Assignment3

Erika Fox

9/19/2021

Summary:

In the following experiment, we were interested in determining if whether or not a mother is a smoker impacts the odds that her baby will be born prematurely (gestational age less than 270 days). Additionally, we were interested in finding evidence that the association between smoking and the odds of a premature birth differs by the mother's race. We utilized exploratory data analysis (EDA) and selected logistic regression model using AIC stepwise selection to pursue these research goals. According to our model, the relationship between maternal smoking behavior and the odds of premature births is not significant. Additionally, we found the interaction effects of maternal races and smoking to be insignificant. However, we did find the odds of a premature birth increase if the mother's race is Black, and that a higher pre-pregnancy weight decreases the odds of a woman having a premature birth, according to our model.

Introduction:

In this experiment, we will be using birth data in order to explore and see if smoking increases the odds of a mother birthing her child prematurely. Additionally, we will explore if other factors, such as race, amplify or diminish the effects of smoking on premature births.

Questions of interest: 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? 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. Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

Data:

In order to explore our questions of interest, we have a data from a study that originally included 15,000 families. However, the data that we actually ended up using has significantly less observations (869) after the modifications that had to be made to best suit our research goals. First, in order to simplify analyses, we are only going to be comparing babies whose mothers smoke to babies whose mothers have never smoked. This elimination was our first major change we made to the data. The second major cut we had to make has to do with the observations that were lacking in data about the baby's father. As it is often difficult to get information on fathers when collecting birth data, many of our observations had missing values such as the father's height and weight, so when we went to cut rows with missing observations, we were losing more rows than necessary as we do not plan to work with the father data at all. Thus, we removed the columns that had to do with the father entirely in order to preserve as many rows as possible. We then opted to drop any remaining rows with missing values as our next cleaning step, but found that this was unnecessary as after removing our father data, our new data set was now as full as could be. Finally, the *mrace* column needed an adjustment, as 0-5 all represented white mothers. I modified the *mrace* column to combine all the white mothers into a single label, White, and reassigned each of the other factor labels to correspond

with the English name of the race the numbers corresponded to. I also relabeled the factors of *med* to have English labels rather than numbers.

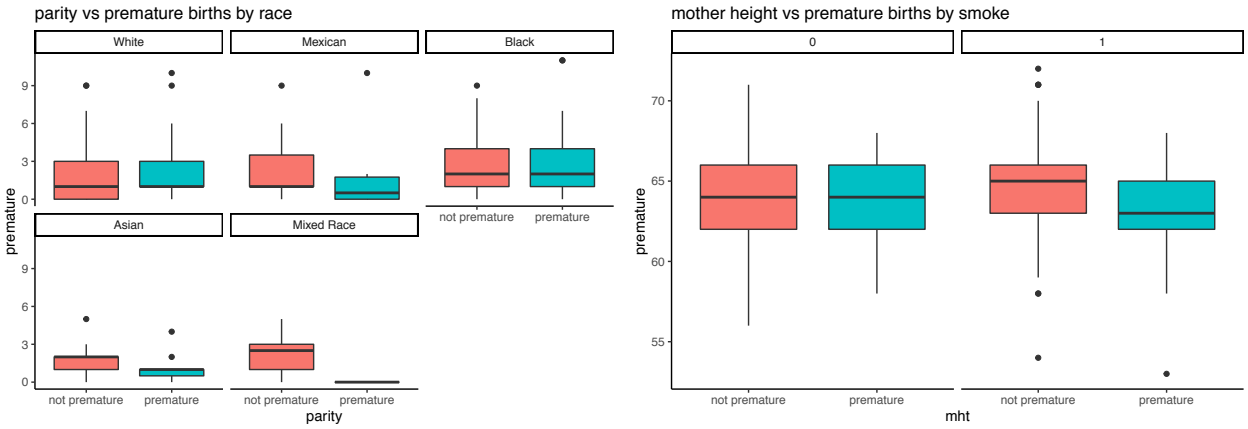
Our response variable for this study is *premature*, which is a measure of whether or not the baby the baby was born at a gestational age of less than 270 days. I added this variable to the data set smoking by creating a new column using the variable *gestation*, recording a "1" if the baby was born prematurely at a gestational age of less than 270, and a "0" otherwise.

EDA: Let's see what this data can tell us before we start modeling...

First we made a table that compares the factored version of the *premature* variable we made to *smoke*, to get an idea of what our numbers in this variable look like. We found that for mothers who have never smoked, our dataset records that 16.5% of their births were premature. And for mothers that smoke, our dataset records that 21.6% percent of babies were born premature. These numbers suggest that smoking might make it more likely for a premature birth to occur, however, after running a chi-squared test for this table to check for independence, it is questionable if we should deem these results as significant with a p-value of 0.069, which is a bit too far over 0.05 for comfort.

Next, I made two subsets of our data set, smoke_0 and smoke_1, which contained only mothers who don't and do smoke respectively. I made conditional probability tables for each of these subsets for each of the categorical variables against *premature* to begin looking into interactions.

Finally, I also made a plot for each of these interactions plus many more in order to get a glimpse of which ones might be worth including in the model. Here are a few of those plots that I thought stood out:



The left plot is parity vs. premature, facet wrapped by race. Although the effects of parity on the odds of having a premature don't seem to change dramatically by race, there does appear to be enough of an interaction based on this plot that it is worth considering for our model. The right plot is premature vs mht facet wrapped by smoke. Although this interaction doesn't come to mind as one that I would hypothesize as significant, the plot definitely suggests that there is an interaction going on, so I want to consider this one as well. Some other interactions that I considered based on my plots and tables include but are not limited to: $smoke * parity$, $smoke * mpregwt$, $smoke * med$, $mht * mrace$, and $mpregwt * mrace$

Model

In order to choose a model for this analysis, I first ran and assessed a Full Model, including all of the possible predictor variables (I centered the numeric variables) as well as the interactions that seemed like they could be significant during EDA. This model left us plenty of room to try and beat it with an AUC of 0.640.

Model selection: I tried running AIC using the initial model as the Full Model input. For the null model input, we used a very simple model that only had $smoke : mrace$ as a predictor for our response variable, *premature*, as this is as simplest that could help us reach our research goals.

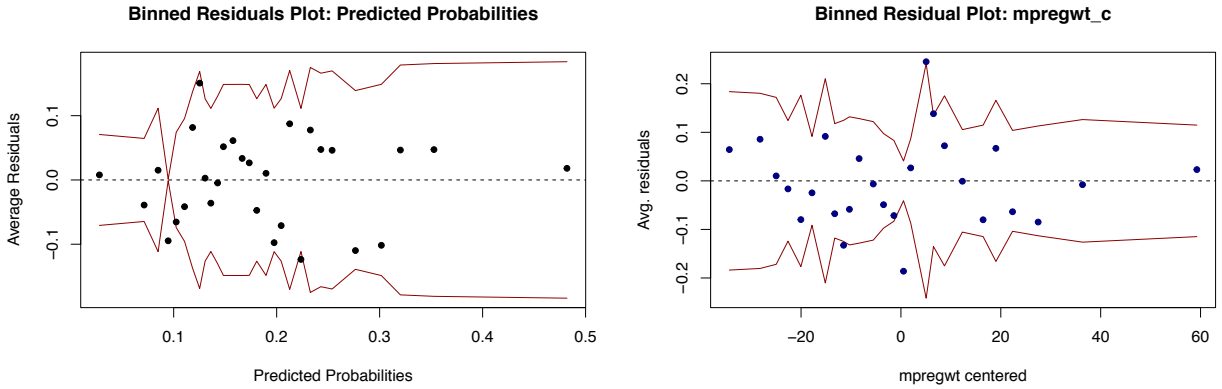
The final model I ended up using here is the following:

$y_i|x_i \sim Bernoulli(\pi_i) \log(\frac{\pi_i}{1-\pi_i}) = x_i\beta,$

where y_i is the binary variable indicating whether or not a baby was born prematurely. x_i includes the predictors variables: factor(smoke), factor(med), factor(mrace), mpregwt_c, smoke:factor(mrace). β is a vector representing the predictor coefficients.

Transformations: Once we selected a model with AIC, I still had some remaining steps to do to make sure the model was ready to go. I assessed the model to see how well it fit the assumptions for logistic regression, and saw the assumptions were met reasonably well (explained in detail in the next section). However, even though I was satisfied that the assumptions were met, I saw that there was room for improvement, as our AUC was still not as high as it could be at 0.664. As $mpregwt_c$ is the only continuous variable that made it into my AIC selected model, I decided to try and add a transformation to that predictor, by taking the log of the column *mpregwt*, and re-centering. I found this transformation to not make any significant changes to my model's performance, so I decided to keep my model transformation free.

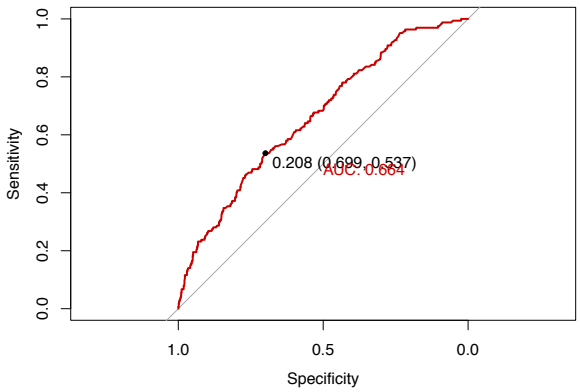
Final model assessment:



This model satisfactorily meets the assumptions for logistic regression. In the right plot above, Binned Residuals Plot: Predicted Probabilities, we can see that 95% of points are within the bands, and there does not seem to be a trend to be concerned about. On the right is a similar plot with our predictor, mpregwt_c, the only numeric variable that made it in my final model. Just like the previous plot, we can see that 95% of points are within the bands, and there does not seem to be a trend to be concerned about.

Using the mean of our *premature* response variable column as the cut-off threshold, a mother is predicted to give birth prematurely if the predicted probability is greater than or equal that *premature* column mean. The model achieves 0.60 accuracy, 0.61 sensitivity, and 0.60 specificity. The accuracy tells us that the model predicted 60% of the data correctly. 0.61 sensitivity means that given a baby was born prematurely, the model has 61% probability of predicting it was born prematurely. 0.60 specificity means that given a baby was not born prematurely, the model has 60% probability of predicting that baby was not born prematurely.

Below is the roc plot for this model, reporting AUC of 0.664. While this AUC is not the best, it is still a significant improvement from the AUC of the original full model we ran, which was 0.640.



Finally, it must be noted that multicollinearity could be a concern for our model, as the VIF score for both the interaction between smoke and race $smoke * mrace$ and $mrace$ on its own came out on the high side, each at 5.23. However, the other VIFs don't cause alarm.

Model summary/interpretation:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8301	0.2033	-9.00	0.0000
factor(smoke)1	0.3971	0.2279	1.74	0.0814
factor(mrace)Mexican	0.1874	0.6292	0.30	0.7658
factor(mrace)Black	1.0552	0.3058	3.45	0.0006
factor(mrace)Asian	0.8273	0.4947	1.67	0.0944
factor(mrace)Mixed Race	-13.5150	413.9505	-0.03	0.9740
factor(med)8th-12th Grade	0.3472	0.2535	1.37	0.1708
factor(med)College	-0.1567	0.2677	-0.59	0.5582
factor(med)High School plus Trade School	0.1652	0.3825	0.43	0.6658
factor(med)Less than 8th Grade	0.9064	0.9602	0.94	0.3452
factor(med)Some College	-0.6680	0.2687	-2.49	0.0129
factor(med)Trade School	2.7456	1.1762	2.33	0.0196
mpregwt_c	-0.0127	0.0048	-2.62	0.0087
factor(smoke)1:factor(mrace)Mexican	-0.0325	1.1125	-0.03	0.9767
factor(smoke)1:factor(mrace)Black	-0.5652	0.4241	-1.33	0.1826
factor(smoke)1:factor(mrace)Asian	0.3170	0.8451	0.38	0.7076
factor(smoke)1:factor(mrace)Mixed Race	14.4624	413.9524	0.03	0.9721

Table 1: Final Regression Model

As this summary indicates, the only significant predictors are $mraceBlack$ (when the mother's race is black), $mpregwt_c$, and a couple of the med factors, $medSomeCollege$ (when the mother has graduated high school and has done some college) and $medTradeSchool$ (the mother has done trade school but it's unclear if she graduated from high school). The coefficient for $mraceBlack$ is 1.06, which means that if the mother's race is black, the odds of her having a premature birth increase by 100*(exp(1.06)-1) = 187%. The coefficient for $mpregwt_c$ is -0.01, which means for each pound increase in the mother's pre-pregnancy weight, the odds of her having a premature birth decrease by 100*(exp(-0.01)-1) = 1%. The coefficient for $medSomeCollege$ is -0.67, which means that for a mother who has graduated high school and has done some college, the odds of her having a premature birth decrease by 100*(exp(-0.66)-1) = 48.8%. The coefficient for $medTradeSchool$ is 2.75, which means that for a mother who is in trade school, the odds of her having a premature birth increase by 100*(exp(2.75)-1) = 1464%.

We interpret the -1.83 intercept to mean that for a mother where all of our predictor variables are at their baseline, the odds of her having a premature birth are about exp(-1.83)= 0.16. For this model, this would

be a mother who is white, was of average weight before she got pregnant, has a high school education with no additional schooling, and has never smoked.

Our model raises concern as there are some coefficients here that don't make sense and a very small number of our predictor variables ended up being significant. More data would likely help us improve this model if we were to continue working to improve it. For instance, the most extreme coefficient let us know that the the odds of a mother having a premature birth increase 1464% if the mother is in trade school. We only have 4 data points for mother's in trade school, which likely explains why this value is so large.

Because the predictor variables *smoke*, *mrace*, and the interaction $smoke * race$ came out to be insignificant, answering our questions of interest for this analysis becomes difficult. Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? According to this analysis, whether or not a mother smokes is not a significant predictor of pre-term births, as it's p-value is over 0.05 at 0.08. However, the coefficient for this relationship is 0.397, which we would interpret as a 49% increase in the odds of a mother giving birth prematurely given that she smokes. So yes, if this relationship was significant, we could conclude that smoking mothers do tend to have higher chances of pre-term birth than mothers who do not smoke. A likely range for the odds ratio of pre-term birth for smokers is [0.95, 2.34], the 95% confidence interval for *smoke*. There is not substantial evidence that this relationship between smoking behavior in mothers and pre-term births changes based on race either. Each of the p-values for the $smoke * mrace$ interactions are quite high, ranging from 0.18 (Black mothers) to 0.97 (Mexican mothers). However, if we were to deem them as significant, we would find that given that the mother smoked, the odds of a premature birth would decrease by 3.2% for Mexican mothers, decrease by 43% for Black mothers, increase by 37% for Asian mothers, and increase 190963700% for mixed race mothers. This extreme value can be again be explained by a lack of data for mixed race mothers. Additionally, it doesn't make sense that smoking would decrease the chance of a premature birth for some races. I believe more data would help resolve these issues, or perhaps at least create a model where this relationship is considered significant. However it is interesting and worth discussing the few predictors that our model did find to be significant. Although the smoke:mrace interaction is not significant according to our model, we did find that the odds of a premature birth increase by 2.87% for black mothers, which shows that even though the effects of smoke don't necessarily change by race, race does seem to have an effect on its own, at least for Black vs White mothers. Additionally, our model reports that a higher pre-pregnancy weight decreases the odds of a premature birth.

Conclusion:

In this analysis, we sought out to determine whether or not smoking behavior in mothers increases the odds of them having a pre-term birth compared to mothers who do not smoke. The logistic model we built did not find the relationship between maternal smoking behavior and the odds of a premature birth to be significant. Additionally, our model did not provide compelling evidence that the effects of maternal smoking behavior on the odds of having a premature birth differ by the mother's race. However, we did learn some other interesting things from our model, such that both race on its own (at least if the mother's race is Black) and pre-pregnancy rate have an effect on the odds of a premature birth.

Limitations:

- This is an observational study, so we cannot make causal inference statements from the results of a regression model.
- Our model is lacking in significant predictor variables, however we keep them to control effects. This lack of significant predictors prevented us from really answering our questions of interest for this study.
- Our model has some extreme coefficients. As mentioned several times in the previous section, many of the this could probably be resolved with more data.
- The AUC is not super high despite trying many different logistic regression model versions. A better model could possibly be fit with more data.
- Similarly to the AUC, our model's accuracy, sensitivity and specificity leave more to be desired.
- Multicollinearity is a concern for our model, as we had some VIF scores above 5, for the variables *mrace* the $smoke * mrace$ interaction.