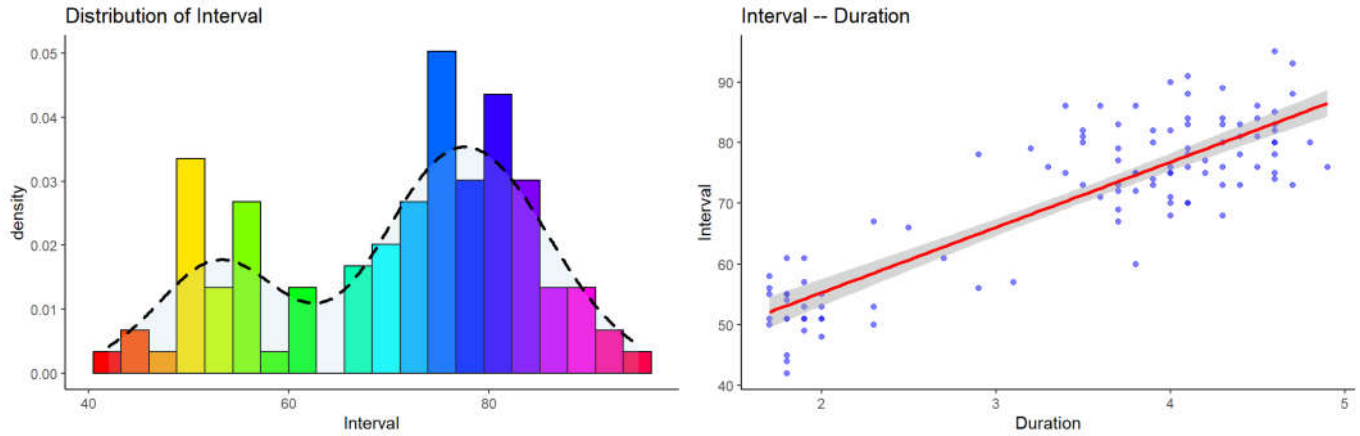
Data Analysis Assignment 2

Yuanjing Zhu

10/05/2022

Question 1: Old Faithful Geyser in Yellowstone National Park

EDA



The histogram of Interval shows that it is not normally distributed, which may cause problem later. From the plot of interval and duration on the right, we can see potential linear relationship between them.

Model Fit

SLR Model Regressing Duration on Interval

Predictor	Estimate	SE	t	p-value
(Intercept)	33.83	2.26	14.96	0
Duration	10.74	0.63	17.15	0

[a] Multiple R-squared: 0.7369, Adjusted R-squared: 0.7344

95% Confidence Interval

Predictor	2.5%	97.5%
Intercept	29.34	38.31
Duration	9.50	11.98

1. p-value is extremely small at the $alpha = 0.05$ significance level, indicating strong evidence that duration is significant in determining interval.
2. The fitted linear regression model can be written as: $Interval = 33.83 + 10.74 * Duration + \epsilon_i$. Interval between eruptions will increase 10.74 minutes as the duration of the previous interval increases by one unit.
3. R^2 value is around 0.74, which means that 74% of the variance for interval between eruptions can be explained by the duration of the previous one
4. The 95% confidence interval of Duration is (9.50, 11.98), which means that we are 95% confidence that the true value of the slop will be in this range.

Model Assessment

1. From the residuals vs fitted plot, we can see two clusters. If this is caused by lack of data in the middle, then linearity holds, but if not, then linear assumption is violated. So we need to check it later for potential violation.
2. The independence assumption is satisfied for this dataset since there is no discernible pattern in the residual plot.
3. The normality assumption holds because most points fall near the 45° line in the qq plot.
4. For constant variance assumption, there are still two clusters on the left and right side of the plot, and the LOESS curve is not a flat line. It may be because of less data in the middle, but we need to check it later for potential violation.

Predicting interval from duration and day

SLR Model Regressing Duration on Interval

Predictor	Estimate	SE	t	p-value
(Intercept)	32.88	3.07	10.72	0.00
Duration	10.88	0.66	16.43	0.00
date_fac2	1.33	2.72	0.49	0.63
date_fac3	0.78	2.70	0.29	0.77
date_fac4	0.16	2.65	0.06	0.95
date_fac5	0.25	2.65	0.09	0.93
date_fac6	1.99	2.66	0.75	0.46
date_fac7	-0.17	2.70	-0.06	0.95
date_fac8	-0.69	2.70	-0.26	0.80

[a] Multiple R-squared: 0.7408, Adjusted R-squared: 0.7196

In this multilinear regression model, the baseline of date is $date = 1$, but date doesn't seem to be significant since p-values for all date variables are larger than 0.4.

k-fold cross validation (with k=10)

RMSE for regressing Interval on Duration

SLR with 10-fold cv	SLR
6.61	6.62

With random seed as 12, RMSE after 10 fold cross validation is 6.61 while the RMSE of the previous linear regression model is 6.62. The model with cross validation has a smaller RMSE thus is slightly more accurate than the previous one, but the difference between the two models is not very not huge.

Question 2: Maternal smoking and birth rate

Summary

In this project, I built a multi-linear regression model to investigate what characteristics will affect baby's birth weight. I started with data pre-processing and exploratory data analysis, where I found that mom smoking will have a negative impact on birth weight and birth weight varies among different race groups, mom's height and pre-pregnancy weight. There are also association between smoking and birth weight differs by mother's race. Then I applied backward model selection using AIC as criterion and performed F tests to evaluate the significance of interaction terms. Our final model aligns with what we inferred from EDA, indicating that smoking mothers tend to give birth to babies with lower weights. The predictors in our final model include smoking, race, height, pre-pregnancy weight and the interaction term of smoking and race.

Introduction

In this data analysis, we used a subset of the dataset from a research study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. The objective of this research is to investigate whether there is an association between mom smoking and birth weight: Do smoking mothers have a tendency to have babies with lower birth weight than non-smoking mothers? What is the likely range of the weight difference between smoking and non-smoking mothers? We are also interested to know whether babies' birth weight is related to other socioeconomic and demographic characteristics of their mother, including race, total number of previous pregnancies, age, education, height, and weight before pregnancy.

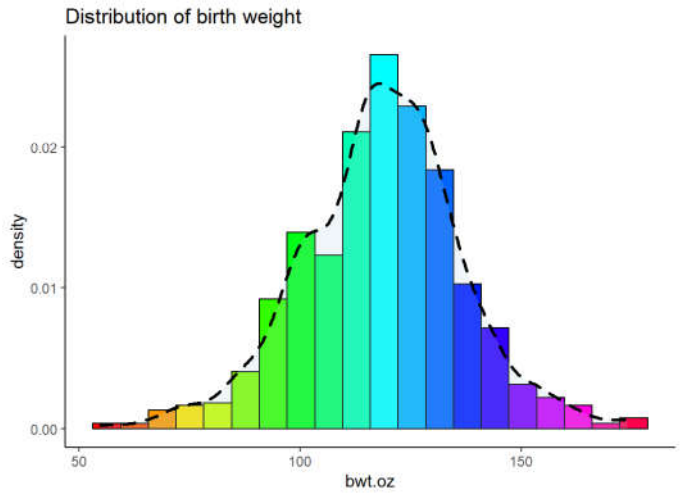
Data

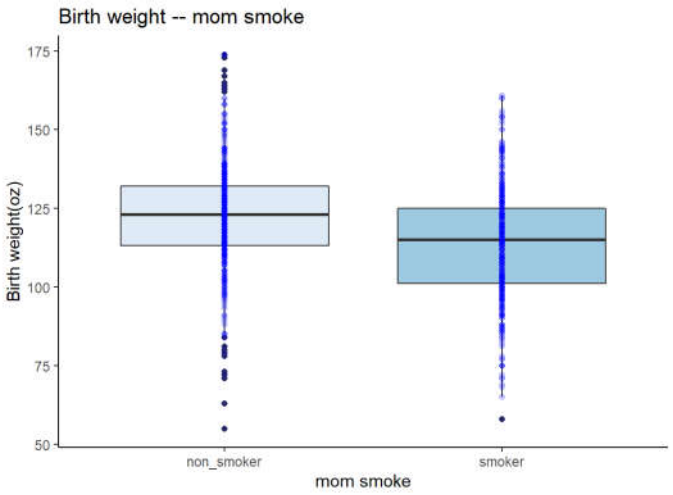
Pre-processing

The data prep-processing contains three steps: check missing values, convert and collapse categorical variables and remove variables that will not be used in later analysis.
1. There are no missing values in this subset of dataset.
2. We consider mother's smoke, race, income and total number of previous pregnancies as categorical variables, so we convert them from integer to factors. Additionally, mrace has originally 11 distinct values, we collapse it into 6 categories, to 0-5: white, 6: mexican, 7: black, 8: asian, 9: mix and 99: unknown.
3. We do not have to use Id and birth in our model, and gestation is another outcome variable like birth weight, so we remove these 3 variables.

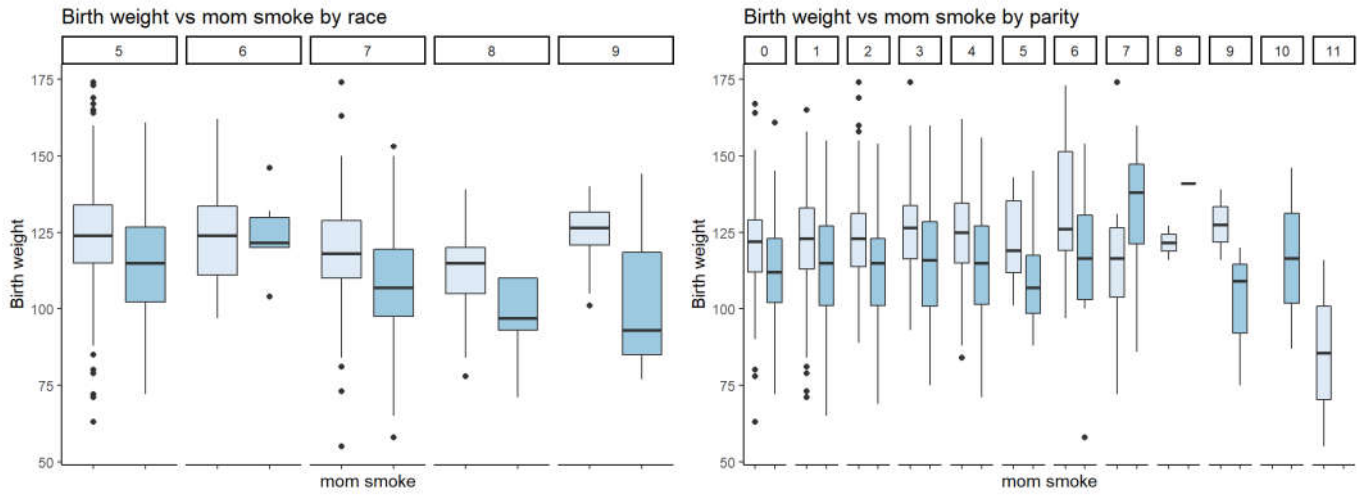
Exploratory Data Analysis

Then we performed exploratory data analysis to better understand data set variables and the relationship among them before making any assumptions. Table 1 is included in the appendix due to the page limit.





From the histogram of the birth weight, we can see that our response variable is normally distributed. Since we were primarily interested in whether smoking is associated with birth weight, we plotted a boxplot of birth weight vs mom smoking. It gives us an intuition that mother smoking tend to have a negative impact on baby's birth weight. From calculation, the average birth weight from non-smoking mother is 122.54 oz while that from smoking mother is 113.53 oz, which confirms the result from the boxplot.
After plotting response variable and each predictor, we found that baby's birth weight is also potentially associated with mom's race, height, pre-pregnancy weight, and total number of previous pregnancies
In order to explore the effect of interaction terms on birth weight, we plotted response variable vs two predictors using *facet_wrap* in ggplot. It turned out there exist two potential significant interaction terms: $smoke * mrace$ and $smoke * parity$.
Since we are primarily interested in the association between smoking and birth weight differs by mother's race, I did some calculations about the average birth weight among different race groups. From table 1, nearly 90% of the interviewees are white and black, and 53.6% of them are smokers. It also turned out that the average birth weight for black and asian mothers is 113.20 oz and 109.44 oz respectively, which is much lower than white mothers(120.00 oz). In addition, compared with smoker and non-smoker within the same race, asian and mix mothers tend to have lighter-weight babies when they are smokers.
For parity, the right figure shows that smoking mothers who have experienced more previous pregnancies are likely to have lower birth weight babies.



Models

Model Selection

Firstly, I started with a baseline MLR model without interaction terms based on the previous EDA. The model can be written as:

$$bwt.oz = \beta_0 + \beta_1 * smoke + \beta_2 * mrace + \beta_3 * parity + \beta_4 * mht + \beta_5 * mpregwt + \epsilon_i;$$

The adjucted R^2 is 0.1528, which means that only about 15.28% of variation in the response variable can be explained by variation in the predictors. P-value for individual predictor indicates that smoke, mrace, mht and mpregwt are highly significant with $p - value < 0.001$ while parity doesn't seem to be significant. The overall p-value based on F-statistic is less than 0.05, indicating that this model is significant.

Then I applied backward selection using AIC as criterion because backward selection is the most poular method and AIC is the more preferred criterion. From the backward selection using AIC, the result aligns with the baseline MLR model. Using 4 predictors: smoke, mrace, mht, mpregwt brings lowest AIC value.

Based on EDA, we also want to explore whether there is relationship between birth weight and interaction terms: $smoke * mrace$ and $smoke * parity$. Here I conducted F test on two pairs of models
Pair1:

$$bwt.oz \sim smoke + mrace + mht + mpregwt$$

$$bwt.oz \sim smoke + mrace + mht + mpregwt + smoke : mrace$$

Pair2:

$$bwt.oz \sim smoke + mrace + mht + mpregwt + parity$$

$$bwt.oz \sim smoke + mrace + mht + mpregwt + parity + smoke : parity$$

The p-value of F test for both pairs are larger than 0.1, indicating the interaction terms are not significant. However, since we consider $smoke : race$ vital prior to modeling using our domain knowledge, we will keep this term while dropping $smoke : parity$

Final Model Interpretations

MLR Model Regressing birth weight

Predictor	Estimate	SE	t	p-value
(Intercept)	49.86	15.39	3.24	0.00
smoke_facsmoker	-9.56	1.34	-7.13	0.00
mrace_fac6	0.19	3.97	0.05	0.96
mrace_fac7	-8.92	1.99	-4.47	0.00
mrace_fac8	-6.30	3.54	-1.78	0.08
mrace_fac9	0.77	4.92	0.16	0.88
mht	0.93	0.26	3.56	0.00
mpregwt	0.12	0.03	3.70	0.00
smoke_facsmoker:mrace_fac6	14.56	7.98	1.83	0.07
smoke_facsmoker:mrace_fac7	1.63	2.92	0.56	0.58
smoke_facsmoker:mrace_fac8	-6.65	6.64	-1.00	0.32
smoke_facsmoker:mrace_fac9	-12.38	10.88	-1.14	0.26

[a] Multiple R-squared: 0.1530
[b] Adjusted R-squared: 0.1421

95% Confidence Interval

	2.5 %	97.5 %
(Intercept)	19.66	80.06
smoke_facsmoker	-12.20	-6.93
mrace_fac6	-7.59	7.98
mrace_fac7	-12.83	-5.01

	2.5 %	97.5 %
mrace_fac8	-13.26	0.65
mrace_fac9	-8.89	10.43
mht	0.42	1.44
mpregwt	0.06	0.18
smoke_facsmoker:mrace_fac6	-1.10	30.22
smoke_facsmoker:mrace_fac7	-4.10	7.37
smoke_facsmoker:mrace_fac8	-19.67	6.38
smoke_facsmoker:mrace_fac9	-33.74	8.98

The final model can be written as:

$$bwt.oz = \beta_0 + \beta_1 * smoke + \beta_2 * mrace + \beta_3 * mht + \beta_4 * mpregwt + \beta_5 * smoke * mrace + \epsilon_i;$$

P-values from the summary table show that smoke, mht, mpregwt, difference between white and black are significant at the significant level of 0.05. The adjusted R^2 is 0.1421, meaning about 14.21% of the variance of birth weight can be explained by this model.
The coefficient of smoke is -9.56, which means that if mother is a smoker, then her baby's birth weight will decrease 9.56oz on average. The baseline for mom's race is 5(white), so compared with white mothers and keeping all other variables constant, average birth weight of mexican mothers' babies will be 0.19oz heavier, black mothers' babies' average birth weight will be 8.92oz lighter, asian mothers will have average 6.3oz lighter weight babies, and mix mothers will give birth to babies who are 0.77oz heavier on average. The coefficient of mht means that for every 1 inch increase in mom's height, her baby's weight tend to increase 0.93oz while the coefficient of mpregwt indicates that for every 1 pound increase in mom's pre-pregnancy weight, her baby's weight would increase 0.12oz. To interpret the interaction term, compared with non-smoking white mothers keeping other variables constant, the average birth weight from smoking mexican mothers would be 14.56oz heavier, from smoking black mothers will be 1.63oz heavier, from smoking asian mothers will be 6.55oz lighter and from smoking mixed-race mothers will be 12.38oz lighter. However, their p-values are larger than 0.05, and their confidence intervals contain 0, indicating they are not statistically significant. But we still include the interaction term in our model due to prior research interest.
The confidence interval of smoking is [-12.20, -6.93], indicating that we are 95% confident that the decrease of birth weight from a smoking mother will fall in this range. We are 95% certain that one unit increase of height will increase 0.42 - 1.44oz birth weight and one unit increase of pre-pregnancy weight will lead to 0.06-0.18oz increase in birth weight. We are also 95% confident that the average weight difference between white and black mother is between -12.83oz and -5.01oz and other factors of race can be interpreted similarly. Since null value is contained within the 95% confidence interval of smoke*mrace, it is not statistically significant.

Model Assessment

From the plot of residuals vs fitted values, there is no discernible pattern and the points are scattered randomly, so linearity and independence of errors assumptions hold. The LOESS curve in residual plot is primarily a flat line around zero, so heteroscedasticity assumption is satisfied. The QQ plot indicates that normality assumption holds since most points are clustering around 45° line. The VIFs of all variables, including the interaction terms between smoking and race are less than 5 and the correlation between mom's height and pre-pregnancy weight is 0.46, so multicollinearity is not a big concern.

Conclusion

Findings: From our final multi-linear regression model, we can conclude that smoking mother will give birth to babies with 9.56oz lower birth weight on average compared to non-smoking mothers. Besides smoking, mom's race, height, pre-pregnancy weight will also affect the baby's birth weight.

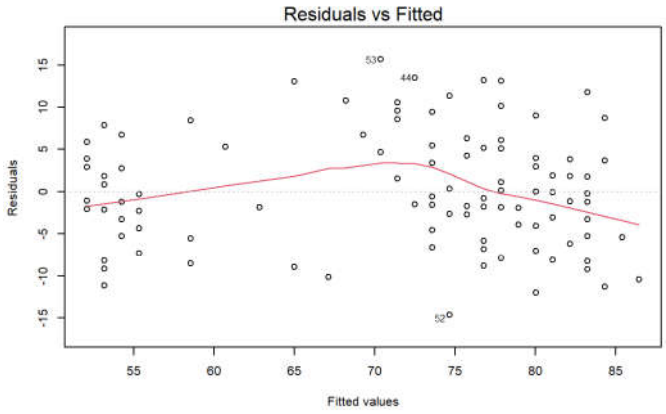
Limitations: Even though our model discovers some significant factors that will affect baby's birth weight, there are still some limitations in our model.

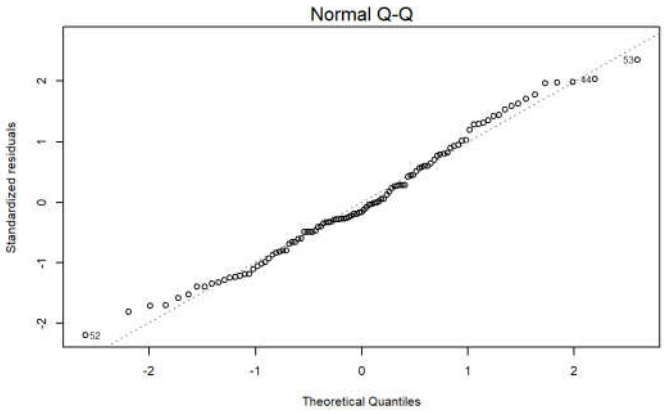
1. The adjusted R^2 value is pretty low, only about 14% of the variance in birth weight can be explained. One potential reason is that we dropped some variables such as the height and weight of the father since these data are missing quite frequently. To improve our model, we should also incorporate some other predictors via survey or interview to better interpret the variation of pre-mature and birth weight.
2. The original dataset has about 15,000 rows while our dataset contains only 869 rows, which is a quite small subset of the original data. The conclusions drawn from this subset may not be able to capture the overall pattern of the entire data. So we could either investigate the original dataset(since it is not very huge) or perform permutation tests.
3. The dataset contains another response variable called gestation. We can build models using gestation as outcome variable to double check our existing model.

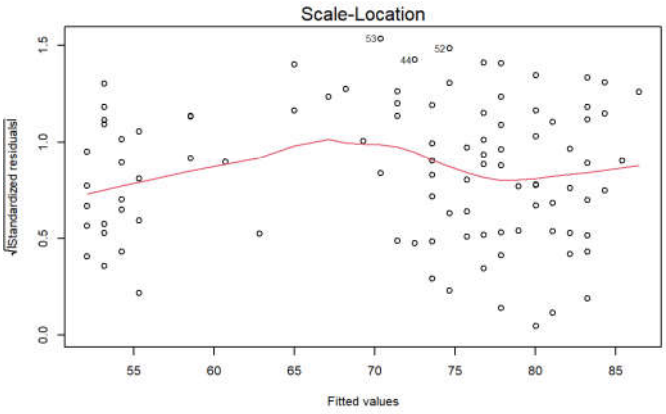
Appendix

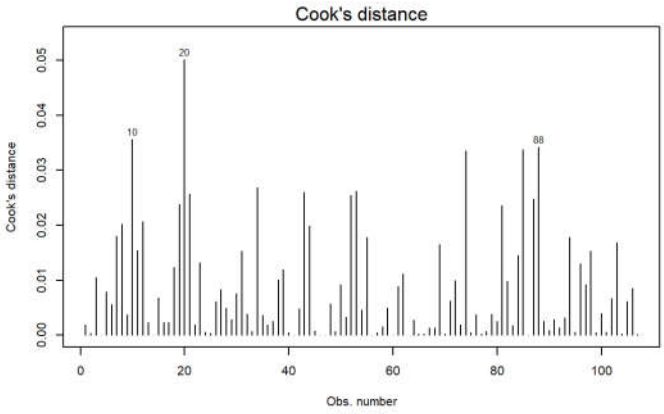
Question 1

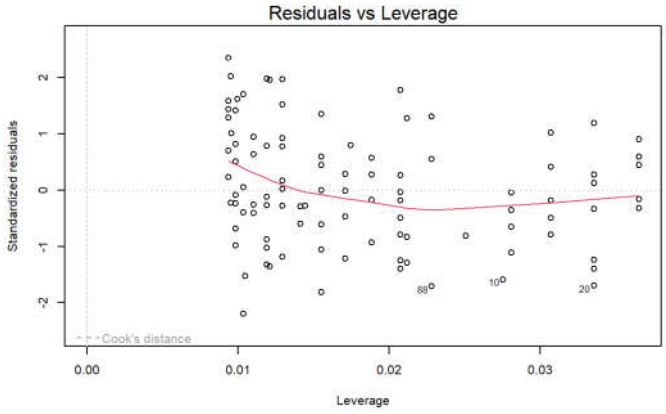
Check assumptions for SLR model regressing interval on duration











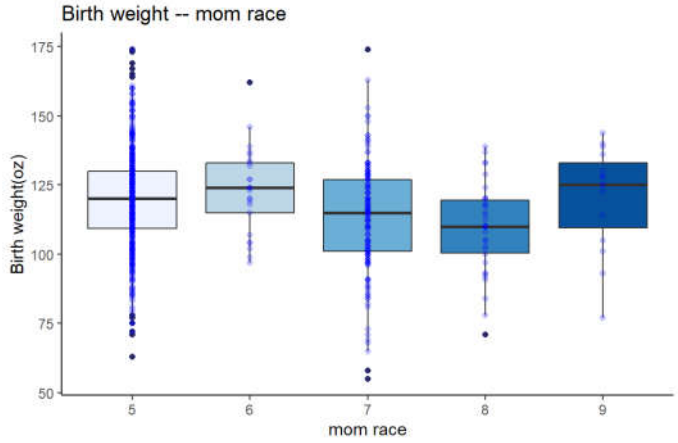
Question 2

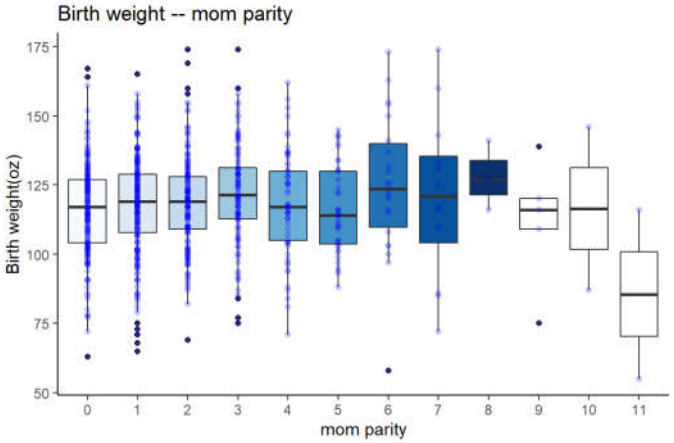
Table 1

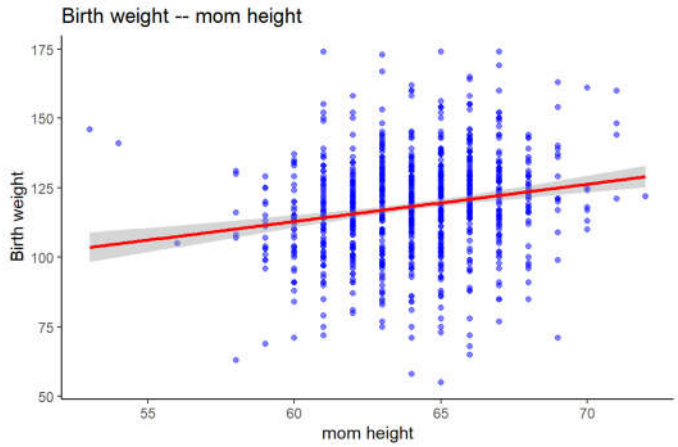
	non_smoker (N=466)	smoker (N=403)	Overall (N=869)
bwt.oz			
N	466	403	869
Q1	113	101	108
Q3	132	125	129
Mean	123	114	118
SD	17.0	18.0	18.1
Median [Min, Max]	123 [55.0, 174]	115 [58.0, 161]	119 [55.0, 174]
mage			
N	466	403	869
Q1	23.0	23.0	23.0
Q3	31.8	30.0	31.0
Mean	27.7	26.9	27.3
SD	5.85	5.51	5.71
Median [Min, Max]	27.0 [17.0, 45.0]	26.0 [15.0, 43.0]	26.0 [15.0, 45.0]
med			
N	466	403	869
Q1	2.00	2.00	2.00
Q3	4.00	4.00	4.00
Mean	3.12	2.72	2.93
SD	1.43	1.41	1.43
Median [Min, Max]	3.00 [0, 7.00]	2.00 [0, 7.00]	2.00 [0, 7.00]
mht			
N	466	403	869
Q1	62.0	63.0	62.0
Q3	66.0	66.0	66.0
Mean	64.0	64.2	64.1
SD	2.47	2.60	2.53
Median [Min, Max]	64.0 [56.0, 71.0]	64.0 [53.0, 72.0]	64.0 [53.0, 72.0]
mpregwt			
N	466	403	869
Q1	115	112	113
Q3	140	138	140
Mean	129	127	128
SD	21.0	20.4	20.8
Median [Min, Max]	126 [90.0, 220]	125 [87.0, 215]	125 [87.0, 220]
mrace_fac			
5	316 (67.8%)	310 (76.9%)	626 (72.0%)
6	19 (4.1%)	6 (1.5%)	25 (2.9%)

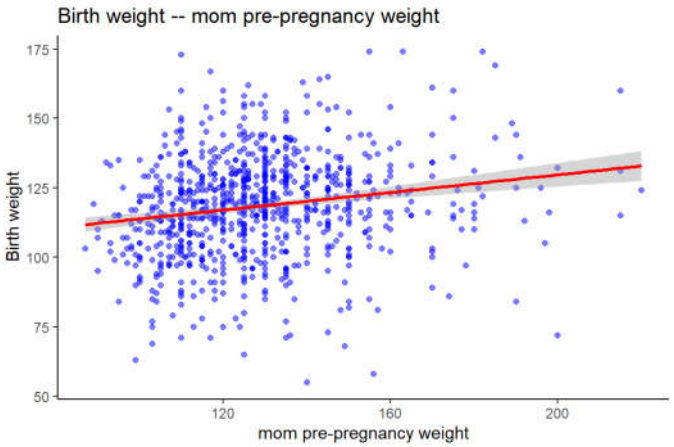
	non_smoker (N=466)	smoker (N=403)	Overall (N=869)
7	94 (20.2%)	75 (18.6%)	169 (19.4%)
8	25 (5.4%)	9 (2.2%)	34 (3.9%)
9	12 (2.6%)	3 (0.7%)	15 (1.7%)
99	0 (0%)	0 (0%)	0 (0%)
inc_fac			
0	16 (3.4%)	10 (2.5%)	26 (3.0%)
1	79 (17.0%)	74 (18.4%)	153 (17.6%)
2	78 (16.7%)	68 (16.9%)	146 (16.8%)
3	82 (17.6%)	54 (13.4%)	136 (15.7%)
4	51 (10.9%)	54 (13.4%)	105 (12.1%)
5	50 (10.7%)	48 (11.9%)	98 (11.3%)
6	28 (6.0%)	29 (7.2%)	57 (6.6%)
7	62 (13.3%)	49 (12.2%)	111 (12.8%)
8	8 (1.7%)	8 (2.0%)	16 (1.8%)
9	12 (2.6%)	9 (2.2%)	21 (2.4%)
parity_fac			
0	117 (25.1%)	92 (22.8%)	209 (24.1%)
1	119 (25.5%)	101 (25.1%)	220 (25.3%)
2	84 (18.0%)	89 (22.1%)	173 (19.9%)
3	70 (15.0%)	50 (12.4%)	120 (13.8%)
4	30 (6.4%)	31 (7.7%)	61 (7.0%)
5	20 (4.3%)	20 (5.0%)	40 (4.6%)
6	12 (2.6%)	10 (2.5%)	22 (2.5%)
7	8 (1.7%)	4 (1.0%)	12 (1.4%)
8	2 (0.4%)	1 (0.2%)	3 (0.3%)
9	2 (0.4%)	3 (0.7%)	5 (0.6%)
10	0 (0%)	2 (0.5%)	2 (0.2%)
11	2 (0.4%)	0 (0%)	2 (0.2%)

Other plots of EDA

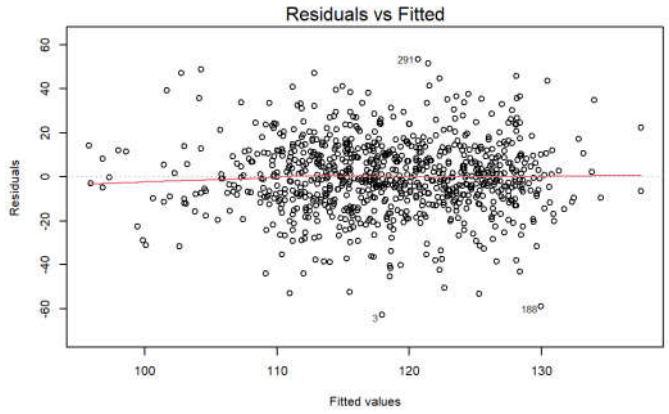


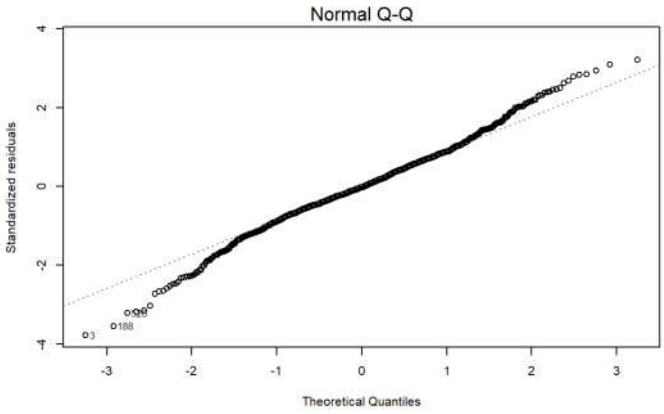


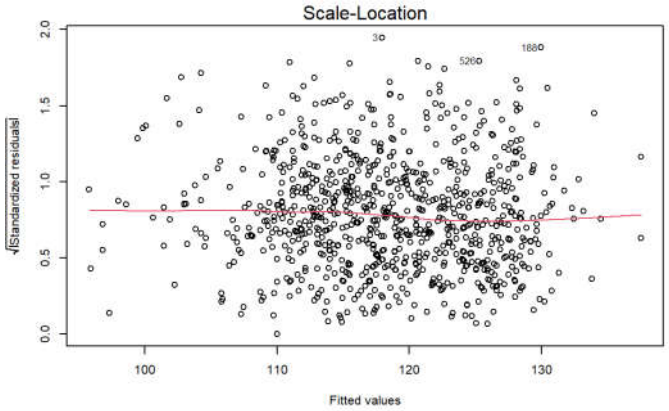


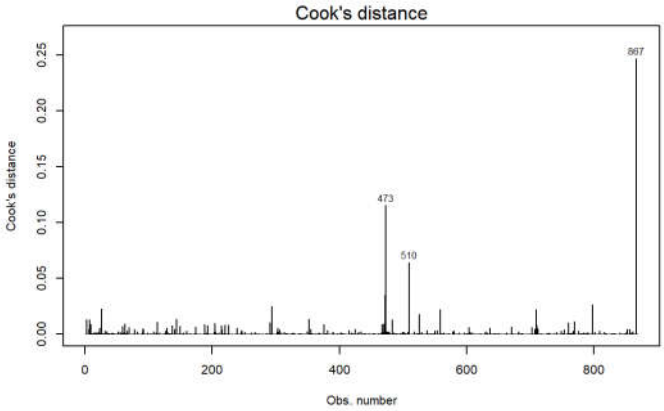


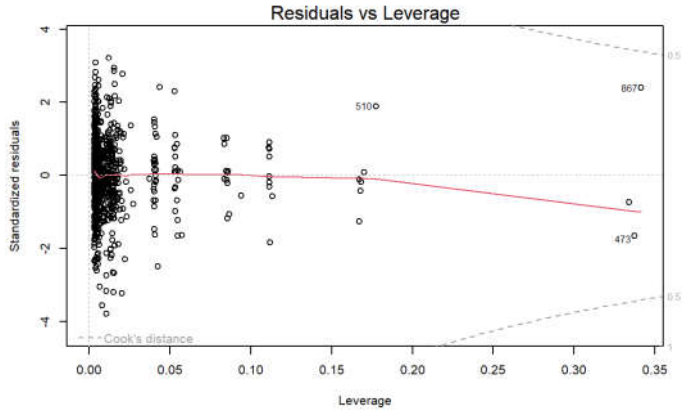
Check assumptions for MLR model regressing birth weight











```
h1, h4 {
  text-align: center;
}

knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyr)
library(dplyr)
library(ggplot2)
library(brew)
library(stargazer)
library(patchwork)
library(corrplot)
library(caret)
library(kableExtra)
library(broom)
library(car)
library(leaps)
library(MASS)
# Data
oldfaithful <- read.csv('OldFaithful.csv', header = 1, stringsAsFactors = TRUE)
dim(oldfaithful)
head(oldfaithful)
summary(oldfaithful)
str(oldfaithful)
# EDA
# Distribution of Interval
gg1 <- ggplot(oldfaithful, aes(x = Interval)) +
  geom_histogram(aes(y=..density..), color = "black", bins = 20, fill = rainbow(20)) +
  geom_density(alpha = 0.2, size = 1, linetype = "dashed", fill = "lightblue") +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Distribution of Interval") +
  theme_classic() + theme(legend.position = "none")

# Interval -- Duration
gg2 <- ggplot(oldfaithful, aes(x = Duration, y = Interval)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method  = "lm", color = "red") +
  theme_classic() +
  labs(title = "Interval -- Duration", x = "Duration", y = "Interval")

gg1 + gg2

# Model Fit
lm_oldf_1 <- lm(Interval~Duration, data = oldfaithful)
summary(lm_oldf_1)
lm_oldf_1 %>%
  tidy() %>%
  # mutate(p.value = c("<.001", "<.001"), term = c("Intercept", "Duration")) %>%
  kable(caption = "<center>SLR Model Regressing Duration on Interval<center>",
        bookstabs = T,
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.7369, Adjusted R-squared: 0.7344")) %>%
  kable_styling(position="center") %>%
  kable_styling(full_width = T)
ci_oldf_i <- confint(lm_oldf_1, level = 0.95)
```

```
ci_oldf_i1_df <- data.frame(ci_oldf_i)
rownames(ci_oldf_i1_df) <- NULL
Predictor <- c("Intercept", "Duration")
cbind(data.frame(Predictor), ci_oldf_i1_df) %>%
  kable( caption="<center>95% Confidence Interval</center>",
         bookstabs = T,
         col.names = c("Predictor", "2.5%", "97.5%"),
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center") %>%
  kable_styling(full_width = T)

plot(lm_oldf_1, which = 1)
plot(lm_oldf_1, which = 2)
plot(lm_oldf_1, which = 3)
plot(lm_oldf_1, which = 4)
plot(lm_oldf_1, which = 5)
plot(1:dim(oldfaithful)[1], lm_oldf_1$residuals, main = "Residuals vs [1:n]", cex.main = 1.2)

oldfaithful$date_fac <- factor(oldfaithful$Date)
lm_oldf_2 <- lm(Interval ~ Duration + date_fac, data = oldfaithful)
summary(lm_oldf_2)
lm_oldf_2 %>%
  tidy() %>%
  # mutate(p.value = c("<.001", "<.001"), term = c("Intercept", "Duration")) %>%
  kable(caption = "<center>SLR Model Regressing Duration on Interval<center>",
        bookstabs = T,
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.7408, Adjusted R-squared: 0.7196")) %>%
  kable_styling(position="center") %>%
  kable_styling(full_width = T)
# set seed to generate reproducible random sampling
set.seed(12)

# define training control and set the value of k = 10
train_oldf <- trainControl(method = "cv", number = 10)

# train the dataset
cv_lm <- train(Interval ~ Duration,
               data = oldfaithful,
               method = "lm",
               trControl = train_oldf)
summary(cv_lm)
print(cv_lm)

# RMSE of the last model
sqrt(mean(lm_oldf_1$residuals^2))

# convert to table
cbind(data.frame(c(6.6099)), data.frame(c(6.6199))) %>%
  kable( caption="<center>RMSE for regressing Interval on Duration<center>",
         bookstabs = T,
         col.names = c("SLR with 10-fold cv", "SLR"),
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center") %>%
```

```
  kable_styling(full_width = T)
smoking_raw <- read.csv("smoking.csv", header = 1, stringsAsFactors = TRUE)
dim(smoking_raw)
head(smoking_raw)
summary(smoking_raw)
str(smoking_raw)
sum(is.na(smoking_raw))
# Convert numeric value to categorical value
## smoke
smoking <- smoking_raw
smoking$smoke_fac <- factor(smoking$smoke,
                            levels = c("0", "1"),
                            labels = c("non_smoker", "smoker"))
## mrace
### (<5], (5,6], (6,7], (7,8], (8,9], (9,99]
smoking$mrace_fac <- cut(smoking$mrace, breaks=c(-Inf,5,6,7,8,9,Inf),
                         labels = c("5", "6", "7", "8", "9", "99"))
## income
smoking$inc_fac <- factor(smoking$inc)

## parity
smoking$parity_fac <- factor(smoking$parity)

# remove 3 variables
smoking <- subset(smoking, select=-c(id, date, gestation))

head(smoking)
summary(smoking)
str(smoking)

# Distribution of birth weight
gg3 <- ggplot(smoking, aes(x = bwt.oz)) +
  geom_histogram(aes(y=..density..), color = "black", bins = 20, fill = rainbow(20)) +
  geom_density(alpha = 0.2, size = 1, linetype = "dashed", fill = "lightblue") +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Distribution of birth weight") +
  theme_classic() + theme(legend.position = "none")

## birth weight -- smoke
gg4 <- ggplot(smoking, aes(x = smoke_fac, y = bwt.oz, fill = smoke_fac)) +
  geom_boxplot() +
  geom_point(alpha = 0.2, color = "blue") +
  scale_fill_brewer(palette = "Blues") +
  #scale_fill_discrete(labels = c("non_smoker", "smoker")) +
  theme_classic() +
  labs(title = "Birth weight -- mom smoke", x = "mom smoke", y = "Birth weight(oz)") +
  theme(legend.position = 'none')

gg3 + gg4

# interaction terms

## smoke_fac -- mrace_fac
gg5 <- ggplot(smoking,aes(x=smoke_fac, y=bwt.oz, fill=smoke_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Birth weight vs mom smoke by race",x="mom smoke",y="Birth weight") +
  theme_classic() + theme(legend.position="none", axis.text.x=element_blank()) +
```

```
  facet_wrap( ~ mrace_fac,ncol=6)

## smoke_fac -- parity
gg6 <- ggplot(smoking,aes(x=smoke_fac, y=bwt.oz, fill=smoke_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Birth weight vs mom smoke by parity",x="mom smoke",y="Birth weight") +
  theme_classic() + theme(legend.position="none", axis.text.x=element_blank()) +
  facet_wrap( ~ parity_fac, ncol=12)

gg5 + gg6
# baseline
mlr_smoking_2 <- lm(bwt.oz ~ smoke_fac + mrace_fac  + mht + mpregwt + parity_fac, data = smoking)
summary(mlr_smoking_2)

# backward selection
mod <- lm(bwt.oz~.,data=smoking)
mod.back <- stepAIC(mod,direction="backward")
summary(mod.back)
# use F-TEST to see whether include interaction term or not

mlr_smoking_no_inter1 <- lm(bwt.oz ~ smoke_fac + mrace_fac  + mht + mpregwt, data = smoking)
mlr_smoking__inter1 <- lm(bwt.oz ~ smoke_fac + mrace_fac + mht + mpregwt + smoke_fac : mrace_fac,
data = smoking)

anova(mlr_smoking_no_inter1, mlr_smoking__inter1)

mlr_smoking_no_inter2 <- lm(bwt.oz ~ smoke_fac + mrace_fac  + mht + mpregwt + parity_fac, data =
 smoking)
mlr_smoking__inter2 <- lm(bwt.oz ~ smoke_fac + mrace_fac + mht + mpregwt + parity_fac + smoke_fac
: parity_fac, data = smoking)

anova(mlr_smoking_no_inter2, mlr_smoking__inter2)

## p-value > 0.19, so the interaction term is not significant
mlr_smoking_final <- lm(bwt.oz ~ smoke_fac + mrace_fac  + mht + mpregwt + smoke_fac:mrace_fac, da
ta = smoking)
summary(mlr_smoking_final)
mlr_smoking_final %>%
  tidy() %>%
  kable(caption = "MLR Model Regressing birth weight",
        col.names = c("Predictor", "Estimate", "SE", "t", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Multiple R-squared: 0.1530","Adjusted R-squared: 0.1421"))%>%
  kable_styling(position="center", full_width = T)
confint(mlr_smoking_final, level = 0.95) %>%
  kable(caption="95% Confidence Interval",
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center", full_width = T)
par(mfrow = c(3,2))
plot(lm_oldf_1, which = 1)
plot(lm_oldf_1, which = 2)
plot(lm_oldf_1, which = 3)
plot(lm_oldf_1, which = 4)
plot(lm_oldf_1, which = 5)
library(table1)
```

```
table1(~ bwt.oz + mage + med+ mht + mpregwt + mrace_fac + inc_fac + parity_fac | smoke_fac, data
 = smoking,
       render.continuous=c(.="N", .="Q1", .="Q3",.="Mean", .="SD", .="Median [Min, Max]"))

# Explore birth weight and each predictor

## Categorical variables
## birth weight -- race
gg7 <- ggplot(smoking, aes(x = mrace_fac, y = bwt.oz, fill = mrace_fac)) +
  geom_boxplot() +
  geom_point(alpha = 0.2, color = "blue") +
  scale_fill_brewer(palette = "Blues") +
 # scale_fill_discrete(labels = c("white","mexican","black","asian","mix","unknown")) +
  theme_classic() +
  labs(title = "Birth weight -- mom race", x = "mom race", y = "Birth weight(oz)") +
  theme(legend.position = 'none')


## birth weight -- parity
gg8 <- ggplot(smoking, aes(x = parity_fac, y = bwt.oz, fill = parity_fac)) +
  geom_boxplot() +
  geom_point(alpha = 0.2, color = "blue") +
  scale_fill_brewer(palette = "Blues") +
  theme_classic() +
  labs(title = "Birth weight -- mom parity", x = "mom parity", y = "Birth weight(oz)") +
  theme(legend.position = 'none')

## numeric variables
gg9 <- ggplot(smoking, aes(x = mht, y = bwt.oz)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method  = "lm", color = "red") +
  theme_classic() +
  labs(title = "Birth weight -- mom height", x = "mom height", y = "Birth weight")

gg10 <- ggplot(smoking, aes(x = mpregwt, y = bwt.oz)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method  = "lm", color = "red") +
  theme_classic() +
  labs(title = "Birth weight -- mom pre-pregnancy weight", x = "mom pre-pregnancy weight", y = "B
irth weight")

gg7 + gg8 + gg9 + gg10
vif(mlr_smoking_final)
# check assumptions
par(mfrow = c(3, 2))
plot(mlr_smoking_final, which = 1)
plot(mlr_smoking_final, which = 2)
plot(mlr_smoking_final, which = 3)
plot(mlr_smoking_final, which = 4)
plot(mlr_smoking_final, which = 5)
#smoking_num <- smoking %>% select(c('mht', 'mpregwt'))
#corrplot(cor(smoking_num), method = "number", type = "upper")
```