STAT151A Final Project, Fall 2021

Wenhao Pan, Rachel Chen, Richard Shuai

December 17, 2021

Contents

1 Introduction

Baby's mass is correlated with mortality risk and potential future developmental problems. For example, researchers in Denmark found that babies with birth weights of less than 5 pounds are more likely to experience health complications and even a lower intelligence quotient as children. Thus, it makes sense for healthcare workers and parents to want to predict a baby's weight based on current information. Intuitively speaking, a baby's mass could be predicted by a lot of factors such as the health of the parents, the sex of the baby, the mother's pregnancy records, etc. In this project, we aim to use linear models to answer the following two questions regarding the baby's weight.

1. How does intervening a pregnant woman's living habits or behaviors affect her baby's birth weight in the future?

2. Given the information about an expecting family, what is our best prediction of their baby's weight? The first question is more related to causal inference, and its answer could help doctors give suggestions to a pregnant woman for delivering a normally weighted baby. The second one is more related to prediction, and its answer could help doctors conjecture a baby's weight right before delivery.

2 Data Description

This dataset was taken from the National Center for Health Statistics, and contains information about 3.8 million childbirths in the US in 2018. There are 55 columns, so we grouped them into the following categories:

- Delivery situation ex) place of birth, number of people around, birth time
- The baby's health information ex) period of gestation, birth weight
- Parents information ex) marital status, education, race
- Parents health records ex) smoking history, age
- Mother's pregnancy records ex) number of prenatal visits, prior births

The User Guide on the website contains the detailed explanations of each column. We will use the baby birth weight column (`DBWT`) as the response variable, and all other variables will be used as explanatory variables.

```
##   ATTEND BFACIL  BMI CIG_0 DBWT DLMP_MM DLMP_YY DMAR DOB_MM DOB_TT DOB_WK
## 1      1      1 30.7     0 3657       4    2017    1      1   1227      2
## 2      1      1 33.3     2 3242      99    9999    2      1   1704      2
## 3      1      1 30.0     0 3470       4    2017    1      1    336      2
## 4      3      1 23.7     0 3140       5    2017    2      1    938      2
## 5      1      1 35.5     0 2125      99    9999    1      1    830      3
## 6      4      2 31.3     0 4082       3    2017    1      1     28      2
##   DOB_YY DWgt_R FAGECOMB FEDUC FHISPX FRACE15 FRACE31 FRACE6 ILLB_R ILOP_R
## 1   2018    231       31     3      1       1       1      1     16     33
## 2   2018    185       35     4      0       3       3      3    180    888
## 3   2018    273       31     4      0       1       1      1    999    888
## 4   2018    138       26     2      0       3       3      3     43    888
## 5   2018    219       35     3      0       2       2      2    999    999
## 6   2018    247       28     6      6       1       1      1     39    888
##   ILP_R IMP_SEX IP_GON LD_INDL MAGER MAGE_IMPFLG MAR_IMP MBSTATE_REC MEDUC
## 1    16      NA      N       N    30          NA      NA           1     6
## 2   180      NA      N       N    35          NA      NA           1     9
## 3   999      NA      N       N    28          NA      NA           1     6
```

```
## 4    43      NA      N       N    23          NA      NA           1     2
## 5   999      NA      N       N    37          NA      NA           1     4
## 6    39      NA      N       N    26          NA      NA           1     6
##   MHISPX MM_AICU MRACE15 MRACE31 MRACEIMP MRAVE6 MTRAN M_Ht_In NO_INFEC
## 1      0       N       1       1       NA       1     N      66        1
## 2      0       N       3       3       NA       3     N      63        1
## 3      0       N       1       1       NA       1     N      71        1
## 4      0       N       3       3       NA       3     N      64        1
## 5      0       N       1       1       NA       1     N      66        1
## 6      0       N       1       1       NA       1     N      67        1
##   NO_MMORB NO_RISKS PAY PAY_REC PRECARE PREVIS PRIORDEAD PRIORLIVE PRIORTERM
## 1        1        1   2       2       3      8         0         1         2
## 2        1        0   1       1       3      9         0         2         0
## 3        1        0   5       4       5     17         0         1         0
## 4        1        1   1       1       5      6         0         2         0
## 5        1        1   1       1       5     15         0         1         4
## 6        1        1   2       2       2     13         0         1         0
##   PWgt_R RDMETH_REC RESTATUS RF_CESAR RF_CESARN SEX WTGAIN
## 1    190          1        2        N         0   M     41
## 2    188          4        2        Y         2   F      0
## 3    215          1        1        N         0   M     58
## 4    138          1        2        N         0   F      0
## 5    220          3        1        N         0   M      0
## 6    200          1        1        N         0   F     47
```

```
## [1] 3801534
```

3 Data Preprocessing

```
## [1] 2354840
```

We first propose that due to the excessive size of the original dataset, 3.8 million observations, we plan to randomly subsample three subsets, one with 5000 observations and two with 100000 observations, with replacement as datasets for EDA, training, and testing. This plan balances the computational cost of the analysis and the complexity of our dataset well. We conduct this subsampling plan at the end of data preprocessing.

The priority of our data preprocessing is to reduce the dimensionality of our dataset by filtering out unuseful features. We have 54 explanatory variables, but it is not efficient to analyze each of them evenly. We suspect that some variables can be combined and condensed into a new variable. To systematically filter the features, we split the explanatory variables into five exclusive categories: Homogeneous: Variables of which more than 99% of entries have the same values. Minor: We do not consider interaction terms involving these variables. Major: We do consider interaction terms involving these variables. Obsolete: After we create a new variable based on these variables, they essentially do not provide enough extra information to be kept. The new variable belongs to "Major". Redundant: After we select one from a group of variables including similar information, the rest become redundant or unnecessary to be kept. The selected variable belongs to "Major". See the appendix for the names of the variables in each category.

Next, we drop all the observations including any missing value in the columns of "Minor", "Major", and "Obsolete" categories. After dropping, we still have about 2.8 million observations left, which are sufficient for subsampling. The missing values in our dataset are not left blank or NA. Instead, they are recoded into values such as '9' or '999', depending on the variable. Thus, we manually look up the recodings from the User Guide and drop the missing values for each feature. Imputing those missing values might be a better

approach since if the missing values exemplify a systematic pattern, we might introduce bias by removing all the missing values. However, given the already complicated structure of our dataset, we choose the simpler approach-dropping missing values, noting that the imputation approach is still valuable to be explored.

Cleaning up missing values allows us to conduct feature engineering, which aims to use domain knowledge to simplify the dataset while maintaining the original information. For example, we create the feature `PREG_LEN` which estimates the pregnancy length by computing the number of months between the last normal mense and delivery date. Then, we categorize `PREG_LEN` into `Early`, `8`, `9`,`10`,`Late` by common sense. Thus, `PREG_LEN` becomes a "Major" variable, and variables about the last normal menses and delivery dates become "Obsolete" variables. `FRACE6`, `FRACE15`, `FRACE31`, and `FHISPX` all describe a father's race but with different granularity levels. We select `FRACE6` to solely describe a father's race since we think six racial categories already sufficiently differentiate people. Thus, `FRACE6` is a "Major" variable, and other father's race variables become "Redundant". See the code appendix for the complete work of feature engineering.

Finally, we drop "Homogeneous", "Obsolete", and "Redundant" variables. It is obvious that we should drop "Obsolete" and "Redundant" variables to mitigate collinearity and duplication in our dataset. We drop the "Homogeneous" variables because it is highly likely that they essentially have one unique value or a negligible number of other values in subsample datasets. We warn that the entire preprocessing approach is considerably subjective, and the following regression analysis is highly dependent on our approach. It may sound like a compromise to the complexity of our dataset to some audiences. We encourage different preprocessing approaches, but we continue with ours since we think it is still reasonable.

4 Exploratory Data Analysis

To gain a better understanding of the variables and data, we performed EDA on our dataset. First, we plotted the distribution of the response variable to verify that our linear regression assumptions are satisfied. From Figure X, we can see that the distribution is quite symmetric and normal-looking (reference QQ plot?). This is desired because symmetric data makes it easier to model, and also normality is an assumption of statistical inference.

To verify symmetry, we use the formula Upper quartile - Median / Median - Lower Quartile (put this in latex): (insert Wenhao's code for measure of symmetry) Since the ratio is close to 1, we can safely conclude that our response variable is symmetric.

We next examined the bivariate relationships between our response variable and each explanatory variable in our dataset. In Figure X, we see that as the percentage of weight gained due to pregnancy increases, the birth weight tends to increase. The positive linear relationship with the birth weight visualizes this correlation.

We also conclude that CIG_0 is an important explanatory variable we want to include in our model, since from the box plot in Figure X we see that the 1st quartile, median, and 3rd quartile of birth weight are all lower if the mother smokes, than if she does not. When plotting birth weight against the prenatal care status of the mother, we find that the distributions of birth weight seem to be most different between mothers who didn't receive prenatal care and mothers who did. This suggests that the most important difference can be observed when we binarize the explanatory variable for the mother's prenatal care status. From Figure X, we see that this holds, and we therefore binarize the mother's prenatal care status for downstream analysis.

In addition to the main effect terms, interaction terms are visualized using boxplots and scatterplots. The box plot in Figure X shows the interaction between SEX and PRECARE. We notice that the difference in the median birth weight for mothers who received prenatal care and who did not changes across the sex of the baby, which motivates using an interaction term. More specifically, such difference is larger in male babies than female babies.

In Figure X, we observe a possible interaction between the mother's prenatal care status and her BMI when predicting the baby's birth weight. We see that if a pregnant woman does not receive prenatal care, as her BMI increases, she is more likely to have lighter babies, which could be a signal for an unhealthy baby. This

might make sense because obesity is linked to health problems, which consequently affect the health and birth weight of the baby. This, however, may be corrected if the mother received care and possibly took actions to mitigate her health problems, leading to more normal birth weights.

5 Model Selection

We first fit the fullest model with only the main effect terms, which contains 30 explanatory variables or 68 regressors. Such a complicated model with so many regressors is not desired for prediction or causal inference because it will tend to overfit on the training data and therefore generalize poorly on new datasets, even if drawn from the same distribution. Also, such a model would require us to collect lots of information to predict, which will limit the usability of the model in a realistic setting. Moreover, such a model is hard to interpret for causal inference. Thus, we conduct model selection to select a simpler model.

We first remove explanatory variables introducing singularity, which have fitted coefficients `NA`. Next, we construct four models using four different approaches: forward/backward selection with AIC/BIC by `step` function and then select the one with the lowest leave-one-out cross-validation (LOOCV) error. Both AIC and BIC measure the in-sample fitness of a model, while BIC penalizes the model size more when the sample size is large. Lower AIC or BIC in value means a better model. We chose forward and backward selection instead of all subset selection because of the large number of explanatory variables, which makes all subset selection computationally infeasible. We chose the `step` function because it adds or drops categorical variables only as an entire unit instead of splitting them up into unconnected dummy regressors. Ideally, we hope that these four models using different criteria and search strategies would explore diverse model choices, so the final one selected by LOOCV error is the most descriptive.

It turns out that both the models returned by forward and backward selection with AIC have the lowest LOOCV error and are identical in terms of the set of included explanatory variables, so both of them are the best model. We then add the selected interaction terms based on our findings from the EDA to the best model and use the incremental F-test with `Anova` to filter out the insignificant interaction terms. See the code appendix for more details about the entire process. The final model that will be used for both causal inference and prediction is

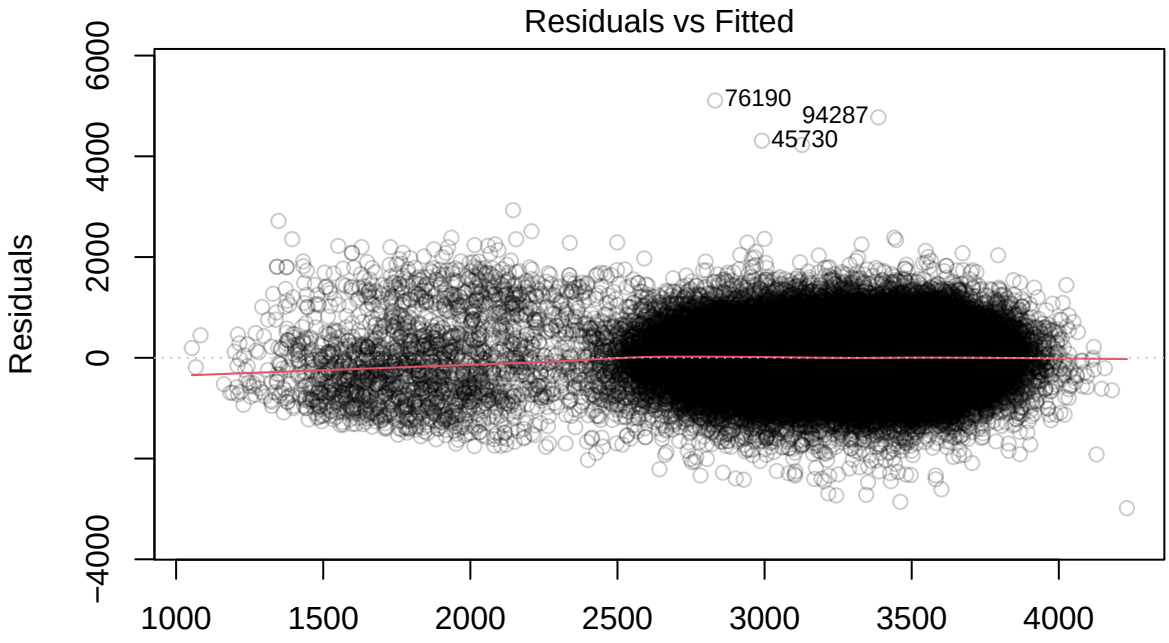
```
formula(final.lm)
```

```
## DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI + WTGAIN_PER +
##     PRIORLIVE + CIG_0 + NO_RISKS + RDMETH_REC + PREVIS + ATTEND +
##     MBSTATE_REC + FRACE6 + PAY_REC + LD_INDL + FEDUC + NO_MMORB +
##     BFACIL + FAGECOMB + NO_INFEC + RESTATUS + MEDUC + PRECARE +
##     DMAR + PREVIS * PREG_LEN + PREG_LEN * MEDUC + CIG_0 * PRECARE +
##     PRECARE * PREG_LEN + CIG_0 * PREG_LEN
```

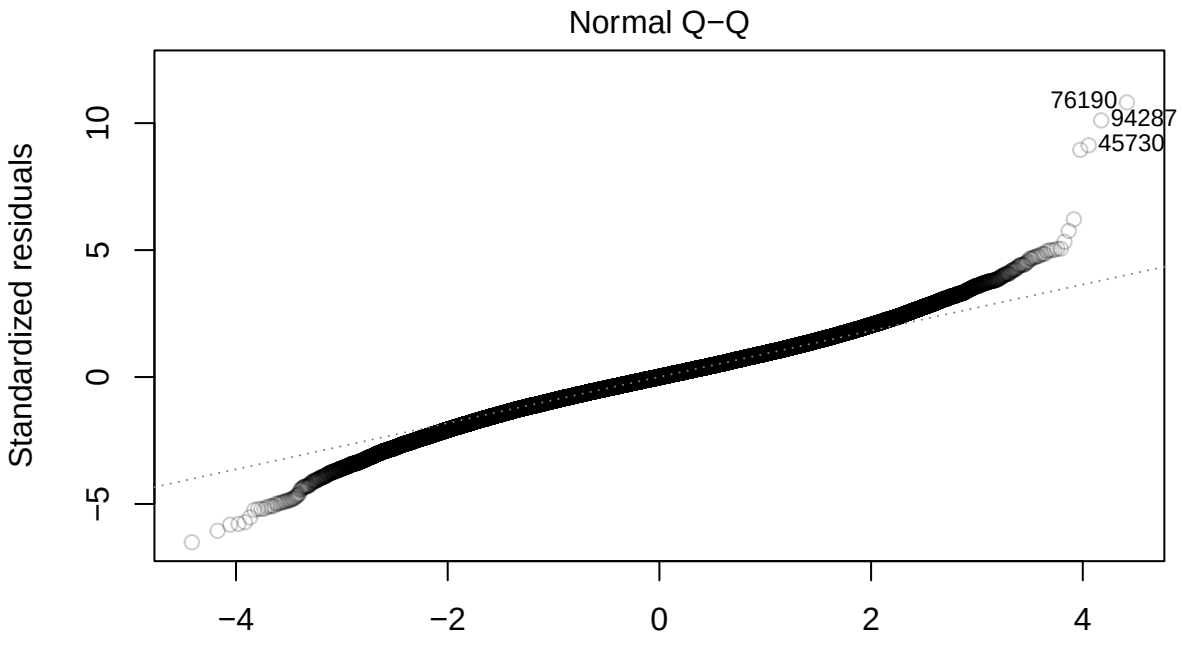
6 Model Diagnostics

Although we require a statistically significant linear model from model selection, the linear model made strong and specific assumptions about the structure of our data (Fox, p266). These assumptions-linearity, constant variance, independent noise, and normality-do not often hold in applications. Moreover, the method of least squares can be very sensitive to unusual or influential data points (Fox, p266). Thus, to examine the credibility and validity of our model, we use a series of model diagnostics techniques to check the model assumptions and identify unusual or influential data points.

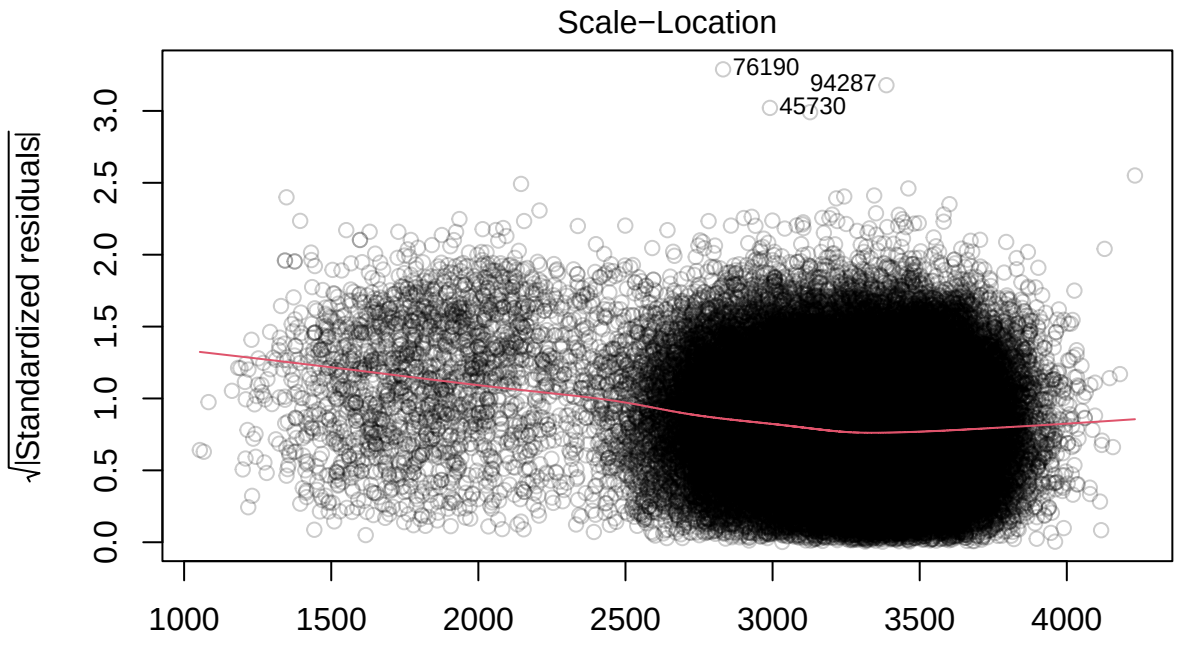
6.1 Linear Modeling Assumptions



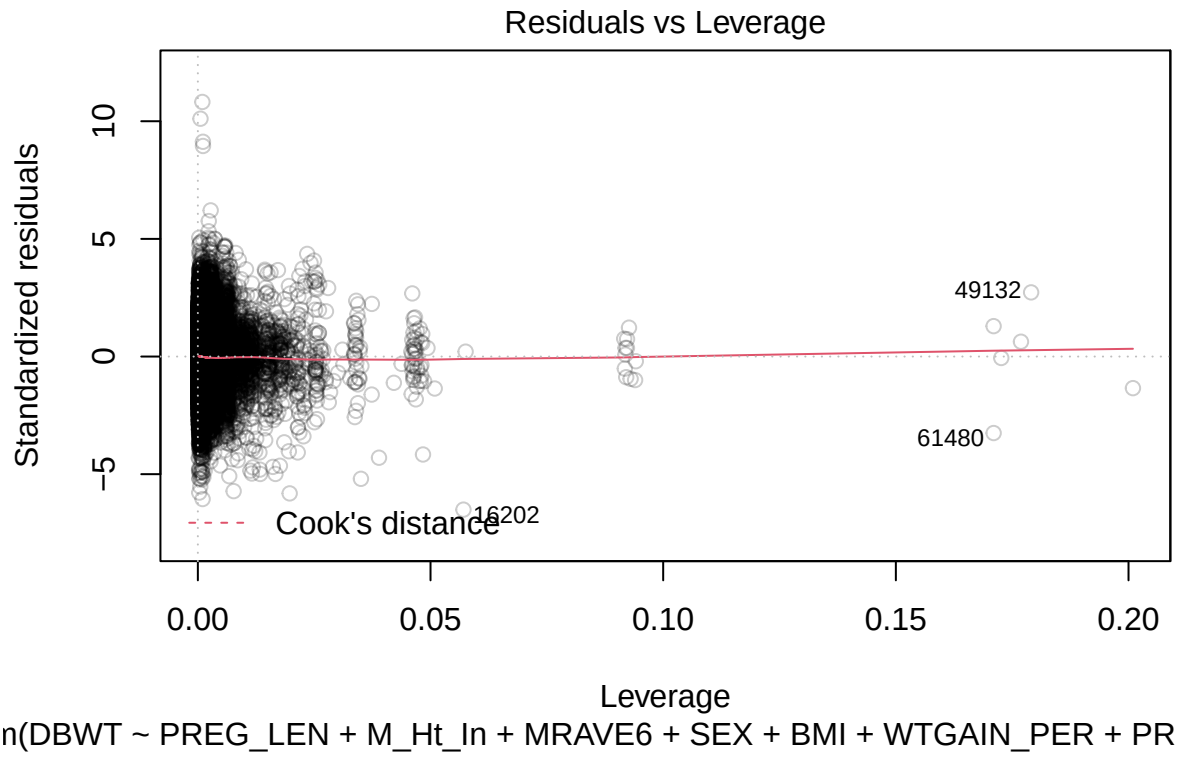
n(DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI + WTGAIN_PER + PRIORL



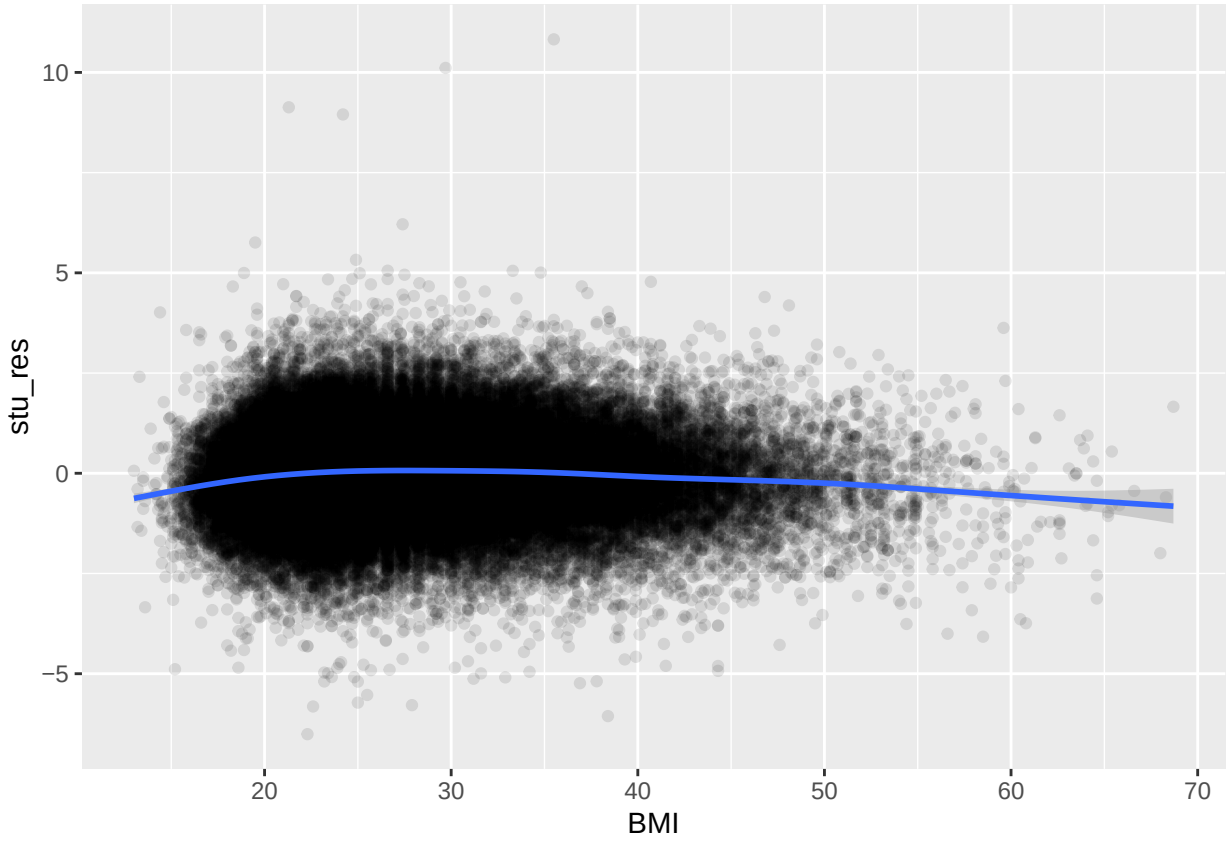
n(DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI + WTGAIN_PER + PRIORL



n(DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI + WTGAIN_PER + PRIORL



n(DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI + WTGAIN_PER + PRIORL



First, we will verify our modeling assumptions using various diagnostic plots. We skip the independent noise assumption because we are not dealing with geospatial and time series data, so we could safely assume that noises are independent of each other.

When plotting the residuals against the fitted values (Figure X), we do not see a clear trend in the spread of residuals as a function of fitted values, which supports our constant variance assumption. Additionally, because the residuals do not show any clear non-linear pattern, the plot (Figure X) supports our linearity assumption. To help us verify normality assumptions, we also plot a quantile-comparison plot of the standardized residuals against the normal distribution (Figure X). Examining the shape of the Q-Q plot, we see that the distribution of the residuals has slightly heavy tails, indicating potential violation of this assumption. Although we can use case bootstrapping to alleviate this issue, we choose to continue with our original data since the issue is not severe. In the scale-location plot, the red line is roughly horizontal, providing additional evidence for the validity of the constant variance assumption (Figure X).

Finally, we plot the studentized residuals versus one of the explanatory variables, BMI, and look for any patterns (Figure X). The studentized residuals appear to be mostly centered around 0 with no clear pattern, which supports our linearity assumption. However, we notice some downward curvature towards the extreme values of BMI, indicating a possible slight violation of our linearity assumption. Additionally, the spread of the studentized residuals does not seem to have a strong dependence on BMI, thus demonstrating homoscedasticity and indicating support for our constant variance assumption.

6.2 Unusual, Influential Data Points

```
## 76190
##  TRUE
```

```
##       ATTEND BFACIL  BMI CIG_0 DBWT DMAR FAGECOMB FEDUC FRACE6 LD_INDL MAGER
## 76190      1      1 35.5 FALSE 7940    1       36     6      1       N    34
## 94287      1      1 29.7 FALSE 8160    1       31     6      1       N    38
## 45730      1      1 21.3 FALSE 7300    1       25     4      2       N    22
## 34402      1      1 24.2 FALSE 7352    2       35     3      1       Y    35
## 16202      1      1 22.3 FALSE 1247    2       22     4      2       N    24
## 82652      3      1 27.4 FALSE 5075    1       36     3      1       N    36
##       MBSTATE_REC MEDUC MRAVE6 M_Ht_In NO_INFEC NO_MMORB NO_RISKS PAY_REC
## 76190           1     7      1      64        1        1        0       2
## 94287           1     8      1      66        1        1        1       2
## 45730           2     4      1      63        1        1        1       3
## 34402           1     5      1      60        1        1        1       2
## 16202           1     4      2      61        1        1        1       1
## 82652           1     6      1      62        1        1        1       2
##       PRECARE PREVIS PRIORDEAD PRIORLIVE PRIORTERM RDMETH_REC RESTATUS RF_CESAR
## 76190    TRUE     15     FALSE     FALSE      TRUE          3        1        N
## 94287    TRUE     12     FALSE      TRUE      TRUE          1        1        N
## 45730    TRUE     10     FALSE      TRUE     FALSE          1        2        N
## 34402    TRUE     14     FALSE     FALSE      TRUE          1        1        N
## 16202    TRUE     60     FALSE      TRUE      TRUE          3        1        N
## 82652    TRUE      8     FALSE      TRUE     FALSE          1        1        N
##       SEX PREG_LEN WTGAIN_PER FIRST_BIRTH   stu_res
## 76190   F        8 0.16425121        TRUE 10.825073
## 94287   F        9 0.08695652       FALSE 10.112046
## 45730   M        8 0.23333333       FALSE  9.130358
## 34402   M        8 0.51612903        TRUE  8.951312
## 16202   M    Early 0.30508475       FALSE -6.507789
## 82652   M    Early 0.36666667       FALSE  6.212057

##       ATTEND BFACIL  BMI CIG_0 DBWT DMAR FAGECOMB FEDUC FRACE6 LD_INDL MAGER
## 16202      1      1 22.3 FALSE 1247    2       22     4      2       N    24
## 61480      1      1 23.8 FALSE 1760    2       32     4      1       N    33
## 49132      1      1 29.6 FALSE 4082    1       28     2      2       N    28
##       MBSTATE_REC MEDUC MRAVE6 M_Ht_In NO_INFEC NO_MMORB NO_RISKS PAY_REC
## 16202           1     4      2      61        1        1        1       1
## 61480           2     4      1      65        1        1        1       1
## 49132           2     2      2      65        1        1        1       1
##       PRECARE PREVIS PRIORDEAD PRIORLIVE PRIORTERM RDMETH_REC RESTATUS RF_CESAR
## 16202    TRUE     60     FALSE      TRUE      TRUE          3        1        N
## 61480   FALSE      0     FALSE     FALSE      TRUE          1        1        N
## 49132   FALSE      0     FALSE      TRUE     FALSE          1        1        N
##       SEX PREG_LEN WTGAIN_PER FIRST_BIRTH   stu_res
## 16202   M    Early 0.30508475       FALSE -6.507789
## 61480   M     Late 0.21678322        TRUE -3.251874
## 49132   M     Late 0.01685393       FALSE  2.726319
```

Next, we detect and analyze unusual data points that may significantly affect the fitted coefficients of our model. In our diagnostic plots, we observed a few unusual data points, possibly outliers, with indices 76190, 94287, and 45730 in Figure X. To identify the influential data points, we check if there is any point outside the contour of the Cook's distance equal to 0.5 in the residuals vs. leverage plot. A larger Cook's distance means a larger influence of a data point on the coefficient estimation. As we cannot even see the contour in the plot, we claim that no data point is highly influential. which shows how the residuals behave as leverage increases, and we see that the spread does not change with leverage.

Mention OVERPLOTTING.

7 Model Interpretation

Our final model, obtained from forward selection with AIC, contains 103 regressors. For model interpretation, to avoid the issue of post-selection inference, we re-fit the model to the test set. This model will be used for interpreting the model's coefficients.

7.1 Causal Inference

```
##
## Call:
## lm(formula = DBWT ~ PREG_LEN + M_Ht_In + MRAVE6 + SEX + BMI +
##     WTGAIN_PER + PRIORLIVE + CIG_0 + NO_RISKS + RDMETH_REC +
##     PREVIS + ATTEND + MBSTATE_REC + FRACE6 + PAY_REC + LD_INDL +
##     FEDUC + NO_MMORB + BFACIL + FAGECOMB + NO_INFEC + RESTATUS +
##     MEDUC + PRECARE + DMAR + PREVIS * PREG_LEN + PREG_LEN * MEDUC +
##     CIG_0 * PRECARE + PRECARE * PREG_LEN + CIG_0 * PREG_LEN,
##     data = Test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3225.4  -286.8     0.0   292.7  4736.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -446.8676    96.3540  -4.638 3.53e-06 ***
## PREG_LEN8                 771.3050    97.7757   7.889 3.09e-15 ***
## PREG_LEN9                1079.8240    89.5016  12.065  < 2e-16 ***
## PREG_LEN10               1167.9886   102.7412  11.368  < 2e-16 ***
## PREG_LENLate              412.3407   241.6996   1.706 0.088010 .
## M_Ht_In                    30.1895     0.5617  53.751  < 2e-16 ***
## MRAVE62                  -115.5742     7.8106 -14.797  < 2e-16 ***
## MRAVE63                    57.8392    19.3229   2.993 0.002761 **
## MRAVE64                   -14.9299    10.9812  -1.360 0.173964
## MRAVE65                   -57.9029    37.2216  -1.556 0.119801
## MRAVE66                   -40.0173    10.1873  -3.928 8.57e-05 ***
## SEXM                      116.8002     2.9905  39.057  < 2e-16 ***
## BMI                        17.6692     0.2887  61.193  < 2e-16 ***
## WTGAIN_PER                886.2944    16.2828  54.431  < 2e-16 ***
## PRIORLIVETRUE             101.8852     3.5075  29.048  < 2e-16 ***
## CIG_0TRUE                 -36.2658    60.5929  -0.599 0.549497
## NO_RISKS1                 135.7091     4.3023  31.543  < 2e-16 ***
## RDMETH_REC2               115.2483    11.2566  10.238  < 2e-16 ***
## RDMETH_REC3               -39.3779     4.1728  -9.437  < 2e-16 ***
## RDMETH_REC4               126.5207     6.1584  20.544  < 2e-16 ***
## PREVIS                     49.2285     2.1659  22.729  < 2e-16 ***
## ATTEND2                     1.0809     5.4209   0.199 0.841957
## ATTEND3                    50.8428     5.4605   9.311  < 2e-16 ***
## ATTEND4                    70.2091    20.5071   3.424 0.000618 ***
## ATTEND5                    43.5789    19.1014   2.281 0.022524 *
## MBSTATE_REC2               53.1461     4.5901  11.578  < 2e-16 ***
## FRACE62                   -57.7418     7.3991  -7.804 6.06e-15 ***
## FRACE63                    29.0755    19.9377   1.458 0.144757
## FRACE64                  -129.7002    11.1709 -11.611  < 2e-16 ***
```

```
## FRACE65                    39.0590    35.7301   1.093 0.274323
## FRACE66                   -33.3422    10.1685  -3.279 0.001042 **
## PAY_REC2                   17.9928     4.0698   4.421 9.83e-06 ***
## PAY_REC3                   34.6201     8.9454   3.870 0.000109 ***
## PAY_REC4                   21.5713     8.3915   2.571 0.010154 *
## LD_INDLY                   35.1811     3.4800  10.110  < 2e-16 ***
## FEDUC2                    -37.7633    12.3375  -3.061 0.002208 **
## FEDUC3                    -16.9855    11.8353  -1.435 0.151244
## FEDUC4                     -3.9415    12.1874  -0.323 0.746385
## FEDUC5                     10.4299    12.9450   0.806 0.420414
## FEDUC6                     18.3435    12.4402   1.475 0.140341
## FEDUC7                     16.4602    13.2590   1.241 0.214446
## FEDUC8                     17.5042    14.7607   1.186 0.235677
## NO_MMORB1                 -83.8594    12.4536  -6.734 1.66e-11 ***
## BFACIL2                    60.1776    19.4558   3.093 0.001982 **
## BFACIL3                    98.9886    20.1652   4.909 9.17e-07 ***
## BFACIL4                  -122.2570    50.5680  -2.418 0.015622 *
## BFACIL5                    7.8084   136.8790   0.057 0.954509
## BFACIL6                   51.5385   101.2064   0.509 0.610584
## BFACIL7                   26.7850    57.1495   0.469 0.639298
## FAGECOMB                  -0.7553     0.2595  -2.911 0.003607 **
## NO_INFEC1                  7.6718    10.6255   0.722 0.470288
## RESTATUS2                 -15.0041     3.3041  -4.541 5.60e-06 ***
## RESTATUS3                 -20.9953     9.1247  -2.301 0.021397 *
## RESTATUS4                 -55.4778    30.4563  -1.822 0.068525 .
## MEDUC2                   -289.5230    74.2383  -3.900 9.63e-05 ***
## MEDUC3                   -356.4660    69.7739  -5.109 3.25e-07 ***
## MEDUC4                   -498.3213    70.5268  -7.066 1.61e-12 ***
## MEDUC5                   -456.3809    74.4826  -6.127 8.97e-10 ***
## MEDUC6                   -588.3729    71.0434  -8.282  < 2e-16 ***
## MEDUC7                   -645.8416    75.6576  -8.536  < 2e-16 ***
## MEDUC8                   -454.7297    89.1494  -5.101 3.39e-07 ***
## PRECARETRUE              -289.1129    59.4746  -4.861 1.17e-06 ***
## DMAR2                     -13.6937     4.0225  -3.404 0.000664 ***
## PREG_LEN8:PREVIS          -40.1639     2.3891 -16.811  < 2e-16 ***
## PREG_LEN9:PREVIS          -43.8478     2.2226 -19.728  < 2e-16 ***
## PREG_LEN10:PREVIS         -42.9118     2.4235 -17.706  < 2e-16 ***
## PREG_LENLate:PREVIS       -39.4345     4.3790  -9.005  < 2e-16 ***
## PREG_LEN8:MEDUC2          236.8374    80.3986   2.946 0.003222 **
## PREG_LEN9:MEDUC2          242.5954    75.3855   3.218 0.001291 **
## PREG_LEN10:MEDUC2         222.2184    79.9226   2.780 0.005430 **
## PREG_LENLate:MEDUC2       484.5443   132.5339   3.656 0.000256 ***
## PREG_LEN8:MEDUC3          284.9772    75.3659   3.781 0.000156 ***
## PREG_LEN9:MEDUC3          314.2224    70.7317   4.442 8.90e-06 ***
## PREG_LEN10:MEDUC3         315.5045    74.7959   4.218 2.46e-05 ***
## PREG_LENLate:MEDUC3       520.8866   123.2808   4.225 2.39e-05 ***
## PREG_LEN8:MEDUC4          419.1204    76.1332   5.505 3.70e-08 ***
## PREG_LEN9:MEDUC4          459.2697    71.4390   6.429 1.29e-10 ***
## PREG_LEN10:MEDUC4         470.4747    75.5225   6.230 4.69e-10 ***
## PREG_LENLate:MEDUC4       623.6254   124.8982   4.993 5.95e-07 ***
## PREG_LEN8:MEDUC5          337.7226    80.4185   4.200 2.68e-05 ***
## PREG_LEN9:MEDUC5          430.1436    75.4569   5.701 1.20e-08 ***
## PREG_LEN10:MEDUC5         416.7550    79.7422   5.226 1.73e-07 ***
## PREG_LENLate:MEDUC5       583.8511   130.9582   4.458 8.27e-06 ***
```

```
## PREG_LEN8:MEDUC6          466.6621    76.5320   6.098 1.08e-09 ***
## PREG_LEN9:MEDUC6          571.3802    71.8651   7.951 1.87e-15 ***
## PREG_LEN10:MEDUC6         572.2985    75.8254   7.548 4.47e-14 ***
## PREG_LENLate:MEDUC6       712.4874   124.7064   5.713 1.11e-08 ***
## PREG_LEN8:MEDUC7          480.9709    81.4108   5.908 3.48e-09 ***
## PREG_LEN9:MEDUC7          626.9303    76.4978   8.195 2.53e-16 ***
## PREG_LEN10:MEDUC7         642.4993    80.5887   7.973 1.57e-15 ***
## PREG_LENLate:MEDUC7       728.1016   131.4678   5.538 3.06e-08 ***
## PREG_LEN8:MEDUC8          289.9933    95.8590   3.025 0.002485 **
## PREG_LEN9:MEDUC8          431.1129    90.1696   4.781 1.75e-06 ***
## PREG_LEN10:MEDUC8         442.0147    94.9694   4.654 3.26e-06 ***
## PREG_LENLate:MEDUC8       517.2541   154.0377   3.358 0.000785 ***
## CIG_0TRUE:PRECARETRUE      89.6994    52.7335   1.701 0.088947 .
## PREG_LEN8:PRECARETRUE     183.8012    70.9970   2.589 0.009631 **
## PREG_LEN9:PRECARETRUE     269.9576    62.6589   4.308 1.65e-05 ***
## PREG_LEN10:PRECARETRUE    281.9624    78.1815   3.607 0.000310 ***
## PREG_LENLate:PRECARETRUE  733.0657   225.4344   3.252 0.001147 **
## PREG_LEN8:CIG_0TRUE      -183.6907    36.2868  -5.062 4.15e-07 ***
## PREG_LEN9:CIG_0TRUE      -164.8395    33.8371  -4.872 1.11e-06 ***
## PREG_LEN10:CIG_0TRUE     -162.2409    36.3335  -4.465 8.00e-06 ***
## PREG_LENLate:CIG_0TRUE   -217.8188    60.0673  -3.626 0.000288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 472 on 99896 degrees of freedom
## Multiple R-squared:  0.3269, Adjusted R-squared:  0.3262
## F-statistic:   471 on 103 and 99896 DF,  p-value: < 2.2e-16
```

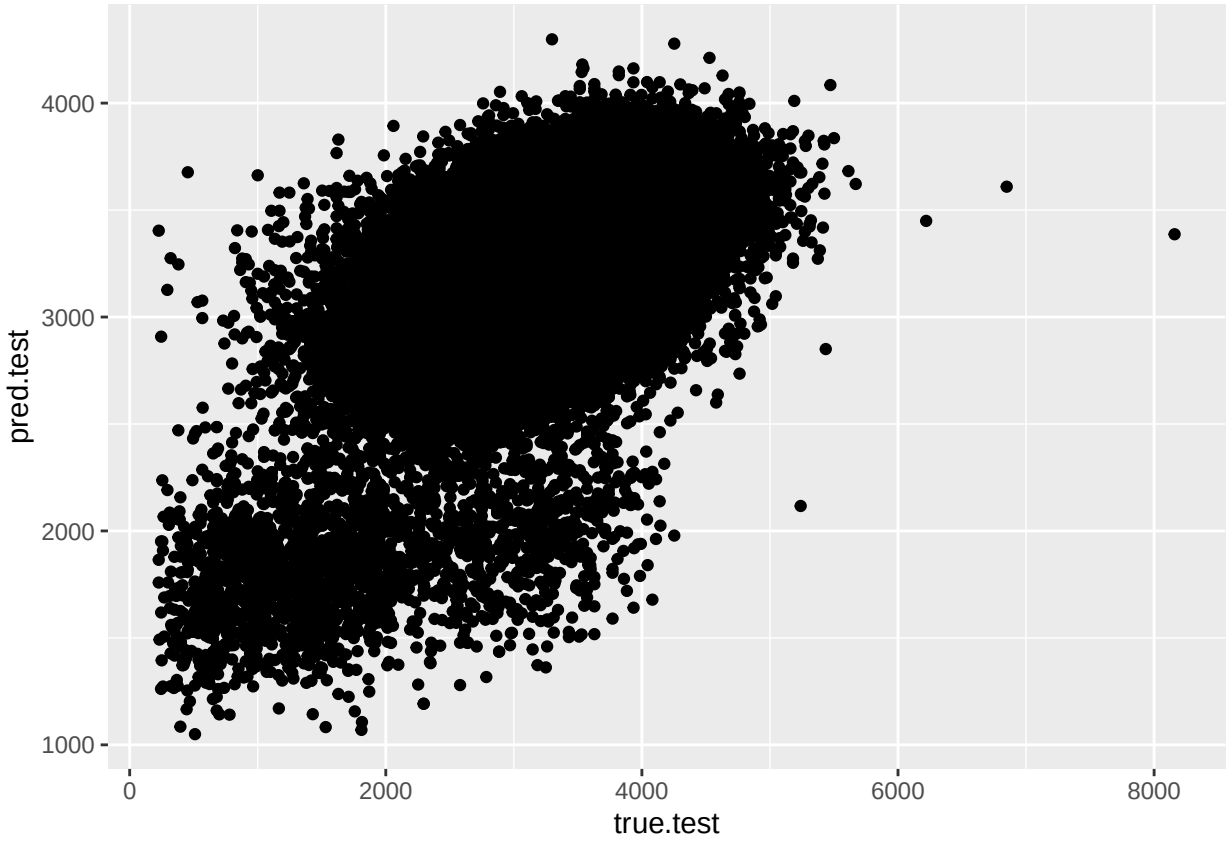
Based on the model's coefficient on `BMI`, we would expect that on average, with all other regressors held constant, a unit increase in a mother's pre-pregnancy BMI will be associated with an increase of baby birth weight by 17.6692 grams. The interpretation of other coefficients is complicated by interaction terms involving that coefficient. For example, because we included an interaction between `PREG_LEN` and `PRECARE`, we interpret the coefficient on `PRECARETRUE` relative to the reference dummy variable for pregnancy length. The coefficient on `PRECARETRUE` tell us that for mothers with a short estimated pregnancy length (less than 8 months), we expect that undergoing prenatal care will be associated with an average decrease in birth weight by 289.1129 grams when holding all other regressors constant. However, interestingly, the model coefficient on `PREG_LENLate:PRECARETRUE` is significant with a value of 733.0657. This means that for mothers with a late estimated pregnancy length (greater than 10 months), we expect that undergoing prenatal care will be associated with an average *increase* in birth weight by (733.0657 -289.1129) = 443.9528 grams.

To interpret the relative significance of explanatory variables in the regression, we could compute the standardized coefficients for all numerical variables and compare the corresponding magnitudes. However, we cannot compare the relative importance of categorical variables or interaction terms using this method. For these coefficients, we can rely on the significance of the coefficients as determined by incremental F-tests. Richard: What else should we add here? I think it's worth computing the actual standardized coeffs and having 1-2 sentences explaining some relative significance. I would be happy to do this. Rachel: should we also add confidence intervals for coefficients?

7.2 Prediction

```
## [1] 222923.6
```

```
## [1] 223176.5
```




```
##        fit      lwr      upr
## 1 3305.784 2380.539 4231.029
## 2 3432.877 2507.518 4358.235
## 3 3058.253 2132.846 3983.660
## 4 3494.928 2568.869 4420.986
## 5 3527.827 2602.576 4453.077
```

This model has an adjusted R^2 of 0.3262, which is very close to the R^2 achieved on the leading board on Kaggle. Our model achieves a test set mean squared error (MSE) of 223183.1, compared with a train set MSE of 222923.6. The relatively small difference between the train and test MSEs indicate that our model is not overfitting to the training set, since the model generalizes fairly well to the test set. To further evaluate whether the model will predict future baby birth weights with high precision, we also examine the 95% prediction intervals of a few randomly selected data points from the test set. Looking at the distribution of baby birth weights in Figure X, we see that the prediction intervals tend to be very wide relative to the entire distribution of values of `DBWT`, indicating that the model's predictions are imprecise. Therefore, the model will likely predict future baby birth weights with low precision.

8 Discussion

The primary purpose of this project was to determine whether we can predict the birth weight of a baby given information about the expecting family, and which factors we can intervene with to change the baby's birth weight. Therefore, we must consider the extent to which our linear model can be used for prediction and for causal inference.

As shown above, although our final model does not show signs of overfitting, the prediction intervals on the test set indicate that the model's predictions of birth weight for new data points are imprecise. This indicates that the model is unsuited for reliable prediction for new data points. Additionally, because the dataset used in this report is specific to US births, it is uncertain how well the model will generalize when predicting the birth weights in other regions of the world. Furthermore, because the test dataset uses only births in 2018, the model's performance has not been evaluated for predicting birth weights for babies born in the current year.

It is difficult to draw definitive applications for causal inference, because, according to a study from the University of Exeter, a baby's weight is mostly determined by his or her genetic code. This suggests that the explanatory variables in our dataset act primarily as proxies for the true cause of a baby's birth weight. Intervening with the explanatory variables in this dataset therefore does not guarantee that the baby's birth weight will change. Furthermore, only certain explanatory variables included within the model are intervenable. For example, it would not make sense for us to recommend expecting parents to change their race or change the sex of the baby, but we can recommend them to stop smoking cigarettes, schedule more prenatal visits, or adjust their BMI by following a weight management program. Thus, we suggest caution when attempting to use this model for causal inference.

9 Conclusion

In this report, we explored data about child births in the United States in 2018 from the National Center for Health Statistics. Given a set of explanatory variables describing information about the baby's parents, such as education status, BMI, and smoking history, we constructed linear models to predict a baby's birth weight. We selected main effects for our model using forward selection with AIC and added interaction terms based on EDA and incremental F-tests. Our final model included 103 regressors, achieving an adjusted R^2 of 0.3262. Our analysis demonstrated shortcomings in terms of precision when predicting the response variable birth weight.

For further analysis, we might explore imputation to minimize bias in our dataset. For numerical variables, we could replace the missing values by the mean. For categorical variables, we could replace the missing values by the mode.

We may also leverage more sophisticated machine learning methods such as random forest, KNN, or deep learning.