A Study On Whether Baby's Birth Weight Is Related To Gestational Factors And Parents' Demographics and Behaviour

By: Ning Wu, Wenbin Zhou, Yuanchi Wei, Neyan Deng
Group 17

Introduction

"Birth weight has significant and lasting effects", and it is even related to adult health and success in many respects such as school behaviour and income [1]. So, the objective of this project is to explore the relationship between demographic and behavioral data of the parents and the birth weight of the infant through regression modeling.

The data selected is a subset of the Child Health and Development Studies (CHDS), a comprehensive investigation of all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco–East Bay area [2]. In particular, the data includes all 1236 male single births where the baby lived at least 28 days during one year of the study. The complete CHDS data was previously used to study the relationship between mother's cigarette smoking habit and the survival of infant [3][4]. The aim of this project is to look at the relationship between birth weight and potential predictors available from the data such as gestational period, parents' height and weight, and family income, and try to answer two questions: 1) which variables have significant impact on babies' birth weight? and 2) can we predict a baby's birth weight with these variables?

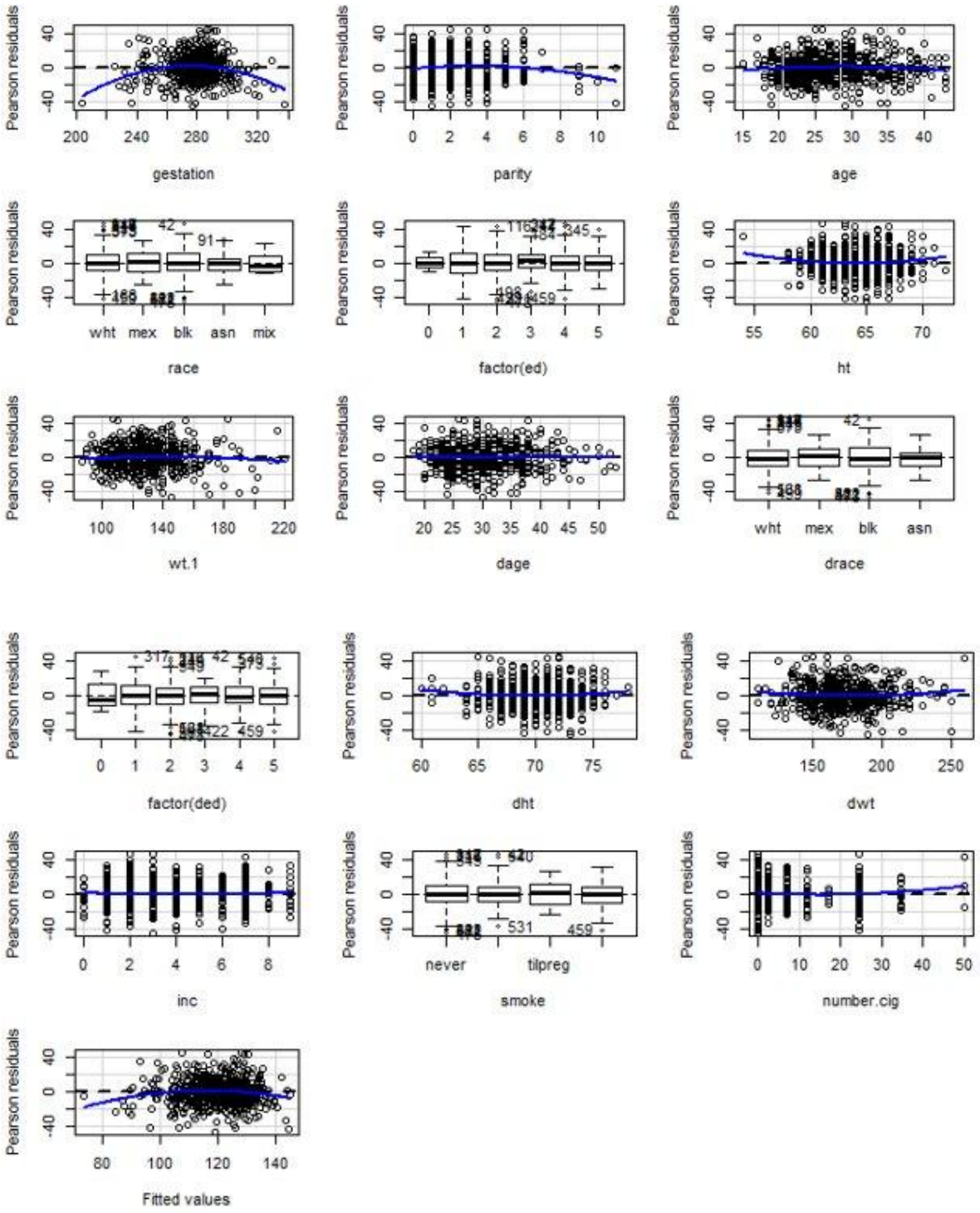
Data

The response variable of the project is "wt", which is the infant birth weight in ounces. In addition, there are 21 other variables available in the data set. Several variables only have one level due to the selection of our particular data set, such as "plurality" (single fetus only), and "sex" (male infants only). The "marital" status variable is dominated by level 1 (married), which account for approximately 98% of the observations, with the other levels typically having 0 to a few observations, too few to produce statistically significant results. These variables are not included in our study. Similarly, there are two levels in "ded", father's education, that only have 1 observation each. To avoid overfitting, these observations are removed from the data. We also noticed that the father's weight and height for many observations are missing from the data. On the other hand, initial exploration of the data shows that these variables may have statistically significant impact on the birth weight of the infant. Therefore we decided to include father's weight and height in the project, even though doing so significantly reduces the sample size. After selecting the initial predictor variables, and deleting incomplete observations and observations with out-of-range data, 586 observations remain out of the original 1236. The 15 predictor variables are described in the table below.

Variable	Description	Type
gestation	Length of gestation in days	Numeric
parity	Total number of previous pregnancies	Numeric
race	Mother's race: "wht', "mex", "blk", "asn", "mix"	Factor
age	Mother's age in years at termination of pregnancy	Numeric
ed	Mother's education: 0= less than 8th grade, 1 = 8th -12th grade - did not graduate, 2= HS graduate–no other schooling, 3= HS+trade, 4=HS+some college, 5=College graduate, 6=Trade school, 7=HS unclear	Factor
ht	Mother's height in inches	Numeric
wt.1	Mother's prepregnancy weight in pounds	Numeric
drace	Father's race, with levels equivalent to mother's race	Factor
dage	Father's age in years	Numeric
ded	Father's education, with same coding as mother's education	Factor
dht	Father's height in inches	Numeric
dwt	Father's weight in pounds	Numeric
inc	Family yearly income in $2500 increments: 0 = under 2500, 1 = 2500-4999, …, 8=12500-14999, 9 = 15000	Numeric
smoke	Does mother smoke? 0 = never, 1 = smokes now, 2 = until current pregnancy, 3 = once did, not now	Factor
number	Number of cigarettes smoked per day for past and current smokers	Numeric

Model Building

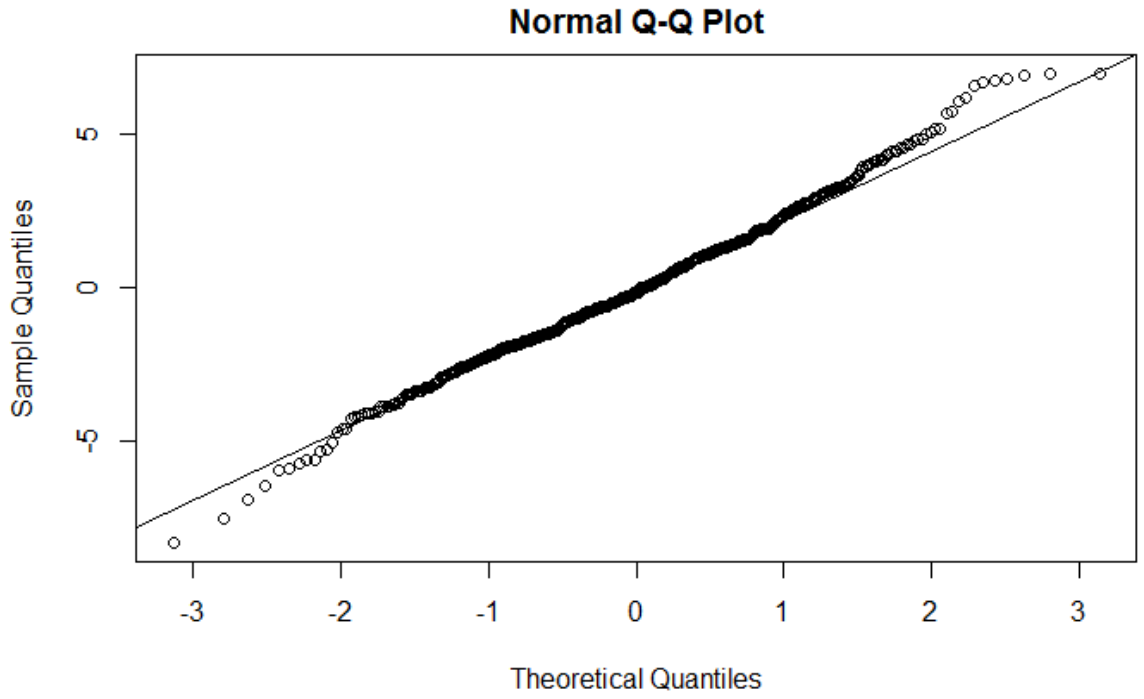
The full model was obtained through an iterative process. First, a model was built with untransformed "wt", and only the first-order terms of the predictors. The residual plots below revealed that the linear relationship with "gestation" may not be appropriate.



Because “gestation” showed the highest correlation with “wt” in the initial data exploration, its correct functional form is critical for building an appropriate model. So a normalized second-order term of “gestation” is added to the model, which is denoted “ges2” in the R code below (ges2 = (gestation - mean(gestation))^2). Further assumption checking showed that there may be normality violation in the residuals. So in the third iteration the Box-Cox procedure was used to determine the optimal transformation of the response

variable. The transformed response is wt'=wt^0.6970. The transformed "wt" is denoted "wtp" in the R code below. Further diagnostics showed that the normality violation was not markedly improved (QQ normal plot and Shapiro-Wilk test results shown below). However, there was not an obvious measure to correct the normality violation, so the full model with the second-order "gestation" term and transformed "wt" is used in the model selection process.

```
##  Shapiro-Wilk normality test
##
## data:  gestation.res
## W = 0.9943, p-value = 0.02713
```



Model Selection

A stepwise model selection process starting from the full model gave the following recommendation:

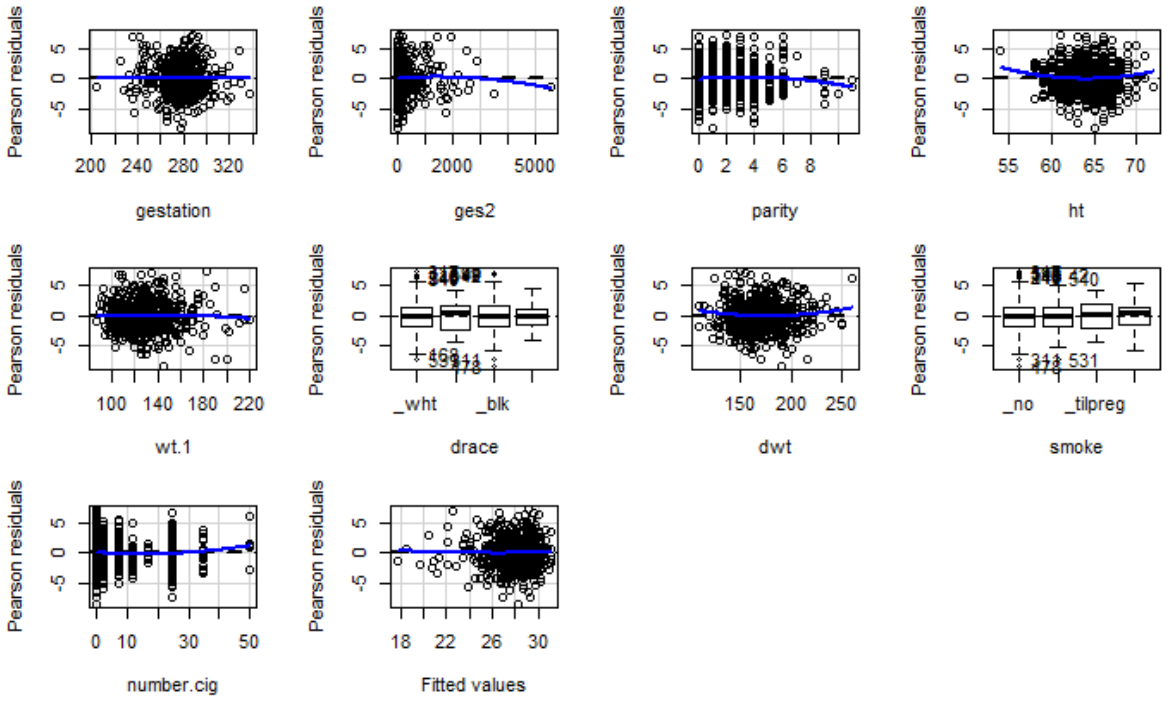
```
## Call:
## lm(formula = wtp ~ gestation + ges2 + parity + ht + wt.1 + drace +
##     dwt + smoke + number.cig, data = gestation)
```

The linear model coefficients and the type I ANOVA results are shown below. From the ANOVA table, all the predictors are significant at least at the 0.05 level. The R2 values is approximately 0.366.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.552416   3.342328  -1.661 0.097213 .
gestation      0.070389   0.007047   9.988  < 2e-16 ***
ges2          -0.001366   0.000218  -6.264 7.37e-10 ***
parity         0.151789   0.056606   2.681 0.007541 **
ht             0.171507   0.045389   3.779 0.000174 ***
wt.1           0.009937   0.005823   1.706 0.088460 .
drace_mex      0.988443   0.620498   1.593 0.111717
drace_blk     -0.980241   0.265717  -3.689 0.000247 ***
drace_asn     -1.148049   0.550226  -2.087 0.037375 *
dwt            0.013369   0.004711   2.838 0.004700 **
smoke_now     -0.728551   0.314855  -2.314 0.021025 *
smoke_tilpreg  0.675196   0.404240   1.670 0.095410 .
smoke_notnow   0.121869   0.407224   0.299 0.764845
number.cig    -0.036354   0.013334  -2.726 0.006599 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.461 on 572 degrees of freedom
Multiple R-squared:  0.3655,   Adjusted R-squared:  0.351
F-statistic: 25.34 on 13 and 572 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: wtp
            Df Sum Sq Mean Sq  F value    Pr(>F)
gestation    1 1010.1 1010.10 166.8295 < 2.2e-16 ***
ges2         1  266.8  266.79  44.0626 7.397e-11 ***
parity       1   31.5   31.48   5.1997 0.0229585 *
ht           1  220.3  220.33  36.3904 2.901e-09 ***
wt.1         1   25.4   25.37   4.1905 0.0411065 *
drace        3  115.1   38.35   6.3347 0.0003136 ***
dwt          1   43.0   43.04   7.1082 0.0078902 **
smoke        3  237.5   79.16  13.0742 2.869e-08 ***
number.cig   1   45.0   45.01   7.4332 0.0065994 **
Residuals  572 3463.3    6.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

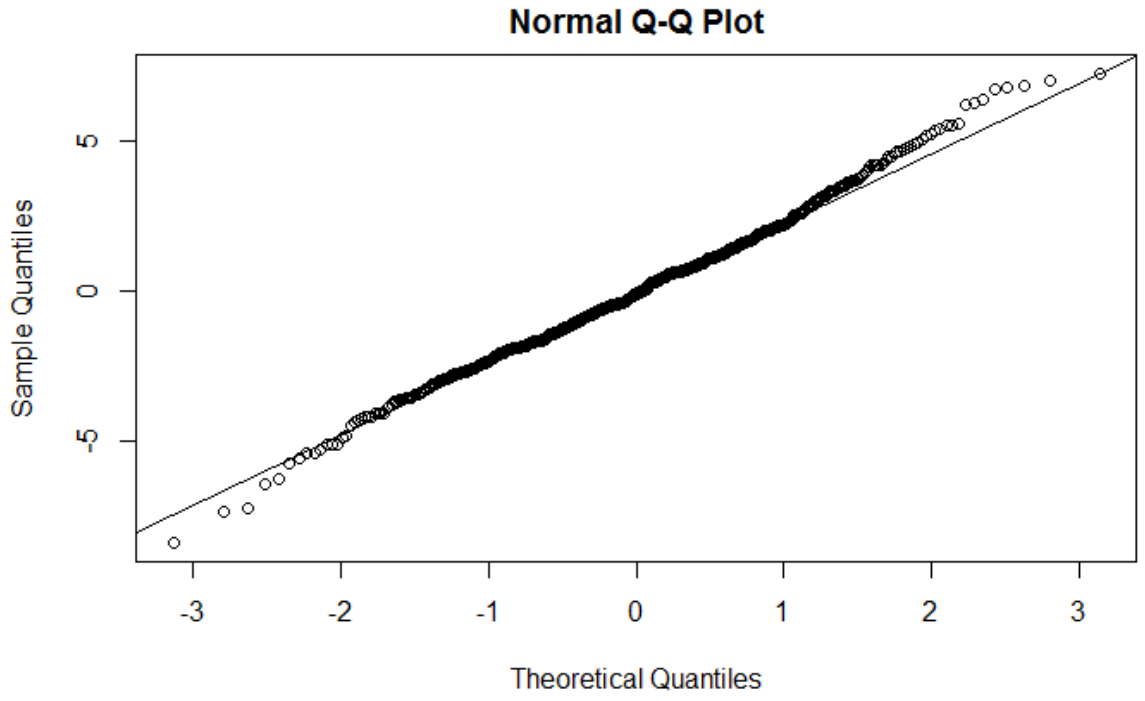
Diagnostics

In the residual plots below, no apparent pattern exists that indicates inappropriate functional form of the predictors.

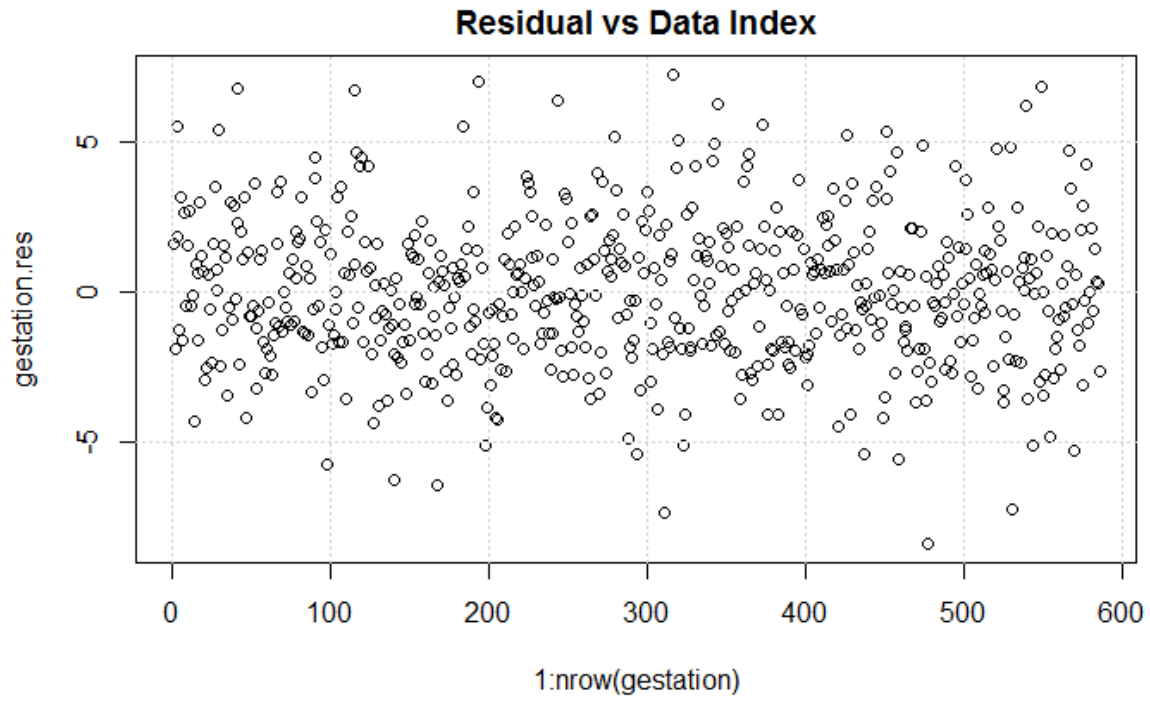


The Brown-Forsythe test (results shown below) did not indicate a violation of the constant variance assumption. The QQ plot still showed some deviation from normal distribution. The Shapiro-Wilk test, on the other hand, gave a p-value of 0.07, which is a moderate improvement from the full model and is above the 0.05 level. This should give us a little more confidence in the estimated mean and new response, as we will discuss below.

```
 Brown-Forsythe Test
------------------------------------------------------------
  data : res and group

  statistic  : 0.3690889
  num df     : 4
  denom df   : 31.99439
  p.value    : 0.8288421

  Result     : Difference is not statistically significant.
------------------------------------------------------------


      Shapiro-Wilk normality test
/*
data:  gestation.res
W = 0.99527, p-value = 0.0709
```



The residual plotted against the data index (below), which is also the order in which the data was collected, did not show any dependence among the data points.



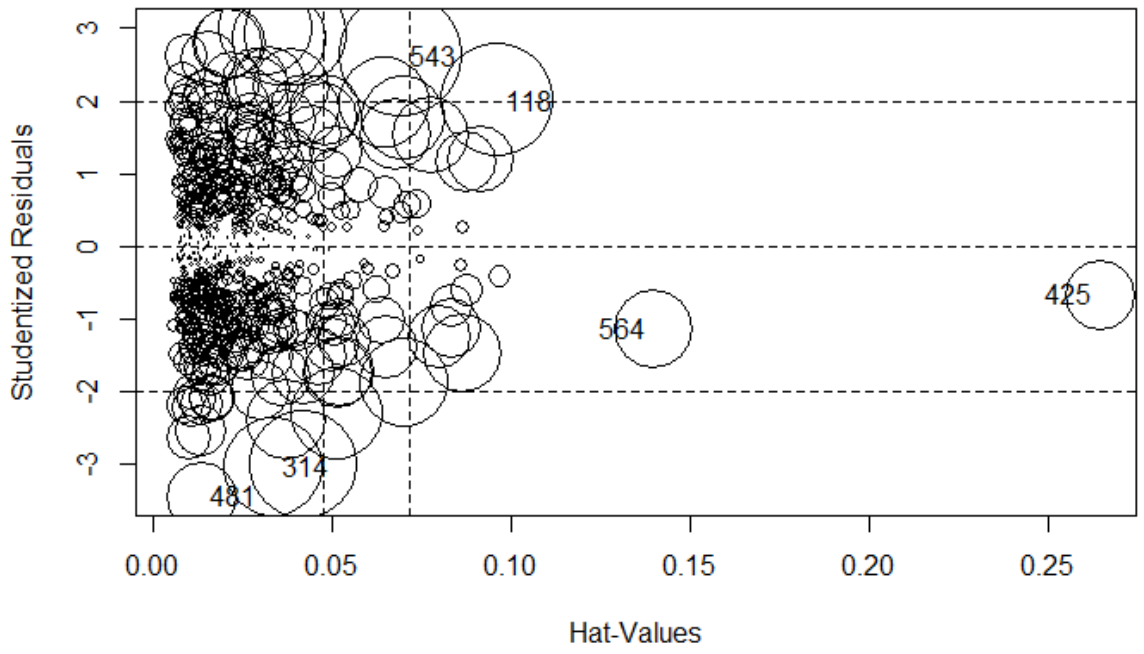
The generalized VIFs (GVIFs) were calculated for the selected predictors, all of which were close to 1. Therefore it can be concluded that multicollinearity is not a problem in the data.

```
               GVIF Df GVIF^(1/(2*Df))
gestation  1.099395  1        1.048520
ges2       1.089404  1        1.043745
```

```
parity      1.124572  1        1.060458
ht          1.340346  1        1.157733
wt.1        1.352055  1        1.162779
drace       1.306275  3        1.045536
dwt         1.124159  1        1.060263
smoke       2.043008  3        1.126449
number.cig 2.012745  1        1.418712
```

The influence plot and statistics of notable points are shown below. There are many points with hat value greater than$2\overline{h} = 2p/n = 0.0478$, including points 118, 425, and 543, and 564 in the plot. The threshold for Studentized deleted residual is $B = t(1 - \alpha/2n;\ n - 1 - p) = 3.96$.None of the observations is considered Y outliers based on this criterion. An observation is considered influential if its Cook's distances is greater than $F(0.2;\ p;\ n - p) = 0.675$, or very influential if greater than $F(0.5;\ n;\ n - p) = 0.954$. Judging from the plot above, none of the observations is influential. Based on these observations, no corrective measure was taken on the selected model.

```
##         StudRes        Hat      CookD
## 118  2.0046133 0.09617821 0.03038375
## 314 -3.0585230 0.03385659 0.02307808
## 425 -0.6665836 0.26441054 0.01141949/
## 481 -3.4489462 0.01363682 0.01152728
## 543  2.6207405 0.06916128 0.03608077
## 564 -1.1391292 0.13974806 0.01504916
```



Discussion

Due to the number of predictors, checking every interaction is computationally intensive, and risks overfitting. On the other hand, we did checked several potential interactions on the selected model based on intuitive possibility

- Smoking status and daily number of cigarettes (smoke : number.cig): p = 0.6553
- Mother's race and her height and weight (race : (ht + wt.1)): p = 0.2408
- Father's race and weight: (drace : dwt): p = 0.5119

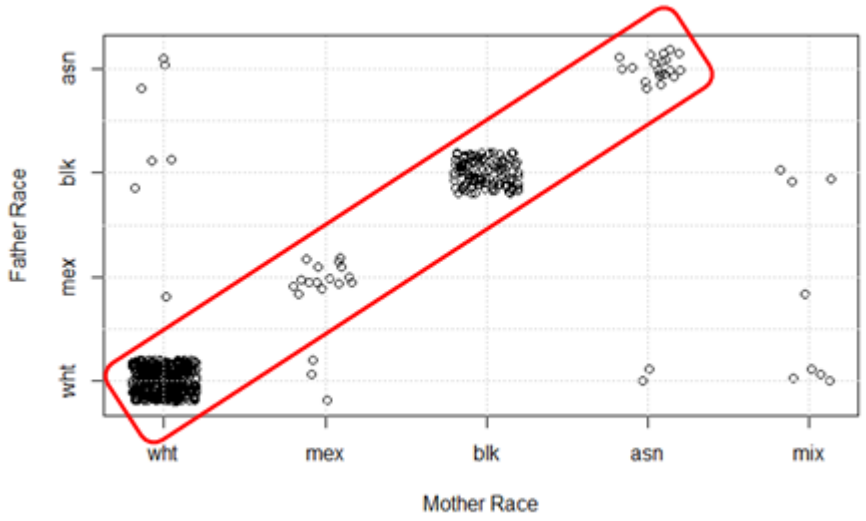
Based on the large values, none of the interactions were significant. So the separate treatment of the predictor was likely appropriate.

Because there is no significant collinearity among the selected predictors, the type I ANOVA table presented before is a good indication of the true impact of the individual predictors on the response variable. The gestation period and its square terms have by far the largest impact on the expected birth weight. Other predictors in the rough order of impact with signs of coefficients noted in parentheses include:

- Smoking status and number of cigarettes per day (-)
- Mother's height (+) and weight (+)
- Father's race
- Father's weight (+)
- and parity (+)

These results make sense. Longer gestation means greater birth weight, as the fetus grows rapidly during the third trimester every day. Smoking cigarettes reduces the baby weight, as nicotine will suppress mother's appetite[7]. Taller and heavier mothers and fathers tend to have heavier babies. Baby born to fathers (and mothers) of different races may have different expected weight.

Only the father's races are included in the selected model, which is not an accident. The reason is that according to the scatter plot of the mother's' and father's races, most of the points concentrates on the diagonal where the mother and the father have the same race. In the early 1960s, when the data was collected, interracial marriages and interracial children were rare. Interracial births account for less than 4% of the observations in the data. So mother's race does not bring in much new information, and including it in the model in addition to the father's race would likely cause collinearity problems. On the other hand, if we were to publish our findings, it may be appropriate to include a version of the model that has the mother's race as a predictor. This way validation can be made with more recent data in which interracial babies are more common.



As mentioned before, the predictors in the model only account for approximately 37% of the total variance in the newborn baby weight. This fact becomes obvious when we use the model to predict the baby weight given the demographical data of the parents. For example, consider a typical birth with a nonsmoking mother of 5'4" height and 130 lb weight, a white father of 170 lb weight, and that this is the mother's first pregnancy and is carried to a full term of 280 days. The confidence interval for the mean newborn weight of such cases is [7.57, 7.87] lb. With an uncertainty of 0.30 lb, this estimate is rather precise. On the other hand, the prediction interval of the baby weight is [5.92, 9.66] lb. With an uncertainty of 3.74 lb, this estimate may be useless for a lot of purposes. This demonstrates the limitation of using parents' demographical data to predict baby weight.

Conclusion

Through model building and model selection process, several variables were identified to significantly impact the expected birth weight. Although the resulting model appear to be appropriate given the data, the birth weight prediction interval based on the model may be too wide to be useful. To arrive at models that give more informative predictions, other predictors may be incorporated in future studies. Such predictors may include [5,6]

- Sex of the baby and birth plurality (only single birth male baby data were available in the data set)
- Exposure to passive smoking
- Maternal education
- Maternal nutrition/weigh gain during pregnancy
- Birth spacing

References

[1] Born To Lose: How birth weight affects adults health and success, June 5, 2007, https://news.umich.edu/born-to-lose-how-birth-weight-affects-adult-health-and-success/

[2] D Nolan and T Speed. *Stat Labs: Mathematical Statistics Through Applications* (2000), Springer-Verlag.

[3] J.Yerushalmy. Mother's cigarette smoking and survival of infant. Am. J. Obstet. & Gynecol., 88:505–518, 1964.

[4] J. Yerushalmy. The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations. Am. J. Epidemiol., 93:443–456, 1971.

[5] Metgud, Chandra S et al. "Factors affecting birth weight of a newborn--a community based study in rural Karnataka, India" *PloS one* vol. 7,7 (2012): e40040.

[6] Neggers, Yasmin & Goldenberg, R.L. & Tamura, T & Cliver, S.P. & Hoffman, Howard. (1997). The relationship between maternal dietary intake and infant birthweight. Acta obstetricia et gynecologica Scandinavica. Supplement. 165. 71-5.

[7] McEwen, Andy, Smoking: How it Affects Diet& Nuitrition, BBC goodfood, https://www.bbcgoodfood.com/howto/guide/effects-smoking-has-diet-nutrition-and-health