Stats 141SL

Rose Din (ID: 804764627), Lisa Kaunitz (ID: 404731988), Claire Markey (ID: 104789928), Allison Ocelo (ID: 404799327), Cristina Sanchez (ID: 805119924), and Kayla Schroeder (ID: 804729436) - Section 1A

03/14/2020

Contents

1 Abstract 2

2 Research Questions 3

3 Variables of the Study and how they were Measured 3
3.1 Final Predata Codebook 4

4 Exploratory Data Analysis 5
4.1 The Five Factors from Literature Review 5
4.2 Drug Use During Pregnancy and Risk Score 7
4.3 Controlling for Sex: 11

5 Statistical analysis used to answer the research questions 11

6 Summary of results 14
6.1 Tables 14
6.1.1 Predata Table 14
6.1.2 Predata Model Diagnostic Plots 14
6.1.3 Maledata Table 15
6.1.4 Maledata Model Diagnostic Plots 15
6.1.5 Femaledata Table 16
6.1.6 Femaledata Model Diagnostic Plots 16

7 Interpretation of Results 16

8 Overall conclusions 18

9 Challenges of the study 18

10 Recommendations for the future 19

11 Bibliography 20

1 Abstract

In this study, we worked to determine factors that predispose children to autistic behaviors. To do this, we explored a dataset obtained from the ChatterBaby application's survey which included input from users across the globe. We began by cleaning the dataset to remove missing values and performed some initial data analysis along with random forest modeling to determine which factors were most important of all prenatal factors. A Poisson model was then used to explain the data. In our literature review, we uncovered that autism diagnoses in males is systematically different from these same diagnoses in females and we want to determine if this difference in sex is statistically significant when looking at predictors for autism in the different sexes. As such, we chose to develop additional Poisson models that were separated by sex in an effort to control for this discrepancy. Through this modeling we found that there are a multitude of different risk factors that are statistically significant, however, these risk factors vary by sex. Consequently, we conclude that there is a significant difference between males and females regarding risk factors that predispose a child to ASD behaviors and that the number of mental health diagnoses for the mother is chief amongst the risk factors as this is the only factor that appears in all three models.

2 Research Questions

In our study, we developed two research questions concentrating on significant predictors that can detect if a child will exhibit behaviors that relate to having Autistic Spectrum Disorder, a developmental disorder that impairs communication and interaction, and if there is a difference in these predictors for males and females. Our first research question focuses on discovering risk factors that could predispose a child's tendency towards exhibiting behaviors that relate to having ASD. We examine the mother's prenatal characteristics in order to see if any of these factors increase their child's ASD behaviors and determine what these factors may be. Our second research question arose after our literature review, which found that the signs for ASD in boys can be diagnosed as early as two years of age while it may take up to 14 years for girls due to the differences in the way that autism presents itself between genders (Geelhand). We also found that there is a stark difference in the proportion of males versus females that have autism with the most recent ratio being 3 males for every female (National Autistic Society). Taking this into account, our second research question focuses on whether there is a significant difference in the predictors of ASD for males versus females and what these specific predictors are.

3 Variables of the Study and how they were Measured

For this study, only prenatal factors were taken into account. Any variable that was not clearly a factor from the time before a child was born was removed and the subsetted dataset "predata" was created. This dataset contains 68 predictors and the outcome variable `risk_score_ASD`. Binary variables coded as numeric were converted to factors where 1 means an attribute is present and 0 means not present. The variable `delivery` was converted to a factor as the numbers indicate a certain type of delivery method. All other variables were kept as numeric.

Nine variables in "predata" were created from other variables in the original dataset; these are `mathealth_count`, `babymortality` and 7 variables beginning with "tri_." The count variable "mathealth_count" created from "matmentalhealth" tells how many different mental health diagnoses the mother had before, during, or after pregnancy. This was done by summing the binary variables under "matmentalhealth." The series of count variables beginning with "tri_" were created from the "drugsalcoholtobacco" categories and they tell how many trimesters a mother used a certain drug for (1, 2, or 3). The variable "babymortality" conveys pregnancies that resulted in a mother not having a live birth so the variable is a difference of "totalpregnacies" and "momslivebirths."

3.1 Final Predata Codebook

Data Codebook		
Number	Attribute	Description
1	momsdeliveryage	Mother's Age at Delivery.
2	momsedu	Mother's Education Level. 1, Less than High School | 2, High School Graduate | 3, Some College, Trade School, or Associates Degree | 4, College Graduate (Bachelors) | 5, Graduate School | 6, Not Sure/Unknown
3	delivery	How was your baby delivered? 1, Unassisted Vaginal Birth|2, Assisted Vaginal Birth (forceps, vacuum)|3, Planned Caesarean|4, Emergency Caesarean
4	babymortality	Misscarriages, or infant death at birth
5	mathealth count	Maternal mental health count
6	sex	Baby's Gender. 1 = Male, 0 = Female
7	tri alcohol	Total number of trimesters alcohol was used.
8	tri tabacco	Total number of trimesters tabacco was used.
9	tri vaping	Total number of trimesters vaping was used.
10	tri cannabis	Total number of trimesters cannabis was used.
11	mom trauma	Did the mother experience any extremely stressful event during pregnancy, such as the death of a family member, divorce, homelessness, living in a war zone, or abuse?
12	asd1	Total number of family members with ASD (experiencing other) 1 = have 1+ family members with ASD, 0 otherwise.
13	szbpfamily1	Total number of family members with SZ/BP (experiencing other). 1 = have 1+ family members with SZ/BP, 0 otherwise.
14	ddfamily1	Did any of the following family members have a history of developmental delay while an infant? 1 = have 1+ family members with developmental delay, 0 otherwise
15	artmethod 1	In vitro fertilization (IVF) assisted in conceiving baby.
16	whenfever 3	Mothers fever during Third Trimester (weeks 28-40).
29	deaf1	Total number of family members Deaf (experiencing other). 1 = have 1+ family members Deaf, 0 otherwise.
17	maternalpregnancyproblems 1	Were there any of the following interventions, complications, or abnormalities during this pregnancy with baby - Gestational diabetes.
18	maternalpregnancyproblems 2	Infections requiring antibiotics .
19	maternalpregnancyproblems 3	Placenta previa.
20	maternalpregnancyproblems 4	Antidepressants to treat depression.
21	maternalpregnancyproblems 7	Preeclampsia/eclampsia .
22	maternalpregnancyproblems 8	Anemia.
23	maternalpregnancyproblems 9	Hypertension (high blood pressure).
24	maternalpregnancyproblems 11	Preterm labor.
25	maternalpregnancyproblems 13	Placental abruption.
26	maternalpregnancyproblems 14	Vaginal bleeding.
27	maternalpregnancyproblems 16	Hyperemesis gravidarum.
28	risk score ASD	Total number of ASD risk behaviors. Large number suggests higher risk.

4 Exploratory Data Analysis

4.1 The Five Factors from Literature Review

To begin our exploratory data analysis, the client provided us with an online article to get familiar with autism spectrum disorder. Literature review from this study revealed that there are certain factors that increase a child's risk for ASD: the child's sex, the child's family history, other disorders present in the child, if the child was born extremely preterm, and the parents' delivery ages ("Autism spectrum disorder"). Males are more likely than females to develop ASD ("Autism spectrum disorder"). A child's risk for autism spectrum disorder increases if a relative, such as a sibling, parent, cousin, etc., has already been diagnosed with the disorder ("Autism spectrum disorder"). Pre-existing health conditions such as fragile X syndrome, tuberous sclerosis, and Rett syndrome, can increase a baby's risk ("Autism spectrum disorder"). Babies born before 26 weeks of gestation are at higher risk compared to those born later ("Autism spectrum disorder"). Finally, children born to older parents at the time of delivery are at higher risk for autism spectrum disorder compared to those born to younger parents ("Autism spectrum disorder").

In order look further into some of these potential risk factors, boxplots were created in order to take a look at the distribution of ASD risk scores in relation to these variables. It should be noted that these plots were created prior any data cleaning. Hence, we saw what needed to be cleaned through some of the following graphs.

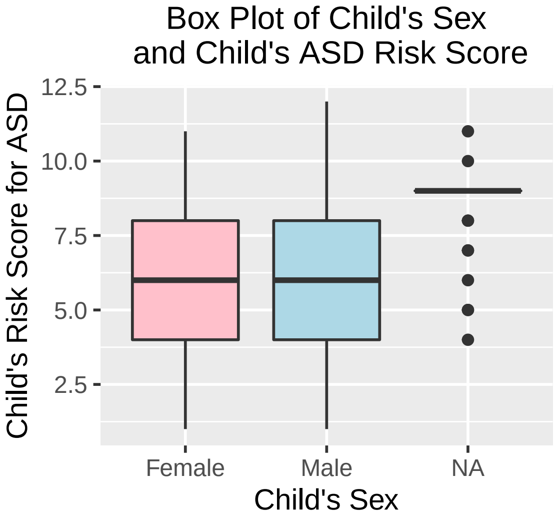


Figure 4.1.1

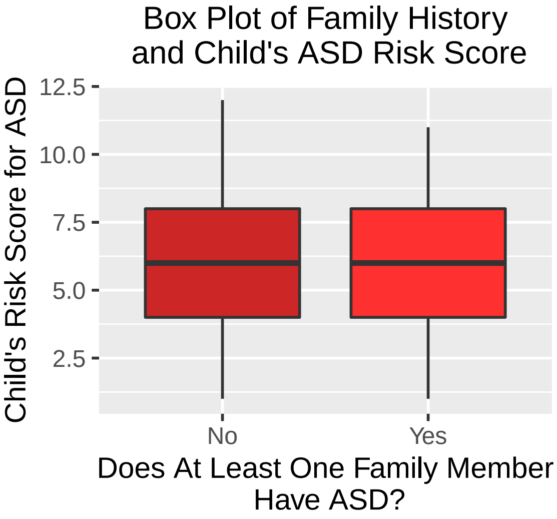


Figure 4.1.2

Figure 4.1.1 consists of the ASD risk score distributions based on males versus females. Figure 4.1.2 separates the visualizations based on family history of ASD diagnosis. For both, the median is around 6 and unfortunately, there is not much difference between the medians and distributions regardless of whatever category the observations fell in.

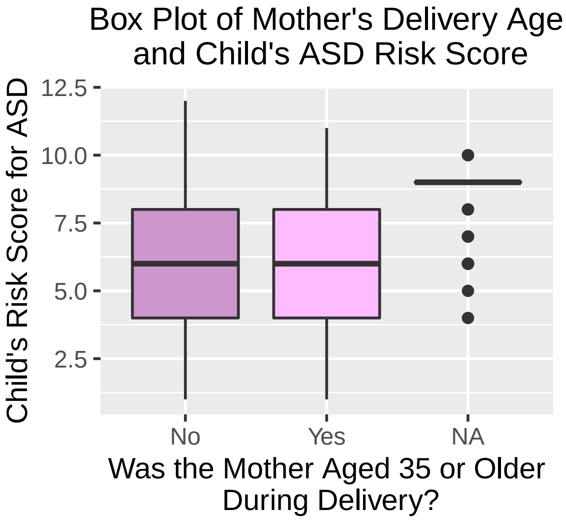


Figure 4.1.3

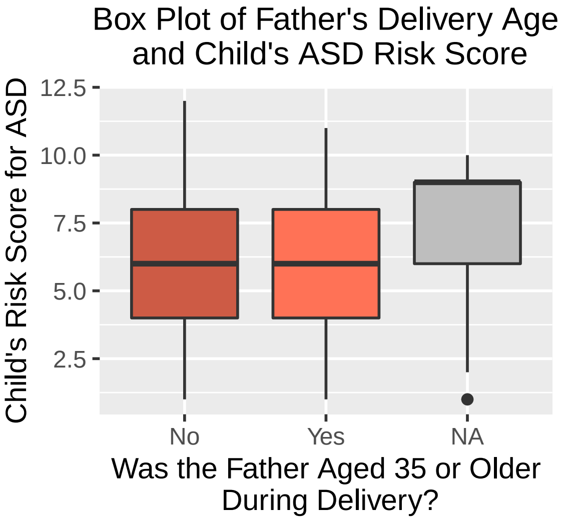


Figure 4.1.4

Figure 4.1.3 visualizes the mother's delivery age and Figure 4.1.4 visualizes the father's delivery age compared to risk score. These boxplots look similar to the previous ones and there is not much difference regardless if the parents were under or over 35 years of age. However, it can be noted that this is an area that still needs to be heavily researched.

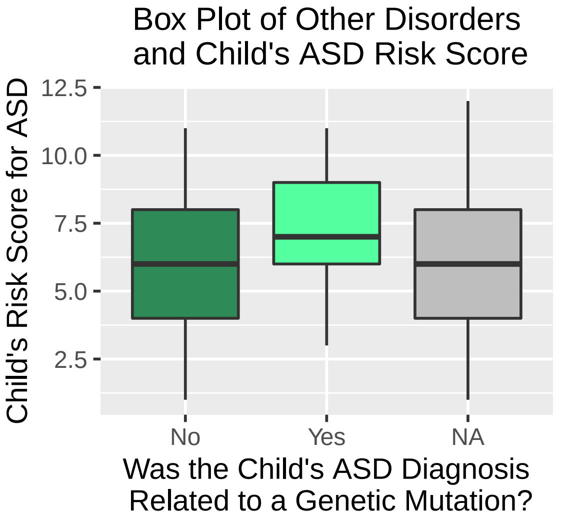


Figure 4.1.5

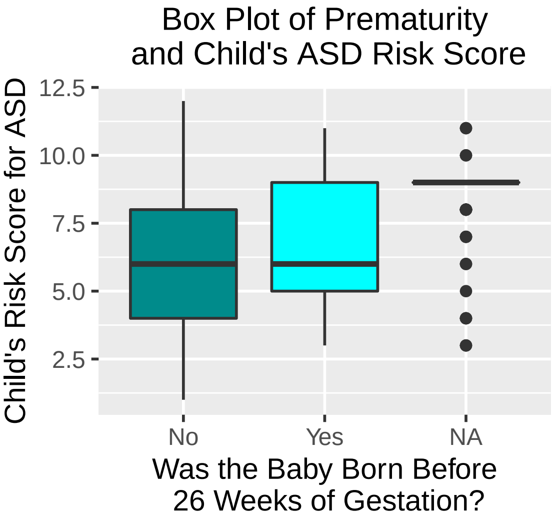


Figure 4.1.6

Figure 4.1.5 consists of distributions based on a child's diagnosis of a genetic mutation, such as fragile X syndrome, and Figure 4.1.6 compares babies born prior to 26 weeks of gestation versus those born later. Those whose ASD diagnosis was related to a genetic mutation have a different distribution and median in risk score compared to those who did not. It can also be observed that the 25th and 75th percentiles of those born before 26 weeks of gestation differs from those who were not.

4.2 Drug Use During Pregnancy and Risk Score

After cleaning the data set, we wanted to look at the relationships between drug use during pregnancy and risk score in order to potentially include these in our models. The seven main drugs listed on the survey were tobacco, vaping, alcohol, cannabis, stimulants (such as cocaine), opioids, and psychoactives (such as hallucinogens). However, one of our main concerns was if there was enough data on each drug. Response bias was potentially present in the survey data and underreporting could have occurred because some of the mothers may not have been truthful about their drug habits. Therefore, jitter plots were made to examine how many data points we had in each category. On the x-axis for all of the jitter plots, we have the number of trimesters a mother used that particular drug. The colors correspond to the risk score and basically follow the colors of a rainbow, beginning with red equaling 1 up to magenta equaling 12.

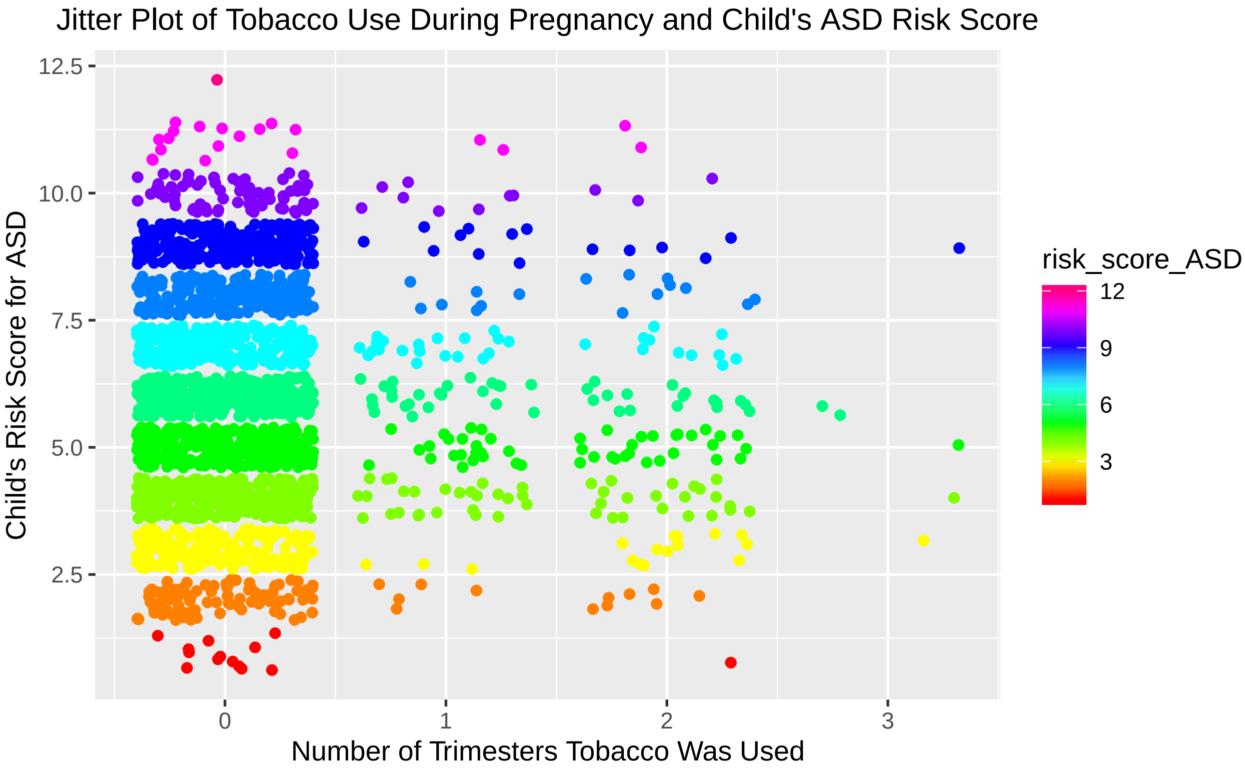


Figure 4.2.1

Figure 4.2.1 is a distribution of risk scores based on the mothers' smoking habits. An overwhelming majority of mothers did not smoke during pregnancy, but there were quite a number who did.

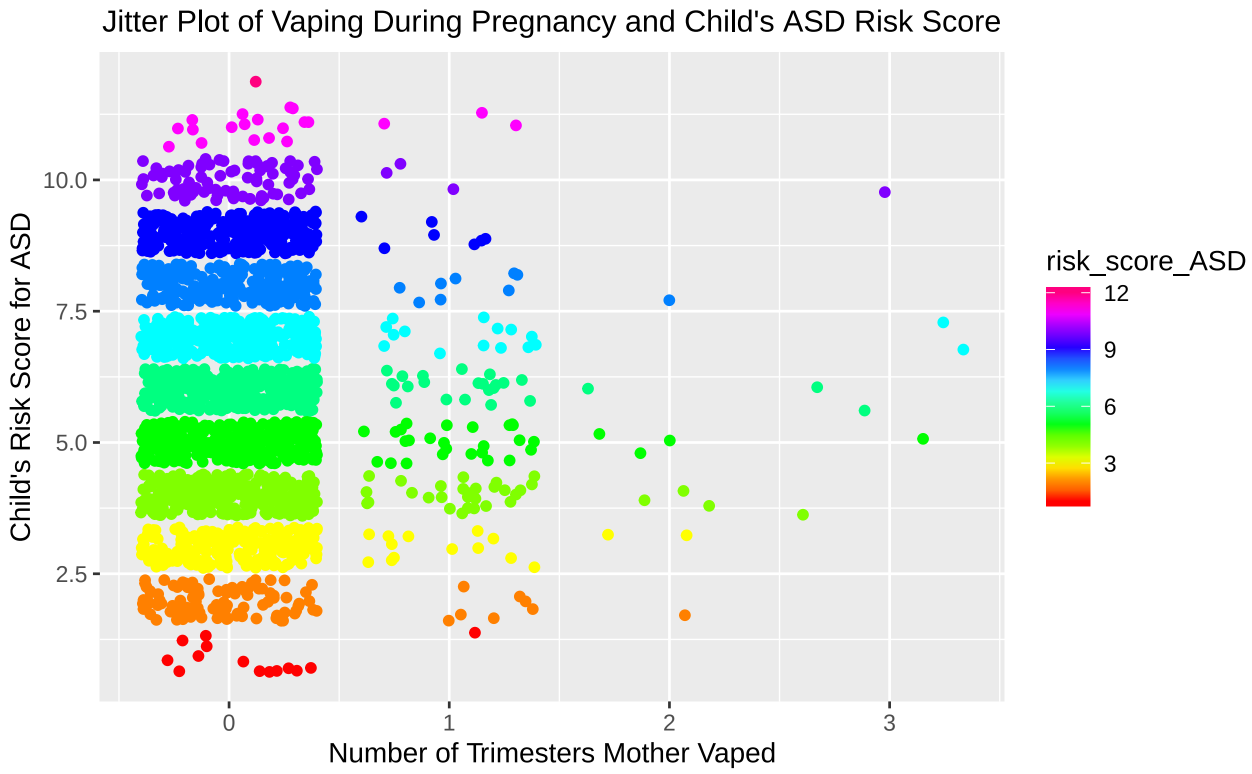


Figure 4.2.2

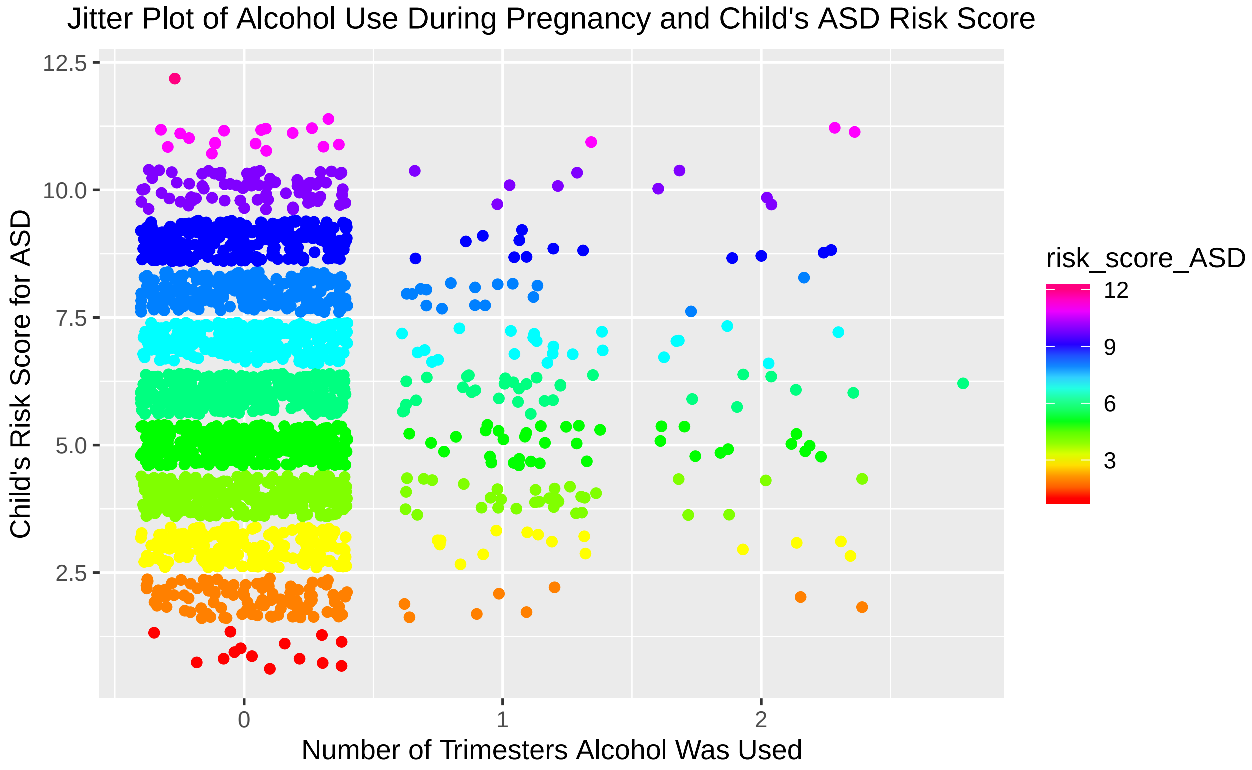


Figure 4.2.3

Figure 4.2.2 is based on a mother's vaping habit and Figure 4.2.3 is based in a mother's drinking habit. Drinking was more popular than vaping during pregnancy. So, it was a bit concerning that the data did not have many mothers who vaped.

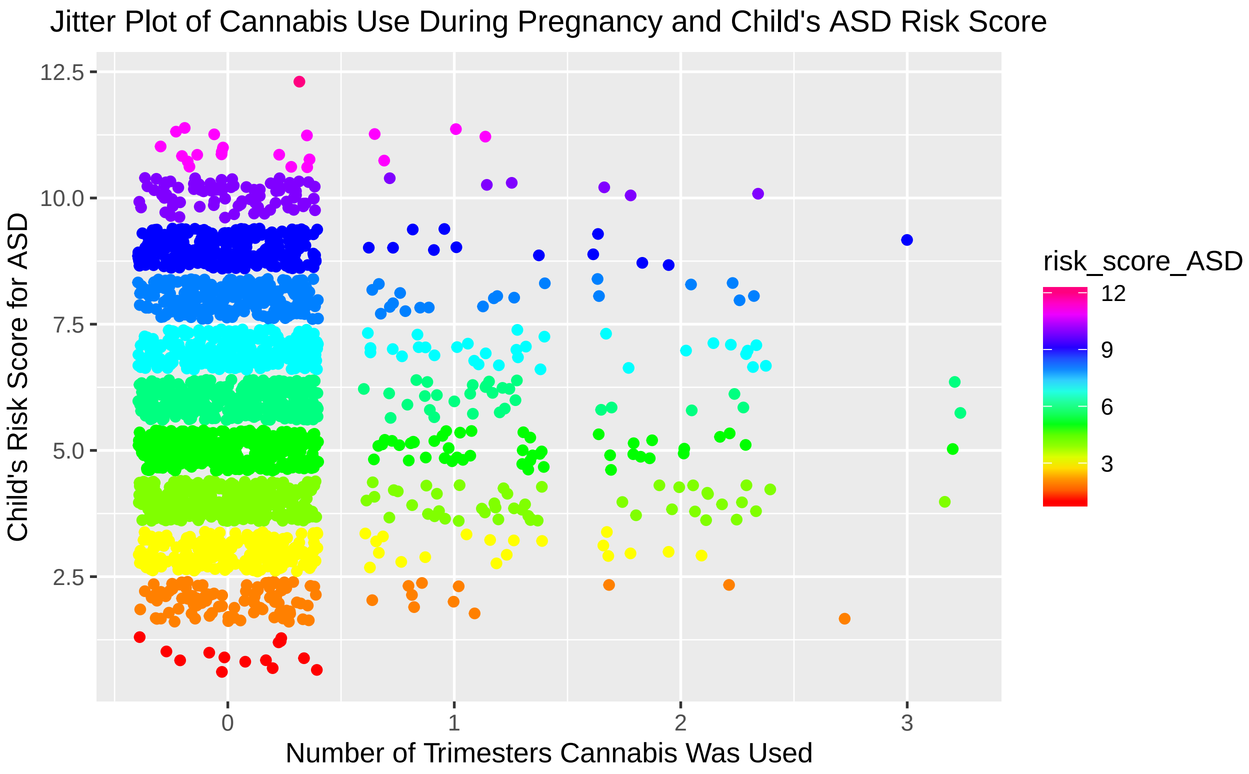


Figure 4.2.4

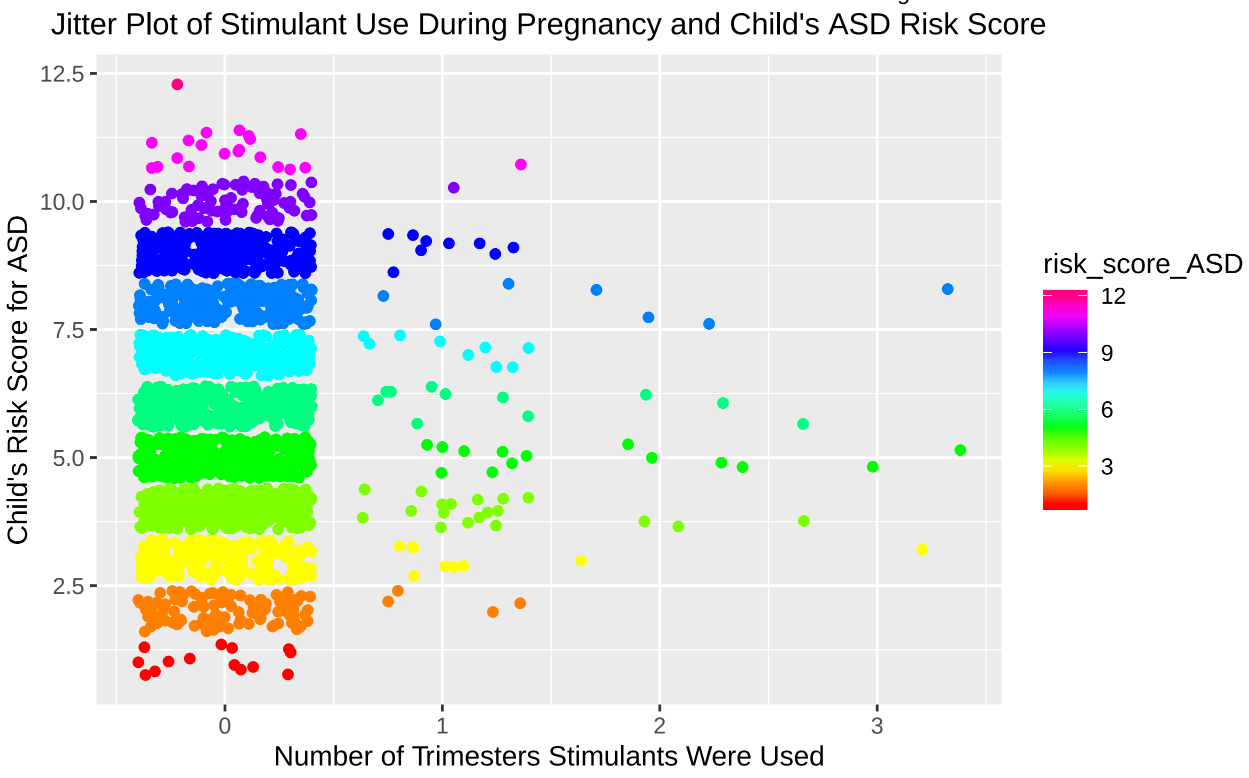


Figure 4.2.5

Figure 4.2.4 is based on cannabis use and Figure 4.2.5 is based on stimulant use. A decent number of mothers used cannabis during pregnancy and not that many used stimulants.

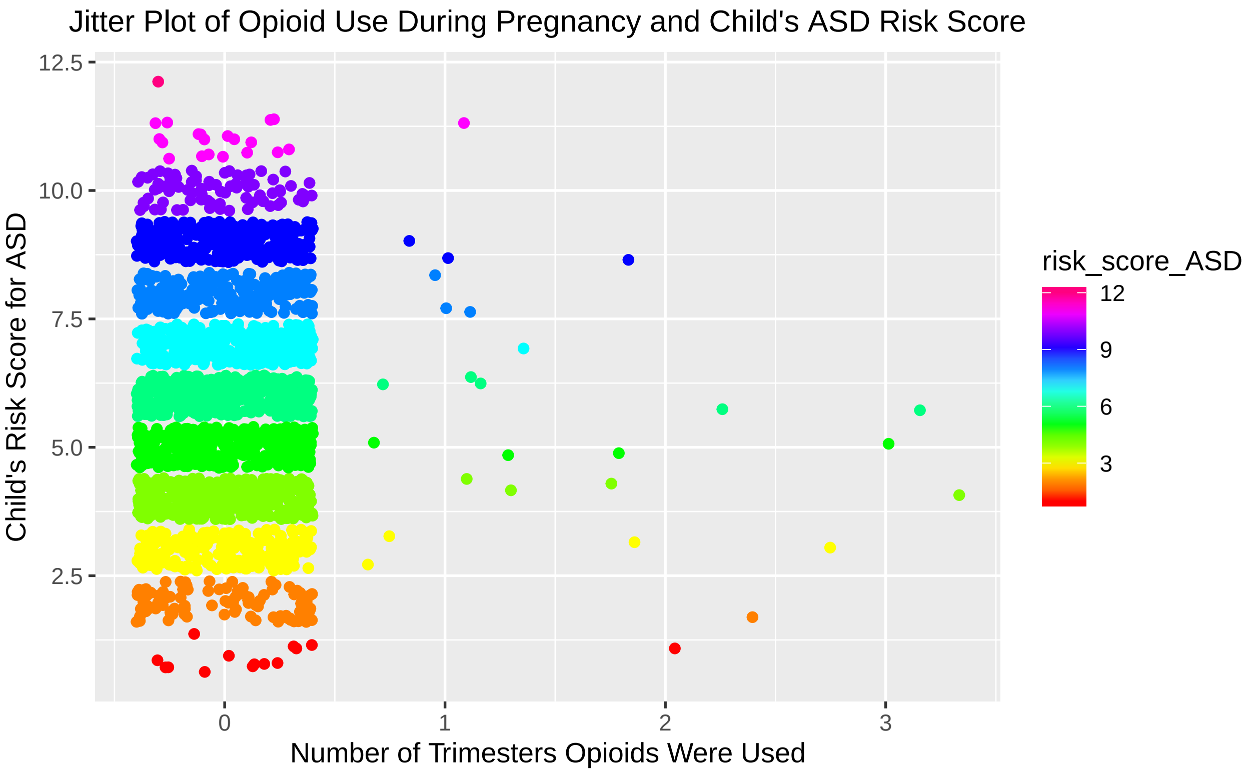


Figure 4.2.6

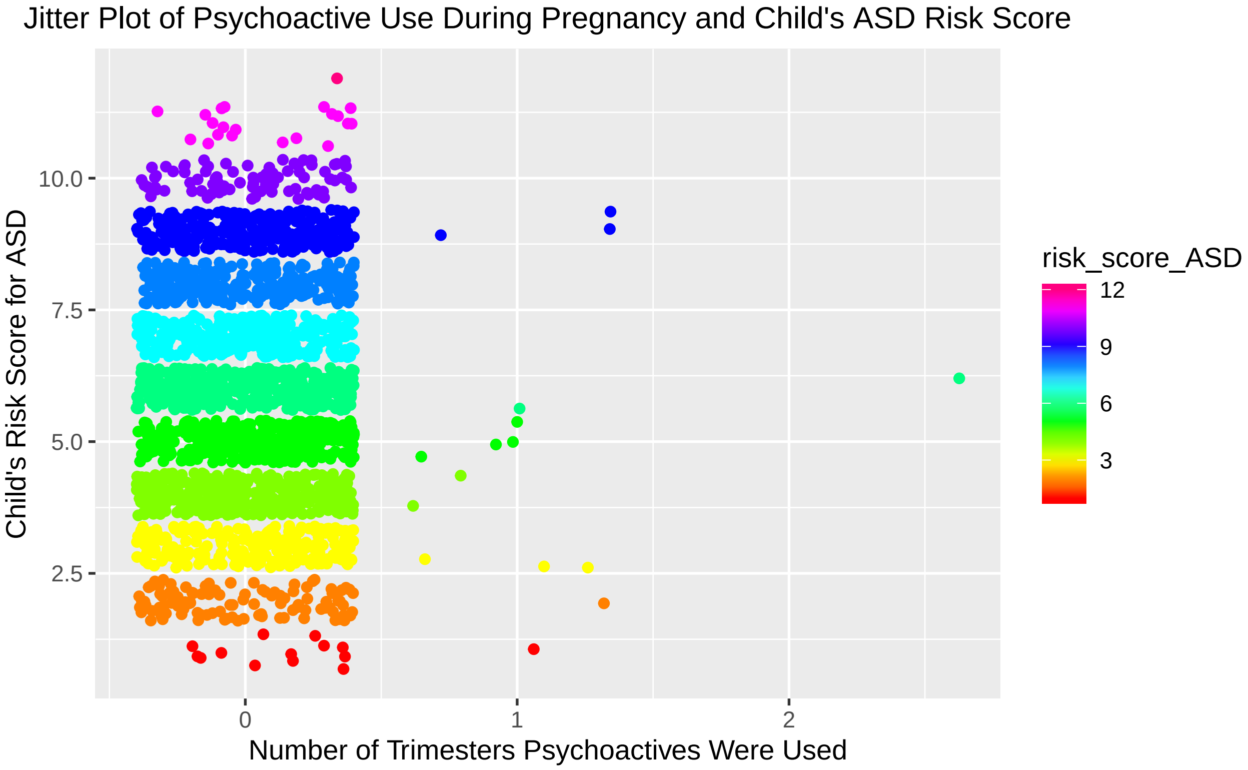
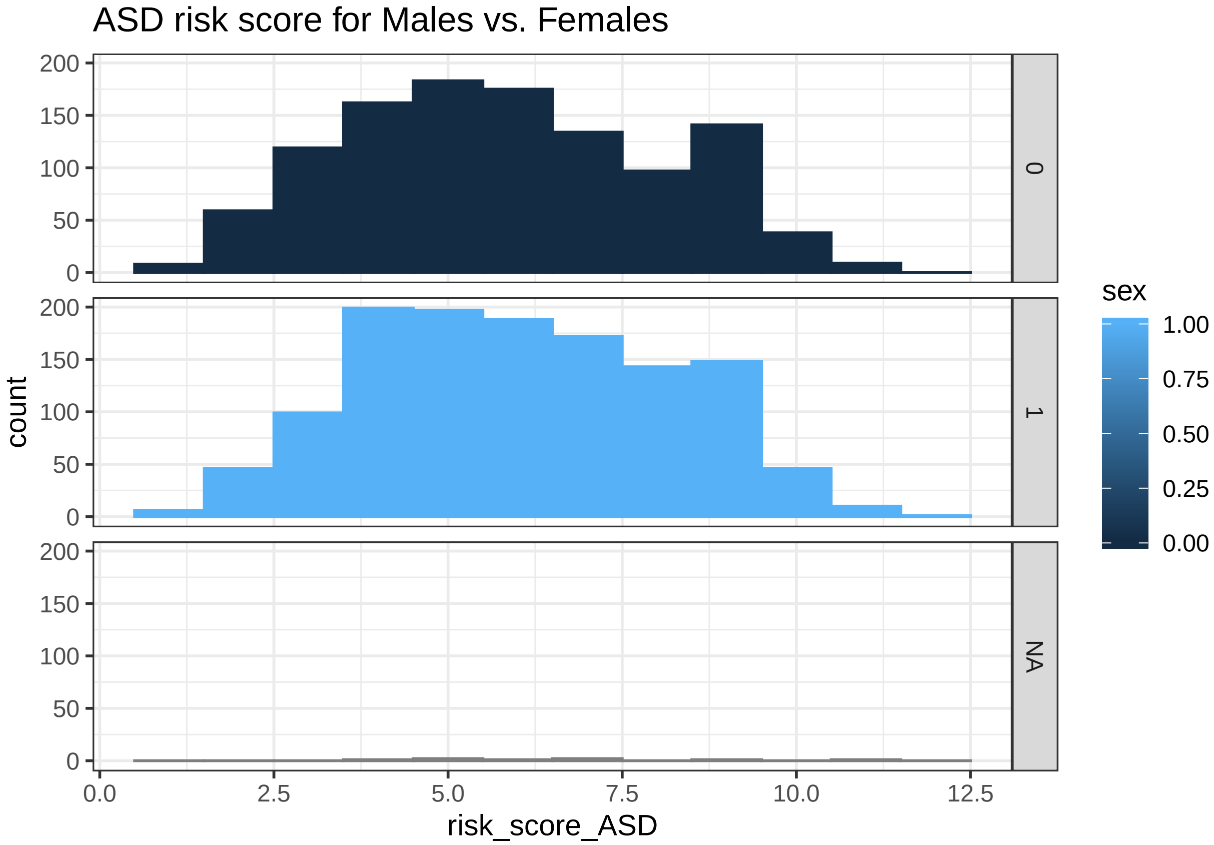


Figure 4.2.7

Finally, Figure 4.2.6 examines the use of opioids and Figure 4.2.7 examines the use of psychoactives. There were only around 18 mothers who used opioids and only 3 mothers who used psychoactives during pregnancy. Hence, there was not enough data present for these drugs to examine them for significance in our models.

4.3 Controlling for Sex:

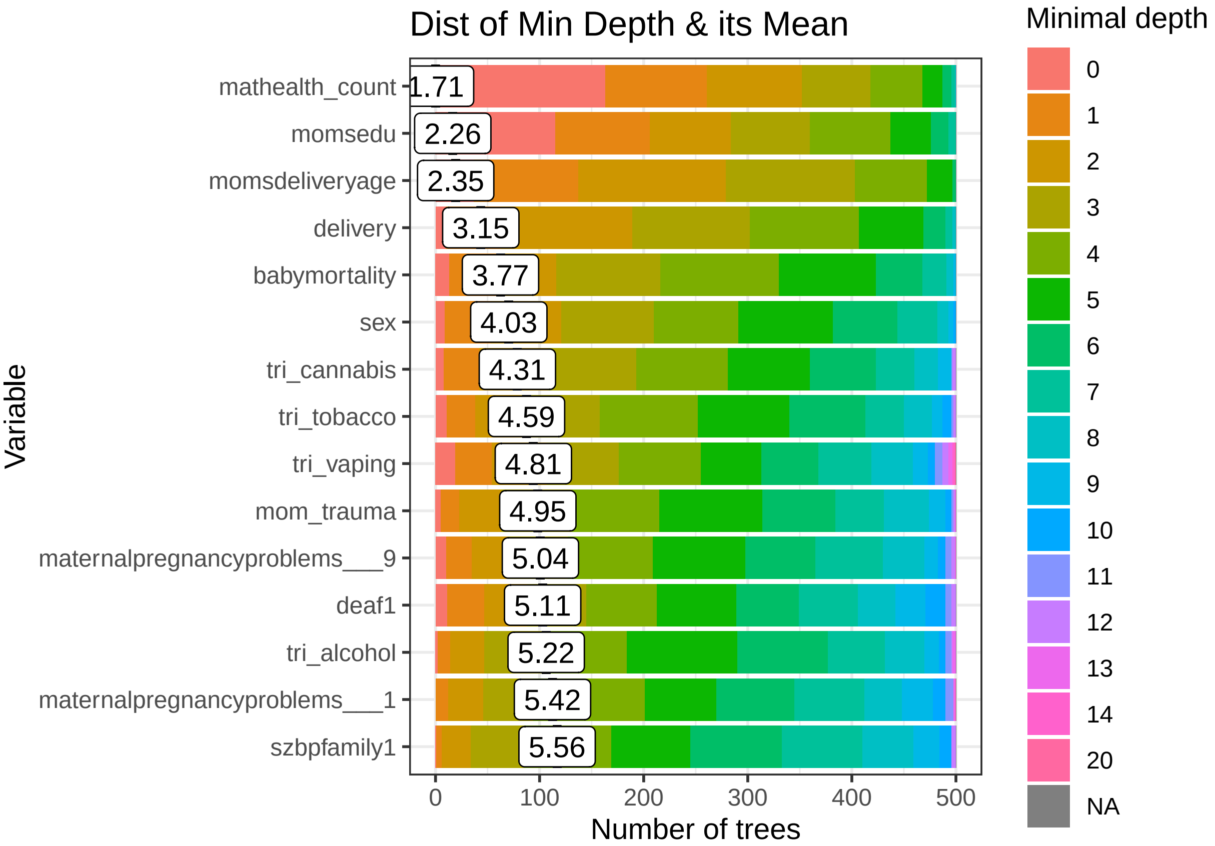
The literature mentions that there are three contributers to sex differences in ASD rates: biologically driven differences, gender socialization, and limited research and measurement tools. Fundamentally, young girls with ASD have proven to achieve more basic milestones on time compared to boys. They have overall better social skills and better peer relationships, which disrupts the typical bias and norms set in the standards for ASD. A recent study by the Children's National Health Systems shows that in current diagnostic practices, in order for females to have an ASD diagnosis they must present more concurrent behavioral, developmental or mental health issues, compared to their male counterparts. A significant factor that plays into these modern discrepancies is that most of the research and testing done on figuring out diagnosis and symptoms of ASD was on a majority of male patients, so many of the benchmarks for the diagnosis will tend to follow that for males, rather than for females. Due to the overwhelming significance and differences highlighted in our outside research of gender and ASD diagnosis, we decided to talk a look if our very own data will follow the literature.



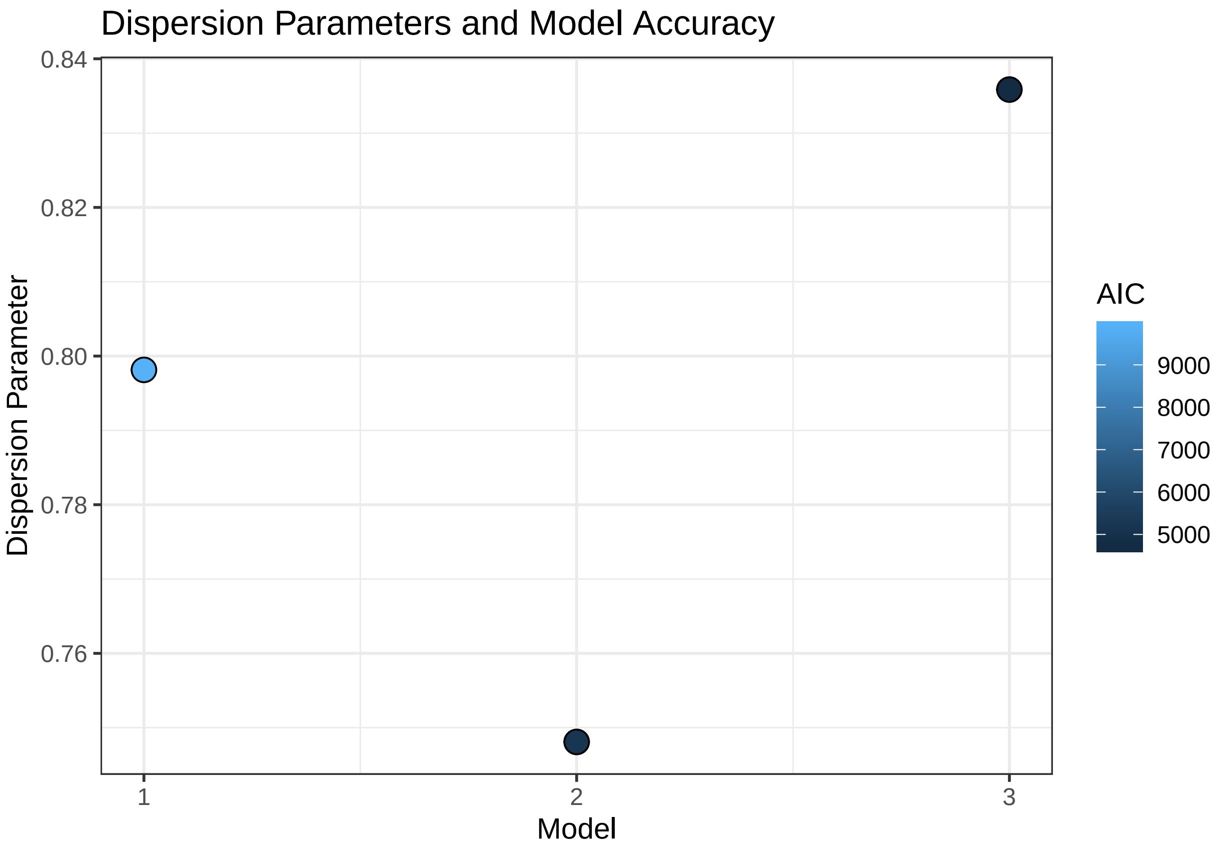
The boxplot above did not follow our expectations at first because it appears that there is no significant difference of the means of the risk scores between males and females. However, it is important to note that the NA's likely play a significant role in these outputs. Moreover, we can see that there tends to be more frequent high risk scores for males as opposed to females, so when we run our final models it is still worth looking into the significance that sex will have on determining the factors of ASD.

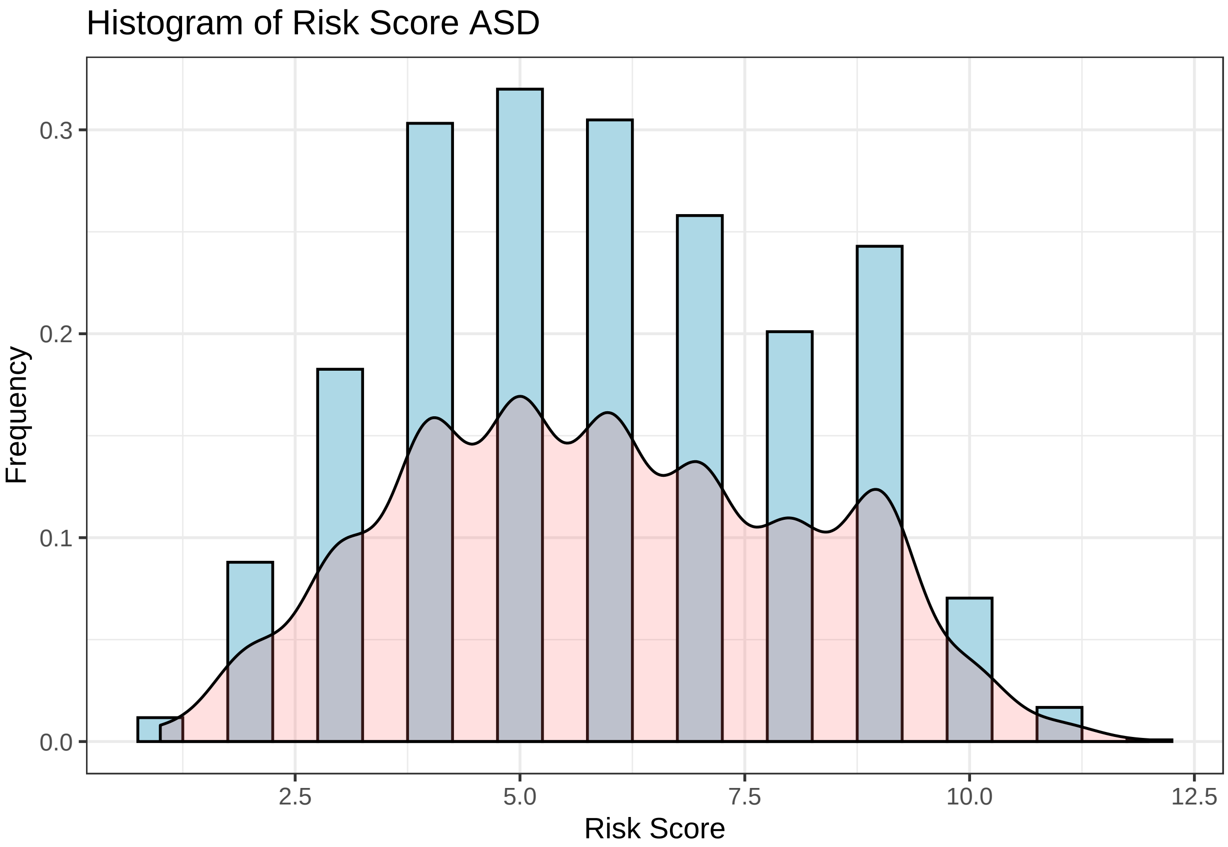
5 Statistical analysis used to answer the research questions

A random forest model was run on prenatal factors to determine which variables were important risk factors for `risk_score_ASD`. Highly correlated variables were dropped from the data before running the random forest model because having too many variables containing similar information will result in inaccurate importance measures due to a model weighing too much on one set of related variables compared to another set of variables. Important variables were determined from the node purity and mean squared error measures. From these measures, 14 total variables were selected as important risk factors from `risk_score_ASD` to be included in our final model.



Poisson regression was used to model the response variable, `risk_score_ASD` which consists of "count data". Three models were created including one full model, one model controlling for males, and one model controlling for females. A Poisson Model was chosen for our statistical analysis rather than a negative binomial model, which also handles count data, since the dispersion parameter for all three models was less than one, indicating the variance for each model was less than the expected value for each model (seen in the dispersion plot below). Assumptions of this model also include independence of observations and the distribution of counts following a Poisson distribution (seen in the density plot below). The AIC measures for all three models also indicate that the models controlling by gender fit the data better than the full model. In particular, the female model appears to fit the data the best due to its small AIC value relative to the other two models. These AIC measures coincide with our literature review and EDA, where males and females have different risk factors associated with ASD risk score.


Dispersion Parameters and Model Accuracy
0.84
0.82
0.80
0.78
0.76
Dispersion Parameter
1
2
3
Model
AIC
9000
8000
7000
6000
5000


Histogram of Risk Score ASD
0.3
0.2
0.1
0.0
Frequency
2.5
5.0
7.5
10.0
12.5
Risk Score

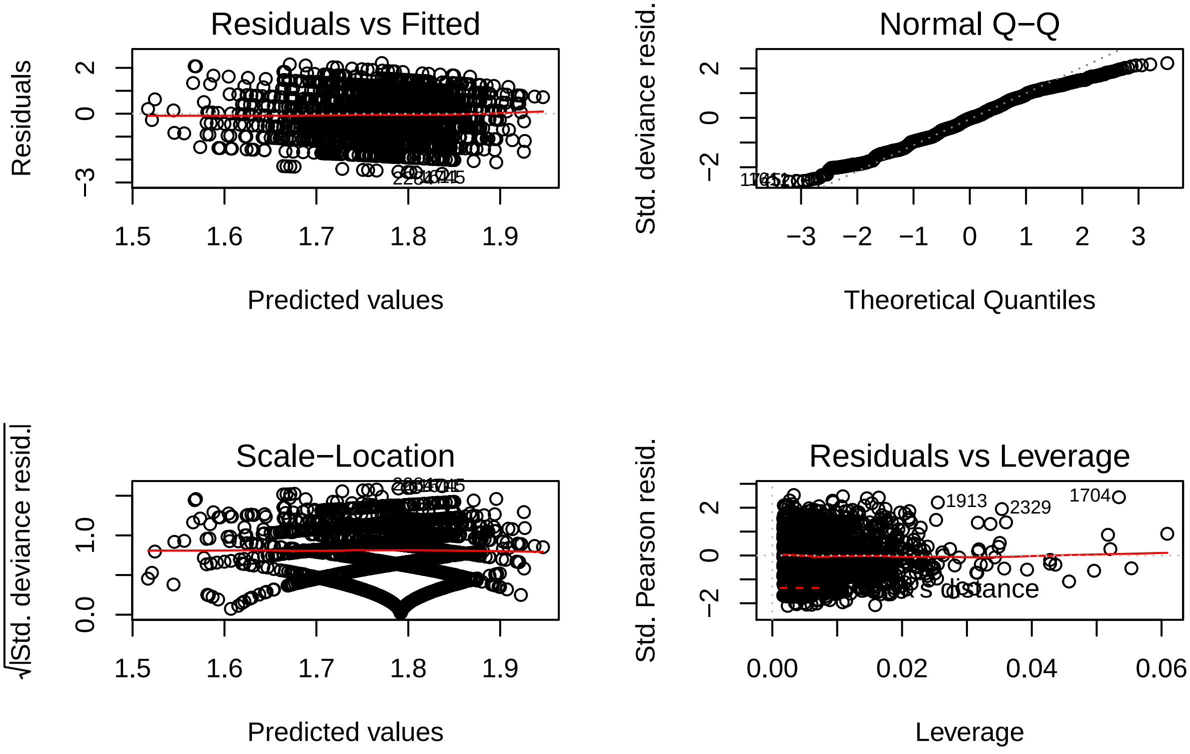
6 Summary of results

6.1 Tables

6.1.1 Predata Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6618	0.0588	28.27	0.0000
momsdeliveryage	0.0010	0.0019	0.54	0.5866
momsedu	0.0208	0.0092	2.27	0.0230
delivery2	-0.0363	0.0305	-1.19	0.2342
delivery3	0.0354	0.0239	1.48	0.1381
delivery4	0.0093	0.0238	0.39	0.6964
babymortality	-0.0123	0.0112	-1.09	0.2745
mathealth_count	-0.0289	0.0095	-3.04	0.0023
sex1	0.0411	0.0175	2.34	0.0191
tri_alcohol	0.0147	0.0268	0.55	0.5833
tri_tobacco	-0.0035	0.0241	-0.15	0.8831
mom_trauma1	-0.0145	0.0265	-0.55	0.5827
maternalpregnancyproblems____81	0.0012	0.0282	0.04	0.9654
artmethod____11	-0.0407	0.0477	-0.85	0.3939
tri_vaping	-0.0255	0.0396	-0.64	0.5208
deaf1	0.0673	0.0284	2.37	0.0177
tri_cannabis	-0.0136	0.0240	-0.57	0.5711

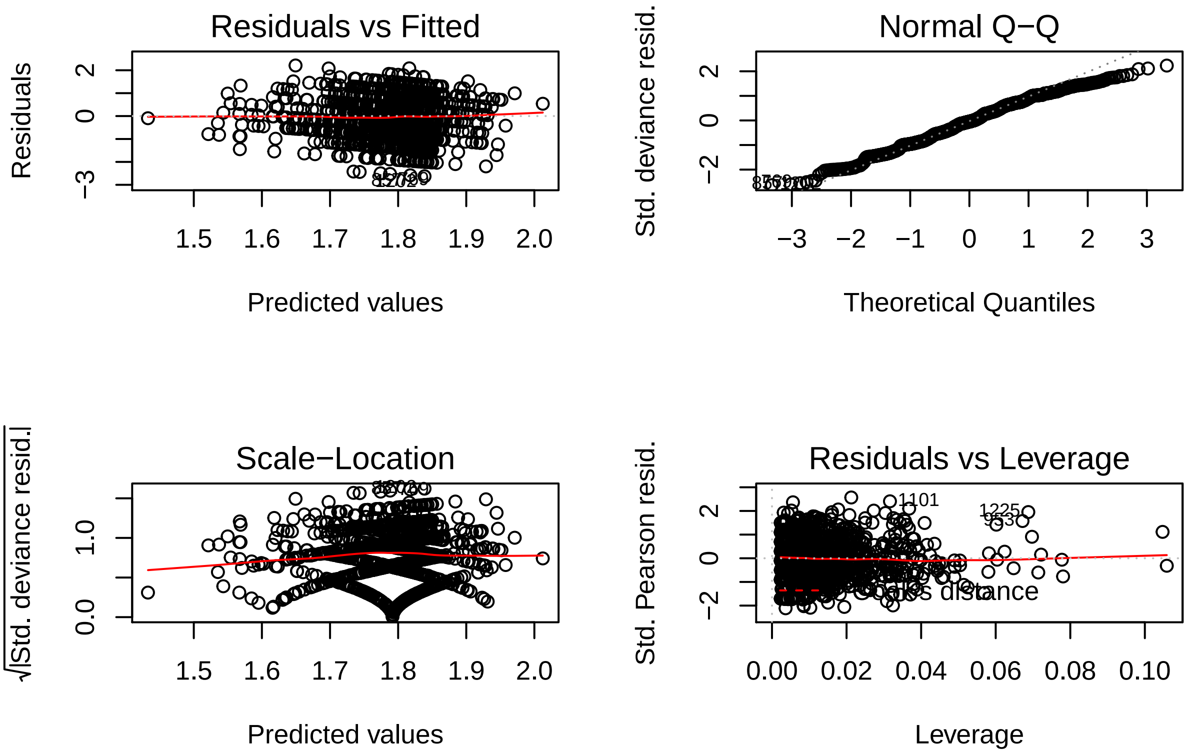
6.1.2 Predata Model Diagnostic Plots



6.1.3 Maledata Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.7743	0.0787	22.56	0.0000
momsdeliveryage	0.0011	0.0026	0.44	0.6574
momsedu	0.0063	0.0125	0.51	0.6123
delivery2	-0.0232	0.0409	-0.57	0.5706
delivery3	0.0192	0.0332	0.58	0.5636
delivery4	-0.0334	0.0323	-1.03	0.3010
babymortality	-0.0072	0.0148	-0.49	0.6273
mathealth__count	-0.0268	0.0130	-2.07	0.0389
tri__alcohol	-0.0268	0.0388	-0.69	0.4893
tri__tobacco	0.0405	0.0310	1.30	0.1922
mom__trauma1	-0.0560	0.0351	-1.59	0.1110
maternalpregnancyproblems_____81	-0.0807	0.0414	-1.95	0.0511
artmethod_____11	-0.0742	0.0729	-1.02	0.3084
tri__vaping	-0.0834	0.0578	-1.44	0.1488
deaf1	0.0859	0.0386	2.23	0.0260
tri__cannabis	-0.0161	0.0353	-0.46	0.6486

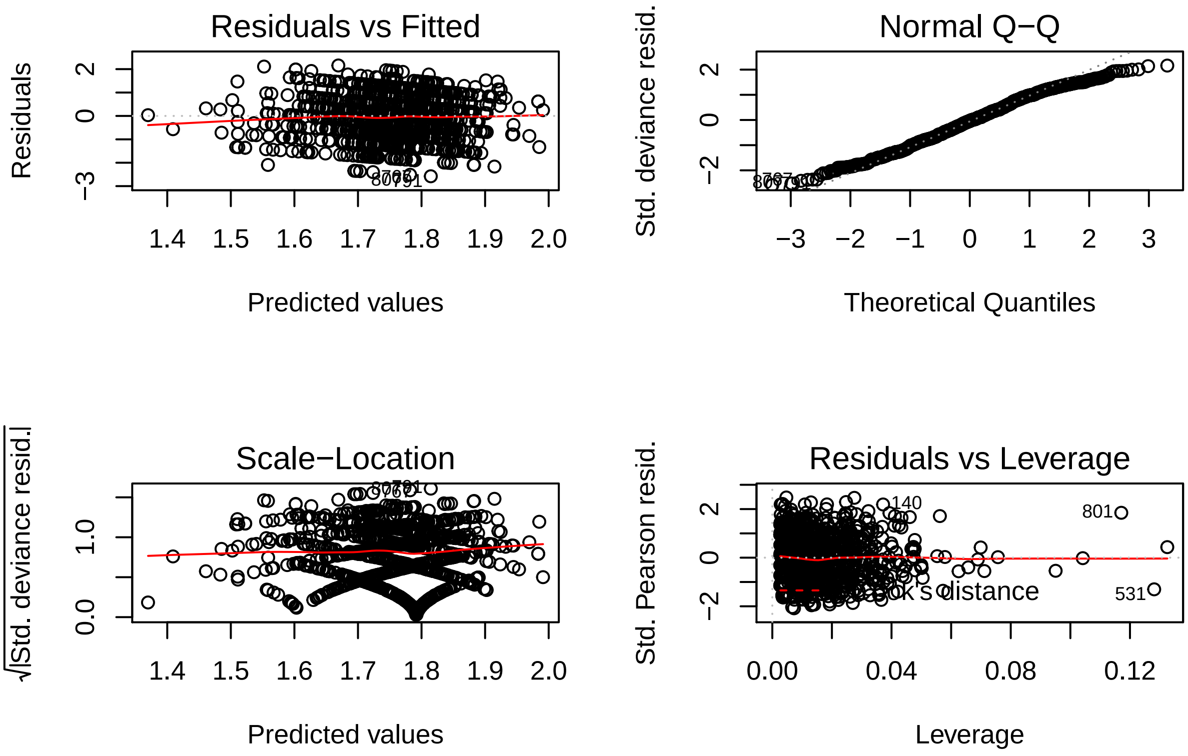
6.1.4 Maledata Model Diagnostic Plots



6.1.5 Femaledata Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5683	0.0852	18.41	0.0000
momsdeliveryage	0.0009	0.0028	0.34	0.7351
momsedu	0.0378	0.0136	2.79	0.0053
delivery2	-0.0548	0.0461	-1.19	0.2346
delivery3	0.0598	0.0345	1.73	0.0833
delivery4	0.0629	0.0354	1.78	0.0757
babymortality	-0.0160	0.0173	-0.93	0.3530
mathealth_count	-0.0287	0.0139	-2.07	0.0382
tri_alcohol	0.0561	0.0374	1.50	0.1335
tri_tobacco	-0.0720	0.0380	-1.90	0.0580
mom_trauma1	0.0445	0.0404	1.10	0.2717
maternalpregnancyproblems___81	0.0924	0.0388	2.38	0.0171
artmethod___11	-0.0185	0.0638	-0.29	0.7717
tri_vaping	0.0464	0.0543	0.85	0.3932
deaf1	0.0493	0.0424	1.16	0.2450
tri_cannabis	-0.0099	0.0333	-0.30	0.7673

6.1.6 Femaledata Model Diagnostic Plots



7 Interpretation of Results

By running a random forest on all of the prenatal variables we found that the significant risk factors that predispose a child to ASD behaviors are: `momsdeliveryage`, `momsedu`, `delivery`, `babymortality`, `mathealth_count`, `sex`, `tri_alcohol`, `tri_tobacco`, `mom_trauma`, `maternalpregnancyproblems___8`, `artmethod___1`, `tri_vaping`, `deaf1`, and `tri_cannabis`. These variables were chosen based on the mean of their minimal depth calculated from our random forest model. In conducting Poisson regression for three models, we used these variables. The first focused on answering our first posed research question while the second and third models were driven towards answering our second research question.

Our first model used the 14 predetermined variables and had four significant factors: mother's education level, a count variable that we created consisting of how many mental health diagnoses the mother has, the sex of the child, and whether the child has more than one deaf family member. The p-values for these predictors are 0.02299, 0.00234, 0.01915, and 0.01774 respectively, which are all significant at a 0.01 alpha level. Based on these findings, we can see that a level increase in the mother's education will increase a child's risk score by a factor of 1.02. We also observed that each additional mental health diagnosis for the mother will decrease the risk score by a factor of 0.03. If a child is male, then the risk score will be increased by a factor of 1.04. Additionally, if a child has more than one deaf family member, then the child's risk score will increase by a factor of 1.07. We were interested by the variables that were not statistically significant in our model as this seemed to be counterintuitive to literature that we had reviewed and our prior knowledge.

Due to the signifcance of the variable ”sex” in our first model, we chose to control for gender in our next two models. Our second model had all of the variables present in the initial model without "sex" but focused on the impact of the gender being male on the statistically significant predictors. In this Poisson model, we saw two predictors significant at the 0.01 level, the number of mental health diagnoses for the mother and whether the child has more than one deaf family member with p-values of 0.0389 and 0.0260 respectively. One of our predictors, `maternalpregnancyproblems___8` which is coded as the mother having anemia, a condition in which the blood does not have enough oxygen to carry to the body's tissues, was significant at the 0.05 level with a p-value of 0.0511. Based on our findings, an increase in mental health diagnoses for the mother decreases their child's risk score by a factor of about 0.03, which is the same value that we observed in our model that did not control for gender. In this case, we can determine that the mental health score does not impact a child's risk score if the child is a male. If the child has more than one deaf family member, their risk score increases by a factor of 1.09, which is slightly higher than our first model. This indicates that, on average, a male child will have a larger risk score if they have more than one deaf family member. If a male child has a mother that has anemia, then their risk score will decrease by a factor of about 0.08. In comparison to our first model, we observe that the variable for mother's education is no longer significant with a p-value calculated of only 0.6389.

For our third model, the data set was filtered by gender being female to focus on the impact of a baby's gender and contained the variables from the first model, except for "sex." This model had the most statistically significant predictors with those being the mother's education level, two of the delivery methods of the pregnancy those of which being planned and emergency C-section, the number of mental health diagnoses for the mother, the number of trimesters a mother used tobacco, and whether the mother had anemia during the pregnancy with p-values of 0.00531, 0.08332, 0.07570, 0.03819, 0.05802, and 0.01713 respectively. We can see that a level increase in the mother's education will increase a child's risk score by a factor of 1.04 given all else constant, which is slightly higher than our first model. A mother having delivered their baby through a planned C-section increases their child's risk score by a factor of 1.06 while delivery through an emergency C-section increases this score by a factor of 1.07. These factors are statistically significant for this model only which controls for female children specifically. Also, based on these findings, an increase in mental health diagnoses for the mother decreases their child's risk score by a factor of about 0.03, which is the same value that we observed in our first model. In this case, we can determine that the mental health score does not impact a child's risk score based on gender as this value is the same for male and female children. We can also see that for an increase in the number of trimesters that a mother used tobacco, their child's risk score decreases by a factor of 0.07. Lastly, if a mother had anemia during their pregnancy, their child's risk score will increase by a factor of 1.1.

8 Overall conclusions

This study was quite informative in discerning which risk factors are statistically significant in predisposing children to ASD behaviors and if a statistically significant difference exists when looking at predictors for autism in males and females. In order to effectively discern the answers to these questions, we look at all three models: one for all children, one for males, and one for females. As mentioned in the interpretation of results section above, we discovered that mother's education level, a count variable that we created consisting of how many mental health diagnoses the mother has, the sex of the child, and whether the child has more than one deaf family member were all statistically significant in the full model. Meanwhile for males the number of mental health diagnoses for the mother, whether the child has more than one deaf family member, and the mother having anemia were statistically significant while for females mother's education level, two of the delivery methods of the pregnancy those of which being planned and emergency C-section, the number of mental health diagnoses for the mother, the number of trimesters a mother used tobacco, and whether the mother had anemia during the pregnancy were statistically significant. These findings show that the models point to diverging risk factors for ASD behaviors between males and females. Clearly, then, there is a statistically significant difference when looking at predictors for autism in males and females.

Using the understanding acquired by modeling the sexes both together and separate, we are then able to come to a more informed conclusion about the statistically significant risk factors that predispose a child to ASD behaviors. Specifically notable among the commonalities for predictors amongst males and females was the presence of statistical significance for both number of mental health diagnoses for the mother and whether or not the mother had anemia during pregnancy. As such, we can conclude that these factors are major risk factors that predispose a child to ASD behaviors, with number of mental health diagnoses for the mother being of extreme importance given that it was a statistically significant factor in the full model as well. Additionally, whether the child has more than one deaf family member was also present as a statistically significant factor in both the full model and the model with just males and also mother's education level is a significant factor in both the full and female models, so these factors are notable as major risk factors as well. Some factors appeared in only one of the models, and thus are statistically significant risk factors but weigh less heavily than the factors in multiple models with regards to predisposition of a child to ASD behaviors. These factors are two of the delivery methods of the pregnancy (those of which being planned and emergency C-section) and the number of trimesters a mother used tobacco.

While we found statistical significance in all the aformentioned cases, it is important to note that these differences may not be of practical significance. Some of these risk factors present themselves as only minor contributiors to an individual's risk of developing autism. As such, determining which factors are truly practically significant should be assessed based on the specific needs of an individual project using this data.

9 Challenges of the study

Many of the challenges we faced during this project occurred when cleaning the dataset. These challenges included determining how to handle NAs, sifting through the `other` variables, finding variables with small counts, removing outliers, and addressing collinearity among variables. In dealing with NAs, we either removed observations with NAs in prenatal columns that we were interested in or set the NAs to 0 (particularly when the variable of interest was binary). Cleaning the 'other' columns proved to be very time consuming with a variety of responses present in these columns. Responses were either categorized into existing columns, combined to create new columns if enough similar responses were found, or removed due to obvious inconsistencies. Outliers were determined by looking at the summary statistics of the columns. Outliers with a large number of NAs and inconsistencies of columns such as mother's age and number of live births were removed. Collinearity was dealt with by plotting correlation plots between variables and either completely removing one variable or combining highly correlated variables to make a new measure. It should be noted that implications involving collinearity, including its effect on the precision of estimates may be due to an unrepresentative sample or insufficient information in the sample [2]. Another challenge unrelated to cleaning the dataset was converting binary columns and non-numeric variables to factors. All of our variables were type “double” so after running a couple models and getting strange results, we solved this issue by converting binary/categorical variables to type “factor”.

Some of the challenges that we encountered during this project were not solved. These unsolvable challenges occured when testing the prediction accuracy of our model. Our models all had a high test error rate and our predictions of the response only ranged from 5 to 7. The actual response variable of `risk_score_ASD` had a range of 1 to 12 with a median of 6, so these results did not match our expectations. We got these results despite checking and meeting the assumptions of our model and are not entirely sure why our poisson model did a poor job predicting our response. Another challenge we faced was that our findings from our literature review did not match our actual findings after running the model. Variables such as `asd’ (number of family members with asd) and ‘mat_pssd` (whether or not mom has psychosis, schizophrenia, or schizoaffective disorder) were not significant, while `momsedu` was highly significant across all three models. Even after running a random forest model to select for important variables to include in our model to refine our model to only including 14 prenatal factors, all three models only had 2 to 3 statistically significant variables.

Overall, challenges that involved cleaning the survey data and converting column types to reflect the type of variable being measured were worked through and solvable issues. Challenges that involved testing the prediction accuracy of our model and some of the findings were unsolvable something to further explore in future work.

10 Recommendations for the future

In the future, we recommend that other risk factors occurring at different times in a mother’s life be explored. In our work, we narrowed the scope of our research question by only exploring prenatal risk factors. We recommend that additional models and analysis be done on postnatal risk factors and their effect on `risk_score_ASD`. Additional analysis could be done on the delivery method and its effect on the autism risk score. According to our models, having a cesarean section significantly increases the risk score for autism. It would be interesting to explore the reasons for having a cesarean section, whether it was planned, due to prolonged labor, abnormal positioning, cord prolapse, etc. and see if any of these effects are significant risk factors for autism.

We also recommend changing some aspects of the survey in order to simplify the data cleaning process. Many of the `other` columns that we cleaned ended up being discarded due to low counts. We recommend that maternal health problems be condensed to include specific categories that you would want to study. This would mean possibly getting rid of some of the fill in “other” options on the survey. In addition, a lot of collinearity between variables were present, so there could be additional condensing of survey questions/variables to avoid these associations.

11 Bibliography

"Autism and Gender." Autism Support - Leading UK Charity - National Autistic Society, www.autism.org.uk/about/what-is/gender.aspx.

"Autism spectrum disorder." Mayo Clinic, 6 Jan. 2018, www.mayoclinic.org/diseases-conditions/autism-spectrum-disorder/symptoms-causes/syc-20352928.

Geelhand, Philippine, et al. "The Role of Gender in the Perception of Autism Symptom Severity and Future Behavioral Development." Molecular Autism, 2019, www.molecularautism.biomedcentral.com.

Halladay, Alycia K, et al. "Sex and Gender Differences in Autism Spectrum Disorder: Summarizing Evidence Gaps and Identifying Emerging Areas of Priority." Molecular Autism, BioMed Central, 13 June 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4465158/.

"Poisson Regression Analysis Using SPSS Statistics." How to Perform a Poisson Regression Analysis in SPSS Statistics | Laerd Statistics, statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php.

Vatcheva, Kristina P, et al. "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies." Epidemiology (Sunnyvale, Calif.), U.S. National Library of Medicine, Apr. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4888898/.