IDS 702 - Data Analysis Assignment 2

Aarushi Verma

11/09/2021

Question 1 - OLD FAITHFUL

In this exercise, we are analyzing data pertaining to geysers in Yellowstone National Park, Wyoming. Our objective is to fit a regression model to predict the interval between geyser eruptions using the measurements collected which includes, duration and date.

First we look at the summary of the data.

Table 1: Summary Statistics

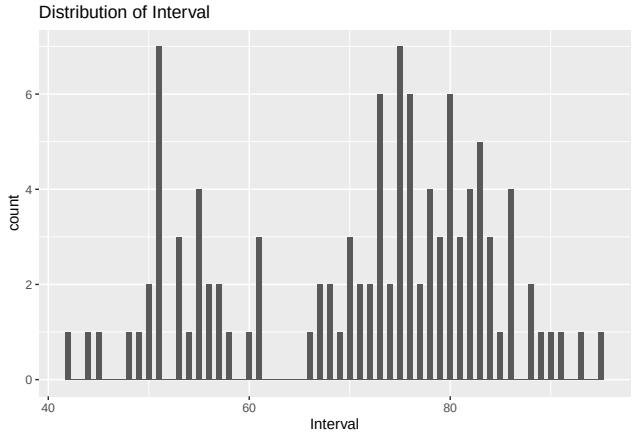
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	107	54.000	31.032	1	27.5	80.5	107
Date	107	4.514	2.275	1	3	6	8
Interval	107	71.000	12.967	42	59	80.5	95
Duration	107	3.461	1.036	1.700	2.300	4.300	4.900

We have 3 variables (X is only the index value) and 107 rows of observations. Table 1 highlights the other key information pertaining to our variables.

We then move on to build a model

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.

Before building a model, we will look at the distribution of our response variable - Interval



By looking at the plot we can see that response variable is not normally distributed. However for our regression assumption we are more concerned with the distribution of the residuals. We can also choose to transform the response variable in the future should we need to.

We fit a simple linear regression model with Interval as the response variable and Duration as predictor.

We center the Duration variable to interpret the intercept better, and then fit a regression equation:

$$Interval \sim \beta_0 + \beta_1 \ Duration(centered) + \epsilon$$

Table 2 (Appendix) gives us the summary of the above regression model. It tells us that on average, an increment in the eruption duration by a minute is associated with a 10.74 minute increase in the interval time for the subsequent duration ($p<0.001$). For an event where the eruption duration is at mean (i.e, 3.46 minutes), the average interval for the next eruption is 71 minutes. According to the adj. R^2, this model explains 73.44% of the variation in interval times.

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

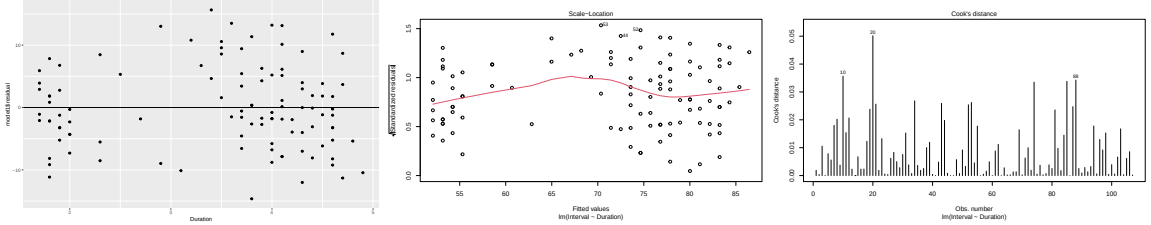
Table 2: CI for Duration

	2.5 %	97.5 %
Duration	9.499	11.983

Table 3 shows that the 95% confidence interval for duration is (9.49,11.98) indicating that all things constant, in 95% of eruptions, with an increment increase in duration, we expect the interval between eruptions to increase by between 9.49 and 11.98 minutes.

3. Describe in a few sentences whether or not you think the regression assumptions are plausible based on residual plots (do not include any plots). Checking model assumptions:
- On examining the normality of the dependent variable, we see an almost bimodal distribution, which could be due to missing data points in the underlying distribution of the Intervals. However in this case, if our residuals follow a normal distribution then there is no violation of our normality assumption. However, we could also consider transforming our response variable, which could decided later in the process. - On examining the residuals of the model against Duration, the points are distributed randomly, but appear to be in clusters. Again, this could be because of missing data therefore, we conclude that the linearity assumption has not been violated.
- Looking at residuals vs fitted plot, the points look randomly scattered in a fixed band around the zero mark. We can conclude that the constant variance and independence assumption are not violated.
- The QQ plot for normality of residuals tells me that most points are on the 45 degree line, although we have outliers. We can conclude that the assumption of normality is not violated. - From the Leverage against Standardized residuals plot and the Cook's distance plot, we can observe see that no points lie beyond Cook's distance of 1. Thus we have no influential points.

All the regression assumptions are met.



4.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).

Fitting regression model:

$$Interval \sim \beta_0 + \beta_1 \ Duration(centered) + \beta_2 \ Days + \epsilon$$

The table (Appendix) gives us the summary of the above regression model. It tells us that on average, an increment in the eruption duration by a minute is associated with a 10.88 minute increase in the interval time for the subsequent duration ($p<0.001$) keeping all other variables constant. According to the adj. R^2, this model explains 71.96% of the variation in interval times. The date variables are not significant.

5.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?

On performing an F-test, we get a p-value that is greater than 0.05. Therefore we conclude that the inclusion of the date variable is not justified as it does not significantly add to the predictive power/ improve the model as a whole.

Table 3: Analysis of Variance Table

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Res.Df	2	101.500	4.950	98	99.8	103.2	105
RSS	2	4,654.587	48.687	4,620.161	4,637.374	4,671.801	4,689.014
Df	1	−7.000		−7.000	−7.000	−7.000	−7.000
Sum of Sq	1	−68.853		−68.853	−68.853	−68.853	−68.853
F	1	0.209		0.209	0.209	0.209	0.209
Pr(>F)	1	0.983		0.983	0.983	0.983	0.983

6.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?

The model with Date as a predictor has an average RSME of 7.27, while that without Date has one of 5.43. Therefore the model which does not include Date as a predictor (with lower RSME values) has a higher predictive power and is more accurate. We can validate that the adj R^2 is higher for this model as well. Hence we conclude that exclusion of Date is justified since it does not improve the predictive power of the model.

Question 2 - MATERNAL SMOKING AND BIRTH WEIGHTS

Summary

This report covers the analysis of data derived from the CHild Health and Development Studies, a comprehensive studies of children born between the years 1960 to 1967 at the Kaiser Foundation Hospital.Our analysis, considers a subset of the original study which covered 15000 families. We removed the observations with missing data (such as information related to the fathers) and our dataset contains 869 observations. We built a linear regression model to study the association between the birthweight of a child and various other parameters - some of which include: Whether the mother smokes, Mother's age, Mother's race, Family income, Mother's height and weight.

Introduction

In this report we perform exploratory data analysis and data processing to determine the relationship of different variables with a child's birth weight. The objective of this study is based on Surgeon General's claim that mother's who smoke have increased rates of premature delivery (before 270 days) as well as lower birth weights among their children. Through this analysis we will check and determine if there is an association between smoking and the child's birth weight. We will also evaluate the other variables such as Mother's age,race, education, income, height and weight etc. We compared the birth weight of children whose mother's smoke to the birth weight of children whose mother's have never smoked.

Data

Data Pre-processing

Before analysis, we performed a few transformations on our data to prepare it for our analysis. The steps taken are as follows:
1. We converted the variables, education, race, income and smoke from integer to factor variables.
2. We collapsed levels for some variables and renamed them based on the labels provided in the data dictionary.

- Race: Levels 0-5 were collapsed to 5 and renamed to 'white'
- Education: Levels 6-7 were collapsed to 7 and renamed to 'trade school'

3. We also dropped the columns - ID, Date and Gestation from our data set. ID and Date do not add any value to our study in this context. Gestation is similar to the birth weight and both can be response variables. In this case we are focusing on birth weight.

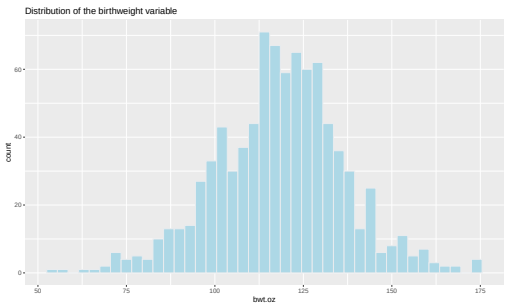
Table 4: Summary Statistics

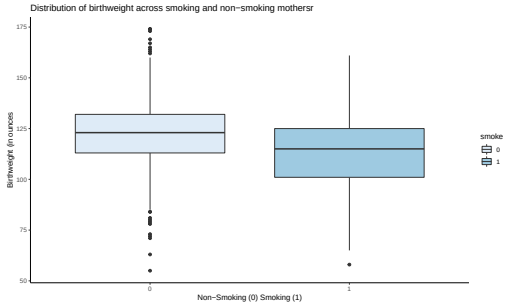
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
bwt.oz	869	118.360	18.051	55	108	129	174
parity	869	1.953	1.882	0	1	3	11
mage	869	27.295	5.708	15	23	31	45
mht	869	64.069	2.534	53	62	66	72
mpregwt	869	128.479	20.778	87	113	140	220

EDA

We perform EDA to understand the underlying relationships between independent and response variables. Firstly we plot our response variable birthweight to see whether it follows a normal distribution or not. From the plot it is apparent that it is normally distributed.

Primarily we are interested in how the smoke variable affects our response variable birthweight. Plotting these variables will help us establish the relationship between these variables.

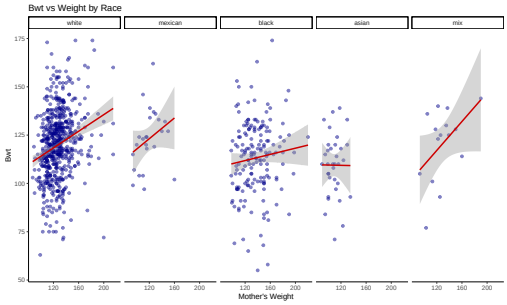


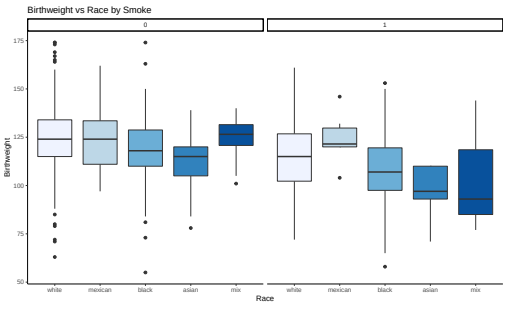


We observe from the boxplot, that there is a difference in medians across the two groups, with a higher spread in birth-weights for babies whose mothers smoke. We may infer from this that mothers who smoke may be giving birth to children with lower birth weight.

We also explore the relationship between our response variable and other predictors to see if there is any significant relationships.We observe that there is linear relationship between birth weight and parity as well as birthweight and mother's height and weight.

We also explore interaction effects and note that theres is an interaction effect on birthweight of race and smoke as well as race and pregweight Race also affects birthweight especially in the smoking population





Model

We begin by constructing a simple regression model with **birth weight** as our response variable, and **smoke, parity, mother's race, mother's age, mother's height, mother's pregnancy weight, mother's education and income** as response variables.

$$Birth.Wt \sim \beta_0 + \beta_1\ smoke + \beta_2\ parity + \beta_3\ mrace + \beta_4\ mage + \beta_5\ mpregwt + \beta_6\ mht + \beta_7\ med + \beta_8\ inc + \epsilon$$

The summary of this table is attached in the appendix. Based on the summary of the base model we can observe the following: 1. The smoke variable is highly significant with p value <0.001. Keeping all other variables constant, on average smoking mothers can give birth to children with weight lower by 9.23 ounces than mothers who do not smoke. 2. Other significant predictors were race (for black and asian mothers), height and pregnancy weight 3. We have an adjusted R^2 value of 0.1386 which means that this model explains 13.86% of variation in our model.

Based on this model, we can remove the variables mage, med and inc as they seem to be insignificant on the basis of their p values. However, looking at this experiment scientifically, these variables are important and would affect a child's birth weight, therefore we can choose to retain them. For the purpose of our experiment, to make out model simpler we will remove these variables. However we will retain the parity variable since its p value (0.0515) is only marginally above the significant value of 0.05.

Based on our EDA we observed some interactions between variables and we will now add those to our model.

$$Birth.Wt \sim \beta_0 + \beta_1(smoke) + \beta_2(parity) + \beta_3(mrace) + \beta_4(mpregwt) +$$
$$\beta_5(mht) + \beta_6(mrace : smoke) + \beta_7(mpregwt : mrace) + \epsilon$$

With this model our Adjusted R^2 improved to 14.88% which means our model now explains 14.88% of the variability in birth weight. The smoke variable is still the most significant with a p-value < 0.001.

We will also perform an F test to see whether our interactions are significant

Based on this F test, we get a p value of 0.77 which means that the interactions are not significant to our model. We will drop the interaction effect of race and weight from our model. However we will retain our interaction between smoke and race since it is vital to our experiment. However, we must acknowledge that by dropping the variables and interaction we may be losing some information about our response variable.

We also look at the VIF to check for multicollinearity between our continuous variables.

For both our models, the vif for all continuous variables is below 3 and therefore we do not need to worry about multicollinearity.

We also use step wise modeling and use AIC and BIC metrics to decide on the best model

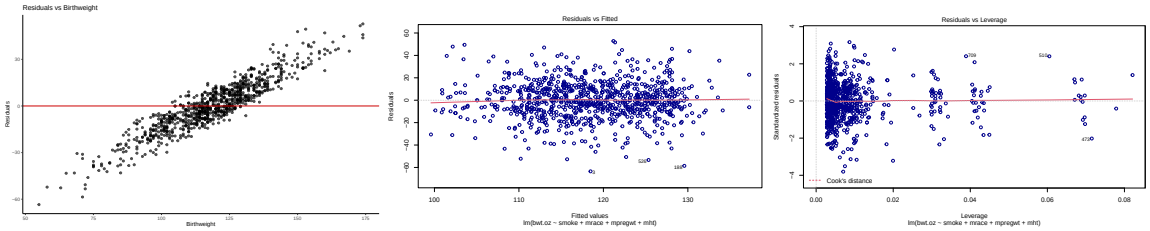
Based on and AIC and BIC neither of the interaction effects significant. Additionally, since BIC lays a higher penalty on the model due the large size of n we have only 4 continuous variables that are significant to our model. Therefore as per BIC our final model is.

Based on our F test as well as AIC and BIC, there is no significance of the interaction effects. However as mentioned previously we will retain our interaction between smoke and race since it is vital to our experiment but drop the remaining variables. Therefore our final model is:

$$Birth.Wt \sim \beta_0 + \beta_1(smoke) + \beta_2(mht) + \beta_3(mrace) + \beta_4(mpregwt) + \beta_5(mrace : smoke)$$

Model Assessment

We assess our model to validate whether it violates any of the assumptions 1. Looking at the residuals vs. fitted plot, we observe that the residuals have a linear trend and the points in the plot do not follow any apparent pattern. Therefore we can conclude that the Linearity assumption is not validated 2. The Q-Q plot shows a fairly linear line excluding the top most and lower most quantiles where the points are slightly away from the 45 degree line. Therefore we can conclude that the normality assumption is met. 3. The residual plot shows apparent pattern hence implying that there is no heteroscedascticity. It is linear and therefore we can conclude that the assumption of Independence and Equal variance is not violated.



Final Model Interpretation

Our final model is :

$$Birth.Wt \sim \beta_0 + \beta_1(smoke) + \beta_2(mht) + \beta_3(mrace) + \beta_4(mpregwt) + \beta_5(mrace : smoke)$$

Table 5: Results

	Dependent variable:
	bwt.oz
smoke1	-9.56^{***}
	(1.34)
mracemexican	0.19
	(3.97)
mraceblack	-8.92^{***}
	(1.99)
mraceasian	-6.30^{*}
	(3.54)
mracemix	0.77
	(4.92)
mpregwt	0.12^{***}
	(0.03)
mht	0.93^{***}
	(0.26)
smoke1:mracemexican	14.56^{*}
	(7.98)
smoke1:mraceblack	1.63
	(2.92)
smoke1:mraceasian	-6.65
	(6.64)
smoke1:mracemix	-12.38
	(10.88)
Constant	49.86^{***}
	(15.39)
Observations	869
R^2	0.15
Adjusted R^2	0.14
Residual Std. Error	16.72 (df = 857)
F Statistic	14.07^{***} (df = 11; 857)
Note:	*p<0.1; **p<0.05; ***p<0.01

The Table gives a summary of this model. According to our model, keeping the mother's height,race and pregnancy weight constant on average, compared to a mother who does not smoke, the child of a smoking mother has a birth weight lower by 9.56 ounces. This result is highly significant with p value <0.01. A baby's birth weight depends on the mother's body characteristics as well. For increase in height by an inch the birthweight increases by 0.93 ounces. Similarly, for an increase in weight by one pound,the baby's birth weight will increase by 0.11 ounces (p<0.01 for both). Compared to mothers of White race, keeping all else constant, Black and Asian mothers give birth to babies that are lighter by 8.92 ounces (p<0.01) and 6.30 ounces (p<0.05) on average. The interaction between smokers and race is not significant. Overall, the model explains 14.21% of the variance in birth weights. The intercept estimate is 49.86 which we can interpret as that the baseline weight of a baby whose mother is not a smoker is 49.86 ounces.

Table 6: CI for Smoke =1

	2.5 %	97.5 %
smoke1	-12.197	-6.931

We also calculated the confidence interval for our most significant variable smoke. With 95% confidence that the true difference of babies mean birth weight across smoke =0 and smoke =1 lies between -12.20 and -6.93 ounces.

Limitations

1. Our model has very low value of adjusted R^2 at 14.21%. This means that our model only explains 14.21% of the variation on our model which is extremely low.
2. The original study includes observations from 15000 families however we have done our analysis only 869 families. This is a very small subset of the population and may not be fully representative of the population.
3. The leverage plot shows that there are some high leverage points. We are yet to investigate those through our model. We can see these values on our leverage plot where some points lie close to the 0.5 cook's distance line. Point 867 has high leverage.

Conclusion

From our model we concluded that Mothers who are smokers give birth to children with lower weight as compared to mothers who do smoke. This is what we inferred from our EDA as well. As per our model this difference is almost 9.56 ounces. Based on our limitations, there are multiple ways of incorporating more variables to improve our model. Currently our dataset does not include the data about a child's father since there were many missing values. Perhaps including that data could show us more insight to how the birth weight is affected. Furhter, we could also incorporate the gestation variable and build a model with 2 response variables. Since our model exaplains very less variability in our response variable, we can further investigate the impact of the categorical variables specifically race and income on the birth weight. This could be part of future potential work for this project.

Appendix

Q1 part Part 1 - Summary of the model

Table 7: Results

	Dependent variable:
	Interval
Duration	10.74***
	(0.63)
Constant	33.83***
	(2.26)
Observations	107
R^2	0.74
Adjusted R^2	0.73
Residual Std. Error	6.68 (df = 105)
F Statistic	294.08*** (df = 1; 105)
Note:	*p<0.1; **p<0.05; ***p<0.01

Q1 part Part 4 - Summary of the model

Table 8: Results

	Dependent variable:
	Interval
Duration	10.88***
	(0.66)
Date2	1.33
	(2.72)
Date3	0.78
	(2.70)
Date4	0.16
	(2.65)
Date5	0.25
	(2.65)
Date6	1.99
	(2.66)
Date7	−0.17
	(2.70)
Date8	−0.69
	(2.70)
Constant	32.88***
	(3.07)
Observations	107
R^2	0.74
Adjusted R^2	0.72
Residual Std. Error	6.87 (df = 98)
F Statistic	35.00*** (df = 8; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01