Assignment 3

Mohammad Anas

Summary

The data set provided to us contains the information about 869 pregnant mothers and their biological and behavioral attributes. The variables of interest in this data are the smoke variable, which indicates whether a mother smoked or not and whether mother gave a premature birth. We will try to address the following questions throughout this report.

1. Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the odds ratio of pre-term birth for smokers and non-smokers?
2. Is there any evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race? If so, characterize those differences.
3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

We conducted a logistic regression analysis to answer the above questions and found out that the smoking does not have a significant effect on the the log of odds of having a premature birth. We also found out that this effect does not vary by race. However, we do note that according to our model mother's weight at pregnancy and previous birth rates can have a significant affect on odds of having premature births.

Introduction

The gestational age variable was used to derive the pre-term birth variable, which was binary in nature. If the baby was born before the gestational period exceeded 270 days, the instance was recorded as a preterm birth and otherwise zero.

Given that we built our model to study the effect of smoking and race on the preterm birth rate, the smoking and race variables were included in our model. We also add other variables based on our exploratory data analysis. This not only allows us to isolate the effect of smoking and the race variable in our model but also helps us measure the effect of other variables on the odds of having a premature births.

Based on our analysis we saw some counter intuitive yet interesting results. Smoke and its interaction with race were found to be insignificant variables. However, we see the the odds ratios of have a premature birth is significant for races like Asian and Black when compared to the white race. The confidence intervals for the statistically significant variables are shown below. These confidence intervals indicate the range of odds ratios of have a premature birth for the categorical variables and the change in odds of having premature birth for continuous variables.

Extremely large confidence intervals for the variable of 'education trade school' is probably because of the low number of data entries in the data for trade school education. Only four variables were recorded, this leads to large standard errors and have huge confidence intervals.

	2.5 %	97.5 %
(Intercept)	0.14	0.40
mother_educcollege	0.32	1.09
mother_educhigh school + college	0.20	0.66
mother_eductrade school	1.28	222.30
mother_raceasian	1.08	5.41
mother_raceblack	1.39	3.34
mpregwtc	0.98	1.00

Exploratory Data Analysis

To asses our model better we conduct the exploratory data analysis in our model. We start by taking the effect of our main variable in the data and that is smoking. We compute the probabilities of have a premature birth rate given the mother smoked or not. We note that the probabilities of having a premature birth varies by mother's smoking habits. The conditional probabilities are given below.

	never	smoke now
0	0.83	0.78
1	0.17	0.22

To ensure that our smoking variable is independent of the premature variable,we conduct a Chi-squared test of independence between the two variables. We note that the p-value for this test is 0.097. Therefore, with a 95% significance level, we can say that the variables are independent and we will not be surprised if the smoke variables comes out to be insignificant in our model. We do this the same for our other variable of interest that is the race of the mother and found out that the p-value for that was 0.04, proving that the variables race and pre-term birthrate are not independent. Following the similar procedure for other categorical variables and found out that income was not independent.

We make box plots of our continuous variables as well and see that their median value does not vary much in case of whether the instance was of a premature birth weight or not. We also explore the affect of interactions and found out a few interesting cases. The box plots show a difference in trend for age and premature births for each value of education. We also a difference in the trend for parity variable on preterm birth rates and saw that the effect varied for various income levels. Therefore we notice through our EDA that it would be interesting to explore the affect of interactions between income and parity and mother's education and age.

We also notice that the effect of smoking on premature births does not vary for different races. Therefore including an interaction of smoking and race in our model will not be good idea. Other interaction that were found to be interesting were mother's education and income, and income and age.

We also explore whether we need to transform our predictor variables. We do this by looking at the binned plots for our binary variable against our predictor variables. We notice randomness in these plots and we are unable to decipher any transformations for our variables. These plots have been included in the appendix section of the report.

Model

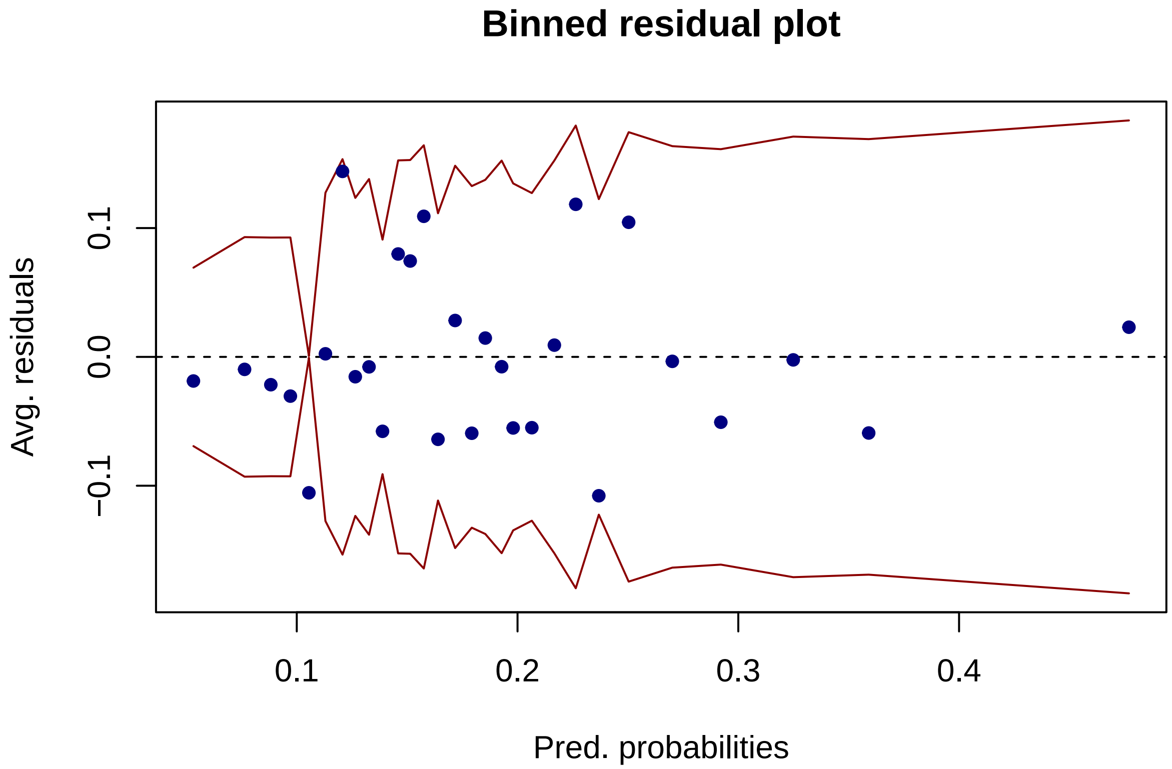
Moving onto our model building, we include all our variables in our logistic regression model regardless of what we found through our EDA. We do this so that we can run them through step wise selection. We do the step wise variable selection using BIC. However, using step wise regression with BIC removes all the variables from my model, leading us to adopt a more lenient approach with AIC. Step wise variable selection with AIC leaves us a few variables in our model namely smoke, race, weight and education. After running the model, we see that most of these variables were significant.

We now include the interactions that were found to be interesting during our EDA. These interactions were income and parity, mother's education and age, mother's income and education, and income and age. As we plan to explore whether the smoking effects on pre-mature birth rate vary by race or not, we also include the interaction of smoke and race. We add each of these interactions of our model that resulted from step wise selection and measure whether the interaction improved the fit of our model by comparing the deviance residuals of both models. We do this by using "Chi-squared" test in R. After repeating this same procedure for every interaction listed above, we noted that the interaction for smoke and race was highly insignificant confirming that we remove the interaction of smoking and race from our model. We also note that mother's education and income interaction was a significant variable. However, when we add this interaction to our model and plot the binned residuals, we see somewhat increasing trend in the plot and plenty of outliers. Given that the independence assumption is not satisfied,we remove this interaction from our model as well.The binned residual plot has been added to the appendix section. Adding the rest of the interactions to our model results does not result in significantly improved fit. Therefore, our final model remains the same as what we get from stepwise selection. The equation of this model is presented below.

$$Pr(Premature_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 MotherSmoke_i + \beta_2 MotherRace_i + \beta_3 WeightX_i + \beta_4 MotherEduc_i + \epsilon_i)}{1 + exp(\beta_0 + \beta_1 MotherSmoke_i + \beta_2 MotherRace_i + \beta_3 WeightX_i + \beta_4 MotherEduc_i + \epsilon_i)} \tag{1}$$

Model Assessment

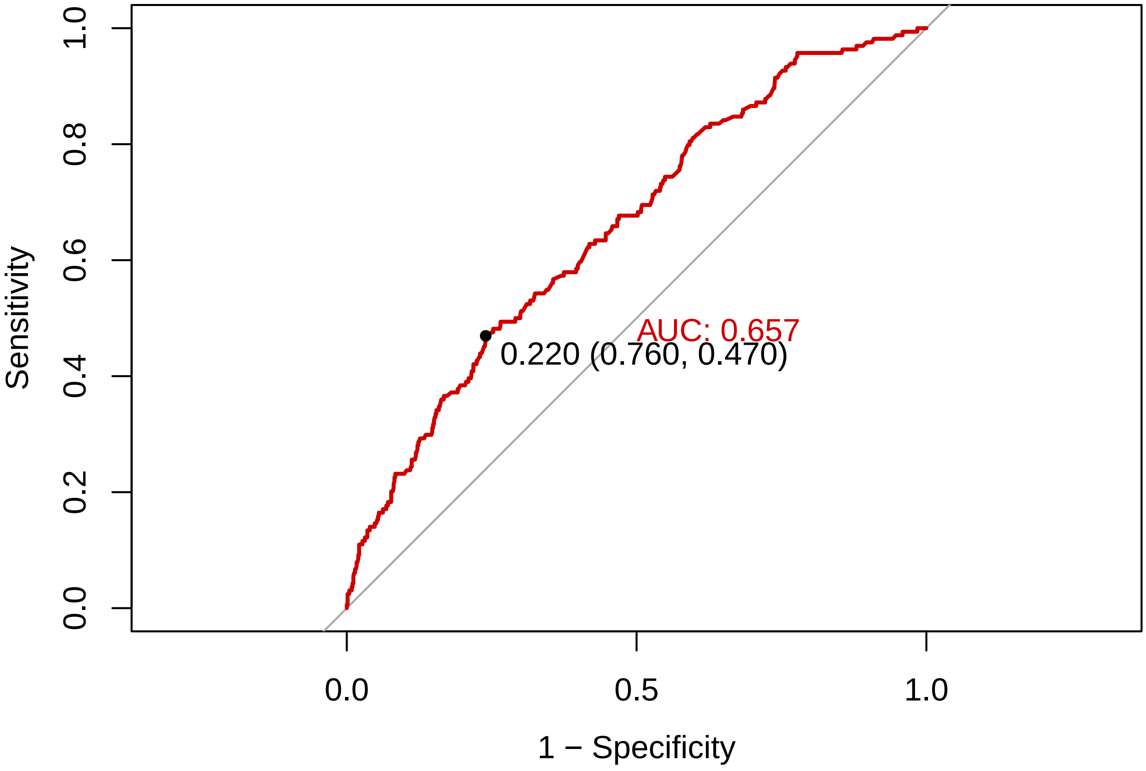
Now we move onto our model assessment. To do this we plot our model residuals against the binned average predicted probabilities. We are unable to see any pattern in our plot confirming that our residuals are independent. We also see the majority of points in our data fall within the 95% confidence interval. Therefore, we can safely assume that there are no outliers.



We also note that our residual deviance of our model is 795.91. Using a chi-squared test we see that the difference between our model residual deviance and null model's deviance is a statistically significant. We also plot the residuals against our predictor variables. We see randomness in these plots confirming that we do not need to transform any of our predictor variables to improve our fit. The plots of our residuals vs our predictor variables are also include in the appendix.

Model Validation

To measure the accuracy, we take a look at the confusion matrix and the ROC curve of our model. We use the ROC curve to optimize our probability threshold that allows us to classify our premature births correctly.



Using the confusion matrix we note that our model achieves an accuracy of 70.5%. We also note our model does a good job at predicting non-premature birthrates achieving 76% specificity. However, it fails at predicting the preterm birth rates correctly. The sensitivity rate of the model is very low and was found out to be 47%. However, if we change the probability threshold that we use for classification we note that we can improve our model's sensitivity. Setting the probability threshold to the mean value of our premature variable, we achieve a sensitivity rate of 58%. However, this can only be achieved by compromising on specificity and overall accuracy. We can improve our overall predicitve power of the model by adding interaction terms.

Interpretation of Coefficients

The coefficients of our model are shown in the table below. Note that the coefficients shown here are on log of the odds scale. However, we will exponentiate them to make interpretation easier.

So for our model, holding all else constant, a non-white mother who does not smoke and has an education of between 8th to 12th grade and a weight of 129 ounces has the odds of 0.24 for having a premature birth. For a mother who goes to college, the odds of having a premature birth will be less than the mother whose education is between 8th and 12th grade as indicated by the negative sign of the coefficient. The odds ratio for mother who went to college are 0.59 as compared to a mother whose education is between 8th and 12th

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.42	0.26	-5.54	0.00
mother_educcollege	-0.52	0.31	-1.69	0.09
mother_educhigh school + college	-1.01	0.30	-3.33	0.00
mother_educhigh school + trade school	-0.17	0.41	-0.41	0.68
mother_educhigh school but no other schooling	-0.35	0.25	-1.38	0.17
mother_educless than 8th grade	0.54	0.95	0.57	0.57
mother_eductrade school	2.37	1.18	2.00	0.05
mother_raceasian	0.91	0.41	2.22	0.03
mother_raceblack	0.77	0.22	3.46	0.00
mother_racemexican	0.15	0.52	0.30	0.76
mother_racemix	-0.75	1.05	-0.72	0.47
mpregwtc	-0.01	0.00	-2.51	0.01
mother_smokesmoke now	0.29	0.18	1.57	0.12

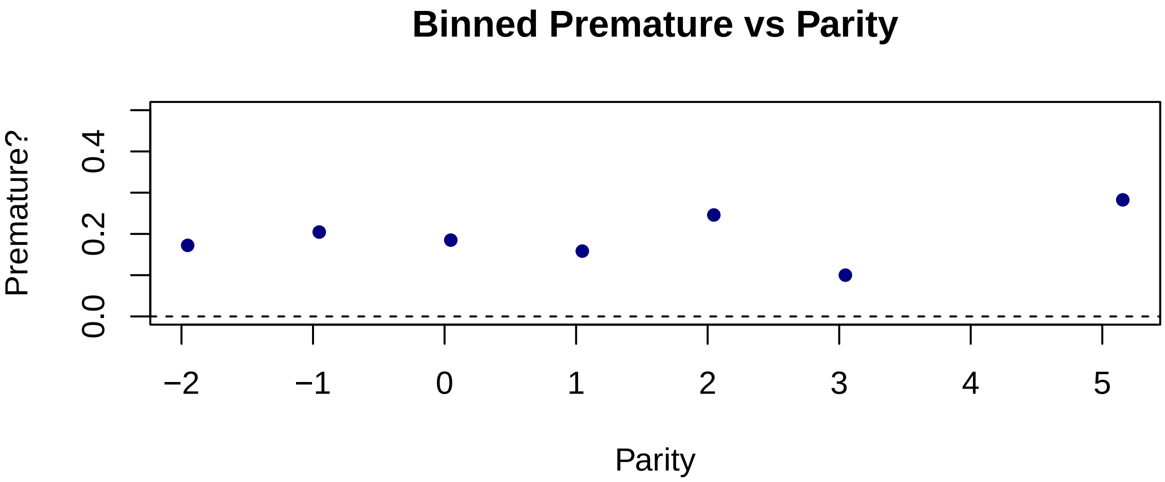
grade. We notice that a mother who went to high school and college had an even lower odds ratio of having a premature birth (0.37) as compared to a mother whose education is between 8th and 12th grade. For mothers who only went to trade school,we see that the odds ratio of having a premature birth is 10.66 as compared to mothers with education between 8th and 12th grade. This means that the odds of having a premature birth are greater for trade school mothers as compared to mothers with education between 8th and 12th grade. This is also indicated by the positive sign of the coefficient. For Asian mother the odds ratio of having a premature birth are 2.47 as compared to white mothers. The odds ratios of having a premature birth for black mothers as compared to white mothers is 2.16. Lastly holding all else constant we note that one unit increase in mother's weight at pregnancy decrease the odds of having a premature birth by a factor of 0.99. We can also say that the odds of premature birth decrease by approximately 0.01%. Please note that the interpretation here for each coefficient has been made holding all other variables constant. The variables that have not been interpreted are statistically insignificant. We also do not need to worry about multicollinearity as the vif of any of our variables does not exceed 10.

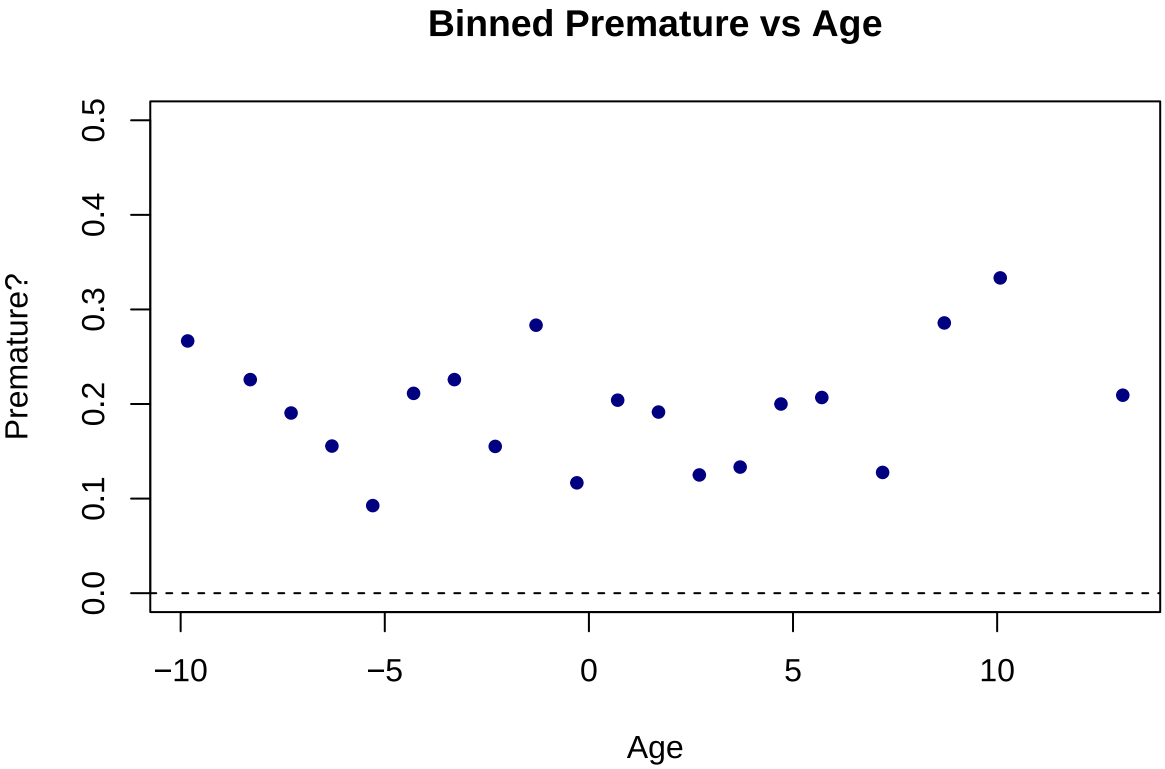
Conclusion

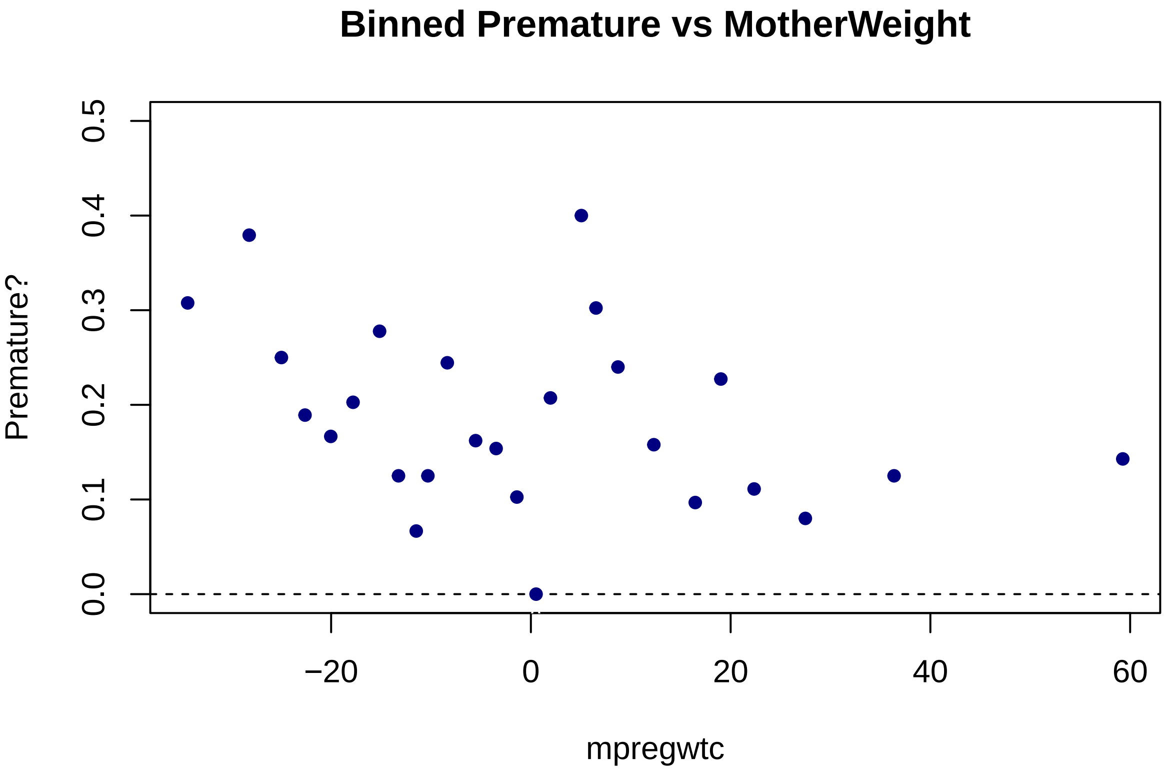
Therefore, we see through our model that the variable smoking does not have a significant affect on the premature births. Moreover, this affect does not vary with race. However, there are a few potential limitations of our model. The distribution is uneven in our data for premature births. We have a very low number of premature births in our model. The trade school coefficient is not reliable as there are only 4 data entries in our model that went to only trade school. We also cannot rely on the race variable as most of the values correspond to the white race. Lastly, the defiance residual is still very high in our model.

Appendix

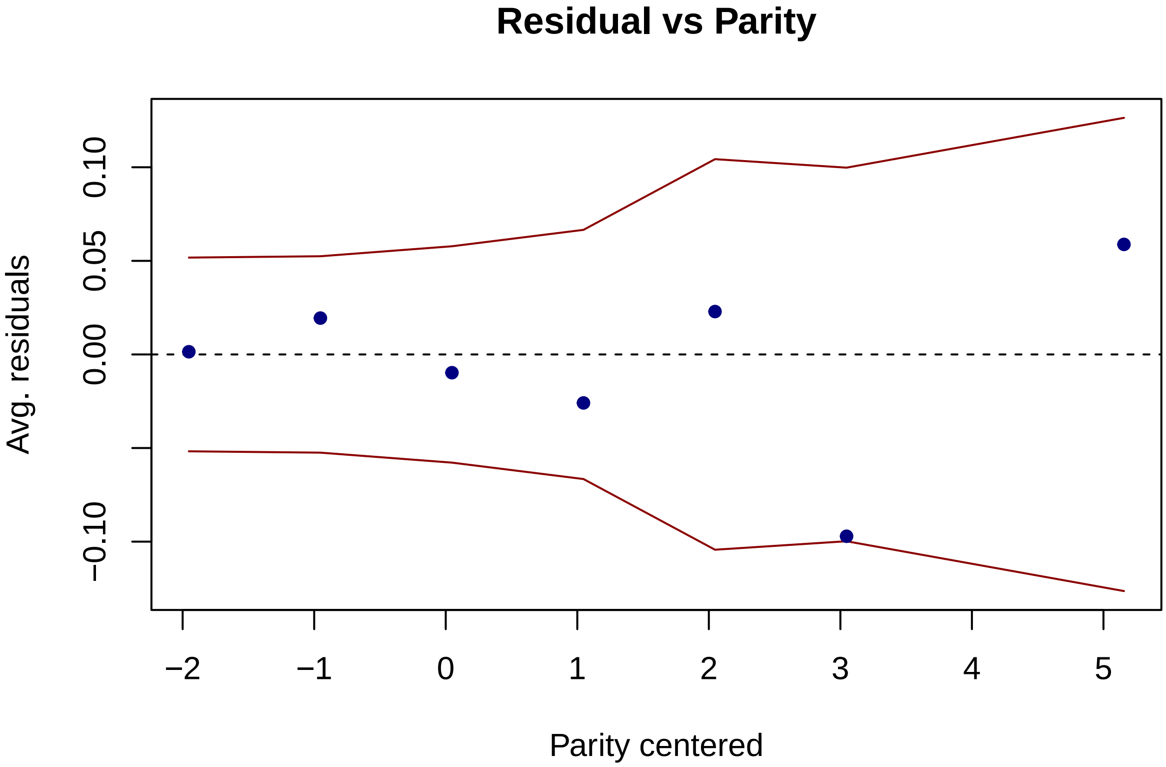
The binned plots our continuous predictor variables against our premature variable are shown below.

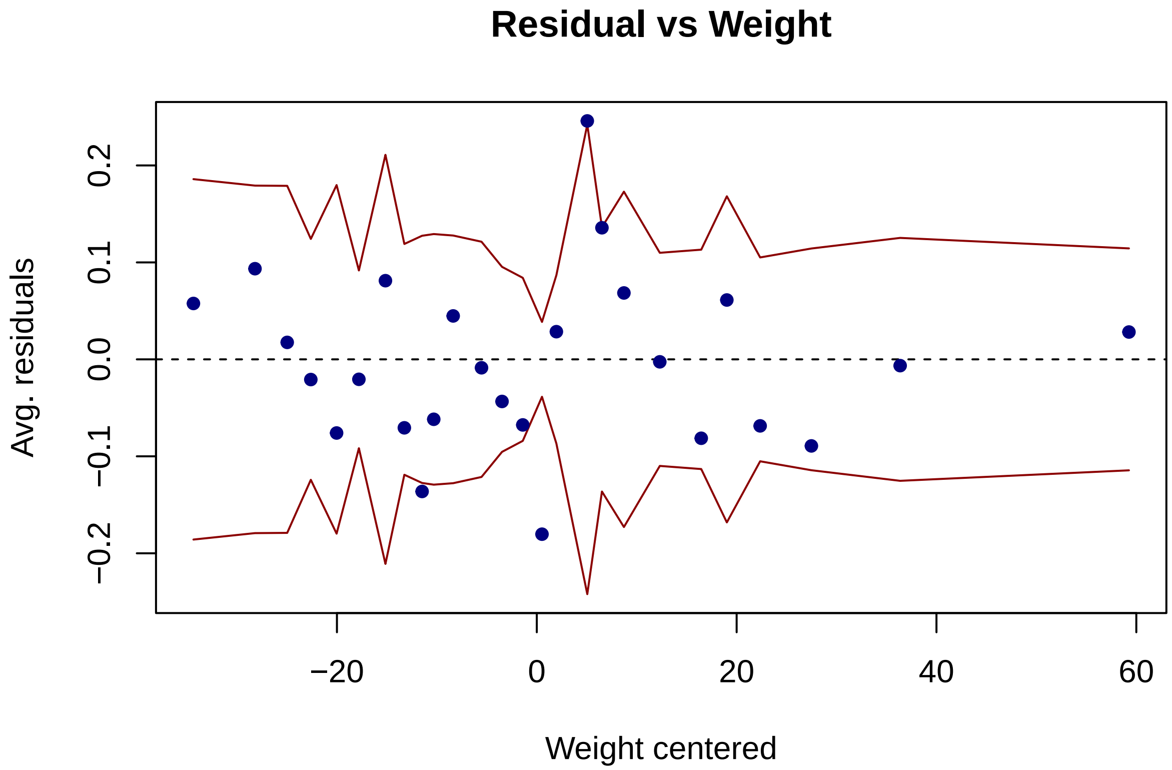

Binned Premature vs Parity
Premature?
0.0
0.2
0.4
−2
−1
0
1
2
3
4
5
Parity

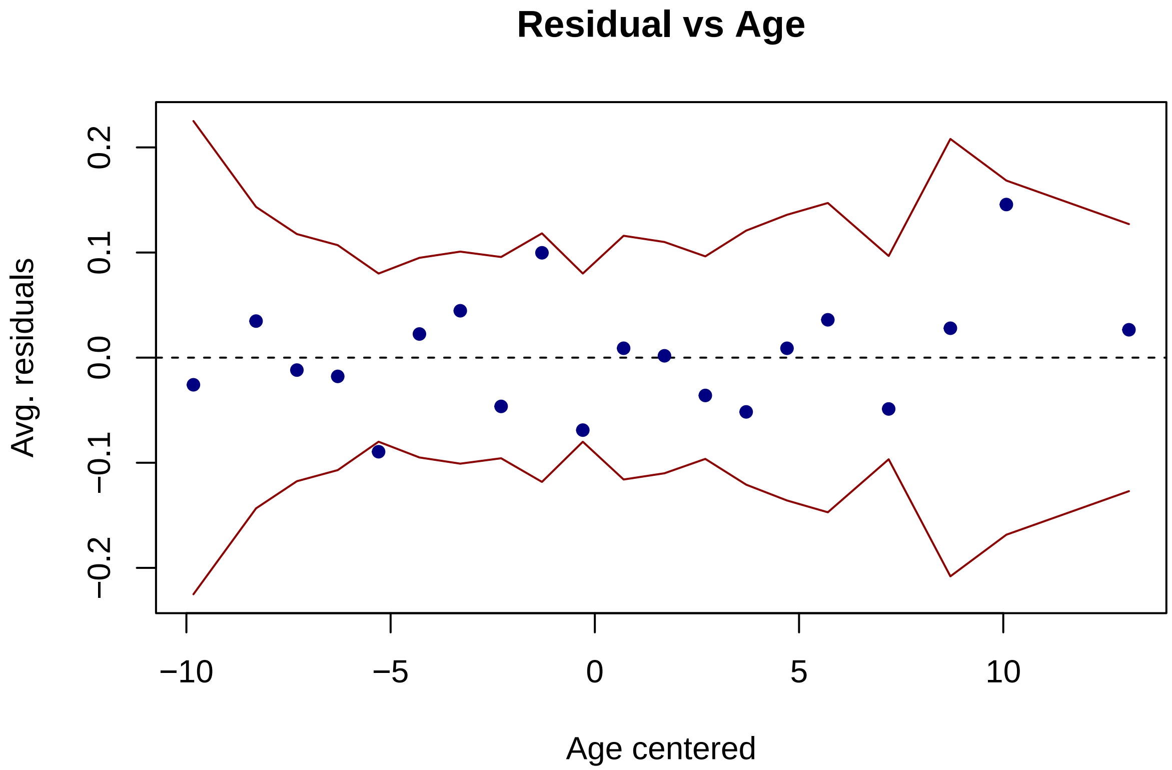

Binned Premature vs Age
Premature?
0.0
0.1
0.2
0.3
0.4
0.5
−10
−5
0
5
10
Age



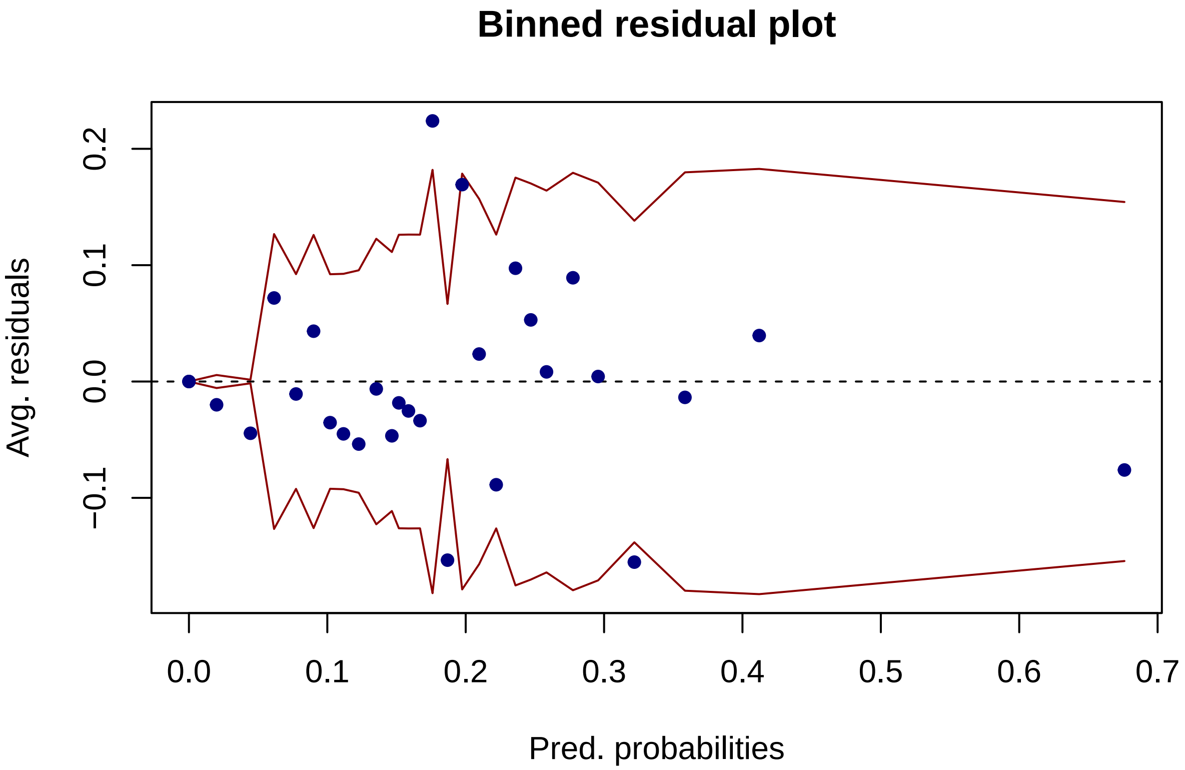
Below shown are the binned plots of our residual against our continuous predictors.







The binned plot of residuals when we create a model for interactions of mother's education and income can be seen below.


Binned residual plot
Avg. residuals
0.2
0.1
0.0
−0.1
Pred. probabilities
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7