Data Analysis Homework #2

[redacted]

October 7th, 2022

Question 1: Old Faithful Geyser

Data Description Question 1 evaluates the interval between eruptions of the Old Faithful Geyser as a function of the duration of the previous eruption. 107 Observations were gathered over eight days, consisting of three measurements: the day of measurement (1-8), eruption duration (minutes), and eruption interval (minutes). The interval observations have a mean of 71 minutes, a median of 75 minutes, and a standard deviation of 12.97 minutes. The distribution of interval observations is roughly bi-modal, with a small peak at 50-55 minutes and a second peak, a more normal peak between 70-85 minutes. Duration observations have a similar, bi-modal distribution with a peak between 1.5-2.0 minutes and a second peak between 3.5-4.5 minutes. The mean eruption duration is 3.46 minutes, the median is 3.8 minutes, and the standard deviation is 1.04 minutes.

Model Fit and Assessment Below is a scatter plot of intervals between eruptions as a function of the duration of the previous eruption and a summary of a linear regression model fit with intervals regressed onto duration.

Fig 1. Eruption duration and interval

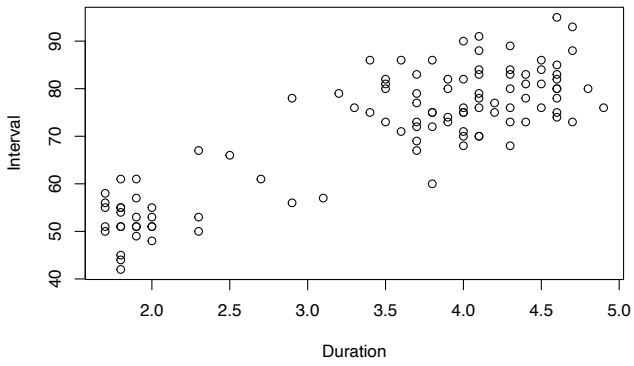


Table 1: SLR Model Regressing Interval onto Duration

	Estimate	SE	t	p-value
Intercept	33.8282	2.2618	14.9562	<.001
Duration	10.741	0.6263	17.1489	<.001

The linear model results indicate that duration is a significant predictor of interval at the $\alpha = 0.05$ significance level ($p < 0.001$). For every minute increase in eruption duration, the interval increases by 10.74 minutes. A 95% confidence interval (Appendix A, Table 4) for duration is (9.499,11.983), meaning we are 95% confident that this interval contains the true value for the slope parameter. Logically, this relationship makes sense. Additionally, the model fit produces an adjusted R^2 value of 0.734, meaning just over 73% of the variation in the interval is explained by duration. Cursory research concludes that a geyser eruption is a release of

sub-surface pressure. Longer eruption times mean that more pressure has escaped, contributing to a more extended 'reset' time before the geyser reaches critical pressure again. In the context of this domain, these results appear to make sense.

Based on the information presented in the residual plots from this regression model, it is reasonable to conclude that the fitted model meets the assumptions reasonably well and can be used with a high degree of confidence.:

- Linearity. As proven by the scatter plot (Fig 1), there is an obvious and discernible linear relationship between the two variables
- Independence. Observing the residual plots, there appears to be no discernible pattern in the data, indicating that the residuals are independent
- Homeoscedasticity. Similar to Independence above, there are no discernable patterns in the graph of the square root of the residuals
- Normality. Based on the theoretical distribution residual plot, the data appears to fit closely to the optimal line. This indicates that the distribution is approximately normal

Model Re-fit with Factored Variable The model was then refitted using a factored variable to represent the day of the observation (1-8). The new variable (day_fac) had no permutations that significantly affected the model's outcome (all p-values were over 0.4, see Appendix A, Table 3). The adjusted R^2 value of the model decreased slightly from 0.734 to 0.720. A visual inspection of a boxplot (Appendix A, Fig 3.) of the interval as a function of day_fac confirms that there is no noticeable effect from the day. This makes sense. The geyser is inanimate and thus unaware of what day of the observation it is erupting on. Introducing the day_fac variable to the model added noise and, therefore, slightly reduced the quality of the overall fit.

K-Fold Cross-Validation Finally, the two models were compared using k-fold cross-validation with ith $k = 10$. The results are as follows (See Appendix):

- *Original model: RMSE of 6.60*
- *Model with day_fac added: RMSE of 7.16*

The model predicting interval as a function of duration is conclusively the better fit for this application. The addition of the day variable had no observable or measurable positive effect on the predictive value of this model.

Question 2: Effects of Smoking During Pregnancy

Summary In this analysis, we attempt to address the significance of several possible predictors for a baby's birth weight. A primary research goal for this analysis is to determine whether mothers who identify as smokers have babies with a lower birth weight than mothers who do not. To accomplish this analysis, multiple linear regression is used. First, a model is fit with nearly all available predictors, then reduced to a final model based on the measured significance of each predictor. In conclusion, this analysis finds that mothers who smoked did indeed have babies with lower birth weights than mothers who did not.

Introduction Today, the general effects of smoking on a person's health are generally well documented. Fifty years ago, however, that was not necessarily the case. In this analysis, we examine data gathered on all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. This data was part of the Child Health and Development Studies, one of the most important studies that sought to address the impacts of a mother's smoking on a baby in utero. The specific research questions that will be addressed in this analysis are:

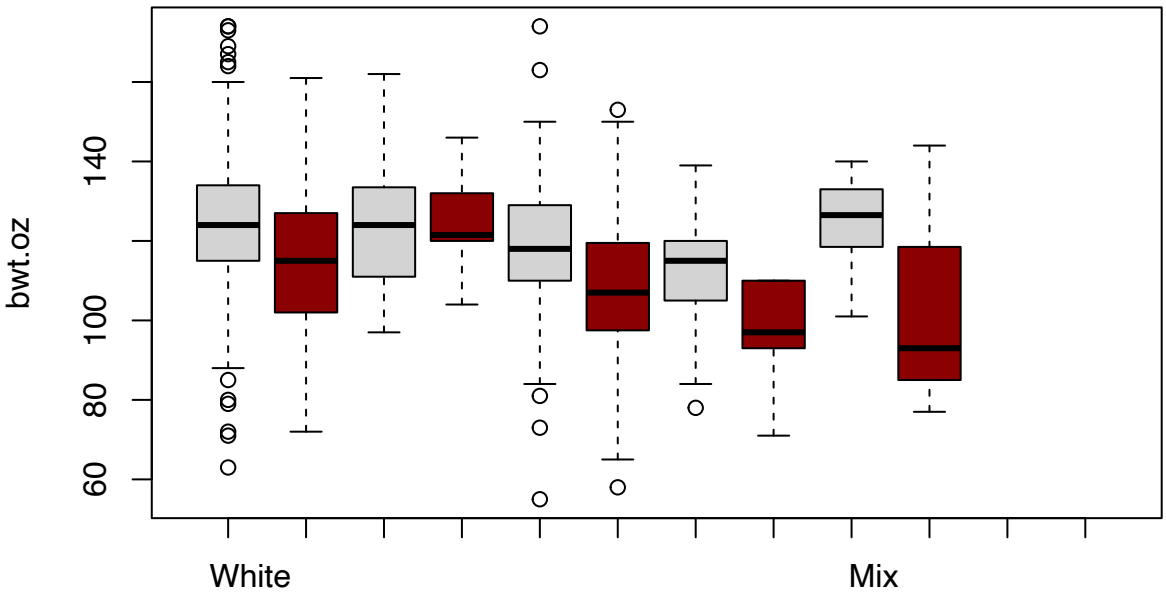
- *Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?*
- *What is a likely range for the difference in birth weights between smokers and non-smokers?*
- *Is there any evidence that the association between smoking and birth weight differs by a mother's race?*
- *Are there other interesting observations worth mentioning?*

Data The dataset used for this analysis consists of 869 observations across 12 variables. One (*bwt.oz* - birth weight (oz)) is the outcome variable for this study. We will drop id, date, and gestation from this analysis to reduce the number of possible predictors to eight. The id variable is a random number assigned to each data set row, and the date is unimportant for our purposes. Gestation is removed because of its obvious correlation to the outcome variable: babies that remain in utero longer will also weigh more. It is worth noting that exploring the impacts of smoking on gestation time could be insightful. The remaining predictors are:

- *parity: a count of a mother's previous pregnancies.*
- *mrace: an integer categorization of a mother's race. Factored in the variable 'race_fac.'*
- *mage: the mother's age.*
- *med: an integer categorization of the mother's education level.*
- *mht: the mother's height, expressed in inches."*
- *mpregwt: the mother's pre-pregnancy weight in pounds.*
- *inc: household income, expressed in increments.*
- *smoke: 0/1 status for whether a mother smoked. Factored into the variable 'smoke_fac.'*

An exploratory data analysis resulted in interesting observations about some variables in this data set. First, we will address the main variables of concern. Crucially, there is a near-even balance between smokers (403) and non-smokers (466). This balance will bode well for our ability to fit a model and make reasonable conclusions from its outcome. The data skew heavily by race, with the significant majority (626 of 869) women being classified as 'white.' The next largest race group was 'black' with 169 mothers. Birth weight (bwt.oz) has a normal distribution when plotted, with a peak spanning the intervals of 120 to 130 ounces. When we examine birth weight, race, and smoking status, the result is Fig. 2. From Fig. 2 we can observe that birth weights appear to be lower for smoking mothers (denoted by red) across all race groups, with the effect being somewhat limited among Mexican mothers in this data set. The limited size of that particular subset (25 total) may explain the observed effects.

Fig. 2: Birth Weight in Ounces



Across the remaining variables, there are other noteworthy observations. 'Parity' shows a noticeable right skew in the distribution. The mean value is 1.95, with a maximum of 11; a histogram supports this observation. Further analysis of this variable found that the 3rd Quartile was three pregnancies, supporting the assumption that many of the mothers in this study have had few pregnancies each. Analyzing the mother's ages (mage) yields a consistent output. The right skew is not as significant here when plotted, but it is still apparent. The average age is 27.29, with a maximum in this set of 45. Several very low values slightly reduce the skew in the data. The minimum age is 15. Together, the distribution of these variables is logical. Younger mothers are likely to have fewer pregnancies, with the count potentially increasing with age.

When examined, the mother's education level ('med') showed a somewhat binomial distribution. The highest observed counts were for mothers who completed high school and mothers who had some college. Including birth weight in plots of this variable yields little insight. There is a minimal logical argument for including this variable in the model, so it will be dropped from further analysis. The remaining variables are height ('mht'), pre-pregnancy weight ('mpregwt'), and household income ('inc'). In height, we find a slight left skew to the data, confirmed by the mean's (64.1) closer proximity to the maximum (72) than the minimum (53). The inverse is true for pre-pregnancy weight. The mean (128.5) is closer to the minimum (87) than the maximum (220). A histogram confirms this observation. A relationship between either variable and birth weight is difficult to determine from simple plotting, so both will be included in the initial model. We will also include income. The relationship with birth weight is difficult to deduce from a simple scatter plot, so the further analysis will help us confirm whether it is valuable.

Model We will use a multiple linear regression for this model because we are analyzing a continuous outcome variable (birth weight) using numerous predictors. An initial model was fit to the complete set of predictor variables. The results can be viewed in Appendix B, Table 5. From this output, we can see that the model produced a R^2 value of 0.145, explaining less than 15% of the variation observed in the model. Age and income were not significant indicators based on the output.

We use backward selection and confirm through a step function that the optimal model for this analysis includes race, mother's height ('mht'), mother's pre-pregnancy weight ('mpregwt'), and 'parity' as the predictors. The output of this multiple linear regression is below. It is important to note that the base values

for factored variables (race and smoking status) are white and non-smoking. The data below will indicate a significant effect from a value not in the base group:

Table 2: Final Regression Summary

	Dependent variable:
	bwt.oz
smoke_facTrue	−9.646 (1.339)
	p = 0.000***
as.factor(race_fac)Mexican	0.211 (3.959)
	p = 0.958
as.factor(race_fac)Black	−9.688 (2.025)
	p = 0.00001***
as.factor(race_fac)Asian	−5.905 (3.543)
	p = 0.096*
as.factor(race_fac)Mix	0.487 (4.914)
	p = 0.922
mpregwt	0.108 (0.032)
	p = 0.001***
parity	0.654 (0.316)
	p = 0.039**
mht	0.986 (0.262)
	p = 0.0002***
smoke_facTrue:as.factor(race_fac)Mexican	13.106 (7.993)
	p = 0.102
smoke_facTrue:as.factor(race_fac)Black	1.954 (2.922)
	p = 0.504
smoke_facTrue:as.factor(race_fac)Asian	−7.391 (6.633)
	p = 0.266
smoke_facTrue:as.factor(race_fac)Mix	−12.546 (10.863)
	p = 0.249
Constant	46.394 (15.448)
	p = 0.003***
Observations	869
R^2	0.157
Adjusted R^2	0.145
Residual Std. Error	16.687 (df = 856)
F Statistic	13.303*** (df = 12; 856)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2 above shows the results of the final model. Results of the multiple linear regression indicate that a mother's smoking status is a statistically significant predictor of newborn baby birth weight (in ounces) at the $\alpha = 0.05$ significance level ($p < 0.001$). The result can be interpreted as mothers who smoke seeing an average decrease in birth weight of 9.64 ounces. A 95% confidence interval for 'smoke' is (-12.28, -7.02), meaning we are 95% confident that this interval contains the true value for the slope parameter (Appendix B, Table 6). Other predictors returned significant p-values in this model: the mother's pre-pregnancy weight, height, and the number of previous pregnancies are all significant in this context. There is a significant interaction between smoking status and race for Black women, and the low, marginally significant p-value for Asian women could be worth noting for further analysis on a larger data set. While the interaction between smoking status and race does not appear to be a significant factor, its presence does improve model fit versus when the two variables are disconnected. The model accounted for approximately 14.5% of the variance in birth weight: $R^2 = 0.157$ and Adjusted $R^2 = 0.145$.

Finally, we can analyze the information presented in the residual plots from this regression model to determine whether the model meets the assumptions. In this instance, it appears to pass the test:

- Linearity. With multiple linear regression it is more difficult to determine the linearity of the entire model. However, from Fig. 2 there is a discernible pattern to the impact of smoking for a portion of the model, so extrapolating that assumption across the entire model is acceptable.
- Independence. Observing the residual plots, there appears to be no discernible pattern in the data, indicating that the residuals are independent
- Homeoscedasticity. Similar to Independence above, there are no discernible patterns in the graph of the square root of the residuals
- Normality. Based on the theoretical distribution residual plot, the data appears to fit closely to the optimal line. This indicates that the distribution is approximately normal

From this model, we can return to the research questions:

- *Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?* ***Yes.***
- *What is a likely range for the difference in birth weights between smokers and non-smokers?* ***Mothers who smoke will have a plausible range for birth weight between roughly 7 and 12 ounces less than mothers who do not smoke.***
- *Is there any evidence that the association between smoking and birth weight differs by a mother's race?* ***Black mothers who smoke have the most significant effect observed in the analysis.***
- *Are there other interesting observations worth mentioning?* ***The significant effect for Black mothers and notable effect for Asian mothers is surprising and could be worth further study. The effects of smoking being more pronounced by race are intriguing, to say the least.***

Conclusion This analysis confirms the now well-established scientific position that the myriad adverse effects of smoking extend to babies in utero. Additionally, there appear to be more significant effects among Black women and a noteworthy effect among Asian women. The small sample size (34 observations) for Asian women may be a factor, thus further analysis is warranted. These conclusions are supported by the plot and table above. Additionally, a mother's height, pre-pregnancy weight, and previous pregnancies appear significant in this context. It is worth pointing out that this analysis is limited, as it only accounts for approximately 14.5% of the observed variance in birth weight. Further studies with a more robust data set that includes more observations from the various non-White racial groups should be conducted to verify this model and determine if further insights can be gained.

Appendix A (Question 1)

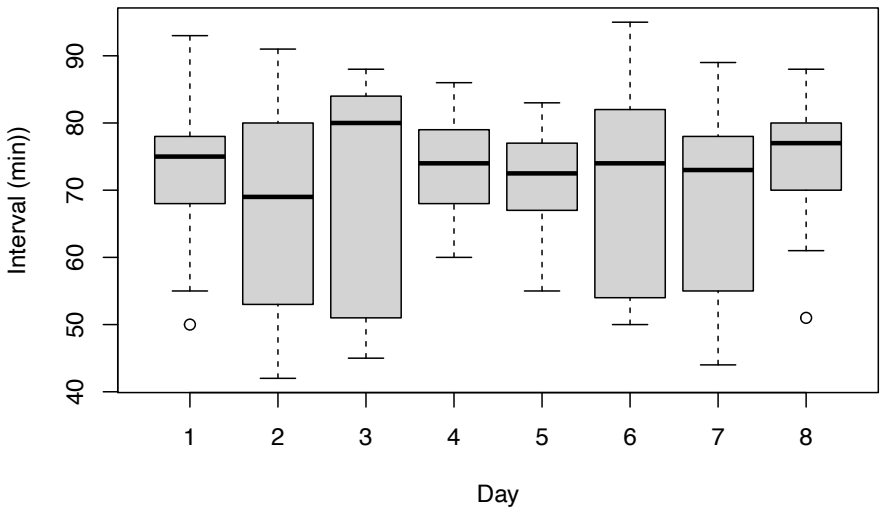
Comparative analysis of the models with and without days factored in.

Table 3: Comparative Regression Summary

	Dependent variable:	
	Interval	
	(1)	(2)
Duration	10.741 (0.626)	10.881 (0.662)
	p = 0.000***	p = 0.000***
day_fac2		1.328 (2.717)
		p = 0.627
day_fac3		0.783 (2.699)
		p = 0.773
day_fac4		0.163 (2.646)
		p = 0.952
day_fac5		0.246 (2.646)
		p = 0.927
day_fac6		1.992 (2.658)
		p = 0.456
day_fac7		−0.170 (2.702)
		p = 0.950
day_fac8		−0.694 (2.696)
		p = 0.798
Constant	33.828 (2.262)	32.877 (3.067)
	p = 0.000***	p = 0.000***
Observations	107	107
R^2	0.737	0.741
Adjusted R^2	0.734	0.720
Residual Std. Error	6.683 (df = 105)	6.866 (df = 98)
F Statistic	294.084*** (df = 1; 105)	35.004*** (df = 8; 98)
Note:		*p<0.1; **p<0.05; ***p<0.01

Boxplots of interval as a function of day.

Fig 3. Intervals as a function of day



Confidence Intervals for Regression Model without days factored.

Table 4: Cofidence Intervals

	2.5 %	97.5 %
(Intercept)	29.343	38.313
Duration	9.499	11.983

Results of the k-fold (k=10) cross-validation on the original model.

Linear Regression

107 samples 1 predictor

No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 96, 97, 96, 96, 97, 96, ... Resampling results:

RMSE Rsquared MAE
6.500372 0.7527839 5.38971

Tuning parameter 'intercept' was held constant at a value of TRUE

Results of the k-fold (k=10) cross-validation on the original model.

Linear Regression

107 samples 2 predictor

No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 96, 97, 95, 95, 97, 97, ... Resampling results:

RMSE Rsquared MAE
7.017037 0.7072188 5.685228

Tuning parameter 'intercept' was held constant at a value of TRUE

Appendix B (Question 2)

Initial Regression Summary

Table 5: Initial Regression Summary

	Dependent variable:
	bwt.oz
smoke_facTrue	−9.706 (1.343)
	p = 0.000***
as.factor(race_fac)Mexican	0.087 (3.962)
	p = 0.983
as.factor(race_fac)Black	−10.123 (2.061)
	p = 0.00001***
as.factor(race_fac)Asian	−5.619 (3.560)
	p = 0.115
as.factor(race_fac)Mix	0.385 (4.918)
	p = 0.938
mage	−0.048 (0.127)
	p = 0.706
mht	1.002 (0.263)
	p = 0.0002***
mpregwt	0.110 (0.032)
	p = 0.001***
inc	−0.243 (0.269)
	p = 0.368
parity	0.755 (0.377)
	p = 0.046**
smoke_facTrue:as.factor(race_fac)Mexican	12.639 (8.006)
	p = 0.115
smoke_facTrue:as.factor(race_fac)Black	2.112 (2.927)
	p = 0.471
smoke_facTrue:as.factor(race_fac)Asian	−7.390 (6.637)
	p = 0.266
smoke_facTrue:as.factor(race_fac)Mix	−12.760 (10.879)
	p = 0.242
Constant	47.342 (15.751)
	p = 0.003***
Observations	869
R^2	0.158
Adjusted R^2	0.145
Residual Std. Error	16.694 (df = 854)
F Statistic	11.488*** (df = 14; 854)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Cofidence Intervals

	2.5 %	97.5 %
(Intercept)	16.073	76.715
smoke_facTrue	-12.275	-7.017
as.factor(race_fac)Mexican	-7.560	7.982
as.factor(race_fac)Black	-13.663	-5.714
as.factor(race_fac)Asian	-12.858	1.048
as.factor(race_fac)Mix	-9.157	10.131
mpregwt	0.045	0.172
parity	0.034	1.275
mht	0.472	1.500
smoke_facTrue:as.factor(race_fac)Mexican	-2.581	28.794
smoke_facTrue:as.factor(race_fac)Black	-3.782	7.689
smoke_facTrue:as.factor(race_fac)Asian	-20.410	5.628
smoke_facTrue:as.factor(race_fac)Mix	-33.866	8.774