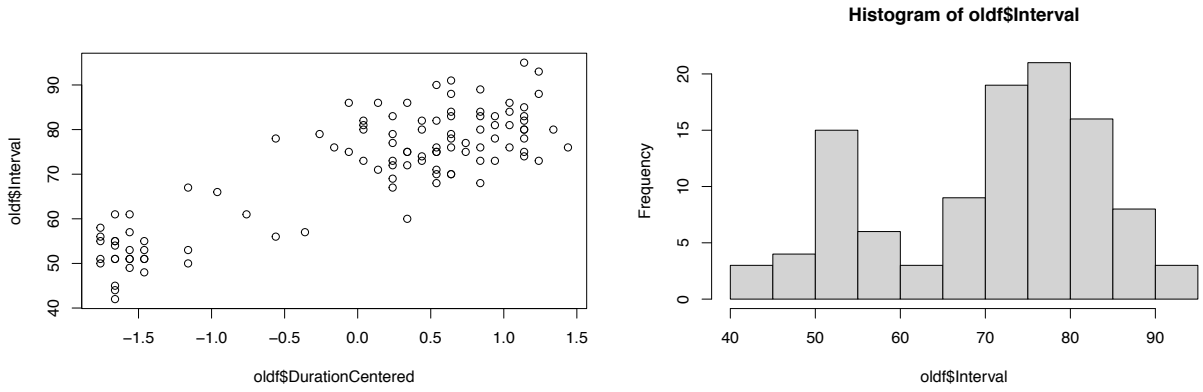
Assignment2

Erika Fox

9/7/2021

Exercise 1

EDA:



This scatterplot tells us before we even begin modeling this data set that there is a general positive correlation between the *Duration* of the previous eruption and the *Interval* between eruptions. This means that the longer the duration of Old Faithful's most recent eruption lasted, the longer the interval between eruptions should be.

This histogram of our response variable interval shows that there is not a normal distribution, which may cause us some problems when we begin modeling.

Regression model for predicting the interval between eruptions from the duration of the previous one:

Model one: $Interval_i = \beta x_i + \epsilon_i$; $\epsilon_i \sim N(0,\sigma^2)$, $i = 1, \ldots, n$

β and x are vectors. β represents coefficients and x represents predictor variables.

Predictor variables: Duration

This model output shows that we have a very small p-value ($p<0.001$), which lets us know that there is in fact a significant statistical relationship between the duration of a volcano's most recent eruption and it's interval between eruptions. The 10.74 coefficient for *Interval* lets us know that per one unit increase in *Duration*, *Interval* is expected to increase by a factor of 10.74.

95% Confidence Interval for slope: 9.50, 11.98) Although the model resulted in a nice p-value ($p<0.001$) and results that make sense, these results should not be relied upon because based on this model's diagnostic plots the regression assumptions do not appear to be met. While the QQ-Plot shows that the normality assumption is met reasonably well with most points hugging the diagonal line, the linearity, equal variance and independence assumptions seem to fail as their plots show a trend that diverges from the randomness we want to see. The trend in these plots with no transformations show an upside down curve. Although the curve is loose, it is still enough of a trend that makes me not want to trust this model as it is.

	Model 1
(Intercept)	71.00***
	(0.65)
DurationCentered	10.74***
	(0.63)
R^2	0.74
Adj. R^2	0.73
Num. obs.	107

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Statistical models

Model two: $Interval_i = \beta x_i + \epsilon_i$; $\epsilon_i \sim N(0,\sigma^2)$, $i = 1, \ldots, n$

β and x are vectors. β represents coefficients and x represents predictor variables. Predictor variables: Duration, factor(Date)

	Model 1
(Intercept)	32.88***
	(3.07)
Duration	10.88***
	(0.66)
factor(Date)2	1.33
	(2.72)
factor(Date)3	0.78
	(2.70)
factor(Date)4	0.16
	(2.65)
factor(Date)5	0.25
	(2.65)
factor(Date)6	1.99
	(2.66)
factor(Date)7	−0.17
	(2.70)
factor(Date)8	−0.69
	(2.70)
R^2	0.74
Adj. R^2	0.72
Num. obs.	107

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Statistical models

According to this model output, there is not a significant difference in mean intervals for any of the days, as each of their p-values are rather large, all over 0.6, where we would need to see them at least at $p<0.05$.

F-statistic:

After running an Anova test comparing model one (without *Date*) and model two (with *Date*) I found that:

Pr(>F): 0.98

This large p-value indicates that our more complex model including *Date* does not fit the data significantly better than the simpler one. Therefore, the simpler model, model one, should be favored.

Using k-fold cross validation, I found the RSME for model one to be 6.69 and the RMSE for model two to be 6.70. These values are very similar. If we were to interpret this as the RMSEs being the same, model one is preferable, as it's the simpler model and we don't want to have extra predictors if they are not needed. However, even though they are close they are not the same, and the RMSE for model one is smaller. So in more ways than one, our model that does not include *Date* is the preferable model.

Exercise 2

Summary:

In the following experiment, we were interested in determining if whether or not a mother is a smoker has an effect on her baby's birth weight in ounces. Additionally, we were interested in finding evidence that the association between smoking and birth weight differs by mother's race. After some EDA, we discovered that there does appear to be a relationship between a mother's smoking habits and her baby's weight, in that the mean birth weight of babies in ounces to mothers who smoke was smaller than the mean birth weight of babies in ounces to mothers who have never smoked. We then selected a model using BIC in order to further pursue answers to our questions of interest, and were able to build on our findings from the EDA, in that we continued to see a decrease in birth weight among babies born to smoking mothers compared to non-smoking mothers. We were unable to find evidence that the association between smoking and birth weight differs by mother's race based on our model. Finally, there are some limitations to our model that are worth discussing, including a low R-squared value, and some coefficients that go against what we expected based on EDA and the rest of our model results.

Introduction:

In this experiment, we will be using birth data in order to explore the effects of smoking mothers on their babies' birth weights. We are interested in proving or disproving the assertion that mothers who smoke have increased rates of low birth weights. Additionally, we will explore if other factors, such as race, contribute to the effects of smoking mothers on their babies' birth weights.

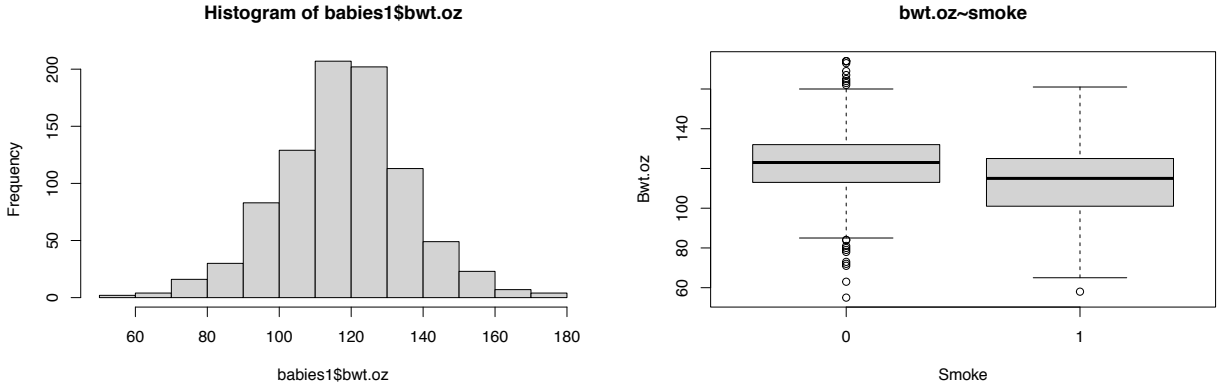
Questions of interest: Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke? What is a likely range for the difference in birth weights for smokers and non-smokers? Is there any evidence that the association between smoking and birth weight differs by mother's race? If so, characterize those differences. Are there other interesting associations with birth weight that are worth mentioning?

Data:

About this dataset:

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, 1.

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



These plots give us some basic information about the data we are working with.

This histogram on the right introduces us to our response variable, bwt.oz, and shows us that it has a nice, centered, normal distribution.

The boxplot on the right gives us a basic introduction to how the predictor variable smoke relates to birth weight in ounces. The plot indicates that when smoke is 1 (meaning that the mother does smoke), the range of the babies birth weights in ounces is lower than when smoke is 0. This shows that, before considering any other predictors, smoking and birth weight are generally negatively correlated.

I calculated several means in order to gain some more preliminary metrics that tell us about how the *smoke* variable and *bwt.oz* are related. I found the mean birth weight for mothers who smoke to be 113.53 ounces, and the mean birth weight of mothers who have never smoked to be 122.54 ounces. These values confirm the conclusion we drew from our boxplot.

Because some of our questions of interest pertain to the race of the mother as well, I did some more mean calculations to help start off when a general understanding of how a mother's race affects birth weight without considering smoke along with it just yet.

I found the average birth weight to white mothers to be 120.00, the average birth weight to black mothers to be 113.20, and the average birth weight for Asian mothers to be 109.44. These means show us that race does effect birth weight, as it seems that minority mothers tend to give birth to smaller babies. It will be interesting to see how this finding exhibits itself when we start modeling and considering the interaction between a mother's race and smoking habits.

Model:

Baseline model/assessment:

In order to choose a model to help us answer our questions of interest, I began by running a very simple baseline model that only had $smoke : mrace$ as a predictor for our response variable, *bwt.oz*, as this is as simple as a model could be and still be sufficient enough to help us reach our research goals. While this model seemed to do reasonably well when checking for assumptions, this model gave us an R-squared value of only 0.08, which shouldn't be hard to beat by adding some more predictors and/or playing with transformations.

Model selection: After choosing to move on from the baseline model, I tried running both BIC and AIC in order to select my model.

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

$bwt.oz_i = \beta x_i + \epsilon_i$; $\epsilon_i \sim N(0,\sigma^2)$, $i = 1, \ldots, n$

β and x are vectors. β represents coefficients and x represents predictor variables.

Predictor variables: smoke, mht, factor(mrace), mpregwt, smoke:factor(mrace)

This is the model that I got when running BIC. The model I got from running AIC is the following:

$bwt.oz_i = \beta x_i + \epsilon_i$; $\epsilon_i \sim N(0,\sigma^2)$, $i = 1, \ldots, n$

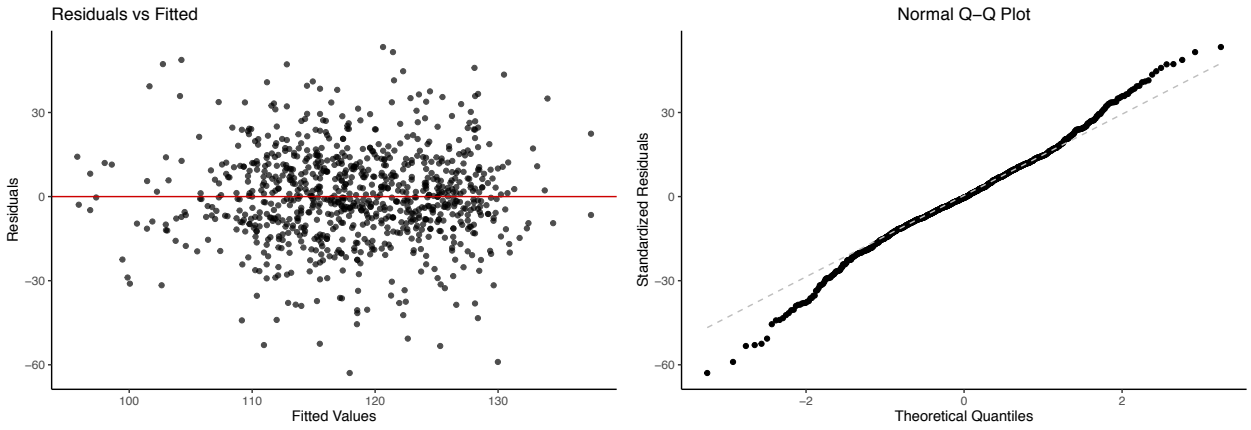
β and x are vectors. β represents coefficients and x represents predictor variables. Predictor variables: smoke, mht, factor(mrace), mpregwt, parity, smoke:factor(mrace)

I ended up choosing my BIC model over AIC because I found it to have a lower RMSE value ($15.79 < 19.14$) using k-fold cross validation with k=10. The BIC model also has the added benefit of using less variables, as we don't want to have extra variables in our model unless it significantly improves it.

Transformations: Once I decided to go with the BIC over the AIC model, 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 linear regression, and saw the assumptions were met reasonably well (explained in detail in the next section, Final model assessment). However, even though I was satisfied that the assumptions were met, I saw that there was room for improvement, especially when it came to the two tail ends of the QQplot for the normality assumption, and the R-squared value, which remained low at 0.15. I continued to try adding in a log transformation on the response variable, and then I re-assessed, and found that the changes were not significant enough for me to sacrifice my simpler, easily interpretable model. And then I added log transformations to each of the non-factored predictor variables in turn, and reached the same conclusion. Thus, I decided to move forward with my model exactly as it was reported by BIC, without transformations.

Final model assessment: This model satisfactorily meets the assumptions for linear regression. When plotting each predictor against the model residuals (see appendix for plots), there does not appear to be a trend, so the model satisfies linearity. When plotting the fitted values against the residuals for the model, we get a nice, evenly scattered result with no trend (the line of best fit is essentially the x-axis), so independence and equal variance are satisfied. Normality is satisfied, as most points cling to the line in the QQ plot. As noted in the previous section, the normality assumption could probably be satisfied better, but there is a trade off between more complex models and meeting the assumptions. Of any model transformations that might have slightly improved these assumptions, they did not improve the model enough to sacrifce the easily interpretable model.

Here are two latter plots:



When analyzing the Residuals vs Leverage plot, there did appear to be a handful of outliers, but I made the decision to not remove them as none appearred to have very high influence. In other words, I did not think any outliers skewed the model enough for there to be a need to remove them.

Finally, we do not have a multicollinearity issue because all of the VIFs for the model are low, in that they are less than 5. This is another benefit of choosing the simpler model that BIC choose over the AIC model.

model VIFs: smoke (1.39), mht (1.36), mrace (4.65), mpregwt (1.36), smoke:mrace (4.94)

Model summary/interpretation:

	Model 1
(Intercept)	49.86**
	(15.39)
smoke	−9.56***
	(1.34)
mht	0.93***
	(0.26)
factor(mrace)6	0.19
	(3.97)
factor(mrace)7	−8.92***
	(1.99)
factor(mrace)8	−6.30
	(3.54)
factor(mrace)9	0.77
	(4.92)
mpregwt	0.12***
	(0.03)
smoke:factor(mrace)6	14.56
	(7.98)
smoke:factor(mrace)7	1.63
	(2.92)
smoke:factor(mrace)8	−6.65
	(6.64)
smoke:factor(mrace)9	−12.38
	(10.88)
R^2	0.15
Adj. R^2	0.14
Num. obs.	869

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Statistical models

According to my model the coefficient for *smoke* is -9.56, which means that if a mother is a smoker, her baby's birth weight will decrease by an average of 9.56 ounces. The coefficient for *mht* is 0.93, which means that for each inch increase in mother's height, her baby's birth weight will increase by an average of 0.93 ounces. The coefficient for *mpregwt* is 0.12, which means that for each pound increase in mother's pre-pregnancy weight, her baby's birth weight will increase by an average of 0.12 ounces.

I found the coefficients for *mrace* to be insignificant (large p-value, p is not less than 0.05) except for when *mrace* is equal to 7, for black mothers. The coefficient for factor(*mrace*)7, is -8.92 with p<0.001, which means that if the mother's race is black, her baby's birth weight will decrease by an average of 8.92 ounces.

However, when interpreting the effects on birth weight based on mother's race assuming that she is a smoker ($smoke : factor(mrace)$), the interaction between smoke and race), I found that none of the races had a significant impact on birth weight in that none of their p-values were less than 0.05. In other words, there is no evidence that the association between smoking and birth weight differs by mother's race. The test also has an f-stat of 14.07, which shows that there is no association between the response and predictor variables. Actual birth weight should deviate from the regression line by about 16.72 on average, according to the RSE. These results allow us to conclude that mothers who smoke do indeed give birth to babies with lower weights, as we found a significant negative correlation with a slope of -9.56 and a p-value $<$ 0.001. Although my model gave -9.56 as the coefficient for *smoke*, the likely range for the difference in birth weights for smokers and non-smokers is [-12.19, -6.93], the 95% confidence interval for *smoke*. Finally, aside from race and smoke as predictors, my model reveals that a mother's height and pre-pregnancy weight have associations with birth weight, as it appears that as mother's increase in their size, so do the babies they

birth. This makes logical sense, but it helps to see it reinforced in the model with positive correlations at slope 0.93 and 0.12 for weight and height, respectively.

Conclusion:

This experiment provides evidence that mothers who smoke tend to give birth to smaller babies. The association between the smoking status of mothers and their babies' birth weights is extremely important information as it will help the medical world accurately inform patients on what to do in order to get the healthiest possible baby in regards to smoking (namely, not to smoke), and generally reinforces the idea that smoking is not a great health choice. However, this experiment should not be taken as the sole proof for the assertion that smoking leads to less healthy babies, as there are limitations in this study. First of all, this is an observational study, so we cannot make causal inference statements from the results of a standard regression model. The results and interpretations from this study can only provide evidence towards the assertion. This study cannot guarantee causation. There are other factors that lead to premature babies that have nothing to do with the mother's smoking status or other predictors used in the model here. For example, we did not consider gestation time at all in this experiment, as that is an outcome variable in our data set. This is because the study was looking to see if smoking caused more babies to be born prematurely, but there are many instances in the real world where babies are born early that have nothing to do with the mother smoking, which is something to consider.

More limitations of this experiment include our model's low R-squared value, the coefficients our model outputted for the $smoke : mrace$ interaction, and the fact that we did not end up adjusting for any outliers.

Even though we did an extensive model selection process including an exploration of a long list of transformation possibilities, we still ended up with a low R-squared value for our model, at a mere 0.15. I am convinced this is the best we could do for this particular experiment with our particular data set, however, we would ideally want to see a higher R-squared value in order to have the most trustworthy model.

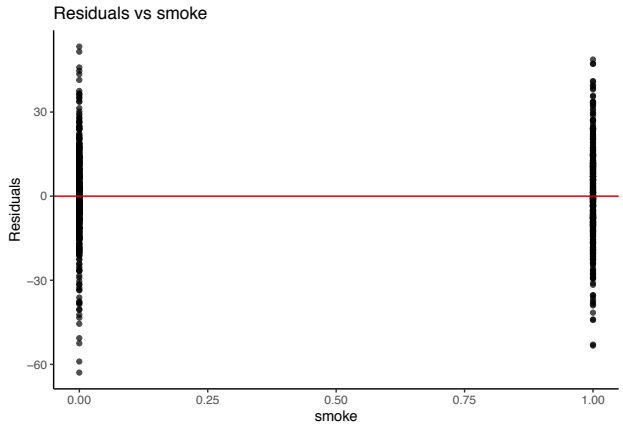
Another glaring limitation this experiment has is that the coefficients our model outputted for the $smoke : mrace$ interaction just don't make any sense. For $smoke:mrace$6 and $smoke:mrace$7, the model reported a positive correlation between $smoke$ and $bwt.oz$ for mother's of those races that smoke (Mexican and Black respectively). This would indicate that smoking actually yielded heavier babies for mother's of those races, which does not align with what we found during our EDA and with the rest of the model's results. This leaves us to doubt the integrity of our model.

Finally, there is the fact that we did not end up removing any outliers from our dataset. Although this was a deliberate choice after seeing that the apparent outliers in the data did not have high influence, there is always the possibility that there is an effect happening because of an outlier that we could not have seen coming through the assessment checks that we did.

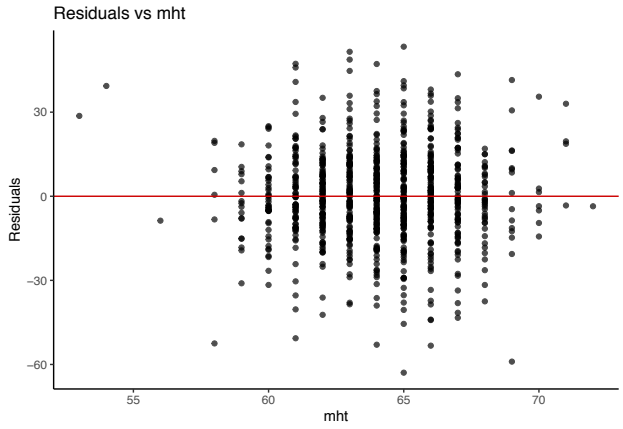
A great next step for our questions of interest would be to run a similar study to this one but on a different dataset, to see how results compare.

Appendix:

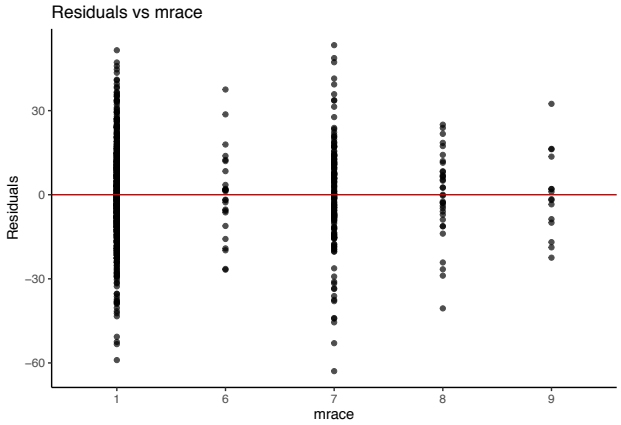
```
model=lm(bwt.oz ~ smoke + mht + mrace + mpregwt + smoke:mrace,
    data = babies1)
ggplot(babies1,aes(x=smoke, y=model$residual)) +
  geom_point(alpha = .7) + geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs smoke",x="smoke",y="Residuals")
```




```
ggplot(babies1,aes(x=mht, y=model$residual)) +
  geom_point(alpha = .7) + geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs mht",x="mht",y="Residuals")
```




```
ggplot(babies1,aes(x=mrace, y=model$residual)) +
  geom_point(alpha = .7) + geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs mrace",x="mrace",y="Residuals")
```




```
ggplot(babies1,aes(x=mpregwt, y=model$residual)) +
  geom_point(alpha = .7) + geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs mpregwt",x="mpregwt",y="Residuals")
```