Data Analysis Assignment 2

Himangshu Raj Bhantana

9/14/2021

Question 1. Old Faithfull

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.00	0.65	109.90	0.00
Durationc	10.74	0.63	17.15	0.00

- After fitting a regression model to the data for predicting the interval between eruptions based on the duration of the preceding one, the final result was that one minute increase in duration leads to 10.7 duration minutes. The interval will last 71 minutes on average.

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)	69.72	72.28
Durationc	9.50	11.98

- Since the p-value is less to significance level, there is a statistically significant association between the intervals and between eruptions from the duration. Looking at the estimated coefficient there is 10 mins intervals and between eruptions from the duration. one unit increase in duration the change in interval will range from 9.5 to 11.9

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

- I think the regression assumption are plausible based on residual plots because, residuals is normally distributed.

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 Ftest?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Durationc	1.00	13132.99	13132.99	278.57	0.00
factor(Date)	7.00	68.85	9.84	0.21	0.98
Residuals	98.00	4620.16	47.14		

- After performing an F-test, adding a day variable suggests that there is no significant improvement in the 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?
[1] 7.037925 [1] 6.635333

- The model only with duration as predictor has low average RSME of 6.64 comparing to 7.04, hence it is a better predictor.

Question.2 Maternal Smoking and Birth Weights

Summary

This study goes into maternal smoking and birth weights. I examined the data to determine see whether there was a relationship between smoking and birth weight in terms of having a premature delivery and low birth rate. I utilized the Backward selection Bic model to see if there was a significant correlation between smoking and birth weight. It provided me with strong predictors and indicated that there is a considerable relationship between smoking and birth weight.

Introduction

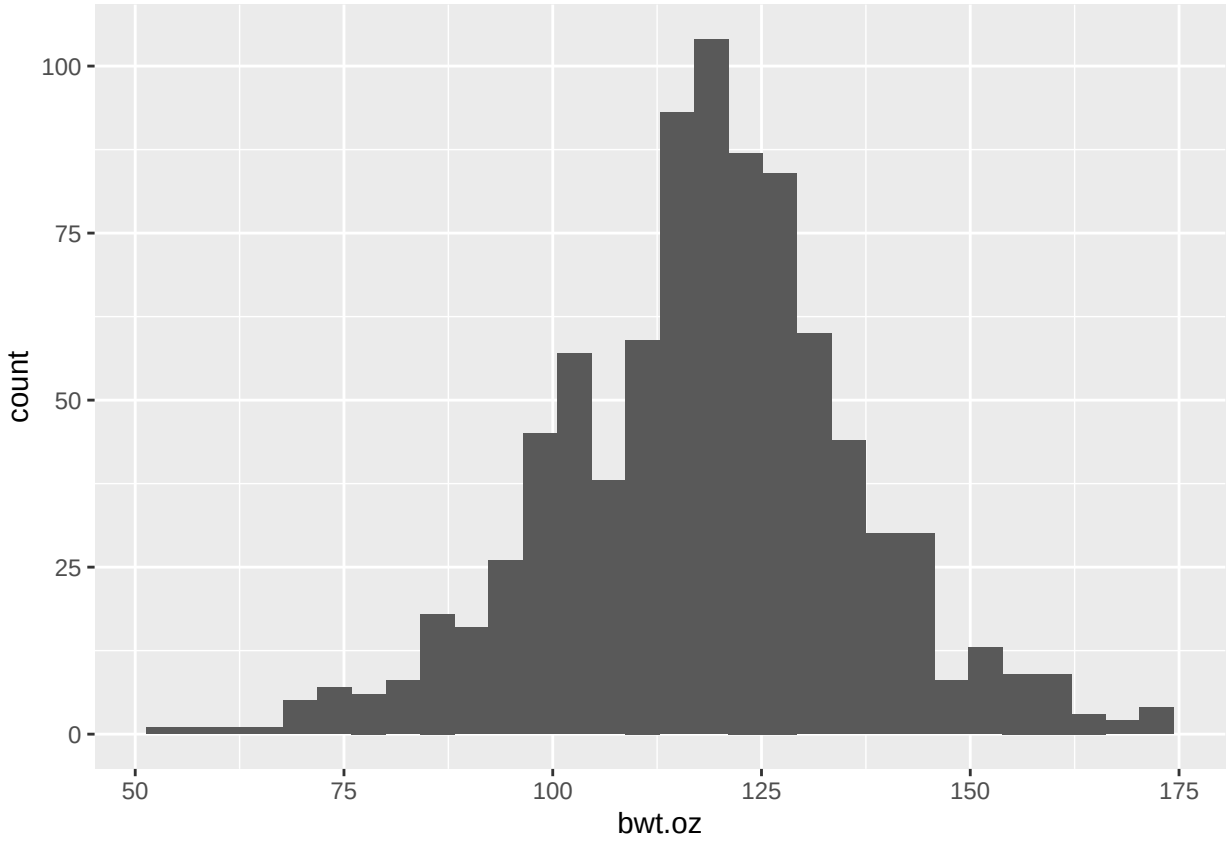
In today's world, it is widely acknowledged that pregnant mothers who smoke risk exposing their infants to a range of health problems. Perhaps not in the early years, because there was no awareness that smoking by pregnant women has a negative impact on their unborn child. So, in this study, we will investigate if a woman who smokes has a negative impact on her unborn kid, as evidenced by the link between smoking and birth weight. The primary goal of this analysis is to determine if:

- Do smoking moms have lower birth weight kids than non-smoking mothers?
- What is the most likely range of birth weight differences between smokers and non-smokers?
- Is there any evidence that the relationship between smoking and birth weight varies depending on the mother's race?
- Are there any more noteworthy birth weight connections worth mentioning?

Data

Before delving into the data analysis for this study, the Smoking Data set, which comprised 869 observations on 12 variables with a mix of categorical and continuous variables, was examined. Because we are primarily interested in smoking and non-smoking, not all factors were utilized in this study, such as gestation, date, id, premature and time. Variables with the most sub levels of data were condensed into one, such as race=white, for our convenience and to keep the data short. For this study our response variable is bwt.oz which is birth weight in ounces.

Starting with exploratory data analysis, the major focus is on the distribution of our response variable, bwt.oz, in order to determine whether or not the bwt.oz distribution is normal, and if not, transformation will occur for our response variable. The following histogram depicts the study of distribution of our response variable, This was really intriguing because the data was mostly in the center, which is difficult to achieve with large datasets at very beginning:



Looking at the histogram distribution of our response variable bwt.oz, we can see that the data is disrupted appropriately and is normal, thus there is no need for modification.

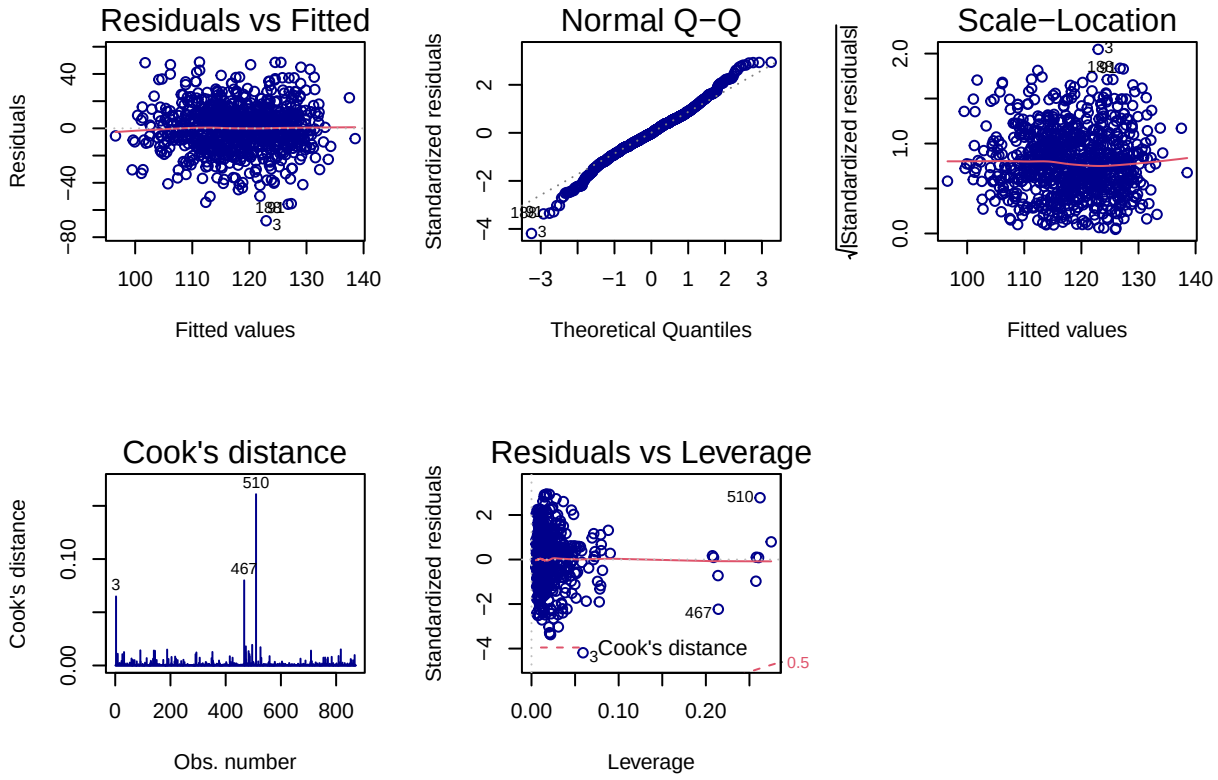
We'll now examine the link between our response variable bwt.oz and each predictor in our data more closely. A scatter plot for continuous variables and a box plot for categorical elements are used to investigate a link. To be clear, numerous factors were omitted from this study. When examining the relationship with the response variable to all of the factors, there was some random distribution of data for the predictors of parity, weight, height, age, and education, but we can also see a straight line emerge at the same time. Since our aim is primarily focused on nonsmokers and smokers, it is better to concentrate on the predictor pertaining to that, which is the interaction between race and smoking, which I will extensively explore in our model section to determine whether or not it has an influence on our study.

Model

Starting with the baseline model using bwt.oz as a response variable, smoke and mrace as major predictors, and additional variables to determine whether they have any bearing on our outcome. Our summary of the model after fitting the baseline model is provided below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.30	17.52	2.53	0.01
parity	0.87	0.40	2.19	0.03
mrace6	4.01	3.50	1.15	0.25
mrace7	-9.16	1.57	-5.84	0.00
mrace8	-7.66	3.10	-2.47	0.01
mrace9	-2.38	4.42	-0.54	0.59
mage	-0.04	0.13	-0.31	0.76
med1	6.18	7.75	0.80	0.43
med2	8.42	7.65	1.10	0.27
med3	6.37	7.95	0.80	0.42
med4	8.72	7.70	1.13	0.26
med5	7.78	7.73	1.01	0.31
med7	-3.91	11.26	-0.35	0.73
mht	0.92	0.27	3.44	0.00
mpregwt	0.11	0.03	3.42	0.00
inc	-0.32	0.27	-1.16	0.25
smoke1	-9.17	1.18	-7.80	0.00

According to the description of this baseline model, it admits that when compared to nonsmokers, women who smoke gives birth to children that weighs 9.17 ounces less than nonsmokers. Similarly, when compared to white moms, black mothers gives birth to children weighing less than 9.16 ounces, which is statistically significant. Looking at the model assumptions of linearity, normality, and equal variance, nothing was violated; it may no longer appear that clean, but it hasn't violated any of the assumptions, for instance we can see it in the plots below:



As a result of the scatter plot's linear pattern, linearity is not broken. Normality is not violated since the

residuals are dispersed in a regular manner. Due to the fact that the experiments are not done in a triangular manner, the equal variance assumption is also not broken. Finally, because the observations in the sample are independent of one another, the assumption of independence was not broken in this case as well. Because there was no violation of assumption, no transformation was performed on this baseline model.

However, we cannot reach a conclusion by assessing the initial baseline. We will begin by developing a second regression model in which we will incorporate the interaction of smoke and race because these are categorical and important aspects of our study; we will then determine whether or not the interaction plays a major role in our analysis. We also want to center the numerical predictors in order to avoid multicollinearity. In addition, we will use the Backward Selection Bic model to get important predictors and provide the best explanation since it provides for consistent estimate of the variables for our analysis. Another benefit for using the bic model is that it can assess the effectiveness of the parameterized model in terms of data prediction.

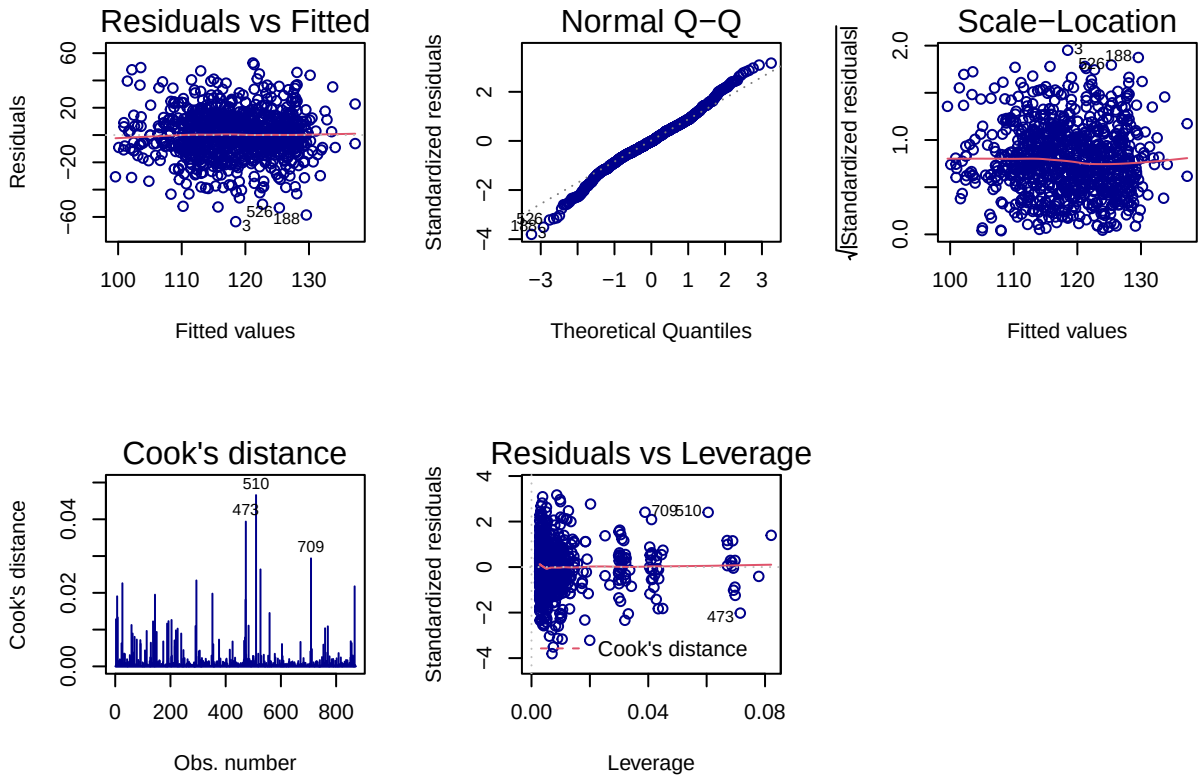
When we call back our backward bic model selection, we obtain the relevant predictors for our study to use as our final model for the study's final conclusion. The model's call back, which reveals the final predictors, is given below, as well as the fact that the interaction variables smoking and race have no meaningful influence because the p value is not significant.: lm(formula = bwt.oz ~ mrace + mht + mpregwt + smoke, data = smoking)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.55	17.71	2.18	0.03
parity	0.87	0.40	2.19	0.03
mrace6	0.68	3.99	0.17	0.87
mrace7	-10.17	2.08	-4.89	0.00
mrace8	-5.52	3.59	-1.54	0.12
mrace9	0.05	4.94	0.01	0.99
mage	-0.04	0.13	-0.31	0.76
med1	8.08	7.82	1.03	0.30
med2	10.71	7.74	1.38	0.17
med3	8.77	8.05	1.09	0.28
med4	10.86	7.78	1.40	0.16
med5	9.92	7.81	1.27	0.20
med7	-0.91	11.37	-0.08	0.94
mht	0.98	0.27	3.64	0.00
mpregwt	0.11	0.03	3.43	0.00
inc	-0.32	0.27	-1.16	0.25
smoke1	-9.55	1.36	-7.01	0.00
mrace6:smoke1	14.34	8.16	1.76	0.08
mrace7:smoke1	2.40	2.94	0.82	0.42
mrace8:smoke1	-7.75	6.64	-1.17	0.24
mrace9:smoke1	-12.44	10.89	-1.14	0.25

As a result of our model selection, our final regression model includes bwt.oz as a response variable and mrace, mht(height), mpregwt(weight), and smoking as predictors. The mathematical regression equation for the final model is:

$$Bwt.oz_i = \beta_1 * mrace_i + \beta_2 * mht_i + \beta_3 * mpregwt_i + \beta_4 * smokei \tag{1}$$

Examining the model assumptions for the final model to see whether there is a violation in normality, linearity, independence or equal variance where the plot reveals no violation and may be ensured by this final model for instance we can see it in the plots below.



Linearity is not violated since the scatter plot follows a linear pattern, indicating that the linear assumption is satisfied. Because the residuals are regularly distributed, normality is not violated. Equal variance is also not violated because the studies are not conducted in a triangular way, implying that the equal variance assumption is satisfied. Finally, because the observations in the sample are independent of one another, the independence assumption was not violated.

We believed we had arrived at an end, but we still need to check to see whether there is any multicollinearity. Checking VIP of our final model for multicollinearity demonstrates that the value for each variable does not pass the value of 10. This ensures that we do not need to be concerned about multicollinearity.

We should also examine whether our final model has any outliers, leverage, or influential points. Despite the presence of outliers, it is reasonable to claim that they are not influential spots. Looking at the cooks distance, no points have crossed the 0.5 threshold, indicating the absence of influential points. When it comes to the leverage point, there is some mention of having points, but it does not appear to be an influential point. As a result, our final model is correct. Summary below is for our final regression model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.25	15.32	3.48	0.00
mrace6	3.63	3.47	1.05	0.30
mrace7	-8.19	1.49	-5.50	0.00
mrace8	-8.15	3.04	-2.68	0.01
mrace9	-1.67	4.39	-0.38	0.70
mht	0.88	0.26	3.37	0.00
mpregwt	0.12	0.03	3.70	0.00
smoke1	-9.27	1.15	-8.04	0.00

Conclusion

Finally, two limitation I discovered is that having a significant correlation does not affect the relationship, as I learned. Also, there are other X variables that we think are unrelated, but it's possible that they aren't.

In line with our primary goal of the analysis, smoking moms had lower birth weight children than non-smoking mothers, according to our final model with relevant variables and its summary. The birth weight of a smoker's kid is anticipated to be 9.3 oz smaller than that of a nonsmoker's child. The significant P value was the primary indicator that the relationship between smoking and birth weight varies depending on the mother's race. When comparing the race of black and Asian women to our base line race, white, it was revealed that the black birth rate of children is 8.2 oz lower. Notable birth weight correlations are that we can see a 0.12 ounce rise in birth weight for every ounce increase in birth weight during pregnancy.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
options(xtable.comment = FALSE )
library(ggplot2)
 library(xtable)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")
odf <- read.csv("Downloads/odf.csv")
library(xtable)
odf$Durationc <-odf$Duration-mean(odf$Duration)
regOld <- lm(Interval~ Durationc, data= odf)
cc<-summary(regOld)
xtable(cc, type='html', title='Duration Summary',header= FALSE, digits=2, no.space = TRUE)
library(xtable)
odf$Durationc <-odf$Duration-mean(odf$Duration)
regOld <- lm(Interval~ Durationc, data= odf)
cd<-confint(regOld)
xtable(cd, type='html', title='95 % Confidence Interval Summary ',header= FALSE, digits=2, no.space = TR
regOldy <- lm(Interval~ Durationc+ factor(Date), data= odf)
go<-anova(regOldy)
xtable(go, type='html', title='Anova F-Test',header= FALSE, digits=2, no.space = TRUE)

#Average RSME-1
set.seed(12) # use whatever number you want
# Now randomly re-shuffle the data
Data <- odf[sample(nrow(odf)),]
# Define the number of folds you want
K <- 10
# Define a matrix to save your results into
RSME <- matrix(0,nrow=K,ncol=1)
# Split the row indexes into k equal parts
kth_fold <- cut(seq(1,nrow(Data)),breaks=K,labels=FALSE)
# Now write the for loop for the k-fold cross validation
for(k in 1:K){
  # Split your data into the training and test datasets
  test_index <- which(kth_fold==k)

  train <- Data[-test_index,]
```

```
  test <- Data[test_index,]
  # Now that you've split the data,
  mod <- lm(Interval ~ factor(Date) + Duration, train)
  predictions <- predict(mod, test)

  RSME[k,] <- sqrt(mean((predictions - (test$Interval))^2))

}
print(mean(RSME))


#Average RSME-2
set.seed(12) # use whatever number you want
# Now randomly re-shuffle the data
Data <- odf[sample(nrow(odf)),]
# Define the number of folds you want
K <- 10
# Define a matrix to save your results into
RSME1 <- matrix(0,nrow=K,ncol=1)
# Split the row indexes into k equal parts
kth_fold <- cut(seq(1,nrow(Data)),breaks=K,labels=FALSE)
# Now write the for loop for the k-fold cross validation
for(k in 1:K){
  # Split your data into the training and test datasets
  test_index <- which(kth_fold==k)

  train <- Data[-test_index,]
  test <- Data[test_index,]
  # Now that you've split the data,
  mod <- lm(Interval ~ Duration, train)
  predictions <- predict(mod, test)

  RSME1[k,] <- sqrt(mean((predictions - (test$Interval))^2))

}
print(mean(RSME1))

library(ggplot2)

smoking <- read.csv("Downloads/smoking.csv")
smoking$mrace<-factor(smoking$mrace)
smoking$med<-factor(smoking$med)
smoking$smoke<-factor(smoking$smoke)
ggplot(smoking,aes(x=bwt.oz)) + geom_histogram()

library(xtable)
remy <- lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc+smoke, data=smoking)
remy1<- summary(remy)
xtable(remy1, type='html', title='Baseline Model Summary',header= FALSE, digits=2, no.space = TRUE)


remy<-lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc+smoke, data=smoking)
attach(mtcars)
```

```
par(mfrow=c(2,3))
plot(remy,which=1:5,col=c("blue4"))

library(xtable)
fullmodel <- lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc +smoke+smoke:mrace, data=smoking)
back<-step(fullmodel, direction = "backward", trace=FALSE,k=log(869))
back$call
j<-summary(fullmodel)
xtable(j, type='html', title='Full Model Summary',header= FALSE, digits=2, no.space = TRUE)

finalmodel<-lm(bwt.oz ~ mrace + mht + mpregwt +smoke, data = smoking)
attach(mtcars)
par(mfrow=c(2,3))
plot(finalmodel,which=1:5,col=c("blue4"))
finalmodel<-lm(bwt.oz ~ mrace + mht + mpregwt +smoke, data = smoking)
b<-summary(finalmodel)
xtable(b, type='html', title='Final model Summary',header= FALSE, digits=2, no.space = TRUE)
library(gridExtra)
library(ggplot2)
library(xtable)

smoking <- read.csv("Downloads/smoking.csv")

#Fiting the model-1
regOld <- lm(Interval~ Durationc, data= odf)
ggplot(regOld,aes(x=Durationc, y=Interval)) + geom_point() + geom_smooth(method="lm")
summary(regOld)
#CEntering

odf$Durationc <-odf$Duration-mean(odf$Duration)


pred<-predict(regOld, interval="prediction");pred


#Confidence Interval
confint(regOld)

#F Test
anova(regOld)

#Analyzing the resudial plots
ggplot(odf,aes(x=Interval, y=regOld$residual)) + geom_point()+geom_smooth(method="lm")
plot(regOld,which=2,col=c("blue4"))

#Fitting the model-2
regOldy <- lm(Interval~ Durationc+ factor(Date), data= odf)
summary(regOldy)


#Confidence Interval
confint(regOldy)
predi<-predict(regOldy, interval="prediction");predi
```

```r
#F Test
anova(regOldy)

#Average RSME-1
set.seed(12) # use whatever number you want
# Now randomly re-shuffle the data
Data <- odf[sample(nrow(odf)),]
# Define the number of folds you want
K <- 10
# Define a matrix to save your results into
RSME <- matrix(0,nrow=K,ncol=1)
# Split the row indexes into k equal parts
kth_fold <- cut(seq(1,nrow(Data)),breaks=K,labels=FALSE)
# Now write the for loop for the k-fold cross validation
for(k in 1:K){
  # Split your data into the training and test datasets
  test_index <- which(kth_fold==k)

  train <- Data[-test_index,]
  test <- Data[test_index,]
  # Now that you've split the data,
  mod <- lm(Interval ~ factor(Date) + Duration, train)
  predictions <- predict(mod, test)

  RSME[k,] <- sqrt(mean((predictions - (test$Interval))^2))

}
print(mean(RSME))


#Average RSME-2
set.seed(12) # use whatever number you want
# Now randomly re-shuffle the data
Data <- odf[sample(nrow(odf)),]
# Define the number of folds you want
K <- 10
# Define a matrix to save your results into
RSME1 <- matrix(0,nrow=K,ncol=1)
# Split the row indexes into k equal parts
kth_fold <- cut(seq(1,nrow(Data)),breaks=K,labels=FALSE)
# Now write the for loop for the k-fold cross validation
for(k in 1:K){
  # Split your data into the training and test datasets
  test_index <- which(kth_fold==k)

  train <- Data[-test_index,]
  test <- Data[test_index,]
  # Now that you've split the data,
  mod <- lm(Interval ~ Duration, train)
  predictions <- predict(mod, test)

  RSME1[k,] <- sqrt(mean((predictions - (test$Interval))^2))
```

```
}
print(mean(RSME1))
#---------------------x--------------------------x--------------------------#
#EDA

ggplot(smoking,aes(x=bwt.oz)) + geom_histogram()

#no transformation needed
#factor=boxplot, continous= scatter

smoking$mrace<-factor(smoking$mrace)
smoking$med<-factor(smoking$med)
smoking$smoke<-factor(smoking$smoke)

#EDA Response Variable VS all th epredictors
ggplot(smoking,aes(x=parity, y=bwt.oz, fill=parity)) +
  geom_point() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Parity vs Bwt.oz",x="Parity",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")


ggplot(smoking,aes(x=mrace, y=bwt.oz, fill=mrace)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mrace vs Bwt.oz",x="Mrace",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")


ggplot(smoking,aes(x=mage, y=bwt.oz, fill=mage)) +
  geom_point() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mage vs Bwt.oz",x="Mage",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")


ggplot(smoking,aes(x=med, y=bwt.oz, fill=med)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Med vs Bwt.oz",x="Med",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")


#Next, `
ggplot(smoking,aes(x=mht, y=bwt.oz, fill=mht)) +
  geom_point() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mht vs Bwt.oz",x="Mht",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")

ggplot(smoking,aes(x=inc, y=bwt.oz, fill=inc)) +
  geom_point() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
```

```
  labs(title="Inc vs Bwt.oz",x="Inc",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")

ggplot(smoking,aes(x=mpregwt, y=bwt.oz, fill=mpregwt)) +
  geom_point() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mpregwt vs Bwt.oz",x="Mpregwt",y="Bwt.oz")  +
  theme_classic() + theme(legend.position="none")

ggplot(smoking,aes(x=smoke, y=bwt.oz, fill=smoke)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Smoke vs Bwt.oz",x="Smoke",y="Bwt.oz") +
  theme_classic() + theme(legend.position="none")

#EDA for interactions race and smoke
ggplot(smoking,aes(x=mrace, y=bwt.oz, fill=mrace)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mrace vs Bwt.oz",x="Mrace",y="Bwt.oz") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ smoke,ncol=4)

ggplot(smoking,aes(x=smoke, y=bwt.oz, fill=smoke)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Smoke vs Bwt.oz",x="Smoke",y="Bwt.oz") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ smoke,ncol=4)

#Ignore Date
remy<-lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc+smoke, data=smoking)
summary(remy)
anova(remy)
plot(remy,which=1:5,col=c("blue4"))
#Ignore Date
#MRACE5 not showing


#Let's mean center the numerical predictors to avoid Mutli


smoking$parity <- smoking$parity - mean(smoking$parity)


smoking$mage <- smoking$mage - mean(smoking$mage)


smoking$mht<- smoking$mht - mean(smoking$mht)


smoking$mpregwt <- smoking$mpregwt - mean(smoking$mpregwt)
```

```r
smoking$inc <- smoking$inc - mean(smoking$inc)

#BIC Model backward selection used
fullmodel <- lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc +smoke+smoke:mrace, data=smoking)
back<-step(fullmodel, direction = "backward", trace=FALSE,k=log(869))
back$call
summary(fullmodel)
anova(fullmodel)

#Final Model
finalmodel<-lm(bwt.oz ~ mrace + mht + mpregwt +smoke, data = smoking)
summary(finalmodel)

#Multicol
vif(finalmodel)

#Checking if there is any violations
plot(finalmodel,which=1:5,col=c("blue4"))
```