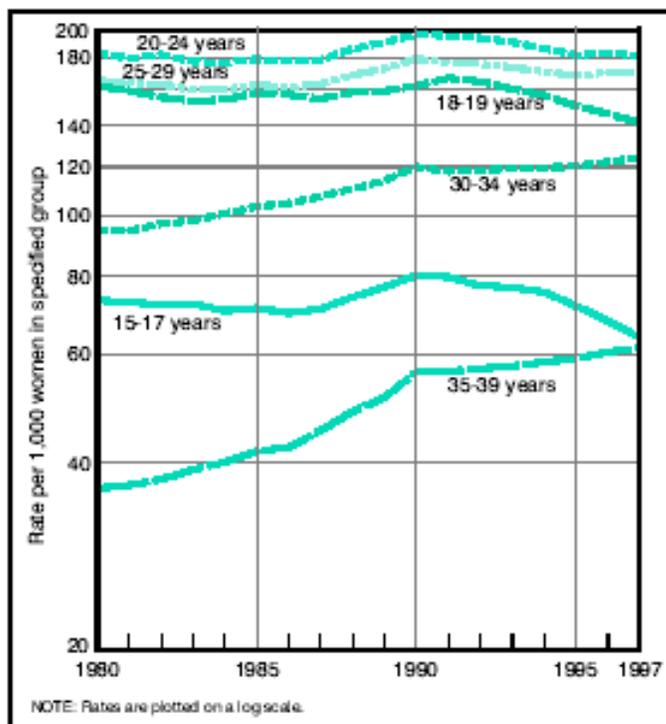


Figure 2: Period of Highest Use for Alcohol, Cigarettes, and Marijuana by Age as a Proportion of All Users²¹

²⁰ Ventura SJ, Mosher WD, Curtin SC, and Abma JC (2001)

²¹ Kandel DB and Logan JA (1984)

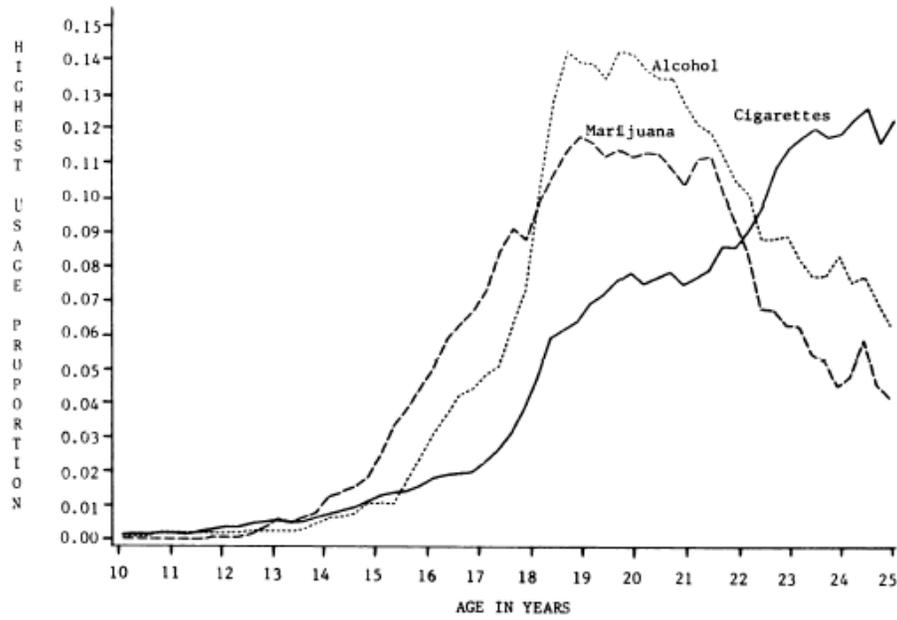


Table 1: The Raw Data

Variable	Observation	Mean	Median	Std. Dev.
pregnancy	1979	-0.74987	N/A	1.664958
education level	1979	6.948459	11	7.09992
contraceptive use	1979	-3.0667	N/A	1.826
no. of cans of beer consumed per week	1979	-1.78979	3	4.577639
family income	1979	12915.02	23000	16891.54
no. of hours worked per week	1979	17.29207	17	18.001
hispanic	1979	0.186963	N/A	0.389981
	1979	0.261243	N/A	0.439423

black

Table 2: The Adjusted Variable for Missing Data

Variable	Observation	Mean	Median	Std. Dev.
pregnancy	1606	0.043587	N/A	0.204237
education level	1979	11.4435	11.4435	0.809217
contraceptive use	443	0.291196	N/A	0.454828
no. of cans of beer consumed per week	1979	3.363787	3.363787	3.050579
family income	1979	23580.9	23580.9	12143.74
no. of hours worked per week	1979	26.6398	26.6398	11.13678
hispanic	1979	0.186963	N/A	0.389981
black	1979	0.261243	N/A	0.439423

Table 3: Pairwise Correlation Results

Variable	pregnancy	education level	no. of cans of beer consumed per week	family income	no. of hours worked per week	hispanic	black	adjusted education by age	college education
pregnancy	1								
education level	0.0148 0.5534	1							
no. of cans of beer consumed per week	0.047 0.0597	0.0317 0.1584	1						
family	0.0286	0.1449	0.009	1					

income	0.2516	(0.02)*	0.6876						
no. of hours worked per week	0.0525 (0.0353)*	0.0301 0.1814	-0.03 0.1826	-0.0249 0.2687	1				
hispanic	-0.0295 0.2378	-0.0555 0.135	-0.0213 0.3429	-0.0627 0.053	0.0669 0.0029	1			
black	0.0068 0.7852	-0.0168 0.4544	-0.0474 0.352	-0.2068 0.426	-0.0229 0.3076	-0.2852 0	1		
adjusted education by age	-0.0279 0.2645	0.7157 (0.001)*	0.0234 0.2973	0.1277 (0.035)*	0.1282 0.215	-0.0779 0.0897	-0.0216 0.2464	1	
college education	-0.0177 0.4775	0.6091 (0.002)*	0.0332 0.1396	0.1186 (0.026)*	-0.0702 0.0265	-0.0479 0.0753	0.0502 0.356	0.4314 (0.001)*	1

pvalues in 2nd line
* denotes 5% significant level

Table 4: The Comparison of the Mean Between Groups of People with Missing and Non-Missing Pregnancy Data

Variable	Mean		p-value
	Missing	Non-missing	
education level	11.34	11.47	0.155
contraceptive use	0.248	0.39	(0.0025)*
no. of cans of beer consumed per week	3.16	3.411	0.391
family income	21175.49	24139.57	(0.0014)*
no. of hours worked per week	28.222	27.272	0.0726
hispanic	0.21	0.18	0.2233

black 0.34 0.24 (0.0002)*
 * denotes 5% significant level

Table 5: The Comparison of the Mean Between Groups of People who Reported being Pregnant and Non-Pregnant

Variable	Mean		p-value
	Non-pregnant	Pregnant	
education level	11.46	11.52	0.2488
contraceptive use	0.413	0.371	0.6276
no. of cans of beer consumed per week	3.378	4.138	0.3455
family income	24060.76	25868.85	0.3801
no. of hours worked per week	26.14	28.12	0.0658
hispanic	0.1842	0.1285	0.2377
black	0.243	0.257	0.7851

* denotes 5% significant level

Table 6: The Comparison of the Mean Between Groups of People with Missing and Non-Missing Contraceptive Data

Variable	Mean		p-value
	Missing	Non-missing	
education level	11.47	11.37	(0.022)*
no. of cans of beer consumed per week	3.39	3.27	0.4665
family income	24163.15	21562.09	(0.016)*
no. of hours worked per week	27.23	28.158	0.2635
	0.1842	0.1963	0.5638

hispanic
 black 0.2596 0.26239 0.4569
 *denotes 5% significant level

Table 7: Sample Regression Results

Variable	Sample Model	
	[1]	[2]
education level	0.024	0.001
	-0.687	-0.834
contraceptive use	-0.056	
	-0.533	
no. of cans of beer consumed per week	0.024	0.003

	-0.07	-0.056
family income	0.0000011 (0.012)*	4.60E-07 -0.262
no. of hours worked per week	0.002 -0.579	0.001 (0.026)*
hispanic	-0.092 -0.445	-0.014 -0.319
black	0.068 -0.524	0.004 -0.737
education level ²		
Income*beer		
Constant	-0.137 -0.842	-0.016 -0.816
Observations	133	1606
R-squared	0.095	0.007
p values in the second line		
* significant at 5%; ** significant at 1%		

Table 8: Linear Probability Model Results

Variable	linear probability model		
	[1]	[2]	[3]
education level	0.024	0.001	0.022
contraceptive	-0.792	-0.81	-0.805

use

no. of cans of beer consumed per week	0.003 -0.055	0.006 -0.155	0.006 -0.156
family income	4.62E-07 -0.261	8.54E-07 -0.184	8.53E-07 -0.185
no. of hours worked per week	0.001 (0.027)*	0.001 (0.029)*	0.001 (0.030)*
hispanic	-0.014 -0.324	-0.014 -0.325	-0.014 -0.33
black	0.005 -0.719	0.004 -0.733	0.005 -0.716
education level ²	-0.001 -0.802		-0.001 -0.817
Income*beer		-1.23E-07 -0.427	-1.22E-07 -0.43
Constant	-0.143 -0.779	-0.028 -0.693	-0.145 -0.776
Observations	1606	1606	1606
R-squared	0.007	0.007	0.007
p values in the second line			
* significant at 5%; ** significant at 1%			

Table 9: Robust Linear Probability Model Results

Variable	robust linear probability model		
	[1]	[2]	[3]
education level	0.024	0.001	0.022

	-0.799	-0.803	-0.812
contraceptive use			
no. of cans of beer consumed	0.003 -0.157	0.006 -0.2	0.006 -0.2
family income	4.62E-07 -0.255	8.54E-07 -0.112	8.53E-07 -0.113
no. of hours worked per week	0.001 (0.019)*	0.001 (0.022)*	0.001 (0.021)*
hispanic	-0.014 -0.262	-0.014 -0.265	-0.014 -0.268
black	0.005 -0.725	0.004 -0.736	0.005 -0.722
education level ²	-0.001 -0.809		-0.001 -0.823
Income*beer		-1.23E-07 -0.318	-1.22E-07 -0.321
Constant	-0.143 -0.789	-0.028 -0.683	-0.145 -0.786
Observations	1606	1606	1606
R-squared	0.007	0.007	0.007

p values in the second line

* significant at 5%; ** significant at 1%

Table 10: Probit Model Results

Variable	probit model		
	[1]	[2]	[3]
education level	0.128 -0.897	0.015 -0.824	0.112 -0.91
contraceptive use			
no. of cans of beer consumed per week	0.024 -0.079	0.049 -0.269	0.049 -0.269
family income	5.03E-07 -0.249	8.27E-06 -0.229	8.27E-06 -0.229
no. of hours worked per week	0.011 (0.027)*	0.011 (0.028)*	0.011 (0.029)*
hispanic	-0.161 -0.332	-0.16 -0.334	-0.159 -0.338
black	0.06 -0.661	0.059 -0.666	0.06 -0.661
education level ²	-0.005 -0.908		-0.004 -0.922
Income*beer		-1.00E-06 -0.553	-9.96E-07 -0.554
Constant	-3.01 -0.594	-2.459 (0.002)**	-3.006 -0.595
Observations	1606	1606	1606

p values in the second line

* significant at 5%; ** significant at 1%

Table 11: Robust Probit Model Results

Variable	robust probit model		
	[1]	[2]	[3]
education level	0.128 -0.905	0.015 -0.824	0.112 -0.917
contraceptive use			
no. of cans of beer consumed per week	0.024 -0.068	0.049 -0.162	0.049 -0.163
family income	5.03E-06 -0.215	8.27E-06 -0.11	8.27E-06 -0.11
no. of hours worked per week	0.011 (0.014)*	0.011 (0.017)*	0.011 (0.015)*
Hispanic	-0.161 -0.325	-0.16 -0.33	-0.159 -0.331
black	0.06 -0.66	0.059 -0.661	0.06 -0.66
education level ²	-0.005 -0.915		-0.004 -0.927
Income*beer		-1.00E-06 -0.36	-9.96E-07 -0.363
Constant	-3.01 -0.627	(0.002)** -2.459	-3.006 -0.626
Observations	1606	1606	1606

p values in the second line

* significant at 5%; ** significant at 1%

Table 12: Dprobit Model Results

Variable	dprobit model		
	[1]	[2]	[3]
education level	0.011	0.001	0.01
	-0.897	-0.824	-0.91
contraceptive use			
no. of cans of beer consumed per week	0.002	0.004	0.004
	-0.079	-0.269	-0.269
family income	4.44E-07	7.30E-07	7.29E-07
	-0.249	-0.229	-0.229
no. of hours worked per week	0.001	0.001	0.001
	(0.027)*	(0.028)*	(0.029)*
hispanic	-0.013	-0.013	-0.013
	-0.332	-0.334	-0.338
black	0.005	0.005	0.005
	-0.661	-0.666	-0.661
education level ²	-0.00044		-0.00037
	-0.908		-0.922
Income*beer		-8.82E-08	-8.78E-08
		-0.553	-0.554
Constant			
Observations	1606	1606	1606
p values in the second line			
* significant at 5%; ** significant at 1%			

Table 13: Robust Dprobit Model Results

Variable	robust dprobit model		
	[1]	[2]	[3]
education level	0.011	0.001	0.01
	-0.905	-0.824	-0.917
contraceptive use			
no. of cans of beer consumed per week	0.002	0.004	0.004
	-0.068	-0.162	-0.163
family income	4.44E-07	7.30E-07	7.29E-07
	-0.215	-0.11	-0.11
no. of hours worked per week	0.001	0.001	0.001
	(0.014)*	(0.017)*	(0.015)*
hispanic	-0.013	-0.013	-0.013
	-0.325	-0.33	-0.331
black	0.005	0.005	0.005
	-0.66	-0.661	-0.66
education level ²	-0.00044		-0.00037
	-0.915		-0.927
Income*beer		-8.82E-08	-8.78E-08
		-0.36	-0.363
Constant			
Observations	1606	1606	1606
p values in the second line			

* significant at 5%; ** significant at 1%

Table 14: Final Robust Probit Model Results

Variable	Robust Probit	Robust Dprobit
adjusted education level	-11.776 -0.55	-1.019 -0.55
adjusted education level ²	7.163 -0.599	0.62 -0.599
no. of cans of beer consumed per week	0.024 -0.066	0.002 -0.066
family income	1.64E-05 -0.225	1.42E-06 -0.225
family income ²	-1.53E-10 -0.427	-1.32E-11 -0.427
no. of hours worked per week	0.029 (0.047)*	0.003 (0.047)*
no. of hours worked per week ²	-3.39E-04 -0.169	-2.93E-05 -0.169
hispanic	-0.187 -0.251	-0.015 -0.251
black	0.048 -0.722	0.004 -0.722
college education	-0.11 -0.641	-0.009 -0.641

Robust p values in second line

* significant at 5%; ** significant at 1%

Table 15: The Comparison Between Hypothesized Effect and Actual Effect for each Variable

Variable	Hypothesized Effect	Actual Effect
adjusted education level	negative	insignificant
adjusted education level ²	positive	insignificant
no. of cans of beer consumed per week	positive	insignificant
family income	negative	insignificant
family income ²	positive	insignificant
no. of hours worked per week	negative	positive
no. of hours worked per week ²	positive	insignificant