Final Project in R

anurag

March 15, 2017

Introduction:

I have chosen the "Birth Weight" data set in the MASS package to analyze. The reason I choose this data was because of the factors considered in the weight of a child. These are interesting insights on factors affecting the weight of a child.

One line of this data set represents the various factors which are recorded for a new born child. This data is collected for Baystate Medical Center, Springfield, Mass during 1986.

Makeup of Data:

The factors considered are as follows: Indicator of baby weight less than 2.5kg, Age of of mother in years, Mothers weight in pounds at the last menstruation period, race, smoking or non-smoking, Previous premature labors, hypertension, Uterine irritability, number of physician visits during the first trimester and lastly the weight of a newborn baby.

In short, some important health aspects of the mother alongwith the weight of a newborn baby are recorded in one row.

Research question:

What effect does age, race, mothers weight, previous labours and number of physician visits have on the weight of a new born child.

This is an interesting relationship since the generalized baby weight is 2.5 kgs or 8 pounds. I wanted to see how any of the above factors affects the weight of baby.

Method of progression:

For each predictor, I have fit a simple linear regression model to predict the response. This is to see the statistical significance of the factors which are present in my research question. Then I have calculated a linear equation for each and procedded to plot the graph for the response.

After the univariate analysis, I have then calculated a multivariate regression analysis to find the find the p-value and how the values overall affect the baby weights. After this analysis, I have then provided my conclusions below.

Analysis:

For each factor in my research question, I have formulated an equation, followed by a plot for the baby weight vs the factor in consideration.

Load Mass package and Birth Weight (birthwt) tables.

```
library(MASS)
?birthwt

## starting httpd help server ...

##  done

data("birthwt")
View(birthwt)
library(ggplot2)
```

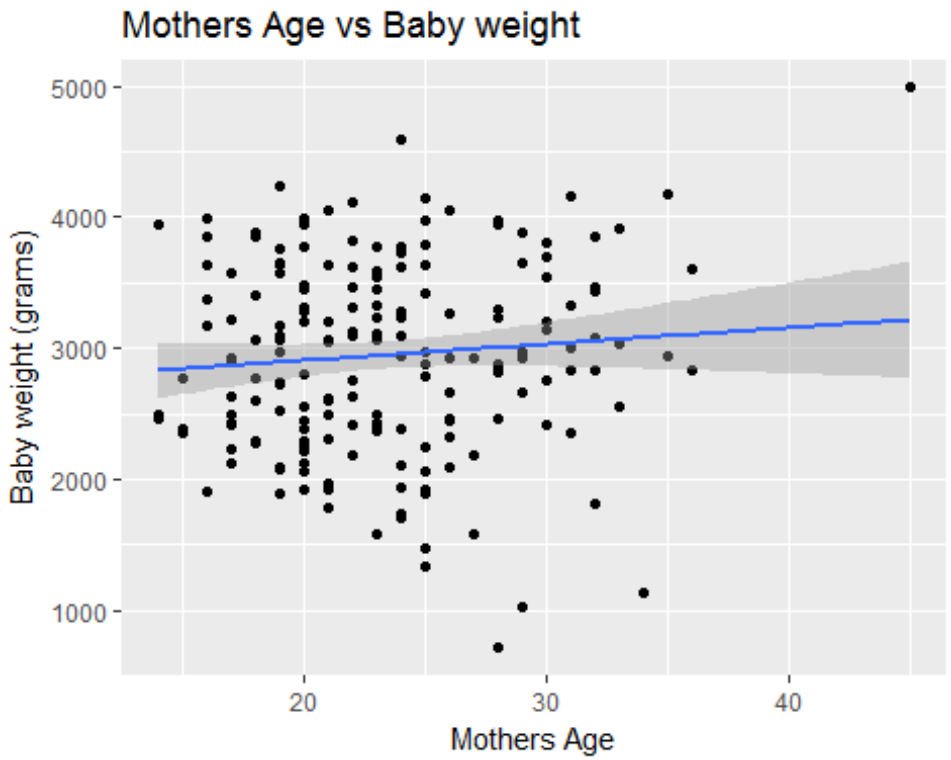
Create a data table for birth weights and view the first few rows of the data.

```
library(data.table)
head(birthwt)

##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622

birthwt <- data.table(birthwt)
View(birthwt)
```

Age

```
lr_age = lm(bwt~age, data=birthwt)
lr_age

##
## Call:
## lm(formula = bwt ~ age, data = birthwt)
##
## Coefficients:
## (Intercept)          age
##     2655.74        12.43
```

```
plot1 <- ggplot(birthwt, aes(y = bwt, x = age)) + geom_point() +
geom_smooth(method = lm) +
labs(x = "Mothers Age", y = "Baby weight (grams)",
title = "Mothers Age vs Baby weight")
plot1
```



As seen above,

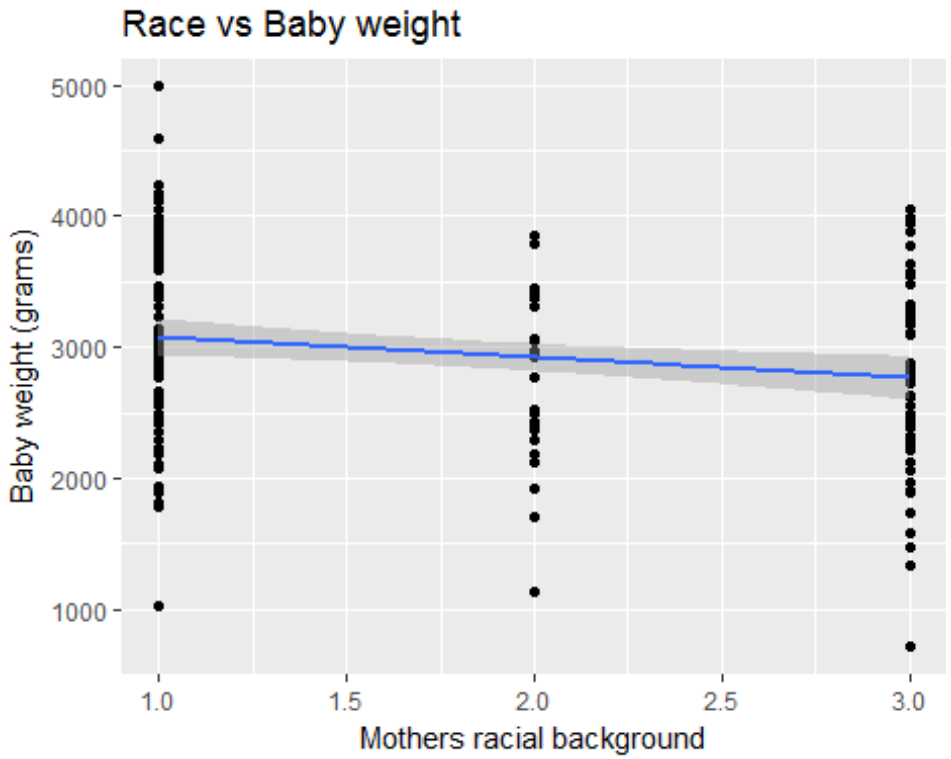
Birth weight of baby = 2655.74 + 12.43(Age of mother)*

For a 1 year increase in age of the mother, baby weight increases by 12.43 grams.

Race

```
lr_race = lm(bwt~race, data=birthwt)
lr_race

##
## Call:
## lm(formula = bwt ~ race, data = birthwt)
##
## Coefficients:
## (Intercept)         race
##      3230.1       -154.6

plot5 <- ggplot(birthwt, aes(y = bwt, x = race)) + geom_point() +
geom_smooth(method = lm) +
labs(x = "Mothers racial background", y = "Baby weight (grams)",
```

```
title = "Race vs Baby weight")
plot5
```



Birth weight of baby = 3230.1 - 154.6(Race of Mother)*

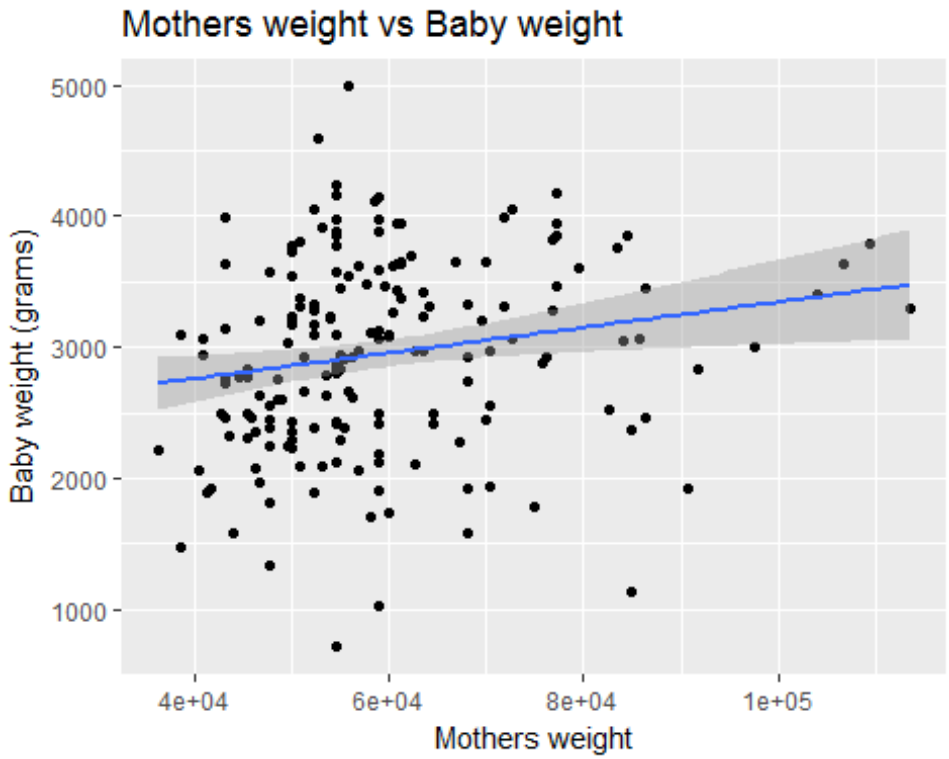
There is a 154.6 gram decrease from women of white racial background to black to other.

Mother's weight in pounds

Converting to grams

```
birthwt$gram_lwt = (birthwt$lwt*454)
View(birthwt)

lr_lwt = lm(bwt~gram_lwt, data=birthwt)
lr_lwt

##
## Call:
## lm(formula = bwt ~ gram_lwt, data = birthwt)
##
## Coefficients:
## (Intercept)     gram_lwt
##   2.370e+03    9.756e-03

plot2 <- ggplot(birthwt, aes(y = bwt, x = gram_lwt)) + geom_point() +
geom_smooth(method = lm) +
labs(x = "Mothers weight", y = "Baby weight (grams)",
```

```
title = "Mothers weight vs Baby weight")
plot2
```



Since mother's weight was expressed in pounds and baby weight in grams, I have calculated a field for converting it to grams so that we can directly compare it to the baby weight.

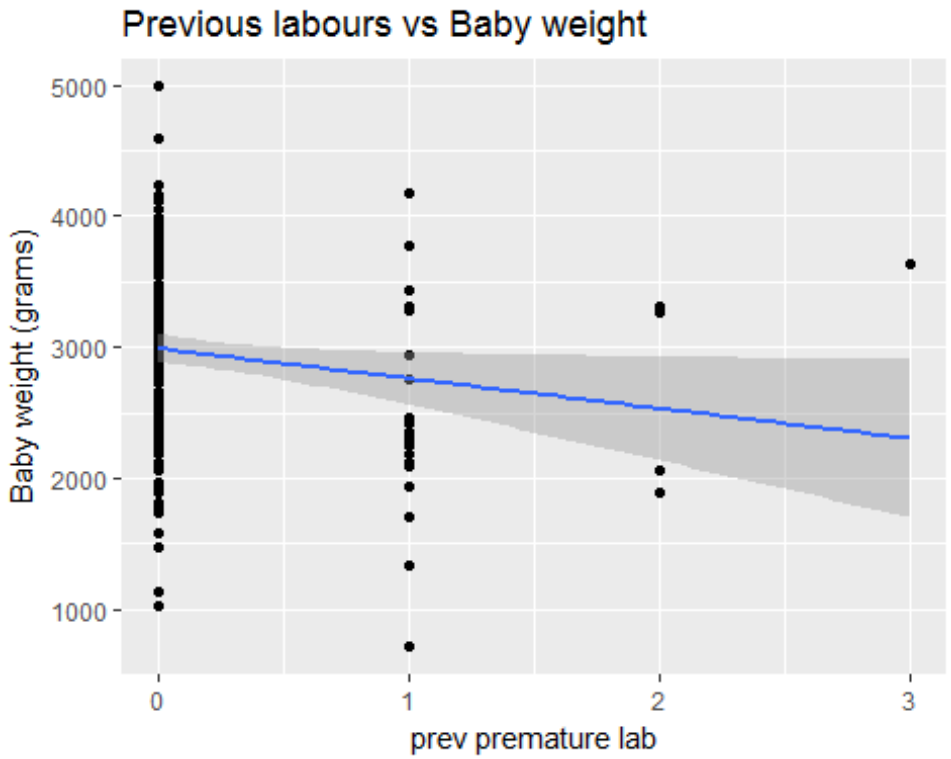
Birth weight of baby = 2.370e+03 + 9.756e-03(Mothers weight at last menstruation)*

For 1 pound increase in mother's weight, baby weight increases by 9.756e-03 grams.

Previous Premature Labours

```
lr_ptl = lm(bwt~ptl, data=birthwt)
lr_ptl

##
## Call:
## lm(formula = bwt ~ ptl, data = birthwt)
##
## Coefficients:
## (Intercept)          ptl
##      2989.3       -228.6

plot3 <- ggplot(birthwt, aes(y = bwt, x = ptl)) + geom_point() +
geom_smooth(method = lm) +
labs(x = "prev premature lab", y = "Baby weight (grams)",
```

```
title = "Previous labours vs Baby weight")
plot3
```

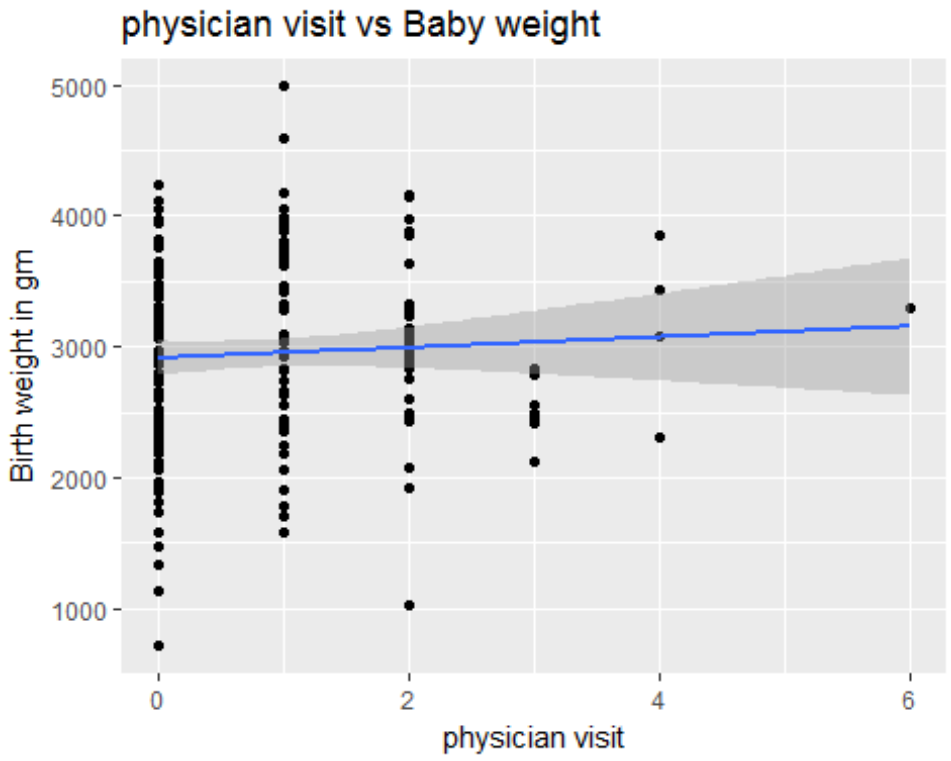


Birth weight of baby = 2989.3 - 228.6(Previous premature labours)*

For each increase premature labours, baby weight reduces by 228 grams

Physician Visits:

```
lr_ftv = lm(bwt~ftv, data=birthwt)
lr_ftv

##
## Call:
## lm(formula = bwt ~ ftv, data = birthwt)
##
## Coefficients:
## (Intercept)          ftv
##     2912.73        40.15

plot4 <- ggplot(birthwt, aes(y = bwt, x = ftv)) + geom_point() +
geom_smooth(method = lm) +
labs(x = "physician visit", y = "Birth weight in gm",
title = "physician visit vs Baby weight")
plot4
```



Birth weight of baby = 2912.73 + 40.15(Number of physician visits)*

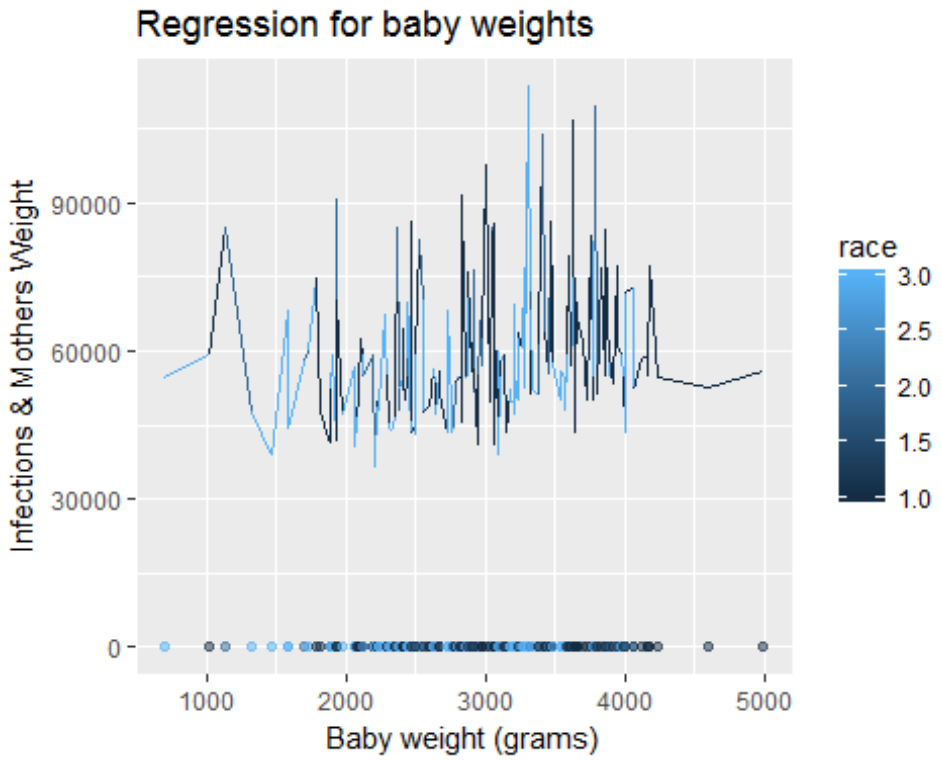
For each visit to physician, baby weight increases 4.4 grams.

Multivariate Regression:

```
model_summ = lm(bwt ~ low + age + gram_lwt + race + smoke + ptl + ht + ui + f
tv, data = birthwt)
summary(model_summ)

##
## Call:
## lm(formula = bwt ~ low + age + gram_lwt + race + smoke + ptl +
##     ht + ui + ftv, data = birthwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -991.22 -300.96   -5.39  277.74 1637.80
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.613e+03  2.295e+02  15.744  < 2e-16 ***
## low         -1.131e+03  7.396e+01 -15.296  < 2e-16 ***
## age         -6.245e+00  6.347e+00  -0.984 0.326416
## gram_lwt     2.314e-03  2.496e-03   0.927 0.355085
## race        -1.009e+02  3.854e+01  -2.618 0.009605 **
## smoke       -1.741e+02  7.200e+01  -2.418 0.016597 *
```

```
## ptl          8.134e+01  6.855e+01   1.187 0.236980
## ht          -1.820e+02  1.377e+02  -1.322 0.187934
## ui          -3.368e+02  9.331e+01  -3.609 0.000399 ***
## ftv         -7.578e+00  3.099e+01  -0.245 0.807118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433.7 on 179 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6462
## F-statistic: 39.16 on 9 and 179 DF,  p-value: < 2.2e-16
```

```
ggplot(birthwt, aes(x=bwt)) +
  geom_point(aes(color=race, y=ui), alpha=0.5) +
  geom_line(aes(color=race, y=gram_lwt)) +
  labs(title="Regression for baby weights",
       x="Baby weight (grams)",
       y="Infections & Mothers Weight")
```



As seen from the co-efficients of P-values, the most statistically signaficant values affecting Weight of a new born baby are the low weights of babies, presence of uterine irritability followed decreasingly by race and smoking.

From the regression, age, previous labours, hypertension and visits to the doctor do not affect the weight of babies.

The Low weight of a baby is directly co-realted with the weight of a baby and hence is by default a part of the analysis.

Uterine irritability is supposedly the main factor affecting weights of babies followed by race and smoking.

My reserch question was based on age, race, previous labours and visits to the physician. From the regression, it is clear that of these, race is the one which is a signifant effector, while the others are not.

One of the major drawbacks here would be that only co-relation may be proven, not causality. This is true for most regressions. It is upto the analyst on what basis to relate the factors.