Assignment 2

Question 1

Part a

term	estimate	std.error	statistic	p.value
(Intercept)	3.4607477	0.0516314	67.02792	0
interval_centered	0.0686062	0.0040006	17.14887	0

Based on the results, we can see that the 'Interval' as a predictor is statistically significnat in response variable 'Duration', and our model predicts that the for a subsequent interval of 71 minutes, the duration of eruption will be 3.46 minutes. For every unit of increase in interval, the time duraiton of eruption will increase by 0.0686 minutes. Besides, this model explains about 73.44% of variation in this data set.

Part b

CI = 0.696 +- 2 * 0.004 We can reject the null hypothesis because our confidence interval does not include 0.

Part c

According to the residual graph, we can see a quadratic pattern, which means that our linearity assumpiton is violated.

Part d

Our results show that, compared to Day1 as the baseline, there is no significant difference in mean intervals for any of the days.

Part e

Using `anova()`, the F-test shows that the difference between these two models are not statistically significnat.

Part f

The result shows that these our second model has decreased RMSE, meaning that compared after adding `Day` as a predictor, the predictive accuracy of the model has been enhanced, although this enhancement may not be statistically significant.

Question 2

Summary

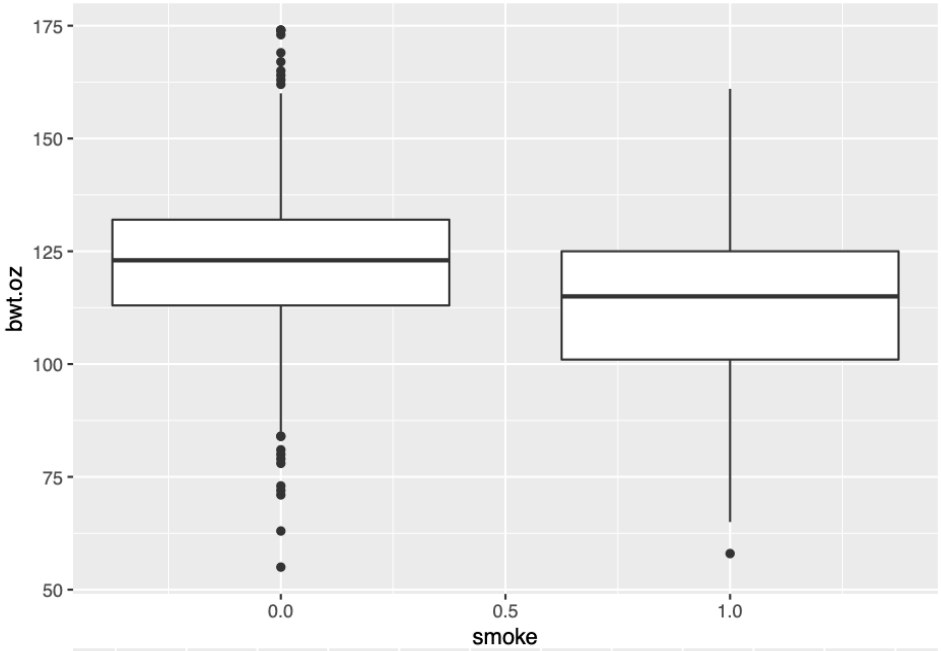
Drug use behavior during pregnancy has been found to be significantly influential to the health of new born. In this report, we look at one particular drug use and the one metrice of the new born, which is whether or not the mother smokes and the weight of the new born. In order to investigte the relationship, we adopt a multiple-linear-regression model, using whether or not the mother smokes, parity, mother's ethnicity, mother's height (in inches), mother's pre-pregnancy weights (in pounds), as well as the interaction between race and smoke as predictors for birth weight (in ounces).The analysis finds that while smoke and race are statistically significant predictors for new born's weight,respectively, the interaction between them is in fact not significant. Other variables such as mother's height and pre-pregnancy weights are also found to be statistically significant.

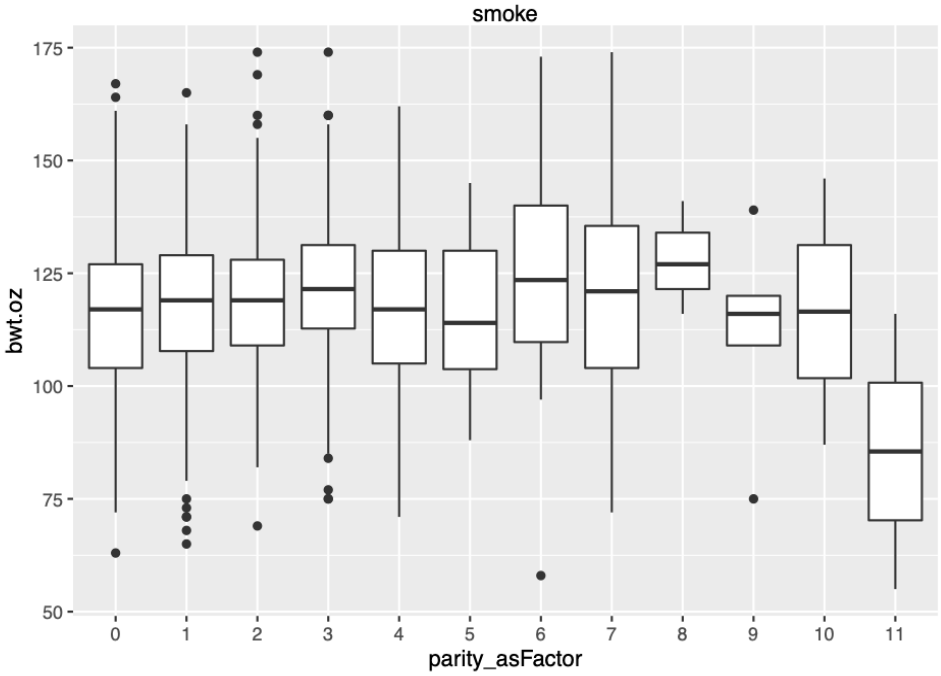
Introduction

In this analysis, we are interested several questions: do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke? What are some other factors that play a role in new borns' weight by themselves or by interacting with the fact that whether the mother smokes or not? If the weight of babies whose mothers smoke are found to be different from those whose mothers who do not smoke, what would be the range of differences? By investigating on these questions, we hope to gain some insights about how cigarette usage during pregnancy can affect the new born, as well as other factors that can potentially exert influence on the baby.

Data

In this dataset, variables such as parity, race, and education are ordinal numerical variables. To improve the interpretability,we have factorized these variables into different levels. Our EDA shows that, first of all, the distribution of birth weight follows a normal curve. When plotting smoke against birth weight using boxplot, we find that the distribution of response variable varies across smoker vs. non-smoker group. However, the significance of this pattern requires further statistical investigation.





Secondly, we find that when plotting parity against birth weight, mothers who have had 11 parities tend to

give births to babies who are significantly lighter than those in other groups.

Model

First of all, we adopt a full linear regression model that includes all variables in the dataset as well as interaction terms that may have a significant effect.

$$y\_birthWeight = \beta_0 + \beta_{1:2}x_{smoke} + \beta_{3:13}x_{mrace} + \beta_{14:25}x_{parity} + \beta_{26:35}x_{med} + \beta_{36}x_{mage} + \beta_{37}x_{mht} + \beta_{38}x_{mpregwt} + \beta_{39}x_{smoke} : x_{mr}$$

Given the large number of predictors we have, backward model selection using AIC is used to screen out variables that are deemed uncessary by the AIC metrice. This method leaves us with variables include: **`smoke`**, **`mrace`**, **`parity`**, **`mht`**,and **`mgregwt`**. Specifically, the interaction between **`smoke`** and **`race`** is significant when smoke level is 1 (i.e, smoker) and race level is 1 (i.e., White). Thus, our new model should also keep the interaction term between the two variables.

Since backward model selection using AIC works as a matrice instead of a statistical test, we need to fit another model using the variables screened by backward model selection and compare our new model with the full model, and choose the one that is more efficient. New model with the interaction of interest can be written as:

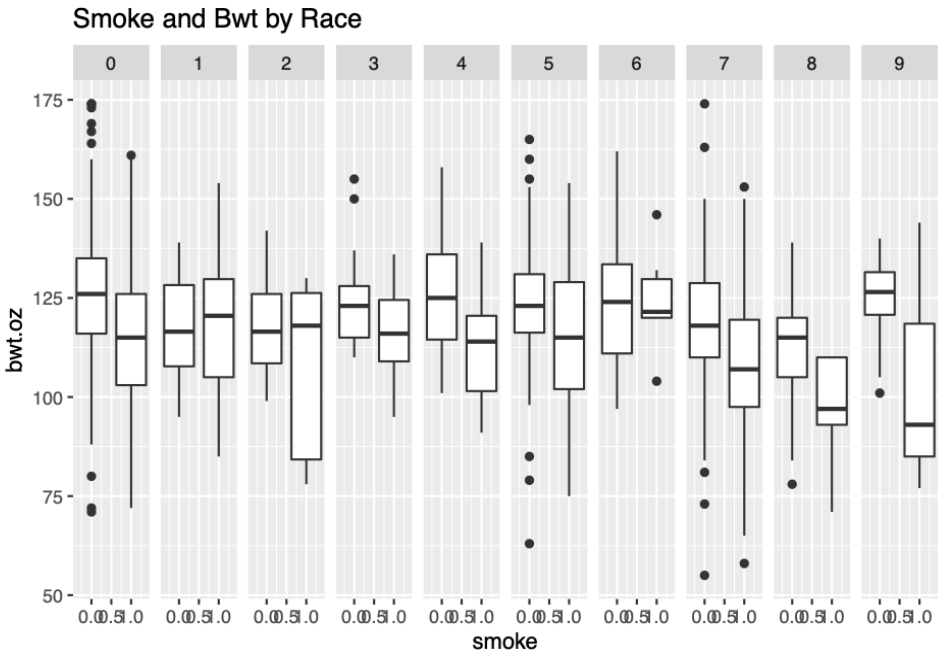
$$y\_birthWeight = \beta_0 + \beta_{1:2}x_{smoke} + \beta_{3:13}x_{mrace} + \beta_{14:25}x_{parity} + \beta_{26}x_{mht} + \beta_{27}x_{mpregwt} + \beta_{28}x_{smoke} : x_{mrace} + \epsilon_i$$

Next, this model is compared with the full model using **`anova()`**. The result shows that in fact, the F value between these two models are 0.56. Consequently, we fail to reject the null hypothesis. This means that both the new model, which has fewer parameters, and the full model, which has more parameters, are equally effective. As a result, the new model is preferred.

term	estimate	std.error	statistic	p.value
(Intercept)	38.2938052	15.9955436	2.3940296	0.0168838
mht	1.1221535	0.2691019	4.1699944	0.0000336
mpregwt	0.0978820	0.0322895	3.0313856	0.0025094
smoke	-10.9899954	1.6922738	-6.4942181	0.0000000
mrace_asFactor1	-8.8918286	4.3336088	-2.0518300	0.0404972
mrace_asFactor2	-5.8215127	6.0290829	-0.9655718	0.3345377
mrace_asFactor3	0.5824243	3.5570449	0.1637382	0.8699768
mrace_asFactor4	2.0253435	4.1121064	0.4925319	0.6224727
mrace_asFactor5	-2.2101032	2.4995084	-0.8842152	0.3768343
mrace_asFactor6	0.4555271	4.0416351	0.1127086	0.9102886
mrace_asFactor7	-10.0953369	2.1749995	-4.6415354	0.0000040
mrace_asFactor8	-6.4054234	3.6195613	-1.7696684	0.0771469
mrace_asFactor9	-0.3805130	4.9627292	-0.0766741	0.9389011
parity_asFactor1	1.9121013	1.6203422	1.1800602	0.2383119
parity_asFactor2	4.2353842	1.7412930	2.4323214	0.0152113
parity_asFactor3	5.5251632	1.9500082	2.8334051	0.0047164
parity_asFactor4	4.6764455	2.4602893	1.9007706	0.0576758
parity_asFactor5	3.3203049	2.9410797	1.1289408	0.2592466
parity_asFactor6	7.9463581	3.7745182	2.1052642	0.0355661
parity_asFactor7	3.3995916	5.0151889	0.6778591	0.4980486
parity_asFactor8	16.8318701	9.7491372	1.7264984	0.0846272
parity_asFactor9	-2.4944389	7.5781712	-0.3291611	0.7421164
parity_asFactor10	4.1754652	12.3638568	0.3377154	0.7356624
parity_asFactor11	-27.2409542	11.9211138	-2.2851014	0.0225565
smoke:mrace_asFactor1	13.0153392	5.9605598	2.1835767	0.0292707
smoke:mrace_asFactor2	-0.5767600	8.0850679	-0.0713364	0.9431471
smoke:mrace_asFactor3	5.4249558	5.3446498	1.0150255	0.3103873

term	estimate	std.error	statistic	p.value
smoke:mrace_asFactor4	-3.5526044	5.3854892	-0.6596623	0.5096522
smoke:mrace_asFactor5	4.0130429	3.8355361	1.0462795	0.2957344
smoke:mrace_asFactor6	13.6801264	8.2763520	1.6529174	0.0987231
smoke:mrace_asFactor7	2.6498875	3.1093916	0.8522206	0.3943358
smoke:mrace_asFactor8	-6.6981192	6.6859679	-1.0018174	0.3167217
smoke:mrace_asFactor9	-10.6611710	10.8660091	-0.9811487	0.3268033

Based on the adjusted-r-squared value, this model explains 15.54% of total variation in this dataset. Specifically, our primary variable of interest: `smoke`, is statistically significant with a p-value less than 0.05. On average, the possible range of difference in new born's weight between smoker and non-smoker according to this model is from 7.3 to 14.68 ounces. The model assessment shows that the linearity assumptions are generally satisfied between each predictor and the response variable. Normality and equal variance assumpitons are also met. Besides, there are not outstanding outliers or influential points that are particularly influential to this model. Based on this model, we can conclude that holding everything else constant, mothers who smoke do tend to give birth to babies with lower weights than mothers who do not smoke by 10.99 ounces. There is no evidence that the association between smoking and birth weight differs by mother's race, except when the mother smokes and the race is at level 1, which is White.



Potential limitation

We checked only male babies.' This analysis helps answer one of the important questions in pregnancy healthcare: will the mother's smoking behavior influence the new born? Based on our model, the answer is positive. In this context, we have found that mothers who smoke tend to give birth to babies who have less weight. Additionally, this effect is generally universal across race, excpet when the mother smokes and the ethnicity falls under level 1. This can be a special case that requires further attention. However, this study is

still limited in two ways: first of all, all the new borns under investigation are male babies. Thus, we have not explored whether or not this effect that smoking has will change if the baby is female. Besides, what does less weight mean to a new born, and to what extent will this difference be dangerous or even fatal for a new born? There are still a lot of questions that are not answered by this analysis, and further investigation is needed.