Project 2: Data Analysis on a Birth Weight Dataset

Trevor Carpenter, Christina De Cesaris, Michelle Tran
STA 135: Multivariate Data Analysis
Professor Maxime Pouokam
University of California, Davis
June 2, 2021

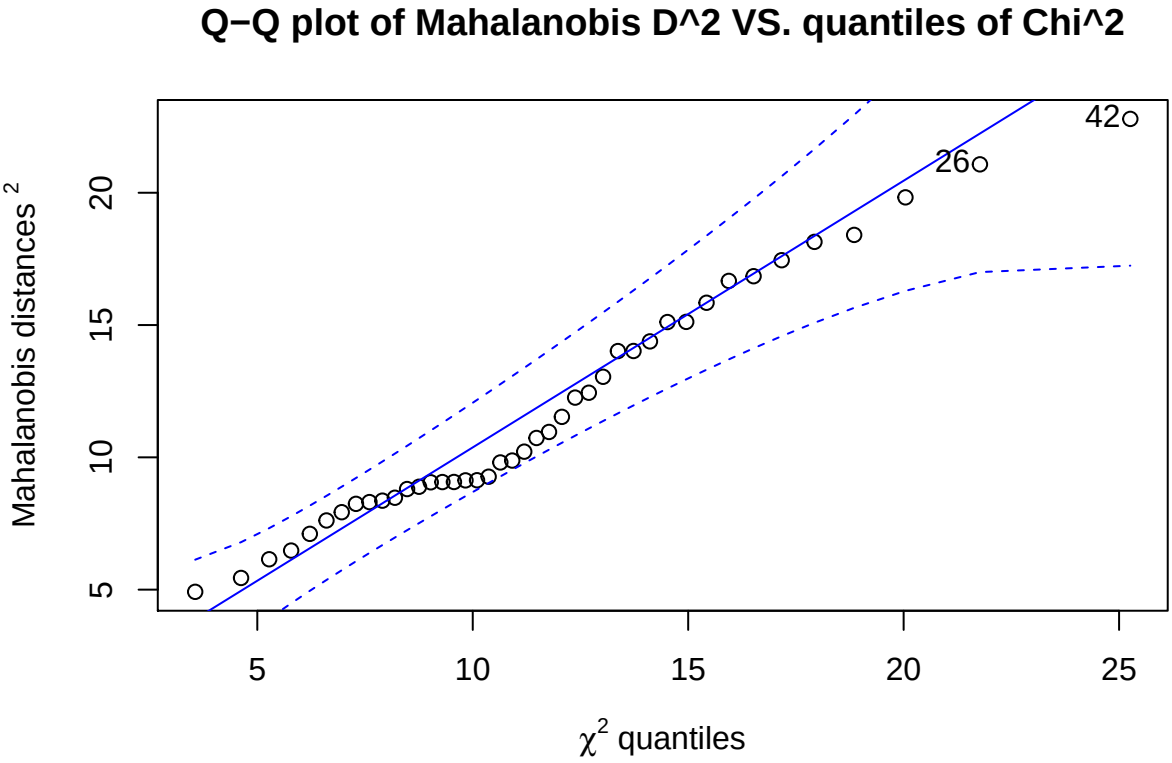
I. Introduction

Although low birth weighted infants can be healthy, under improper care or other circumstances, it can cause serious long-term health problems. A low birth weight is defined as being below 2.5kg. It can be caused by a multitude of factors —poor socioeconomic situation, premature birth, a mother's pre-existing health conditions, and more. In this project, we will highlight how smoking and other factors can cause low birth weights in infants in comparison to non-smoking parents.

II. Data Exploration

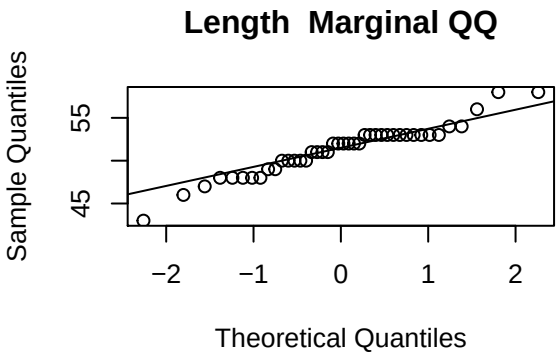
Multivariate Model Assumptions

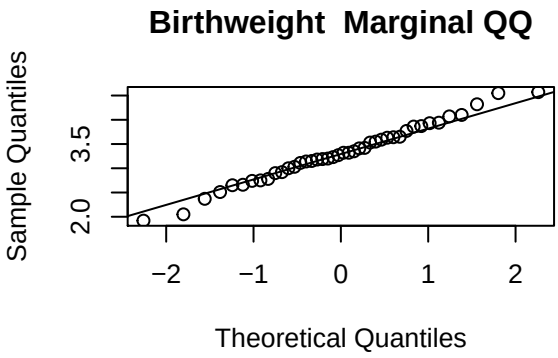
- ASSUMPTIONS

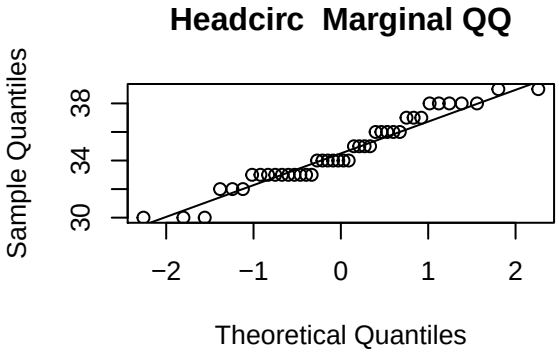


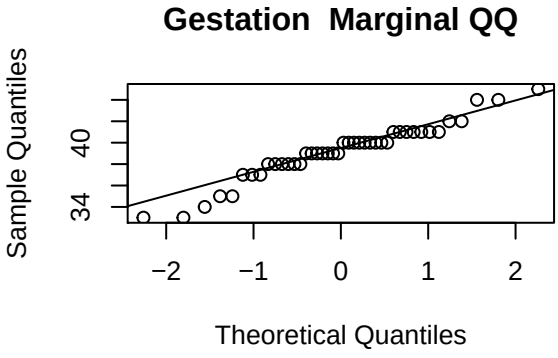
outliers	42	26

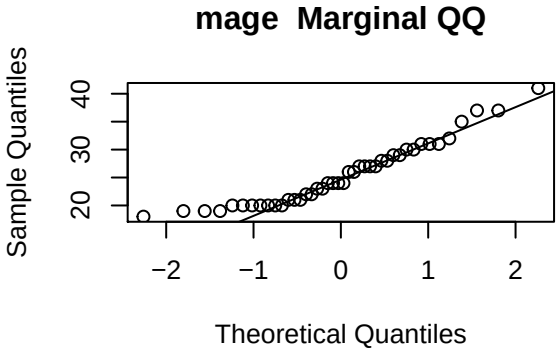
INSERT ANALYSIS HERE

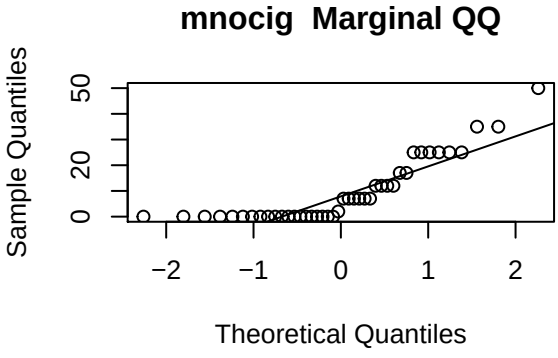

Length Marginal QQ
Sample Quantiles
45
55
−2
−1
0
1
2
Theoretical Quantiles

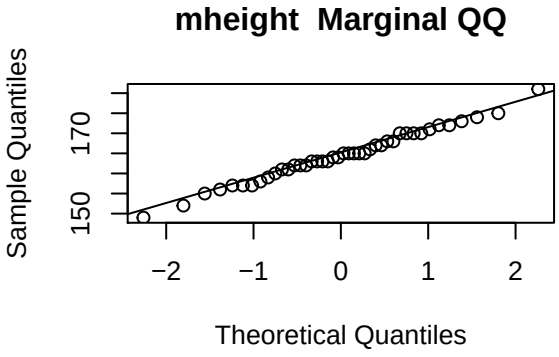

Birthweight Marginal QQ
Sample Quantiles
2.0
3.5
−2
−1
0
1
2
Theoretical Quantiles

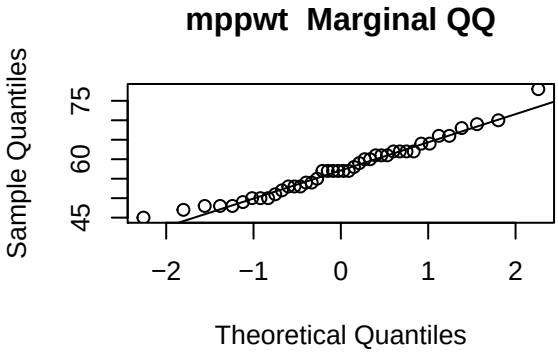

Headcirc Marginal QQ
Sample Quantiles
30
34
38
−2
−1
0
1
2
Theoretical Quantiles

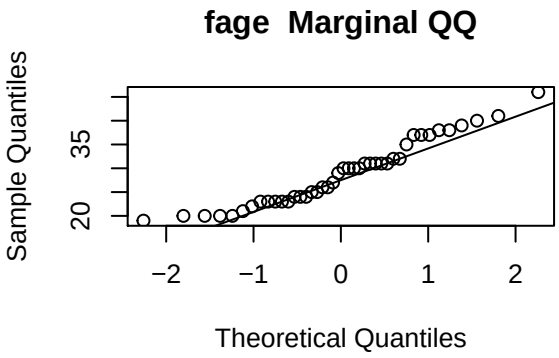

Gestation Marginal QQ
Sample Quantiles
34
40
−2
−1
0
1
2
Theoretical Quantiles

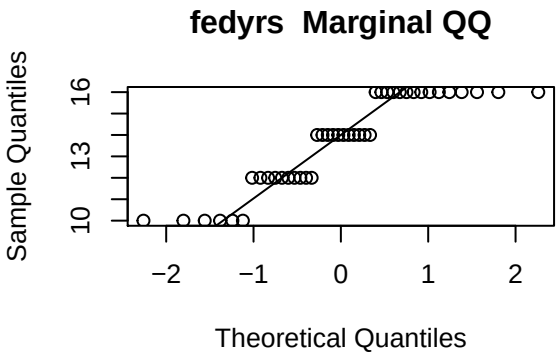

mage Marginal QQ
Sample Quantiles
20
30
40
−2
−1
0
1
2
Theoretical Quantiles

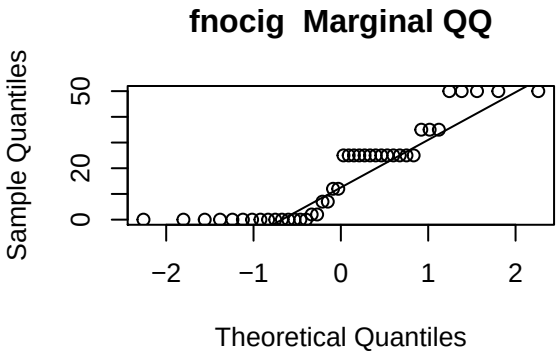

mnocig Marginal QQ
Sample Quantiles
0
20
50
−2
−1
0
1
2
Theoretical Quantiles

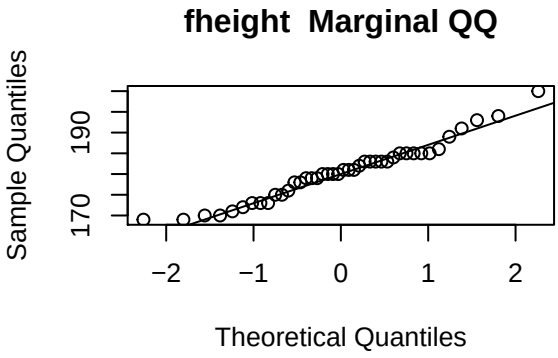

mheight Marginal QQ
Sample Quantiles
150
170
−2
−1
0
1
2
Theoretical Quantiles

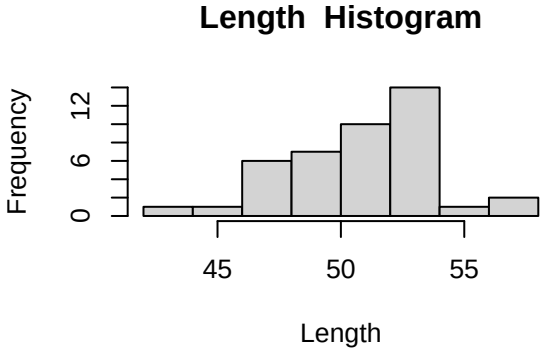

mppwt Marginal QQ
Sample Quantiles
45
60
75
−2
−1
0
1
2
Theoretical Quantiles


fage Marginal QQ
Sample Quantiles
20
35
−2
−1
0
1
2
Theoretical Quantiles

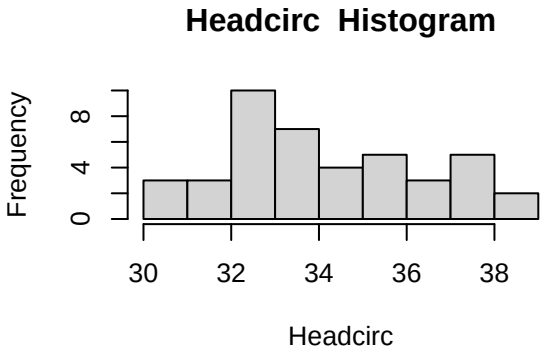

fedyrs Marginal QQ
Sample Quantiles
10
13
16
−2
−1
0
1
2
Theoretical Quantiles

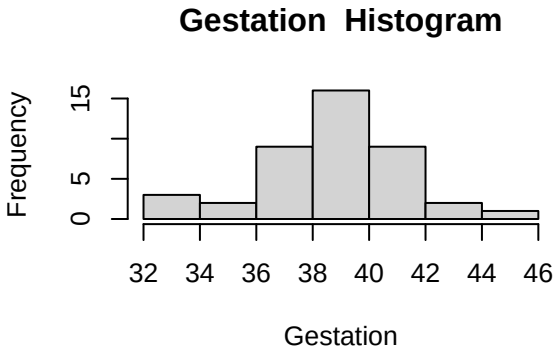

fnocig Marginal QQ
Sample Quantiles
0
20
50
−2
−1
0
1
2
Theoretical Quantiles

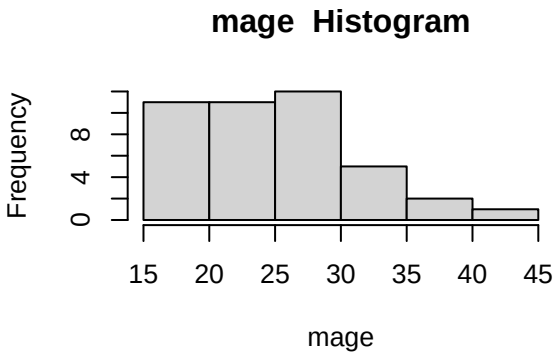

fheight Marginal QQ
Sample Quantiles
170
190
−2
−1
0
1
2
Theoretical Quantiles

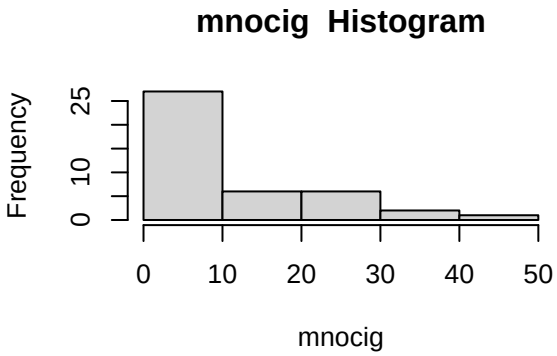

Length Histogram
Frequency
0
6
12
45
50
55
Length

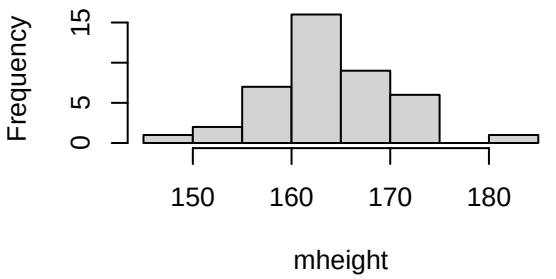

Birthweight Histogram
Frequency
0
6
12
1.5
2.5
3.5
4.5
Birthweight

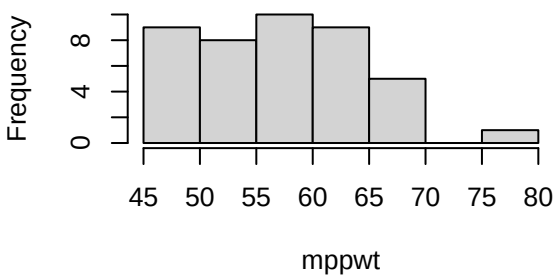

Headcirc Histogram
Frequency
0
4
8
30
32
34
36
38
Headcirc

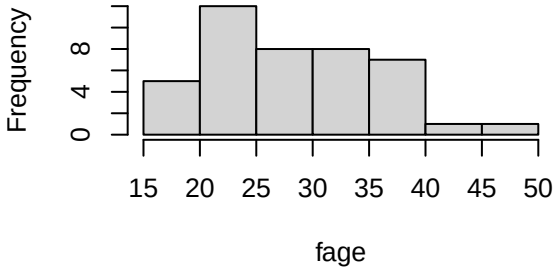

Gestation Histogram
Frequency
0
5
15
32
34
36
38
40
42
44
46
Gestation

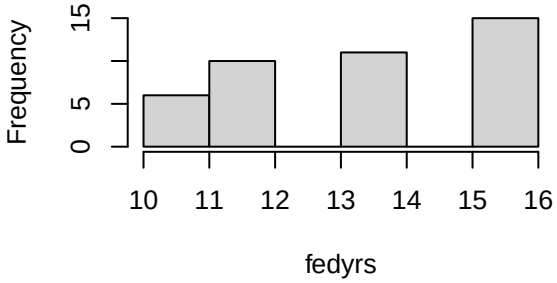


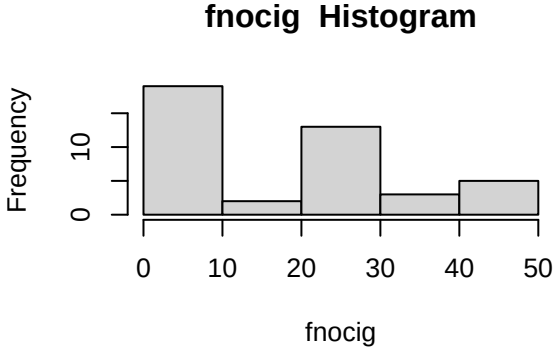


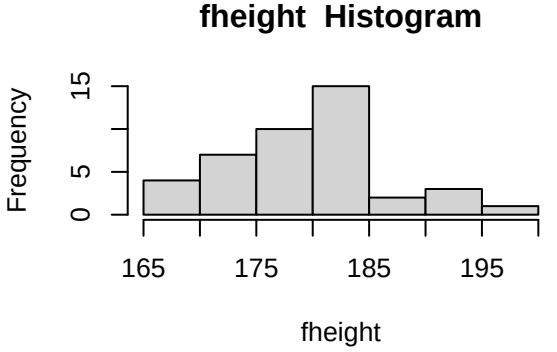












INSERT ANALYSIS HERE

III. Analysis

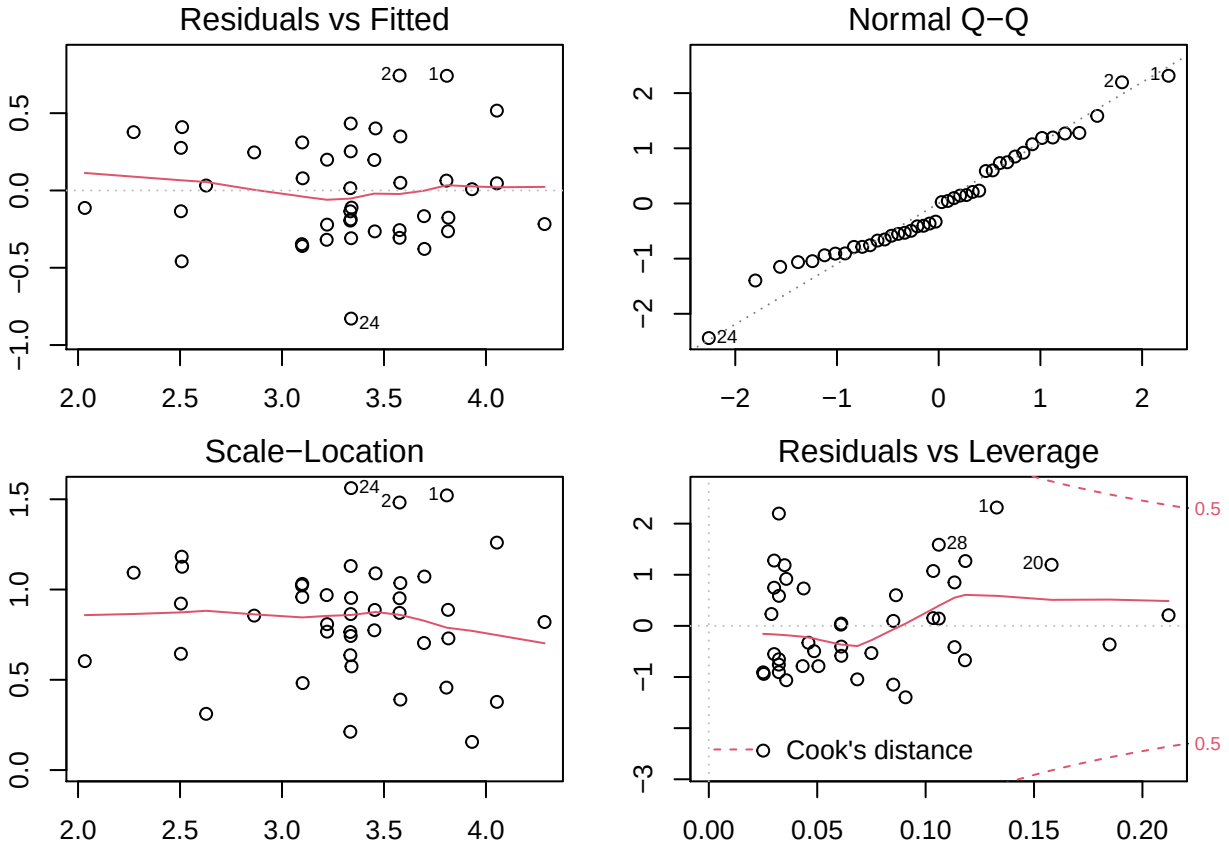
i. Linear Model Fitting

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. In the given data, our response variable is the infant's birth weight and our explanatory variables are factors that supposedly effect birth weight. To get the reduced model, we looked at the full model's coefficients to determine if the predictor variable were significant. If it was significant, we can assume that changes in the predictor variable was significantly associated with changes in birth weight, and therefore its beta coefficient was not 0.

$$\text{Reduced Model: } Y = -5.45\beta_0 + .12\beta_2 + .118\beta_3$$

The reduced model involves the predictor variables: head circumference and gestation period.

After getting the reduced model, we must first check the model assumptions. If we do not check whether the model assumptions are true, it would lead to our model being imprecise due to possible outliers, non-constant variance, and nonlinearity of the outcome.



1) Linearity of Data
   For the Residual vs. Fitted plot, we can assume the linear assumption is reasonable because the line mostly lies on 0 and there are no patterns in the points.

2) Normality of Residuals
   Looking the QQ Plot above, although the model seems a little left skewed, because it is relatively straight, we can assume the reduced model is normal. We can also check normality by also conducting a Shapiro-Wilks test. The result is that the p-value = 0.3, and therefore it fails to reject the normality null hypothesis. The reduced model's errors are normal.

3) Homogeneity of Residual Variance
For the Scale-Location plot, the red line is approximately horizontal and the residuals seems to be randomly scattered around the red line. This means that the spread of the residuals is roughly equal at all fitted values. We can assume that homoscedasticity is likely satisfied for the reduced model.

4) Outliers and High Leverage Points
Looking at the Residuals vs. Leverage plot, it highlights the top three extreme points (#1, #20, and #28). The most obvious outlier is point 20 at about 1.5 standard deviations above/below the mean. These outliers are, however, not influential because they are within the Cook's distance lines.

Overall, the reduced model appears to be adequate because it meets the model assumptions.

To test if our reduced model is correct in assuming that certain predictor variables are not significant, we can use the anova function to compare the full model against the reduced model to perform a partial F test.

$$H_0\text{: } \beta_1 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = 0$$
$$H_A\text{: } \beta_i \neq 0 \text{ for at least one i}$$

Using the partial F test, the F-statistics is F = 1.236 and the p-value = 0.3106. This p-value is not significant at any α value, and therefore we fail to reject the null hypothesis. We can conclude that the other predictor variables do not statistically contribute significantly to an infant's birth weight. The most statistically significant variables in determining an infant's birth weight are: head circumference and gestation period.

ii. LDA & MANOVA

We want to test if the two gropus are significantly different using MANOVA, meaning we want to see if $\mu_{smoker} = \mu_{notsmoker}$ for the two mean vectors. In this case the official test is:

$H_0 : \mu_{smoker} = \mu_{notsmoker}$

$H_a : \mu_{smoker} \neq \mu_{notsmoker}$

The mean vectors for each group are:

smoker	Length	Birthweight	Headcirc	Gestation	mage	mnocig
0	51.80000	3.509500	35.05000	39.45000	24.30000	0
1	50.90909	3.134091	34.18182	38.95455	26.68182	18

mheight	mppwt	fage	fedyrs	fnocig	fheight
164.4500	57.5	27.50000	13.70000	9.7	179.7000
164.4545	57.5	30.18182	13.63636	24.0	181.2273

For the one way MANOVA we can check to see the probability that the smoker variable affects the model. We can calculate 3 different statistics and measure them against their individual F-distributions to test the model.

	Wilks	Pillai	Roy
statistic	0.3500588	0.6499412	1.8566629
approx F	4.4869352	4.4869352	4.4869352
num Df	12.0000000	12.0000000	12.0000000
den Df	29.0000000	29.0000000	29.0000000
Pr(>F)	0.0004592	0.0004592	0.0004592

As seen above, we can reject the null hypothesis for all three tests, thus there is sufficient evidence to indicate a difference among the means for the smoking status.

Because the null hypothesis is rejected, it is of interest to determine what caused the rejection using one way ANOVA. For this we iterate through all of the variables (besides of course our dependent variable "smoker"), and run single factor ANOVA on each. Below are shown the outputs of each one way ANOVA for each.

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Length	1	8.3151515	8.3151515	0.9640247	0.3320766
Birthweight	**1**	**1.4764303**	**1.4764303**	**4.3824556**	**0.0426962**
Headcirc	1	7.8963203	7.8963203	1.3839674	0.2463803
Gestation	1	2.5716450	2.5716450	0.3623253	0.5506145
mage	1	59.4320346	59.4320346	1.8912752	0.1767120
mnocig	**1**	**3394.2857143**	**3394.2857143**	**44.8979592**	**0.0000000**
mheight	1	0.0002165	0.0002165	0.0000050	0.9982284
mppwt	1	0.0000000	0.0000000	0.0000000	1.0000000
fage	1	75.3463203	75.3463203	1.6236045	0.2099434
fedyrs	1	0.0424242	0.0424242	0.0088711	0.9254310
fnocig	**1**	**2142.2761905**	**2142.2761905**	**8.4506270**	**0.0059251**
fheight	1	24.4363636	24.4363636	0.4956506	0.4854961

Based on this, we can see that the mother's and father's respective number of cigarettes are likely the primary mean difference that causes the models for smoker and non-smoker to be different in the MANOVA analysis. We can also see that Birthweight has a low p-value, which fits with earlier analysis suggesting that smoking affects birthweight.

For a new child with the following qualities, we want to determine using LDA if the mother was a smoker or not.

length	birthweight	headcirc	gestation	smoker	motherage	mnocig	mheight	mppwt	fage
61	5.1	36	43	?	43	7	165	64	38

fedyrs	fnocig	fheight
19	45	189

To do this we will calculate with Mahalanobis distance to classify a new observation. We calculate the distance from the centroid of the non-smoker and smoker groups as $D0$ and $D1$ respectively. Whichever value is less is the group that the observation falls closer to. When we do this we see the following:

D0	D1
53.10848	64.4205

Based on the LDA, we would conclude that the mother is NOT a smoker based on $D0 < D1$, implying that the baby would belong in group 0. This is interesting because mnocig $= 7$ which logically would imply the mother is a smoker, however the baby's birthweight is larger than any other birthweights we have seen in our data. Since we saw in single factor ANOVA that smoking has a strong effect on birthweight, the birthweight being so high is likely what placed the baby in the non-smoker group against intuition. This is an interesting example where we see the shortcomings of LDA in prediction power when faced with an outlier.

IV. Conclusion

Based on the Linear Regression model, the two most significant variables in predicting birth weight were head circumference and gestation period and this is proven by our partial F test. The reduced model is valid because we checked the 4 assumptions.

Based on MANOVA and LDA, we see that smoking has a statistically significant relationship with the birthweight, which combined with the Linear Regression Analysis implies that it is likely to have a relationship with the other physical features as well. We also see that the LDA is not very susceptible to outliers, most likely due to the small size of the dataset.

Appendix: R Script

```
knitr::opts_chunk$set(echo = F, warning = F, message = F)
library(knitr)
library(ggplot2)
library(dplyr)
library(mvtnorm)
library(kableExtra)
library(RVAideMemoire)
smok = read.csv("data/Birthweight_reduced_kg_R.csv")
smok = smok[,2:14]
dropped = subset(smok, select = -c(smoker))
x = mqqnorm(dropped,main='Q-Q plot of Mahalanobis D^2 VS. quantiles of Chi^2')
data.frame(outliers = x) %>% t() %>% kable() %>% kable_styling(latex_options = "hold_position")
marginal <-function(df){
  for (i in 1:length(df)){
    qqnorm(df[,i],
         main = paste(names(df[i]),"  Marginal QQ"))
    qqline(df[,i])
  }

}

histogram <-function(df){
  for (i in 1:length(df))
    hist(df[,i],
         xlab = paste(names(df[i])),
         main = paste(names(df[i]),"  Histogram"))


}

par(mfrow=c(2,2))
marginal(dropped)
histogram(dropped)

fmodel = lm(Birthweight ~., data = smok)

#stepwise
#step(fmodel)
#rmodel = lm(Birthweight ~ Headcirc + Gestation + smoker + mppwt,
#data = smok)
#summary(rmodel)

#Under the coefficient method for selecting reduced model, RSE is higher than using the step fcn
summary(fmodel)$coefficient
nmodel = lm(Birthweight ~ Headcirc + Gestation, data = smok)
summary(nmodel)
par(mfrow=c(2,2))
par(mar=c(2, 2, 2, 2))
plot(nmodel)

#normality test
```

```r
ei <- nmodel$residuals
the.SWtest = shapiro.test(ei)

anova(fmodel,nmodel)
save.means<-aggregate(formula = cbind(Length, Birthweight, Headcirc, Gestation, mage, mnocig, mheight, m

kable(save.means[1:7]) %>% kable_styling(latex_options = "hold_position") %>% row_spec(0,bold=TRUE, back
kable(save.means[8:13]) %>% kable_styling(latex_options = "hold_position") %>% row_spec(0,bold=TRUE, ba
save = manova(formula = cbind(Length, Birthweight, Headcirc, Gestation, mage, mnocig, mheight, mppwt, fa
Wilks = summary(save, test = "Wilks")$stats[1,2:6]
Roy = summary(save, test = "Roy")$stats[1,2:6]
Pillai = summary(save, test = "Pillai")$stats[1,2:6]
names(Wilks) = c("statistic", names(Wilks)[2:5])
names(Pillai) = c("statistic", names(Pillai)[2:5])
names(Roy) = c("statistic", names(Roy)[2:5])
data.frame(Wilks, Pillai, Roy) %>% kable() %>% kable_styling(latex_options = "hold_position") %>% row_sp
df = data.frame()
for(i in 1:13){
  if(i != 5) {
    data = data.frame(smok[,i], smok[,5])
    names(data) = c(names(smok[1,])[i], "smoker")
    mod.fit<-aov(formula = data[,1] ~ smoker, data = data)
    s = summary(mod.fit)
    df = rbind(df, data.frame(s[[1]][1,]))
  }
}
row.names(df) = names(smok[1,])[-5]
df %>% kable() %>% kable_styling(latex_options = "hold_position") %>% row_spec(which(df$Pr..F. < 0.05),
kable(matrix(c(61, 5.1, 36, 43,'?', 43, 7, 165, 64, 38), ncol = 10),  col.names = c("length", "birthweig
kable(matrix(c( 19, 45, 189), ncol = 3), col.names = c("fedyrs", "fnocig", "fheight")) %>% kable_styling

new_baby = c(61, 5.1, 36, 43, 0, 43, 7, 165, 64, 38, 19, 45, 189)
names(new_baby) = c("length", "birthweight", "headcirc", "gestation", "smoker", "motherage", "mnocig",
obs = as.matrix(new_baby[-5])
pop0<-smok[smok$smoker == 0,-5]
pop1<-smok[smok$smoker == 1,-5]
N0<-nrow(pop0)
N1<-nrow(pop1)
#head(pop1)
sigma.hat0<-cov(pop0)
sigma.hat1<-cov(pop1)
sigma.hat.p<-((N0 - 1)*sigma.hat0 + (N1 - 1)*sigma.hat1)/(N0 + N1 - 2)
mu.hat0<-as.matrix(colMeans(pop0)) #Force it to be an actual column vector
mu.hat1<-as.matrix(colMeans(pop1))
b<-solve(sigma.hat.p) %*% (mu.hat0 - mu.hat1)
k<-0.5*t(mu.hat0 - mu.hat1) %*% solve(sigma.hat.p) %*% (mu.hat0 + mu.hat1)

D0<-t(obs - mu.hat0) %*% solve(sigma.hat.p) %*% (obs - mu.hat0)
D1<-t(obs - mu.hat1) %*% solve(sigma.hat.p) %*% (obs - mu.hat1)
data.frame(D0, D1) %>% kable() %>% kable_styling(latex_options = "hold_position") %>% column_spec(1, bo
```