Determinant of Birth Weight in the United States in 2018

Stephen Rout

Centre College

12 May 2021

I. introduction

Birth weight is one of most easily accessible traits of newborn children. Speaking from personal experience, I can say that if a child was born unusually large or small, that fact may follow them for years in family stories and gossip. But there is also evidence that a baby's birth weight may correlate with more than just idle chatter. UNICEF reports that children born at lower weights tend to have lower IQ and are more at risk for stunted growth and a variety of health conditions, including, ironically, obesity. In 2014, Figlio et al. found that children who weighed less at birth tended to have poorer school performance, and that this held true across socioeconomic classes, schools, and ethnicities. While more research is needed to confirm these impacts, there is enough evidence to warrant additional interest into the determinants of birth weight itself, which is this question I hope to address. I used data from the Center for Disease Control's National Center for Health Statistics. The dataset includes all documented births in the United States in 2018, along with a wealth of relevant information.

II. Literature Review

Though there is existing research into the determinants of birth weight, it is usually more focused than my project (for example, only seeking to establish a single determinant at a time), or else quite dated. That said, there is a substantial body of research into the impact of more specific factors, as well as on the determinants of birth weight in more specific populations – usually populations outside the US, or populations with specific medical conditions. In some cases, these limitations may reduce the applicability of the studies to my project here today.

Dougherty and Jones (1982) find that the baby's sex is significant to birth weight, with male babies weighing significantly more. The mother's height, weight, marital status, and the

gestation period of the pregnancy all also increase birth weight, while smoking during pregnancy is correlated with lower birth weight. Contrary to expectations, the study also finds that the parent's socioeconomic status does not influence birth weight. This conclusion in particular is not universal. For example, Manyeh et al. (2016) found that poorer women were much more likely to have lighter babies.

Verma et al. (2021) finds that the mother's age correlates positively with birth weight when the mother is over 30 years old. They also find that plurality – that is, the number of babies simultaneously gestating – is impactful, significantly reducing birth weight, and that lack of prenatal care reduces birth weight. The question of the impact of prenatal care is complicated by its self-selected nature: parents who seek it out are more likely to be conscientious of other facts impacting the development of the fetus. Liu (1998) finds that "the significance and scale of the bias depends crucially on specific data and cohorts of the population investigated", being overestimated in some cases and underestimated in others. Conway and Deb (2005) suggest that the difficulty is due to a specification error caused by grouping different types of pregnancies together. They argue that there are "complicated" and "normal" pregnancies, and that these distinct pregnancies will tend to receive different number of prenatal visits, and will end in different types of outcomes, on average. Conceptually, mothers who know their pregnancy will be complicated will likely put a great deal of effort into prenatal care, but in the end, it may not be sufficient, meaning that the care is still correlated with a negative outcome. In support of this theory, Conway and Deb find that prenatal care has a significant impact on the birth weight resulting from "normal" pregnancies, but that in the case of the worst "complicated" pregnancies it had no impact.

There is also evidence suggesting that the mother's relationship status and tendency to engage in risky sexual behavior can influence birth weight. Some sexually transmitted infections seem to cause low birth weight: Johnson et al. (2011) found that chlamydia was directly associated with low birth weight, and that gonorrhea was associated with preterm births, which is associated with low birth weights. A meta-analysis performed by Shah et al. (2011) finds that unmarried mothers are more likely to give birth to low-birth-weight children, but there is some evidence suggesting that this analysis is reductionistic. Bird et al (2000) argue that the critical variable is not marital status per se, but more specific characteristics of the mother's relationship, such as the length of the relationship and cohabitation, both of which influence the amount of support she receives during the pregnancy.

III. Model Specification

Dependent variable:

BIRTHWEIGHT = The weight of baby i at birth, in grams.

Independent variables:

MALE$_i$ = The sex of baby i. 0 if female, 1 if male.

CIG3$_i$ = Cigarettes smoked by the mother of baby i during the third trimmest of pregnancy.

MOTHERHEIGHT$_i$ = The height of the mother of baby i, in inches.

MOTHERWEIGHT$_i$ = The weight of the mother of baby i, in pounds.

MOTHERWEIGHT2$_i$ = The squared weight of the mother of baby i, in pounds.

UNMARRIED$_i$ = Dummy variable for the marital status of the mother of baby i. 0 if the mother is married, 1 otherwise.

GESTATIONTIME$_i$ = Total gestation time of baby i, in weeks.

MOTHERAGE$_i$ = The age of the mother of baby i, in years, with all ages above 50 being counted as 50.

MOTHERAGE2$_i$ = The squared age of the mother of baby i, in years, with all ages above 50 being counted as the square of 50.

PLURALITY$_i$ = The number of babies simultaneously gestating in the mother of baby i.

PRENATALVIS$_i$ = The number of prenatal care visits made by the mother of baby i during the pregnancy.

PRENATALVIS2$_i$ = The square of the number of prenatal care visits made by the mother of baby i during the pregnancy.

STI$_i$ = A dummy variable, equal to 1 if the mother is infected with chlamydia or gonorrhea, 0 if she is not, or if her infection status is unknown.

TABLE 1: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
MOTHERAGE	3,619,321	29.01134	5.797713	12	50
UNMARRIED	3,176,221	.3964919	.4891688	0	1
PRENATALVIS	3,619,321	11.3512	4.15169	0	98
CIG3	3,619,321	.452331	2.626685	0	98
MOTHERHEIGHT	3,619,321	64.12337	2.833879	30	78
MOTHERWEIGHT	3,619,321	188.4282	41.35341	100	400
PLURALITY	3,619,321	1.033198	.1841301	1	5
GESTATIONT~E	3,619,321	38.67274	2.208985	17	47
BIRTHWEIGHT	3,619,321	3278.754	558.3446	227	8165
MALE	3,619,321	.5109229	.4998807	0	1
STI	3,619,321	.0198761	.1395745	0	1

IV. Expected Signs of Coefficients

MALE$_i$: Research shows that male babies tend to weigh more than female babies, so I expect the coefficient to be positive.

CIG3$_i$: Research shows that smoking during pregnancy increases the risk of a variety of health conditions in the fetus, including lower birth weight, so I expect the coefficient to be negative.

MOTHERHEIGHT$_i$: Research shows that taller mothers tend to give birth to heavier babies, so I expect the coefficient to be positive.

MOTHERWEIGHT$_i$ and **MOTHERWEIGHT2$_i$** = Research shows that heavier mothers tend to give birth to heavier babies. However, past a point higher weights are correlated with a variety of health conditions which may negatively impact birth weight. I expect the coefficient of **MOTHERWEIGHT** to be positive, and the coefficient of **MOTHERWEIGHT2** to be negative.

UNMARRIED$_i$ = Research shows that less support from the father is correlated with lower birth weights. While marriage is not a perfect representation of paternal support, it is the best one available to me. So, I expect the coefficient to be negative.

GESTATIONTIME$_i$ = The longer a baby spends in gestation, the more it will grow, and the more it will weigh when it is born. So, I expect the coefficient to be positive.

MOTHERAGE$_i$ and **MOTHERAGE2$_i$** = Research shows that births at particularly young and old ages are riskier and tend to produce lower birth weights. So, I expect the sign of **MOTHERAGE** to be positive, and the sign of **MOTHERAGE2** to be negative.

PLURALITY$_i$ = More simultaneously gestating fetuses means a higher chance of each individual fetus getting less resources, as well as less space for each one. So, I expect the sign of this variable to be negative.

PRENATALVIS$_i$ and **PRENATALVIS2$_i$** = Some researchers have argued that there are two types of pregnancies; “simple” pregnancies, which will likely end well with a modest number of prenatal care visits, and “complicated” pregnancies, which may end poorly even with a very large number of prenatal visits. So I expect the sign of **PRENATALVIS** to be positive, and the sign of **PRENATALVIS2** to be negative.

STI$_i$ = Research shoes that both gonorrhea and chlamydia in the mother can reduce birth weight, so I expect this coefficient to be negative.

The null and alternative hypothesis for each of these variables are:

Positive expected coefficient signs: MALE, MOTHERHEIGHT, GESTATIONTIME, MOTHERAGE, PLURALITY, PRENATALVIS, MOTHERWEIGHT

$$H0: \beta \leq 0$$

$$HA: \beta > 0$$

Negative expected coefficient signs: MOTHERAGE2, PRENATALVIS2, MOTHERWEIGHT2, STI, PLURALITY, UNMARRIED, CIG3

$$H0: \beta \geq 0$$

$$HA: \beta < 0$$

There are no variables with ambiguous expected coefficient signs.

V. Data Collection

The data is a set of approximately 3.8 million observations, collected from the United States' Centers for Disease Control and Prevention's National Vital Statistics System. It consists of every recorded birth in the US in 2018. The advantage of this approach is that the dataset is extremely large and relatively complete, but the downside is that it is impossible to reverse-engineer the dataset and add additional variables to individual observations. This means that I can only work with what the CDC has given me.

VI. Estimate and Evaluate the Equation

Model 1

$$\begin{aligned} BIRTHWEIGHT_i &= \beta_0 + \beta_1 MOTHERAGE_i + \beta_2 UNMARRIED_i + \beta_3 PRENATALVIS_i + \beta_4 CIG3_i + \\ &\beta_5 MOTHERHEIGHT_i + \beta_6 MOTHERWEIGHT_i + \beta_7 PLURALITY_i + \\ &\beta_8 GESTATIONTIME_i + \beta_9 MALE_i + \beta_{10} STI_i + \varepsilon_i \end{aligned}$$

The preliminary model includes the original, linear form of all dependent and independent variables. The regression results are displayed in TABLE 2. All coefficients are significant at well beyond the 1% level, with the "worst" t-statistic having an absolute value of

20. The Adjusted R^2 is not fantastically high, but given the tremendous number of potential determinants of something as complicated as human physiology, and the measure's lack of critical importance, it's value of .3330 is not alarmingly low.

TABLE 2: Model 1 Regression

```
      Source |       SS           df       MS      Number of obs   = 3,176,221
-------------+----------------------------------   F(10, 3176210)  >  99999.00
       Model |  3.3212e+11        10  3.3212e+10   Prob > F        =    0.0000
    Residual |  6.6515e+11 3,176,210  209416.426   R-squared       =    0.3330
-------------+----------------------------------   Adj R-squared   =    0.3330
       Total |  9.9727e+11 3,176,220  313980.976   Root MSE        =    457.62

------------------------------------------------------------------------------
  BIRTHWEIGHT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    MOTHERAGE |   2.630362   .0484563    54.28***0.000     2.535389    2.725334
    UNMARRIED |  -84.06116   .5781589  -145.39***0.000    -85.19433   -82.92799
  PRENATALVIS |   6.023358   .0638068    94.40***0.000     5.898299    6.148417
         CIG3 |  -10.55395   .0931448  -113.31***0.000    -10.73651   -10.37139
 MOTHERHEIGHT |   15.56967   .0977018   159.36***0.000     15.37817    15.76116
    PLURALITY |  -617.2057   1.441529  -428.16***0.000     -620.031   -614.3803
GESTATIONTIME |   104.6381   .1196589   874.47***0.000     104.4036    104.8726
 MOTHERWEIGHT |    2.11267   .0066502   317.68***0.000     2.099636    2.125704
         MALE |    124.888   .5138327   243.05***0.000     123.8809    125.8951
          STI |  -36.02472   1.789907   -20.13***0.000    -39.53288   -32.51657
        _cons |  -1699.334   7.873259  -215.84   0.000    -1714.765   -1683.902
------------------------------------------------------------------------------
```

Model 2

$$BIRTHWEIGHT_i = \beta_0 + \beta_1 MOTHERAGE_i + \beta_2 MOTHERAGE2_i + \beta_3 UNMARRIED_i + \beta_4 PRENATALVIS_i + \beta_5 PRENATALVIS2_i + \beta_6 CIG3_i + \beta_7 MOTHERHEIGHT_i + \beta_8 MOTHERWEIGHT_i + \beta_9 MOTHERWEIGHT2_i \; \beta_{10} PLURALITY_i + \beta_{11} GESTATIONTIME_i + \beta_{12} MALE_i + \beta_{13} STI_i + \varepsilon_i$$

For the second model, I will apply my hypotheses about the behavior of the MOTHERAGE, PRENATALVIS and MOTHERWEIGHT coefficients; that is, that they will be best represented by a polynomial. The regression is presented in Table 3. Like Model 1, all coefficients are

significant beyond the 1% level. The change in form caused a very slight increase in the adjusted R^2; while it is satisfying to see it go above 1/3, I do not think the difference is very significant. My hypothesis for the shape of the function was correct in all cases; initially age, the mother's bodyweight, and prenatal visits are all correlated with increasing birth weight, but they do eventually reach an inflection point, and start to correlate with lower birth weights. For prenatal visits, the maximum point is at 19.56 visits, which suggests that pregnancies requiring more than 20 prenatal care visits tend to be "complicated". Being correlation, this does not necessarily mean that more than 20 visits is not useful for "easy" pregnancies. It is also possible that mothers having "easy" pregnancies simply tend not to attend more than 20 prenatal care visits. For age, the maximum point is at 36.80, suggesting that by the time a mother is 37, age turns from an advantage into a disadvantage. Finally, for weight the maximum is at 282.45 pounds, suggesting that weight above 282 pounds tends to reduce child birth weight. Comparing specification 1 and 2 is made difficult by the degree of significance in both specifications. Given these difficulties, I performed the AIC and BIC tests on the two models.

Model 1 AIC and BIC:

```
Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |  3,176,221  -2.46e+07  -2.40e+07      11   4.79e+07   4.79e+07
------------------------------------------------------------------------------
```

Model 2 AIC and BIC:

```
Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |  3,176,221  -2.46e+07  -2.39e+07      14   4.79e+07   4.79e+07
------------------------------------------------------------------------------
```

The test was not particularly illuminating. However, given the fact that model 2 has stronger theoretical support, I will be moving on with the exponential forms of MOTHERAGE, PRENATALVIS, and MOTHERWEIGHT.

TABLE 3: Model 2 Regression

Source	SS	df	MS	Number of obs	= 3,176,221
Model	3.3895e+11	13	2.6073e+10	F(13, 3176207)	> 99999.00
Residual	6.5832e+11	3,176,207	207267.462	Prob > F	= 0.0000
Total	9.9727e+11	3,176,220	313980.976	R-squared	= 0.3399
				Adj R-squared	= 0.3399
				Root MSE	= 455.27

BIRTHWEIGHT	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MOTHERAGE	11.94914	.3786736	31.56	0.000	11.20695	12.69132
MOTHERAGE2	-.1623465	.0063801	-25.45	0.000	-.1748514	-.1498417
UNMARRIED	-77.23285	.5864747	-131.69	0.000	-78.38232	-76.08338
PRENATALVIS	16.13499	.1531872	105.33	0.000	15.83475	16.43523
PRENATALVIS2	-.4123594	.0055499	-74.30	0.000	-.4232371	-.4014817
CIG3	-10.08453	.0928253	-108.64	0.000	-10.26647	-9.902598
MOTHERHEIGHT	13.75835	.0977695	140.72	0.000	13.56672	13.94997
PLURALITY	-618.2079	1.434727	-430.89	0.000	-621.02	-615.3959
GESTATIONTIME	102.9078	.1196001	860.43	0.000	102.6734	103.1422
MOTHERWEIGHT	8.489022	.0403023	210.63	0.000	8.410031	8.568013
MOTHERWEIGHT2	-.0150275	.0000935	-160.65	0.000	-.0152109	-.0148442
MALE	124.2467	.5112026	243.05	0.000	123.2448	125.2487
STI	-31.10216	1.783318	-17.44	0.000	-34.5974	-27.60692
_cons	-2343.127	9.780065	-239.58	0.000	-2362.296	-2323.959

Model 3:

$$LNBIRTHWEIGHT_i = \beta_0 + \beta_1 MOTHERAGE_i + \beta_2 MOTHERAGE2_i + \beta_3 UNMARRIED_i + \beta_4 PRENATALVIS_i + \beta_5 PRENATALVIS2_i + \beta_6 CIG3_i + \beta_7 MOTHERHEIGHT_i + \beta_8 MOTHERWEIGHT_i + \beta_9 MOTHERWEIGHT2_i \beta_{10} PLURALITY_i + \beta_{11} GESTATIONTIME_i + \beta_{12} MALE_i + \beta_{13} STI_i + \varepsilon_i$$

Model 3 takes the natural log of the dependent variable but is otherwise identical to Model 2. I am not aware of any theoretical reason to do this, but I will do it anyway. The results are presented in Table 4. As with the previous regressions, all variables are significant beyond the 1% level. The adjusted R^2 is higher, but because the functional form of the dependent variable was changed between models 2 and 3, the adjusted R^2s cannot be compared. Model 3 indicates that if the mother of baby i is infected with Chlamydia or Gonorrhea, then baby i's birth weight will fall by .9%, ceteris paribus.

TABLE 4: Model 3 Regression

```
      Source |       SS           df       MS      Number of obs   = 3,176,221
-------------+----------------------------------   F(13, 3176207)  >  99999.00
       Model |  45456.0908        13  3496.62237   Prob > F        =    0.0000
    Residual |  75645.0798 3,176,207  .023816168   R-squared       =    0.3754
-------------+----------------------------------   Adj R-squared   =    0.3754
       Total |  121101.171 3,176,220   .03812745   Root MSE        =    .15432

------------------------------------------------------------------------------
LNbirth weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
    MOTHERAGE |   .0034837   .0001284    27.14***0.000     .0032321    .0037352
   MOTHERAGE2 |  -.0000466   2.16e-06   -21.57***0.000    -.0000509   -.0000424
    UNMARRIED |  -.0240636   .0001988  -121.04***0.000    -.0244533    -.023674
  PRENATALVIS |   .0061288   .0000519   118.03***0.000      .006027    .0062306
 PRENATALVIS2 |  -.0001585   1.88e-06   -84.26***0.000    -.0001622   -.0001548
         CIG3 |  -.0031559   .0000315  -100.30***0.000    -.0032175   -.0030942
 MOTHERHEIGHT |   .0041383   .0000331   124.87***0.000     .0040734    .0042033
    PLURALITY |  -.2241276   .0004863  -460.85***0.000    -.2250808   -.2231744
GESTATIONTIME |   .0397514   .0000405   980.51***0.000     .0396719    .0398309
 MOTHERWEIGHT |    .002688   .0000137   196.76***0.000     .0026612    .0027148
MOTHERWEIGHT2 |  -4.84e-06   3.17e-08  -152.50***0.000    -4.90e-06   -4.77e-06
         MALE |   .0384601   .0001733   221.95***0.000     .0381205    .0387998
          STI |  -.0095556   .0006045   -15.81***0.000    -.0107404   -.0083708
        _cons |   6.064484   .0033152  1829.29   0.000     6.057986    6.070981
------------------------------------------------------------------------------
```

Model 4:

$$LNBIRTHWEIGHT_i = \beta_0 + \beta_1 LNMOTHERAGE_i + \beta_2 UNMARRIED_i + \beta_3 LNPRENATALVIS_i + \beta_4 CIG3_i + \beta_5 LNMOTHERHEIGHT_i + \beta_6 LNMOTHERWEIGHT_i + \beta_7 PLURALITY_i + \beta_8 LNGESTATIONTIME_i + \beta_9 MALE_i + \beta_{10} STI_i + \varepsilon_i$$

For the final model, I chose a double-log form, applying a logarithm to every coefficient for which the resulting interpretation made conceptual sense. This treatment was applied to MOTHERAGE, PRENATALVIS, MOTHERHEIGHT, MOTHERWEIGHT, and GESTATIONTIME. The other variables are either dummies (such as MALE, STI, and UNMARRIED), or discreet with low values (PLURALITY, CIG3), and so would have a nonsensical interpretation. Asking "what is the impact on birth weight if the mother's weight increases by 1%" is entirely logical. Asking "what is the impact on birth weight if the mother has 1% more babies gestating in her womb?" is not. The results are displayed in TABLE 5. As usual, all variables are significant at the 1% level, with the "worst" variable having a t-score of -16, and the best being in excess of 1000. The model suggests that a 1% increase in gestation time will increase the baby's birth weight by 1.53%, ceteris paribus. The adjusted R^2 has increased slightly to an all-time high of .3836, but to get a better sense for which model is superior, I ran the AIC and BIC tests on both models, shown below.

Model 3 AIC and BIC:

```
Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |  3,176,221   681210.5    1428534      14   -2857040   -2856858
------------------------------------------------------------------------------
```

Model 4 AIC and BIC:

```
Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |  3,124,050   694262.4    1450180      11   -2900339   -2900196
------------------------------------------------------------------------------
```

The AIC and BIC are bother lower on Model 4, which lines up with the adjusted R^2. That said, the difference is not very large, and Model 3 has fewer variables. Because AIC and BIC both tend to punish the inclusion of additional variables, I do not believe that the difference between the two is very significant.

TABLE 5: Model 4 Regression

```
      Source |       SS           df       MS      Number of obs   = 3,124,050
-------------+----------------------------------   F(10, 3124039)  >  99999.00
       Model |  44993.2169        10  4499.32169   Prob > F        =    0.0000
    Residual |   72284.575 3,124,039  .023138179   R-squared       =    0.3836
-------------+----------------------------------   Adj R-squared   =    0.3836
       Total |  117277.792 3,124,049  .037540318   Root MSE        =    .15211

------------------------------------------------------------------------------
  LNBIRTHWEIGHT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
    LNMOTHERAGE |   .0229058    .000459    49.90***0.000     .0220061    .0238054
      UNMARRIED |   -.024648   .0001957  -125.98***0.000    -.0250315   -.0242645
  LNPRENATALVIS |   .0232809   .0002151   108.21***0.000     .0228592    .0237026
           CIG3 |  -.0032127   .0000319  -100.65***0.000    -.0032752   -.0031501
 LNMOTHERHEIGHT |   .2765707    .002112   130.95***0.000     .2724312    .2807102
      PLURALITY |   -.217757   .0004831  -450.72***0.000    -.2187039   -.2168101
LNGESTATIONTIME |   1.529033   .0015005  1018.99***0.000     1.526092    1.531974
 LNMOTHERWEIGHT |   .1367207   .0004468   306.00***0.000      .135845    .1375964
           MALE |   .0387081   .0001722   224.76***0.000     .0383705    .0390456
            STI |  -.0097673   .0006019   -16.23***0.000     -.010947   -.0085876
          _cons |   .7126116   .0098746    72.17   0.000     .6932578    .7319654
------------------------------------------------------------------------------
```

Models 3 and 4, while interesting, do not represent a meaningful improvement over Model 2. They have less theoretical support, and no countervailing advantages, beyond the convenience of percentage interpretation. So, I will be discarding them, and focusing on Model 2 from this point out.

VII. Evaluation

Due to the stronger theoretical support, I will be using Model 2 as my final model. Having established this, I conducted further tests on Model 2.

Omitted Variables

In my case, the most important problem to test for is omitted variable bias. Though my coefficients are all of the expected sign, human development is immensely complicated, and my data captures only a small number of the variables that could affect birth weight. So from a purely theoretical perspective, I expect to have significant omitted variable bias. To confirm this, I used the Ramsey RESET test to assess the extent of this problem.

```
Ramsey RESET test using powers of the fitted values of birth weight
       Ho:  model has no omitted variables
             F(3, 3177006) =  25332.89
                  Prob > F =      0.0000
```

Well, that answers that question. It is extraordinarily likely that Model 2 has many omitted variables. One possibility would be any of a variety of birth defects or developmental problems, some of which may not be evident until well past delivery. More complicated problems in the mother's health history, such as a history of low-weight babies, miscarriages, or conditions such as diabetes may also play a roll. It is also possible that a variable I have knowingly excluded based on previous research, such as race or income, is in fact applicable. While this result is disappointing, it is not unsurprising, and I have no option but to move on.

Irrelevant Variables

All included variables were chosen due to previous research finding them impactful on birth weight. The extremely high levels of significance for all my variables further suggests that all included variables are relevant. I may not have as many variables as ideal, but those I do have are impactful, as proven by the extremely high F-statistic.

Serial Correlation

As my model is cross-sectional, serial correlation is not relevant.

Multicollinearity

Multicollinearity occurs when the dependent variables used in a regression correlate with each other. By theory, the model may have some mild multicollinearity. For example, height and weight tend to be correlated, and a young age is likely to be correlated with being unmarried. That said, my sample is sufficiently large, and these links sufficiently variable, that I do not expect to have severe multicollinearity. To test this theory, I used the Variance Inflation Factors test. The results are shown below

Model 2 VIF test:

Variable	VIF	1/VIF
MOTHERAGE	73.16	0.013669
MOTHERAGE2	71.53	0.013980
MOTHERWEIGHT	43.26	0.023116
MOTHERWEIG~2	42.51	0.023523
PRENATALVIS	6.15	0.162514
PRENATALVIS2	6.00	0.166625
UNMARRIED	1.26	0.792875
MOTHERHEIGHT	1.17	0.851454
GESTATIONT~E	1.11	0.899546
PLURALITY	1.08	0.928494
STI	1.03	0.972221
CIG3	1.02	0.978308
MALE	1.00	0.999315
Mean VIF	19.25	

The only variables with VIF values above 5 are the polynomial variables, which is to be expected. To be sure, I re-ran the test on Model 1, which is identical to Model 2, except that it lacks the polynomial terms. The results are shown below.

Model 1 VIF test:

```
    Variable |       VIF       1/VIF
-------------+----------------------
   UNMARRIED |      1.21    0.824306
   MOTHERAGE |      1.19    0.843450
MOTHERWEIGHT |      1.17    0.857768
MOTHERHEIGHT |      1.16    0.861475
GESTATIONT~E |      1.10    0.907979
   PLURALITY |      1.08    0.929288
 PRENATALVIS |      1.06    0.946417
         STI |      1.03    0.975082
        CIG3 |      1.02    0.981681
        MALE |      1.00    0.999366
-------------+----------------------
    Mean VIF |      1.10
```

As expected, with the polynomial terms removed all VIFs are extremely low, indicating that multicollinearity is not a problem.

Heteroskedasticity

Heteroskedasticity is a phenomenon where the variance of error in a sample is not constant. To test for heteroskedasticity I used the White's test. White's test detects more types of heteroskedasticity than the Breusch-Pagan test, and given my large number of observations, the test's variable-intensive nature is not a drawback. The results of the test are shown below.

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(98)     = 262507.42
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |  262507.42     98    0.0000
            Skewness |    1875.64     13    0.0000
            Kurtosis |    2043.40      1    0.0000
---------------------+-----------------------------
               Total |  266426.46    112    0.0000
---------------------------------------------------
```

With a Chi^2 of 262507.42, the White's test shows that it is extremely likely that my regression has heteroskedasticity. As such, I will apply robust standard errors, shown below:

TABLE 6: MODEL 2 Regression with Robust Standard Errors

```
Linear regression                               Number of obs     =  3,176,221
                                                F(13, 3176207)    >   99999.00
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3399
                                                Root MSE          =     455.27

------------------------------------------------------------------------------
             |               Robust
  BIRTHWEIGHT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    MOTHERAGE |   11.94914   .3873791    30.85***0.000     11.18989    12.70839
   MOTHERAGE2 |  -.1623465   .0065659   -24.73***0.000    -.1752155   -.1494776
    UNMARRIED |  -77.23285   .6000464  -128.71***0.000    -78.40892   -76.05678
  PRENATALVIS |   16.13499   .2290809    70.43***0.000       15.686    16.58398
 PRENATALVIS2 |  -.4123594   .0091029   -45.30***0.000    -.4302009    -.394518
         CIG3 |  -10.08453   .1180067   -85.46***0.000    -10.31582   -9.853243
 MOTHERHEIGHT |   13.75835   .0996127   138.12***0.000     13.56311    13.95359
    PLURALITY |  -618.2079   1.404406  -440.19***0.000    -620.9605   -615.4554
GESTATIONTIME |   102.9078   .1766835   582.44***0.000     102.5615    103.2541
 MOTHERWEIGHT |   8.489022    .045214   187.75***0.000     8.400404     8.57764
MOTHERWEIGHT2 |  -.0150275   .0001081  -139.03***0.000    -.0152394   -.0148157
         MALE |   124.2467   .5108423   243.22***0.000     123.2455    125.2479
          STI |  -31.10216   1.801232   -17.27***0.000    -34.63251   -27.57181
        _cons |  -2343.127   11.21489  -208.93   0.000    -2365.108   -2321.147
------------------------------------------------------------------------------
```

Thought the use of Robust Standard Errors does reduce my model's t-statistics, all of the variables remain highly significant, so this is an easy choice to make.

Incorrect Functional Form

While my function's fit is not fantastic, it is not terrible given the circumstances, and the interpretation of the coefficients is smooth and logical, all of which suggests that my functional form is not incorrect. To combat incorrect functional form, I based my model off theory as established by previous researchers. That said, established theory does not always agree with itself, which exposes my model to weakness. I made every effort to perform a thorough literature

review and be aware of these conflicts, but in some cases I simply had to make a choice and live with the consequences, such as my decision to exclude race from consideration.

IIX. Conclusion

My model demonstrates some of the determinants of birth weight in the US. Though the Ramsey Test and (to a lesser extent) the model's Adjusted R^2 indicate that my model is far from complete, those variables that it does include are extremely statistically significant, in most cases far beyond the requirements for significance at the 1% level. The single most impactful variable is plurality: according to my model, an additional baby in gestation will reduce the birth weight of the babies by 618 grams. Overall, the model performed as expected. Although it did encounter severe heteroskedasticity, that has been compensated for with robust standard errors. Otherwise, all variables were highly significant in the expected direction, and there were no major surprises along the way. By far the model's biggest problem is omitted variable bias. Combined with the extremely high levels of significance of the current variables, this model seems to be a strong foundation for further research but needs to be enhanced with additional variables to create a more complete picture of the determinants of birth weight.

Bibliography

Bird, Sheryl Thorburn, Anjani Chandra, Trude Bennett, and S. Marie Harvey. “Beyond Marital Status: Relationship Type and Duration and the Risk of Low Birth Weight.” *Family Planning Perspectives* 32, no. 6 (2000): 281–87.

Conway, Karen Smith, and Partha Deb. “Is Prenatal Care Really Ineffective? Or, Is the ‘Devil’ in the Distribution?” *Journal of Health Economics* 24, no. 3 (May 2005): 489–513.

Dougherty, C R, and A D Jones. “The Determinants of Birth Weight.” *American Journal of Obstetrics and Gynecology* 144, no. 2 (September 15, 1982): 190–200.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth. “The Effects of Poor Neonatal Health on Children's Cognitive Development.” *American Economic Review* 104, no. 12 (December 2014): 3921–55.

Liu, Gordan G. “Birth Outcomes and the Effectiveness of Prenatal Care.” *Health Services Research* 32, no. 6 (February 1998): 805–23.

“Low birth weight.” UNICEF DATA. UNICEF, July 13, 2020. https://data.unicef.org/topic/nutrition/low-birth weight/.

Manyeh, Alfred Kwesi, Vida Kukula, Gabriel Odonkor, Rosemond Akepene Ekey, Alexander Adjei, Solomon Narh-Bana, David Etsey Akpakli, and Margaret Gyapong. “Socioeconomic and Demographic Determinants of Birth Weight in Southern Rural Ghana: Evidence from Dodowa Health and Demographic Surveillance System.” *BMC Pregnancy and Childbirth* 16, no. 1 (July 15, 2016).

Nakamuro, Makiko, Yuka Uzuki, and Tomohiko Inui. “The Effects of Birth Weight: Does Fetal Origin Really Matter for Long-Run Outcomes?” *Economics Letters* 121, no. 1 (October 2013): 53–58.

Shah, Prakesh S., Jamie Zao, and Samana Ali. “Maternal Marital Status and Birth Outcomes: A Systematic Review and Meta-Analyses.” *Maternal and Child Health Journal* 15, no. 7 (October 2010): 1097–1109.

Verma, Nehar, Suprava Patel, Phalguni Padhi, Tripty Naik, Rachita Nanda, Gitismita Naik, and Eli Mohapatra. “Retrospective Analysis to Identify the Association of Various Determinants on Birth Weight.” *Journal of Family Medicine and Primary Care* 10, no. 1 (January 2021): 496–501.