Assignment 2

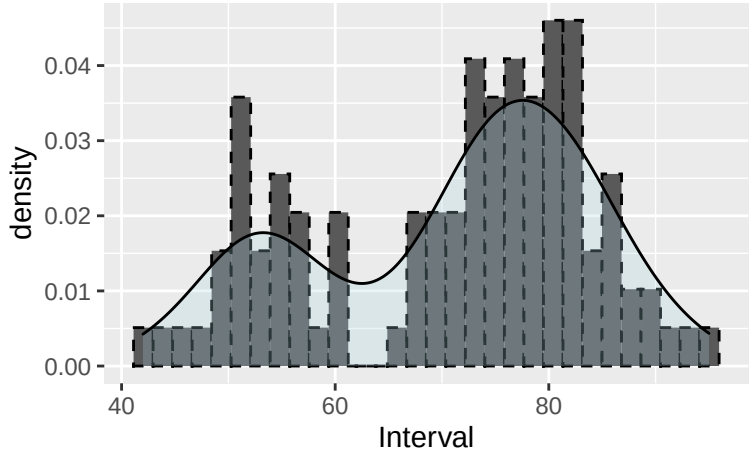
Pranav Manjunath

9/5/2020

Question 1

Fit a regression model for predicting the interval between eruptions from the duration of the previous one, to the data, and interpret your results.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



With the assumption that if there was more data between 55 and 70 the above histogram could be considered as a normal distribution, I have considered to take Interval as the response variable, without undergoing any transformation. When comparing the response variable to it's log transformation, I noticed that the response variable followed a relatively better normal distribution.

Regression Model Formula:

$\hat{Interval_i} = \hat{\beta}_0 + \hat{\beta}_1 * Duration_i$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.83	2.262	14.96	8.98e-28
Duration	10.74	0.6263	17.15	3.249e-32

Table 2: Fitting linear model: Interval ~ Duration

Observations	Residual Std. Error	R^2	Adjusted R^2
107	6.683	0.7369	0.7344

From the above model summary, it is shown that duration is a statistically significant variable (p value 3.249e-32 « 0.05) in predicting the intervals of earthquakes. The intercept of the model is 33.83 indicating that if the duration of the earthquake is 0 minutes (no earthquake), the interval between the earthquake is 33.83 minutes (conceptually does not make sense as the model is not centered). The coefficient of the Duration is 10.74, indicating that a unit increase of duration will increase the interval by 10.74 minutes. This model has an adjusted R-squared value of 0.7344 which shows that 73.44% proportion of variation in the response variable is being explained by the regression model.

Include the 95% confidence interval for the slope, and explain what the interval reveals about the relationship between duration and waiting time.

	2.5 %	97.5 %
(Intercept)	29.34	38.31
Duration	9.499	11.98

It can be seen that in a 95% confidence interval, the Duration interval is 9.50 to 11.98 minutes. In a 95% confidence interval, a unit change in duration will result in an increase in the interval by minimum of 9.50 minutes and a maximum of 11.98 minutes.

Describe in a few sentences whether or not you think the regression assumptions are plausible based on residual plots (do not include any plots).

While investigating the residuals vs fitted plot for independence, I noticed a slight pattern, like a quadratic curve. A reason this might exist is that there is less data for intervals between 55 and 70 minutes and hence the middle portion of the plot does not have enough data points. The equal variance assumption is met. With respect to the normality and linearity, the QQ plot showed signs of normality as majority of the points seemed to lie on the 45 degree line and the model residuals vs predictor plot seems to have randomness.

Fit another regression model for predicting interval from duration and day. Treat day as a categorical/factor variable. Is there a significant difference in mean intervals for any of the days (compared to the first day)? Interpret the effects of controlling for the days (do so only for the days with significant effects, if any).

Linear Regression Model:

$\hat{Interval}_i = \hat{\beta}_0 + \hat{\beta}_1 * Duration_i + \hat{\beta}_{2:8} * Date_i$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.88	3.067	10.72	3.356e-18
Duration	10.88	0.6622	16.43	6.543e-30
Date2	1.328	2.717	0.4885	0.6263
Date3	0.7825	2.699	0.2899	0.7725

	Estimate	Std. Error	t value	Pr(>|t|)
Date4	0.1625	2.646	0.06143	0.9511
Date5	0.2463	2.646	0.0931	0.926
Date6	1.992	2.658	0.7494	0.4554
Date7	-0.17	2.702	-0.06292	0.95
Date8	-0.6944	2.696	-0.2576	0.7973

Table 5: Fitting linear model: Interval ~ Duration + Date

Observations	Residual Std. Error	R^2	Adjusted R^2
107	6.866	0.7408	0.7196

The baseline group for the factor variable Date is Date1. When noticing just the coefficients for the various Date values, Date6 has the highest difference from the baseline Date1, with a coefficient of 1.992. However, since the p values of each date is above 0.05 (not statistically significant), there is no significant difference in mean intervals for any of the days when compared to the first day.

Perform an F-test to compare this model to the previous model excluding day. In context of the question, what can you conclude from the results of the F-test?

Table 6: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
98	4620	NA	NA	NA	NA
105	4689	-7	-68.85	0.2086	0.9828

As the P Value is significantly large (0.9828) when compared to 0.05, we would fail to reject the Null Hypothesis. This concludes that adding the extra parameter (Date) does not add statistical significance to the original model

Using k-fold cross validation (with k=10), compare the average RMSE for this model and the average RMSE for the previous model excluding day. Which model appears to have higher predictive accuracy based on the average RMSE values?

```
## RSME of Model 1 (without Date):  6.561417
```

```
## RSME of Model 2 (with Date):  6.977411
```

The model without Date has a RSME value of 6.56 while the model with date has a RMSE value of 6.98. By comparing the RSME values, The original model (without DATE) seems to have a higher predictive accuracy than the model with Date as it has a smaller RSME value.

Question 2

SUMMARY

The goal of this analysis is to understand the relationship between the birth weight of a child and the mother's smoking habits along with other attributes. To understand these relationships, I performed EDA on the data and then fitted a multiple linear regression model (Formula: bwt.oz ~ smoke + parity + mht + mpregwt +mrace + smoke:mrace) to answer the questions of interest. After creating the model, it can be statistically concluded that mothers who smoke tend to give birth to babies with lower birth weight and that the mother's pre-pregnancy weight, height, race, previous pregnancies, and smoking habits play an important role in predicting the newborn's weight (birth weight).

INTRODUCTION

The Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA addressed the issue of pregnancy and smoking. The researchers interviewed mothers early in their pregnancy to collect information on socioeconomic and demographic characteristics, including an indicator of whether the mother smoked during pregnancy. This data is now used to help analyze and answer the following questions.

- Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke? Ideally this question is looking at the relationship between the birth weight of the child and the mother's smoking habit.
- What is a likely range for the difference in birth weights for smokers and non-smokers? Using the 95% confidence interval, we can determine the 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. This question mainly looks at determining whether the interaction between race and smoke and it's effect on birth weight is statistically significant.
- Are there other interesting associations with birth weight that are worth mentioning? This question looks at understanding if there are other variables in the data that are statistically significant with the response variable birth weight (bwt.oz).

To help answer these questions, EDA plots and a multiple linear regression model are built incorporating the attributes that are statistically significant with the response variable (bwt.oz).

DATA

The data I have considered for the analysis consists of 869 observations and 12 variables. The following columns are: id, date, gestation, bwt.oz, parity, mrace, mage, med, mht, mpregwt, inc, and smoke.

The columns id, date, and gestation will not be used in this analysis.

When plotting the histogram for the response variable birth weight (bwt.oz), we can conclude that the birth weight of a newborn (bwt.oz) follows a normal distribution. This will be the response variable in the linear regression model. No transformations will be done on the response variable. I did not consider the rows with missing values. Used the smoking.csv dataset for the entire analysis.

Transformations

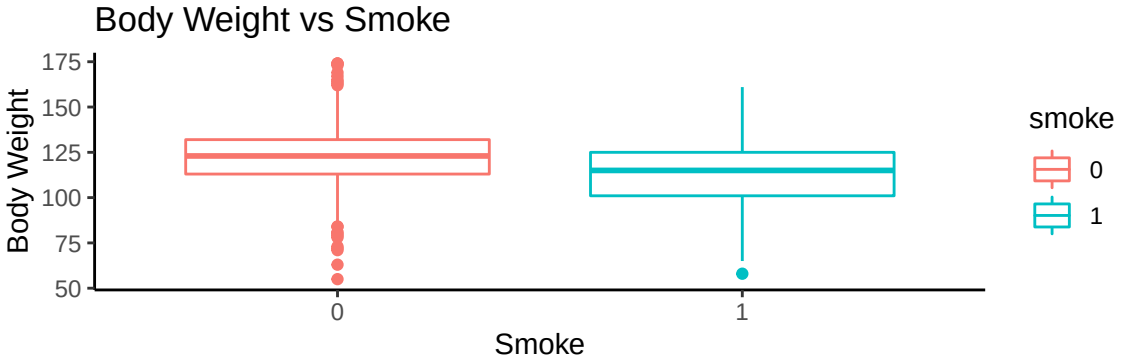
I did not perform any log transformation to any of the variables. The data transformations performed was

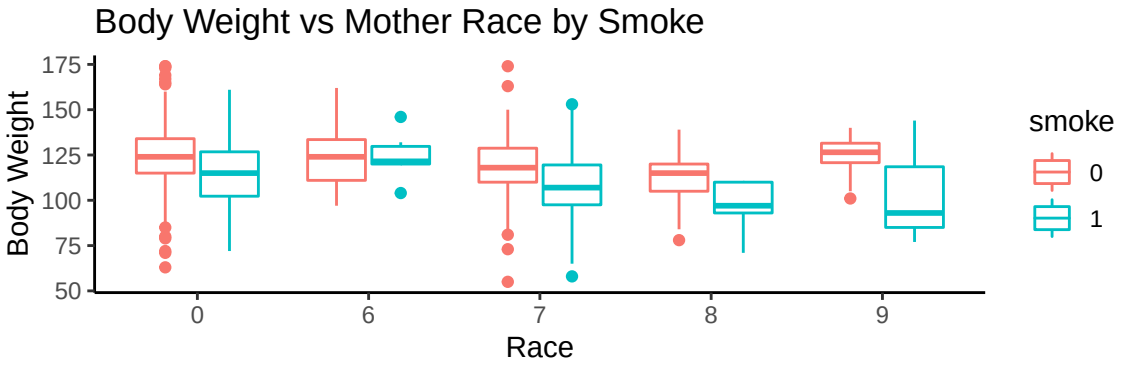
1. Converting the mrace variable values 0,1,2,3,4,5 to 0 (White race)
2. Converting categories 6 and 7 of the med variable into one category, Trade School

The variables Smoke, inc, med, and race variables have been converted into factor variables.

Exploratory Data Analysis

The two plots I have included in the report are plots I felt were important to help answer the questions. The first plot is a box plot of Birth Weight vs Smoking and this plot clearly indicates the effects of mothers smoking on the newborn birth weight. The second plot describes the relationship between Race, Smoking and Body Weight and identifies the interaction between race and smoking. From the second plot, we can identify that Race plays a significant role in determining the birth weight along with understanding that the interaction between smoking and race does not seem to be significant (the trend that mothers who smoke tend to give birth to new born babies of lesser weight remains similar across various races). However as we are aiming to understand the importance of this interaction, I have included it in the final model.





Observations from EDA:

- *Smoke vs Body Weight:* - There is a difference in median baby's body weights of mother who smoke and who do not smoke. When observing the Birth Weight vs Smoking box plot above, it is seen that there is a shift in the boxes. The median birth weight from mothers who smoke are less than the median birth weight from mother who do not smoke. As the questions of interest look at this relationship, I have considered this variable in the regression model.
- *Parity vs Body Weight:* - There is a difference between the median baby's body weight among categories of Parity.
- *Mother's Race vs Body Weight:* - There is a significant difference between median baby's body weight amongst categories of Race.

- *Mother's Age vs Body Weight:* - There does not seem to be much difference between the mother's age and the birth weight (indicated by a slope near 0) indicating that an increase in mother's age will not affect the birth weight tremendously.
- *Mother's Education vs Body Weight:* - There is not a significant difference in the body weight among the categories of Education implying that birth weight does not depend significantly on the mothers education status.
- *Mother's Height vs Body Weight:* - There seems to be a positive relationship between the mother's height and the weight of a newborn (indicated by the positive sloping curve).
- *Mother's Income vs Body Weight:* - There seems to be a very minute difference between the various income groups and the newborn's body weight.
- *Mother's Pre Pregnancy Weight vs Body Weight:* - There seems to be a positive relationship between the mother's pre pregnancy weight and the birth weight of the new born baby. Mothers who weigh more before pregnancy tend to give birth to babies of more body weight. I have included this in the model

MODEL

Upon performing EDA, I noticed that variables such as smoke, mrace, mpregwt, mht, and parity seem to have some relationship with the response variable. Looking at interactions, it is important for the final model to include the interaction between smoke and race as it is one of our question of interest.

I used forward AIC for the Model Selection Criterion.

The null model used includes smoke, mrace, and the interaction between smoke:mrace as the predictor variables. According to the questions that needs to be addressed, it is necessary that the final model must contain these predictors. The full model used contains all the variables in the data (excluding id, gestation, and date). As the study mainly looks at the smoking patterns, I have also included all the interactions with smoke in the full model.

The full model has an AIC score of 7410.71.

After performing forward AIC, the model outputted included mht, mpregwt, race, smoke, parity, and smoke:race as the predictor variables.

The AIC model above has an AIC score of 7372.87.

The final regression model used is:

$\hat{bwt.oz_i} = \hat{\beta}_0 + \hat{\beta}_1$ * smoke + $\hat{\beta}_2$ * parity + $\hat{\beta}_3$ * mht + $\hat{\beta}_4$ * mpregwt +$\hat{\beta}_{5:8}$ * mrace + $\hat{\beta}_{9:12}$ * smoke:race

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.39	15.45	3.003	0.002749
smoke1	-9.646	1.339	-7.201	1.306e-12
parity	0.6544	0.3162	2.07	0.03878
mht	0.9862	0.262	3.764	0.0001783
mpregwt	0.1085	0.03216	3.373	0.0007759
mrace6	0.2106	3.959	0.0532	0.9576
mrace7	-9.688	2.025	-4.784	2.019e-06
mrace8	-5.905	3.543	-1.667	0.09593
mrace9	0.4871	4.914	0.09912	0.9211
smoke1:mrace6	13.11	7.993	1.64	0.1014
smoke1:mrace7	1.954	2.922	0.6686	0.504
smoke1:mrace8	-7.391	6.633	-1.114	0.2655
smoke1:mrace9	-12.55	10.86	-1.155	0.2484

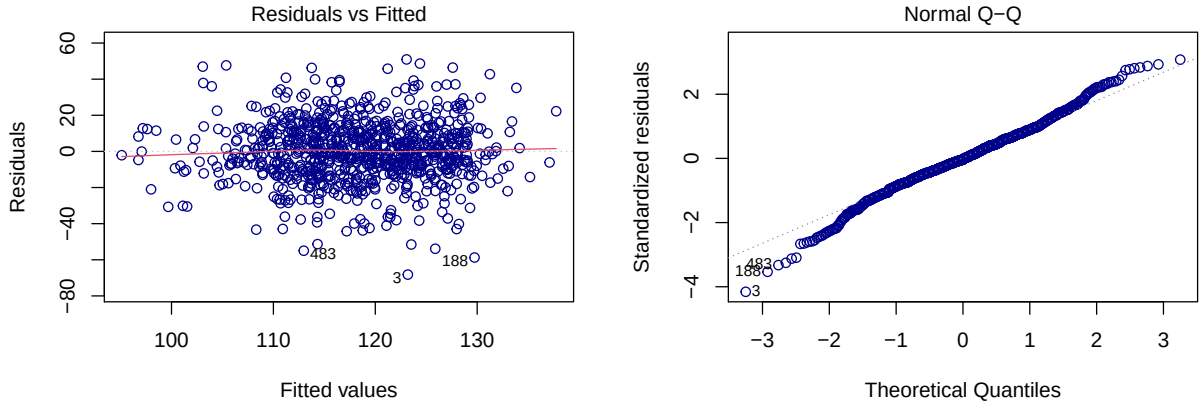
Table 8: Fitting linear model: bwt.oz ~ smoke + parity + mht + mpregwt + mrace + smoke:mrace

Observations	Residual Std. Error	R^2	Adjusted R^2
869	16.69	0.1572	0.1454

The baseline values taken in the intercept is smoke=0 and race=0 (White). Keeping all predictors at 0, the birth weight of the newborn would be 46.39 ounces (does not make any sense as centering is not done). The main statistically significant variables in this model are smoke, mpregwt, race, mht, and parity as the p values are significantly less than 0.05 (Reject the Null Hypothesis). Keeping the other variables constant,

- If the mother smokes, the birth weight decreases by 9.65 ounces.
- An unit increase of parity increases the birth weight by 0.65 ounces.
- An unit increase in mother's height results in an increase of birth weight by 0.99 ounces.
- An unit increase in mother's pre-pregnancy weight increases the birth weight by 0.11 ounces.
- If the mother is an Asian, the birth weight decreases by 9.688 ounces.

The Adj R^2 value is 0.1454 indicating that it only 14.54% proportion of variation in the response variable is being explained by the regression model.



Model Assumptions

When observing the model residuals against each of the predictor variables, there were no evident patterns found (high degree of randomness) indicating that the model is linear. When observing the QQ plot, most of the points were on the 45 degree line. However, the tails of the line seem to slightly deviate from the 45 degree line. The Normality assumption is true in this model.There are no evident pattern present in the residuals vs fitted values plot (fairly random) and the data points seem to be constant across the x axis, indicating that the equal variance and independence assumptions hold.

There are also no leverage points (points have leverage score below 0.5), outliers (majority of points between 3 and -3 range of the standardized residuals) and influence points (none of the points have a cook's distance above 0.5). The VIF values of each variable is range between 1-5 (moderately correlated). As the value are not above 5, we do not have to worry about multicollinearity in this data.

The answers to the questions from the multiple linear regression model:

- Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?
  - **Yes, mothers who smoke tend to give birth to babies with lower weights**. This question can be answered by the EDA and justified by the regression model. The p value of smoke is 1.65e-15 (statistically significant) and keeping the other variables constant, mothers who smokes tend to decrease the newborn's birth weight by 9.65 ounces.
- What is a likely range for the difference in birth weights for smokers and non-smokers?
  - In a 95% confidence interval, **the birth weight of babies from mothers who smoke tend to be lower by a maximum of 12.28 ounces and a minimum of 7.02 ounces.**
- Is there any evidence that the association between smoking and birth weight differs by mother's race? If so, characterize those differences.
  - The regression model summary indicates race and smoke are important predictors (performed F test and noticed a p value below 0.05). However, as shown below, **the interaction between race and smoking does not seem to be statistically significant (performed F test and noticed a p value 0.2064 larger than 0.05).**

Table 9: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
856	238366	NA	NA	NA	NA
860	240014	-4	-1648	1.479	0.2064

- Are there other interesting associations with birth weight that are worth mentioning?
  - By noticing the regression model output, **along with race and smoke, previous pregnancies (parity), mother's height (mht), and mother's pre pregnancy weight (mpregwt) seem to have a strong statistic association with birth weight**. This is noticed by the small p value of each of the variables in the regression model.

CONCLUSION

The study uses EDA and a multiple linear regression model to understand the relationship between the birth weight of children and the mother's smoking habits, along with identifying other predictors and interactions that are statistically significant to the response variable birth weight (bwt.oz). The conclusions to the inferential questions could be answered using the regression model constructed. The analysis concluded that mothers who smoke tend to give birth to newborns with a lower birth weight. Variables such as mother's race (mrace), height (mht) and pre- pregnancy weight (mpregwt), smoking habits (smoke), and previous pregnancies (parity) have statistical significant associations with birth weight.

Limitations: Noticed that the adjusted R squared value of the fitted model is only 0.1454, indicating that it only 14.54% proportion of variation in the response variable is being explained by the regression model. This is clearly not the best fit and we must use better variables and models to improve the predictive accuracy of the model. The model also depends on understanding the context of inferential question asked. For example, actually removing the interaction term race:smoke gives a lower AIC model but as it is required in context, the interaction cannot be removed.