Data Analysis Assignment 2

Surabhi Trivedi

9th Sept 2021

Maternal Smoking and Birth Weights

Summary

This report investigates if mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke. To infer this I have used multiple linear regression. The data used contains multiple variables about the mother such as height, level of education, mother's ethnicity etc. The final model reports the variable *smoke* (binary variable indicating whether the mother smokes or not) as a statistically significant variable. The average weight of babies born by mother's who smoke is **9.27 ounces** less than babies whose mother's do not smoke, keeping all the other variables constant (this result is statistically significant at a significance level of 0.05).

Introduction

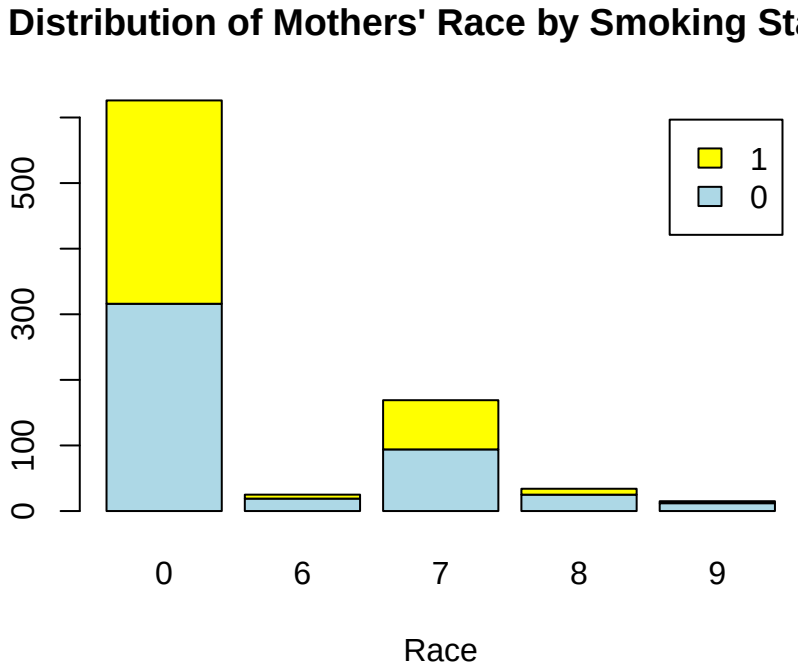
In this report I am primarily interested in establishing if smoking mothers give birth to babies with lower body weights. The secondary goal of this report would be to quantify the effect of smoking on a baby's weight — what is the difference in weight of babies whose mothers smoke vs babies whose mothers do not. Additionally, I am also interested in reporting if the effect of smoking on a baby's weight is disproportionately increases or decreases based on the mother's race/ ethnicity. Furthermore, I also report findings that I found interesting but were not directly related to the mother's smoking status.

Data

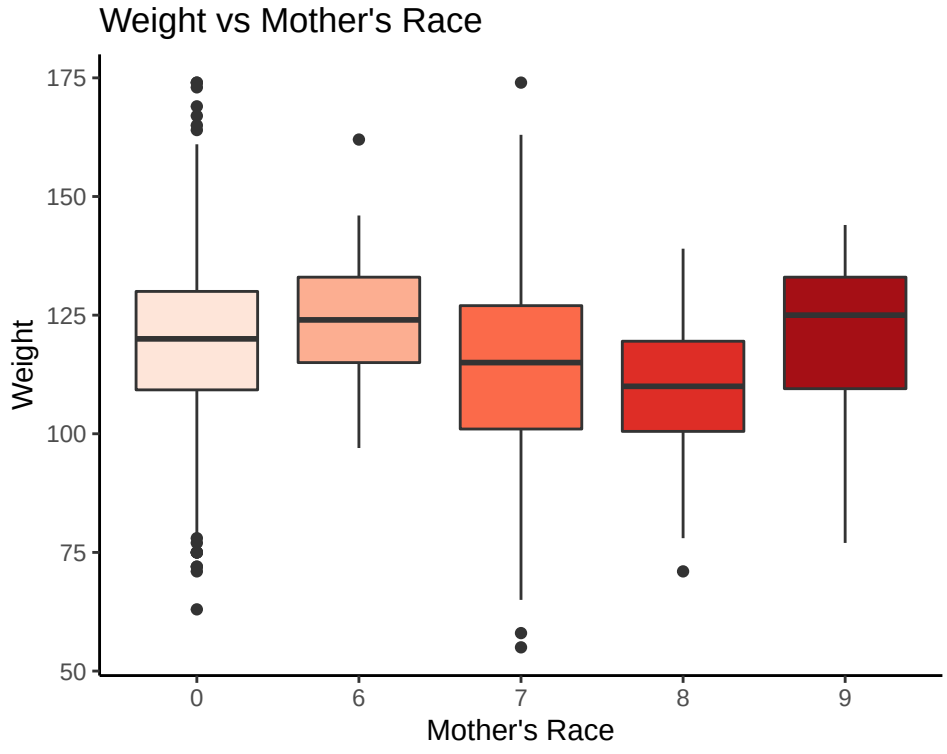
The dataset of interest contains data for 869 male single births where the baby lived up to atleast 28 days. The response variable *bwt.oz* is a continuous variable indicating the new born baby's weight in ounces. There are 9 predictor variables in the dataset of which 4 are categorical and 5 are continuous/ discreet. This dataset contains no missing value.

The original dataset encoded all variables as either num or int. I converted 4 of these variables to factor variables in order to make the analysis more meaningful and for the ease of interpretation. These variables were *mrace, med, inc, smoke*. I would also like to mention that for mrace, levels 0-5 were collapsed to one level: 0. Similarly, for med, levels 6-7 were collapsed to level: 6.

Initial EDA of the data revealed that the data collected is imbalanced with respect to race. Of the 869 mothers present in the dataset, 626 were white and 169 were black (as can be seen in the plot below). The other ethnicities are underrepresented in the dataset – this may make it harder for us to measure and interpret interaction between the variables *smoke* and *mrace*. All the other variables are well balanced in their distribution. Further, the response variable *bwt.oz* is reasonably normal in it's distribution.



On assessing the linearity between the predictors and the reponse variable I noticed that a lot of discreet variables were behaving like categorical variables which made it harder for me to assess their relationship with the response variable. To solve for this, I took the average of bwt.oz for each level and the plotted this average with the predictor. This made it considerably easy for me to visually understand the relationship between the predictors and the response variable. On visually analysing the categorical predictors *mrace* and *smoke* I could see a significant difference in the median for bwt.oz for both these predictors indicating that these two variables might be good predictors of our response variable. Below is a comparative boxplot for *mrace*. Black and Asian mothers have given birth to babies with lower weights.



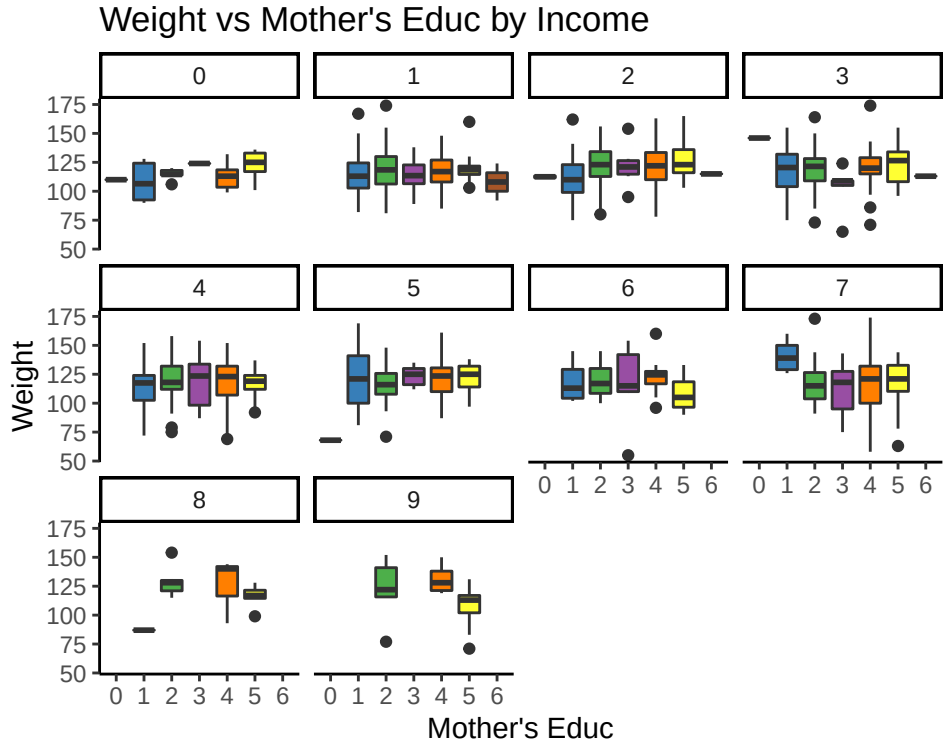
I explored variable interactions only with *smoke* as that is the variable I am interested in inferring and the number of combinations possible is really large. On visually exploring interactions with *smoke*, I did not find any variable interaction worth including in my model. However, while model building I consider all interactions possible with stepwise linear regression and do an F-test to understand if any of the interactions were significant.

Model

I started with a very basic model with all the predictors. Model assumptions were reasonably satisfied for this model. No outliers, leverage points or influential points were detected. Mother's height, weight pre pregnancy, smoking status and some levels of *mrace* were significant. The second step was to include the interaction between *smoke* and *race* as we are interested in inferring how the relationship between *bwt.oz* and *smoke* changes based on the mother's race. The F-test between my initial model and this model was not significant. Following this, I'd like to make the inference that there is no statistically significant interactions between the variables *race* and *smoke*. However, I would like to point out that the dataset is not representative of all races and this inference may change if we could collect more data for asian, mexican, and mixed race mothers.

Next, I checked for multicollinearity between variables and the VIF for *med* were > 30. Based on this insight, I proceeded to check for variable interactions between *med* and the other variables. Intuitively I checked for interaction between *med* and *inc* as logically these two variables can be correlated. The F-test for this interaction was significant as can be seen below:

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
844	236891.2	NA	NA	NA	NA
806	220962.0	38	15929.14	1.529065	0.0227778



However, I find it very hard to conclude anything meaningful from this interaction. I tried plotting box plots for Weight with Mother's Education with Income. But I do not see any trend that can interpreted in the

context of this data. Since the income being referred to is the family income, it might help to have more data about the baby's family – for example the father's data might help us in interpreting this interaction better. There is also a possibility that there is another confounding variable which is effecting both *med* and *inc* in whose absence it might be harder to draw conclusions regarding this interaction.

For model selection, I used stepwise linear model regression with BIC as we're interested in inferring the *smoke* variable and our goal is not to predict *bwt.oz*. Below are the results from the step-wise regression, which I also chose as my final model for the following reasons: 1. The regression output does not include *med* or *inc* and I am happy with that as *med* had a very high variance inflation factor and the interaction for *med* and *inc* was also significant. Were we interested in the inference of *med* or *inc* I would have kept these variables. However since that is not the case, I proceeded with this model. 2. My variable of interest is statistically significant in this model 3. This nodel satisfies all the model assumptions reasonably well 4. The asjusted R squared for the model is not very high but since prediction is not our goal we can ignore this statistic

Below is the linear regression equation and summary of my final model:

$$BirthWeight_i = \beta_0 + \beta_1 smoke_i + \beta_2 mrace_i + \beta_3 mpregwt_i + \beta_4 mht_i + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Observations	869
Dependent variable	bwt.oz
Type	OLS linear regression

F(7,861)	21.19
R^2	0.15
Adj. R^2	0.14

	Est.	S.E.	t val.	p
(Intercept)	53.25	15.32	3.48	0.00
smoke1	-9.27	1.15	-8.04	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
mpregwt	0.12	0.03	3.70	0.00
mht	0.88	0.26	3.37	0.00

Standard errors: OLS

From my final model I can infer that mother's who smoke tend to give birth to babies with lower weights. My results are statistically significant at 0.05. Mothers who do not smoke give birth to babies who are 9.27 ounces heavier in comparison to mother's who smoke. The likely range for difference in birth weights for mothers who smoke vs who do not is **(7 ounces, 11.54 ounces)** i.e. for mothers of same height, pre pregnancy weight, same race, we expect mother's who smoke to give birth to babies which will on average weight 9.27 ounces less than mother's who do not smoke, with 95% confidence in ounces = **(7 ounces, 11.54 ounces)**.

I also find it very interesting to see that mothers of black and asian ethnicities give birth to babies of relatively lower weights (~8 ounces) as opposed to white mothers. Again these results are statistically significant. Mother's height and pre pregnancy weight are also significant and have a positive slope with respect to *bwt.oz*.

Conclusion

In conclusion, I would like to say that women who are pregnant should refrain from smoking as it might affect the health of their child.

The final model and the analyses reported above do suffer from some drawbacks however. Firstly, the variable *mrace* is not well represented which put's in question the significance of this variable in our model. Secondly, *med* has a very high VIF and would be a very major problem if we're interested in predicting *bwt.oz*. Thirdly, I feel the most major drawback of this model is that it does not take into account the health statistics of the father at all. The linear relationship between smoke and weight that we're seeing above could very well be correlated with the health of the father.

R Code Appendix

```
##############################################################################
##############################################################################
################## Maternal Smoking and Birth Weights ########################
##############################################################################
##############################################################################

###### Clear environment and load libraries
rm(list = ls())
library(knitr)
library(ggplot2)
#library(kableExtra)
#library(lattice)
library(dplyr)
library(rms) #for VIF
library(MASS)
library(jtools)


#Loading Data
smoking <- read.csv("smoking.csv")
babies <- read.csv("babiesdata.csv")
summary(smoking)
```

```
##        id             date        gestation         bwt.oz          parity
##  Min.   :  15   Min.   :1350   Min.   :148.0   Min.   : 55.0   Min.   : 0.000
##  1st Qu.:5477   1st Qu.:1444   1st Qu.:272.0   1st Qu.:108.0   1st Qu.: 1.000
##  Median :6734   Median :1540   Median :279.0   Median :119.0   Median : 2.000
##  Mean   :6032   Mean   :1536   Mean   :278.5   Mean   :118.4   Mean   : 1.953
##  3rd Qu.:7587   3rd Qu.:1627   3rd Qu.:286.0   3rd Qu.:129.0   3rd Qu.: 3.000
##  Max.   :9263   Max.   :1714   Max.   :338.0   Max.   :174.0   Max.   :11.000
##      mrace            mage            med             mht
##  Min.   :0.000   Min.   :15.00   Min.   :0.000   Min.   :53.00
##  1st Qu.:0.000   1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00
##  Median :2.000   Median :26.00   Median :2.000   Median :64.00
##  Mean   :2.995   Mean   :27.29   Mean   :2.932   Mean   :64.07
##  3rd Qu.:7.000   3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00
```

```
##  Max.   :9.000   Max.   :45.00   Max.   :7.000   Max.   :72.00
##     mpregwt           inc             smoke
##  Min.   : 87.0   Min.   :0.000   Min.   :0.0000
##  1st Qu.:113.0   1st Qu.:2.000   1st Qu.:0.0000
##  Median :125.0   Median :3.000   Median :0.0000
##  Mean   :128.5   Mean   :3.681   Mean   :0.4638
##  3rd Qu.:140.0   3rd Qu.:5.000   3rd Qu.:1.0000
##  Max.   :220.0   Max.   :9.000   Max.   :1.0000
```

```
str(smoking)
```

```
## 'data.frame':    869 obs. of  12 variables:
##  $ id       : int  4604 7435 7722 2026 3553 3491 6757 6153 8187 8403 ...
##  $ date     : int  1598 1527 1563 1503 1638 1705 1444 1405 1669 1669 ...
##  $ gestation: int  148 181 204 225 233 234 234 235 236 241 ...
##  $ bwt.oz   : int  116 110 55 132 105 85 97 129 63 128 ...
##  $ parity   : int  7 7 11 4 4 7 0 3 0 0 ...
##  $ mrace    : int  7 7 7 7 7 7 6 7 5 7 ...
##  $ mage     : int  28 27 35 28 34 33 26 24 24 17 ...
##  $ med      : int  1 1 3 2 3 1 5 4 5 1 ...
##  $ mht      : int  66 64 65 67 61 67 65 66 58 64 ...
##  $ mpregwt  : int  135 133 140 148 130 130 112 135 99 126 ...
##  $ inc      : int  2 1 6 3 3 2 6 1 7 2 ...
##  $ smoke    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
#subsetting the data and removing unwanted vars
smoking <- smoking[c(-2,-3)]
summary(smoking)
```

```
##        id           bwt.oz          parity           mrace
##  Min.   :  15   Min.   : 55.0   Min.   : 0.000   Min.   :0.000
##  1st Qu.:5477   1st Qu.:108.0   1st Qu.: 1.000   1st Qu.:0.000
##  Median :6734   Median :119.0   Median : 2.000   Median :2.000
##  Mean   :6032   Mean   :118.4   Mean   : 1.953   Mean   :2.995
##  3rd Qu.:7587   3rd Qu.:129.0   3rd Qu.: 3.000   3rd Qu.:7.000
##  Max.   :9263   Max.   :174.0   Max.   :11.000   Max.   :9.000
##       mage            med             mht          mpregwt
##  Min.   :15.00   Min.   :0.000   Min.   :53.00   Min.   : 87.0
##  1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00   1st Qu.:113.0
##  Median :26.00   Median :2.000   Median :64.00   Median :125.0
##  Mean   :27.29   Mean   :2.932   Mean   :64.07   Mean   :128.5
##  3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00   3rd Qu.:140.0
##  Max.   :45.00   Max.   :7.000   Max.   :72.00   Max.   :220.0
##       inc             smoke
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :3.000   Median :0.0000
##  Mean   :3.681   Mean   :0.4638
##  3rd Qu.:5.000   3rd Qu.:1.0000
##  Max.   :9.000   Max.   :1.0000
```

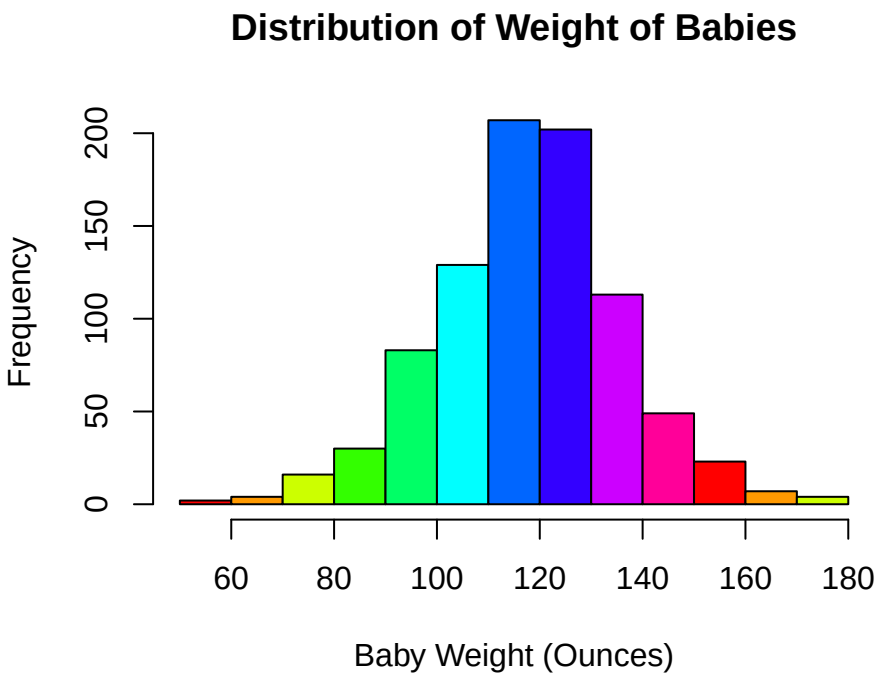
```
describe(smoking)
```

```
## smoking
##
##  10  Variables      869  Observations
## ---------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0      869        1     6032     2344     1087     1997
##      .25      .50      .75      .90      .95
##     5477     6734     7587     8070     8375
##
## lowest :   15   20   58   72  129, highest: 9153 9163 9213 9229 9263
## ---------------------------------------------------------------------------
## bwt.oz
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0      100        1    118.4    20.05     88.0     96.0
##      .25      .50      .75      .90      .95
##    108.0    119.0    129.0    139.0    147.6
##
## lowest :  55  58  63  65  68, highest: 165 167 169 173 174
## ---------------------------------------------------------------------------
## parity
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       12    0.959    1.953    1.968        0        0
##      .25      .50      .75      .90      .95
##        1        2        3        4        6
##
## lowest :  0  1  2  3  4, highest:  7  8  9 10 11
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency    209   220   173   120    61    40    22    12     3     5     2
## Proportion 0.241 0.253 0.199 0.138 0.070 0.046 0.025 0.014 0.003 0.006 0.002
##
## Value         11
## Frequency      2
## Proportion 0.002
## ---------------------------------------------------------------------------
## mrace
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       10    0.901    2.995    3.387        0        0
##      .25      .50      .75      .90      .95
##        0        2        7        7        8
##
## lowest : 0 1 2 3 4, highest: 5 6 7 8 9
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency    389    34    18    44    44    97    25   169    34    15
## Proportion 0.448 0.039 0.021 0.051 0.051 0.112 0.029 0.194 0.039 0.017
## ---------------------------------------------------------------------------
## mage
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       29    0.997    27.29    6.421       19       20
```

```
##      .25      .50      .75      .90      .95
##       23       26       31       36       38
##
## lowest : 15 17 18 19 20, highest: 40 41 42 43 45
## ---------------------------------------------------------------------------
## med
##        n  missing distinct     Info     Mean      Gmd
##      869        0        7    0.927    2.932    1.585
##
## lowest : 0 1 2 3 4, highest: 2 3 4 5 7
##
## Value          0     1     2     3     4     5     7
## Frequency      5   130   321    47   203   159     4
## Proportion 0.006 0.150 0.369 0.054 0.234 0.183 0.005
## ---------------------------------------------------------------------------
## mht
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       18    0.986    64.07    2.839       60       61
##      .25      .50      .75      .90      .95
##       62       64       66       67       68
##
## lowest : 53 54 56 58 59, highest: 68 69 70 71 72
##
## Value         53    54    56    58    59    60    61    62    63    64    65
## Frequency      1     1     1     6    17    43    73    96   113   127   132
## Proportion 0.001 0.001 0.001 0.007 0.020 0.049 0.084 0.110 0.130 0.146 0.152
##
## Value         66    67    68    69    70    71    72
## Frequency    113    81    38    15     7     4     1
## Proportion 0.130 0.093 0.044 0.017 0.008 0.005 0.001
## ---------------------------------------------------------------------------
## mpregwt
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       97    0.999    128.5    22.56      100      105
##      .25      .50      .75      .90      .95
##      113      125      140      155      170
##
## lowest :  87  89  90  91  92, highest: 197 198 200 215 220
## ---------------------------------------------------------------------------
## inc
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      869        0       10     0.98    3.681    2.581        1        1
##      .25      .50      .75      .90      .95
##        2        3        5        7        7
##
## lowest : 0 1 2 3 4, highest: 5 6 7 8 9
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency     26   153   146   136   105    98    57   111    16    21
## Proportion 0.030 0.176 0.168 0.157 0.121 0.113 0.066 0.128 0.018 0.024
## ---------------------------------------------------------------------------
## smoke
##        n  missing distinct     Info      Sum     Mean      Gmd
##      869        0        2    0.746      403   0.4638   0.4979
```

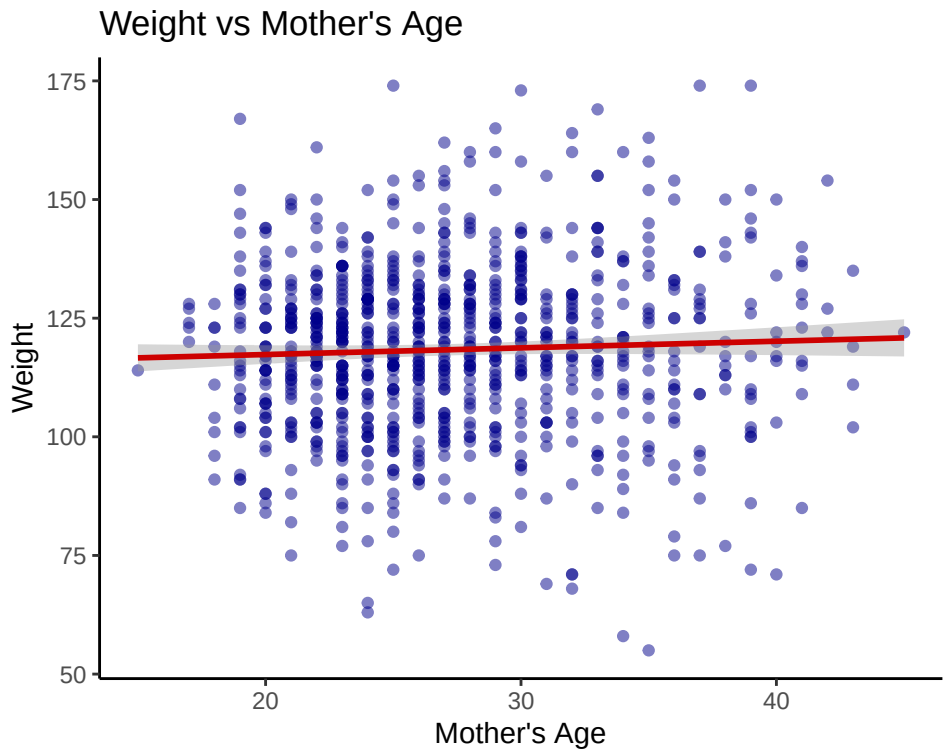
```
##
## ------------------------------------------------------------------------
```

```
#Collapsing race and education categories for easier analysis
smoking$med[smoking$med == 7] <- 6
smoking$mrace[smoking$mrace == 1 | smoking$mrace == 2 | smoking$mrace == 3 | smoking$mrace == 4 | smoki

#converting vars from num to factor
smoking[,'mrace']<-factor(smoking[,'mrace'])
smoking[,'med']<-factor(smoking[,'med'])
smoking[,'inc']<-factor(smoking[,'inc'])
smoking[,'smoke']<-factor(smoking[,'smoke'])

####### EDA
#Checking if the distribution of the response variable is normal
hist(smoking$bwt.oz,xlab="Baby Weight (Ounces)",main="Distribution of Weight of Babies",col=rainbow(10)
```



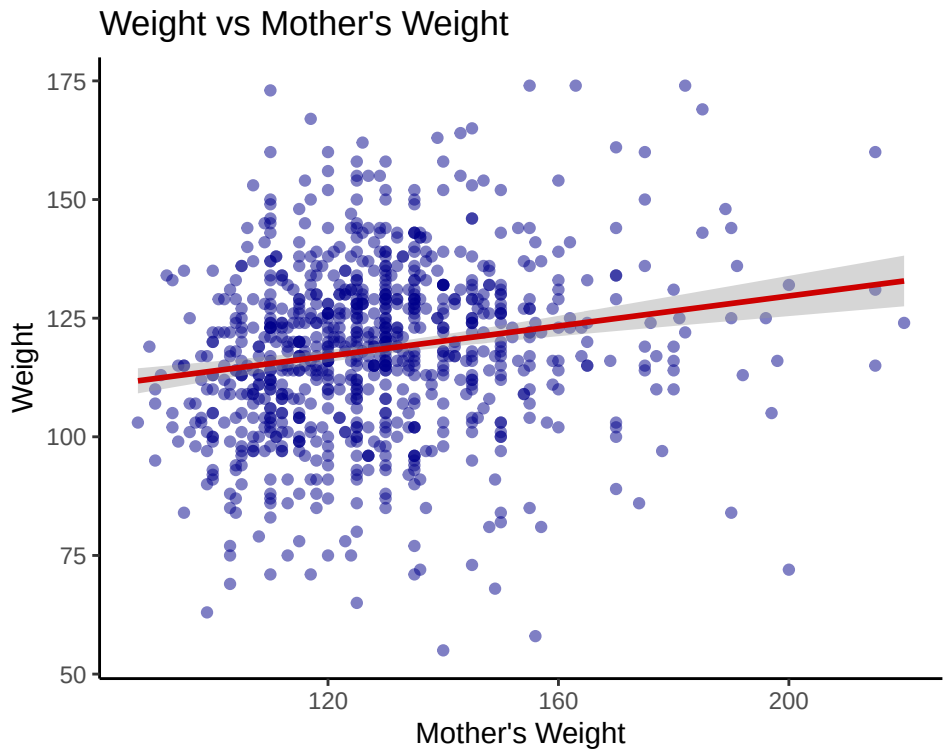

```
#Exploring the relationship b/w weight and predictors
#mage
ggplot(smoking,aes(x=mage, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Age",x="Mother's Age",y="Weight")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



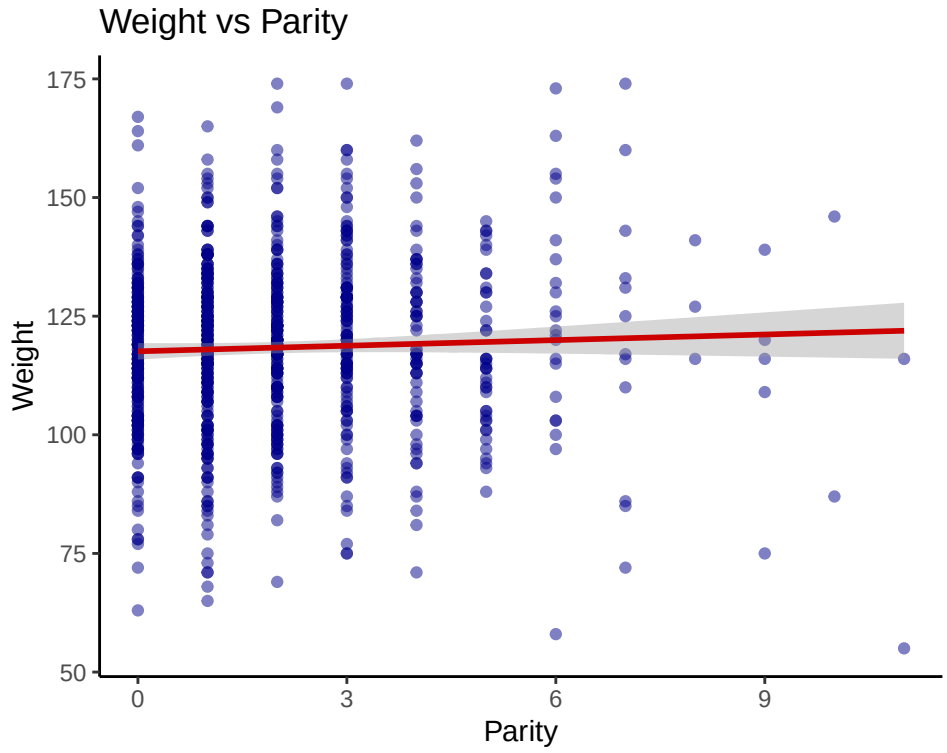

```
#not linear - maybe we can try converting mage to categorical
#you can do the median thing

#mpregwt
ggplot(smoking,aes(x=mpregwt, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm", col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Weight",x="Mother's Weight",y="Weight")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



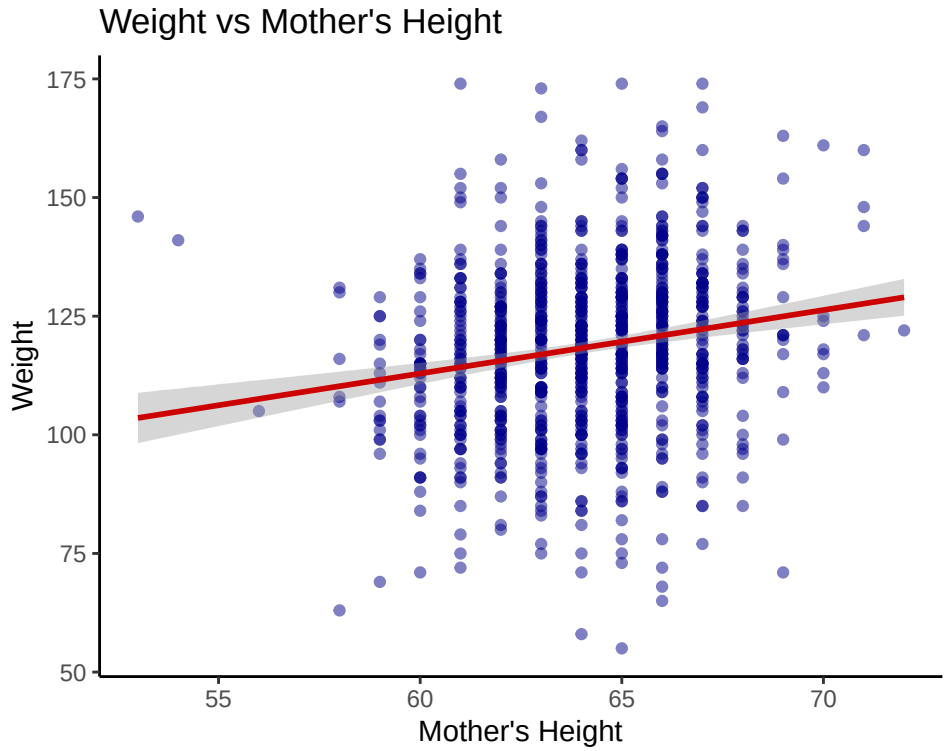

```
#somehwat weakly linear
#you can do the median thing

#parity
ggplot(smoking,aes(x=parity, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm", col="red3") + theme_classic() +
  labs(title="Weight vs Parity",x="Parity",y="Weight")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



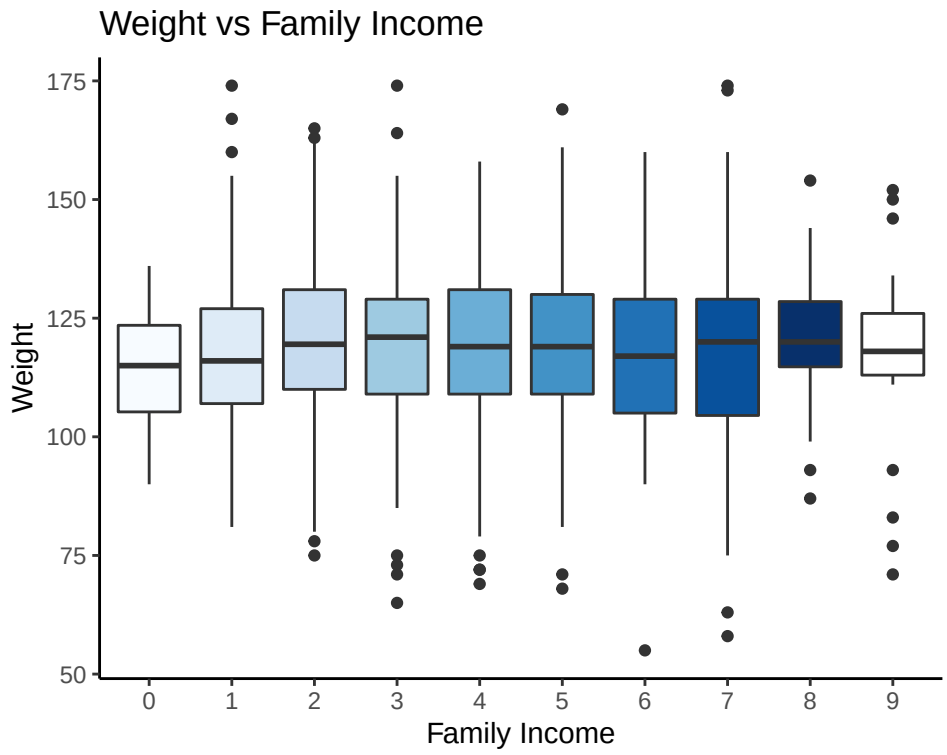

```
#not linear

#mht
ggplot(smoking,aes(x=mht, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm", col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Height",x="Mother's Height",y="Weight")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



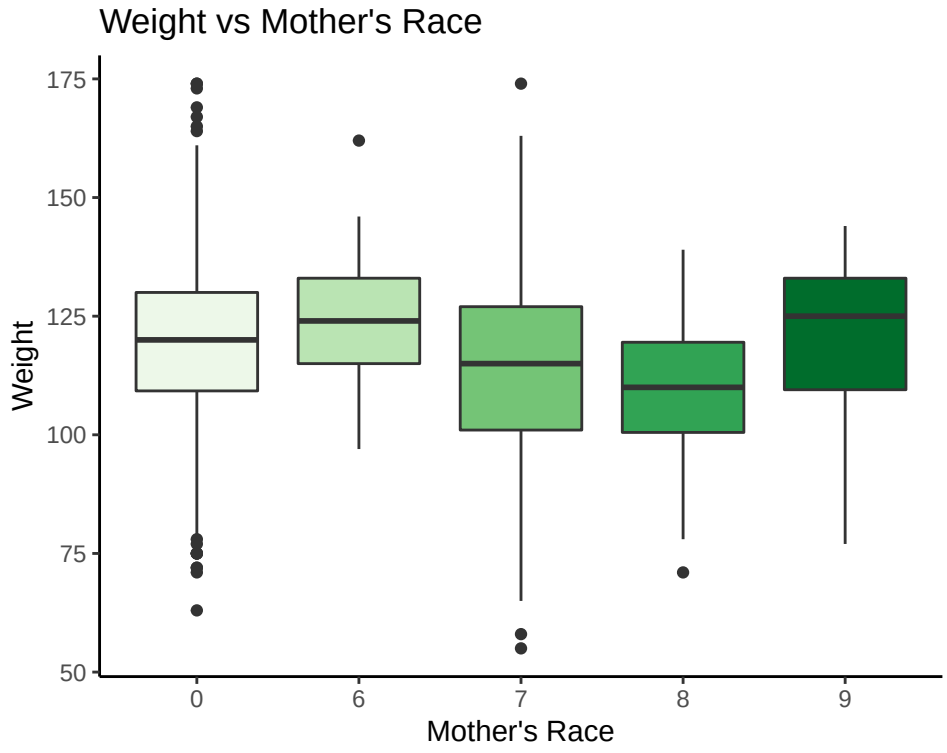

```
#not linear
#you can do the median thing

#inc
ggplot(smoking,aes(x=inc, y=bwt.oz, fill=inc)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Weight vs Family Income",x="Family Income",y="Weight") +
  theme_classic() + theme(legend.position="none")
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Blues is 9
## Returning the palette you asked for with that many colors
```

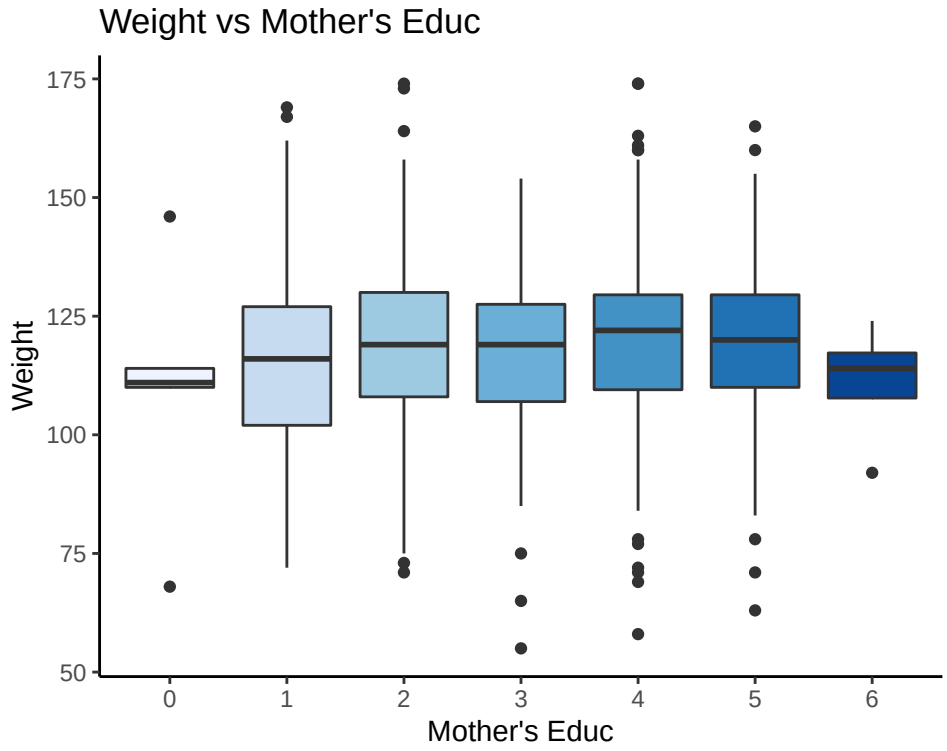



```
#median is somewhat similar

#mrace
ggplot(smoking,aes(x=mrace, y=bwt.oz, fill=mrace)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Red") +
  labs(title="Weight vs Mother's Race",x="Mother's Race",y="Weight") +
  theme_classic() + theme(legend.position="none")
```

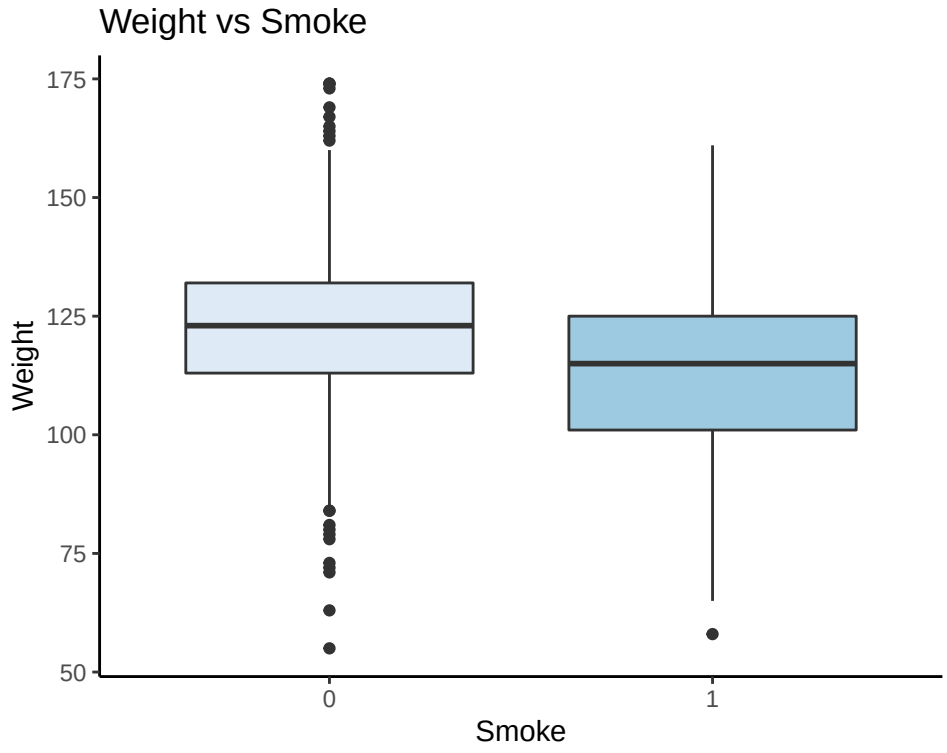
```
## Warning in pal_name(palette, type): Unknown palette Red
```




```
#median is different

#med
ggplot(smoking,aes(x=med, y=bwt.oz, fill=med)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Weight vs Mother's Educ",x="Mother's Educ",y="Weight") +
  theme_classic() + theme(legend.position="none")
```

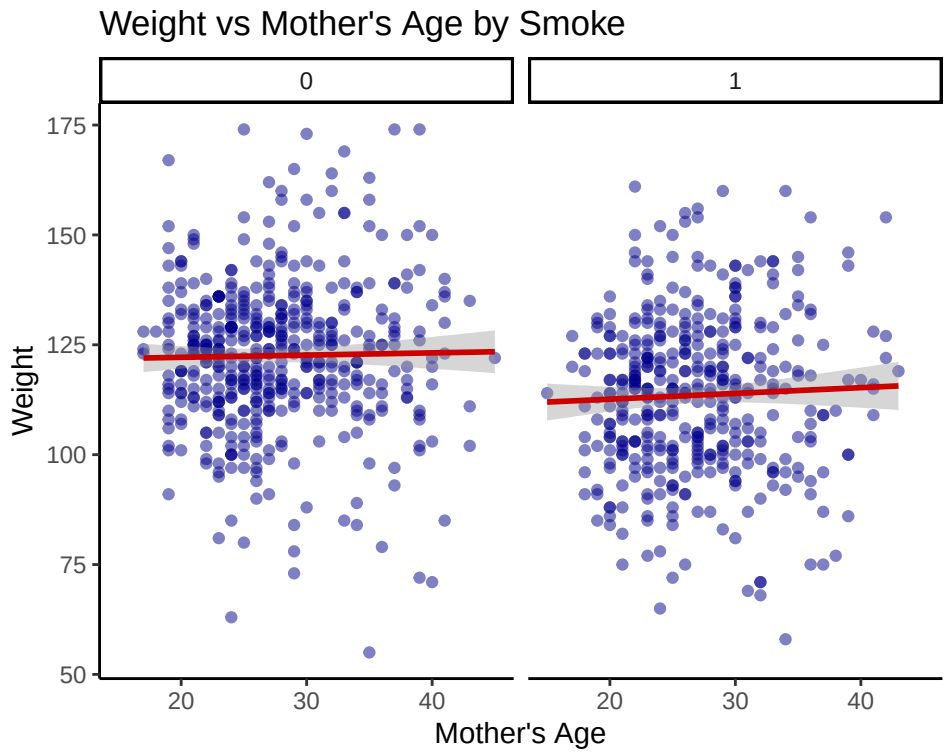



```
#median is similar

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



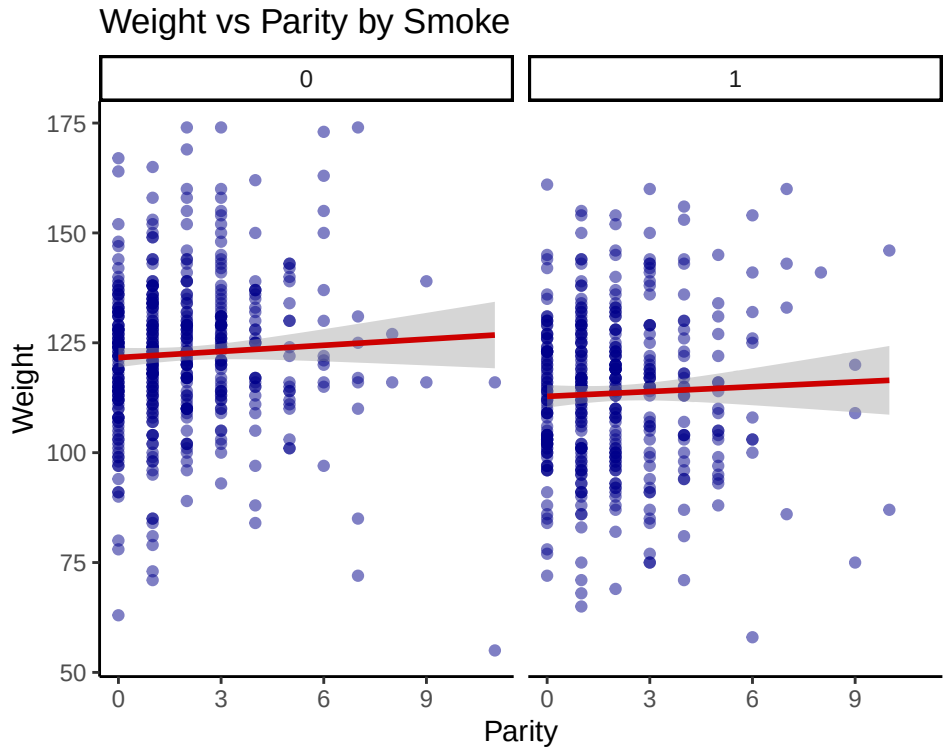

```
#median is different

####### Let's explore interactions
#Weight with mage by smoke
ggplot(smoking,aes(x=mage, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Age by Smoke",x="Mother's Age",y="Weight") +
  facet_wrap( ~ smoke,ncol=4)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



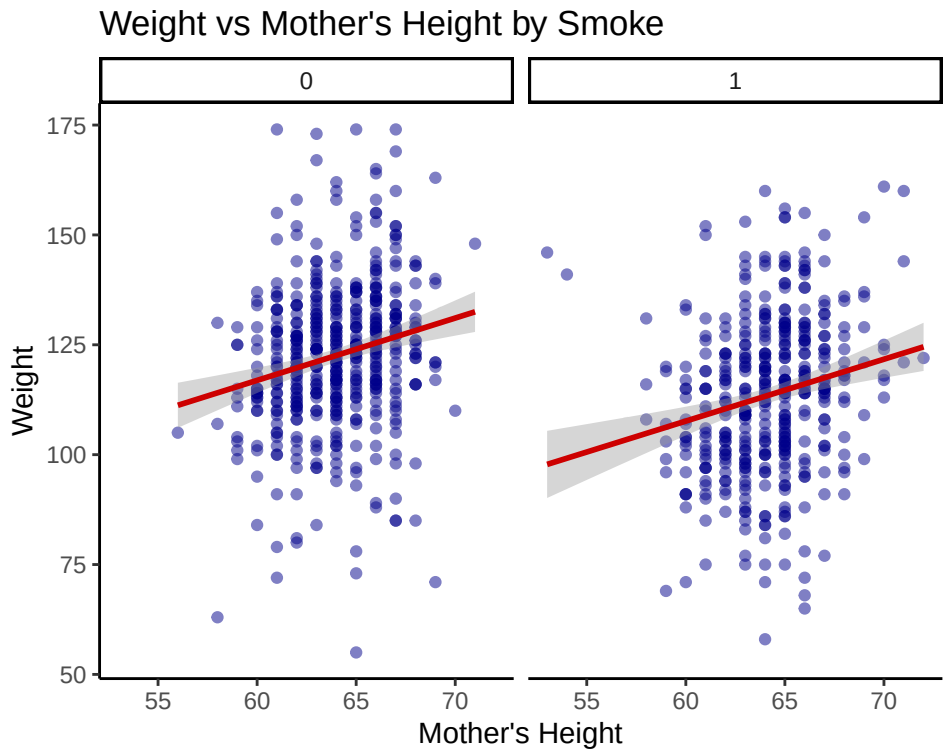

```
#no interaction

#Weight with parity by smoke
ggplot(smoking,aes(x=parity, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Weight vs Parity by Smoke",x="Parity",y="Weight") +
  facet_wrap( ~ smoke,ncol=4)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



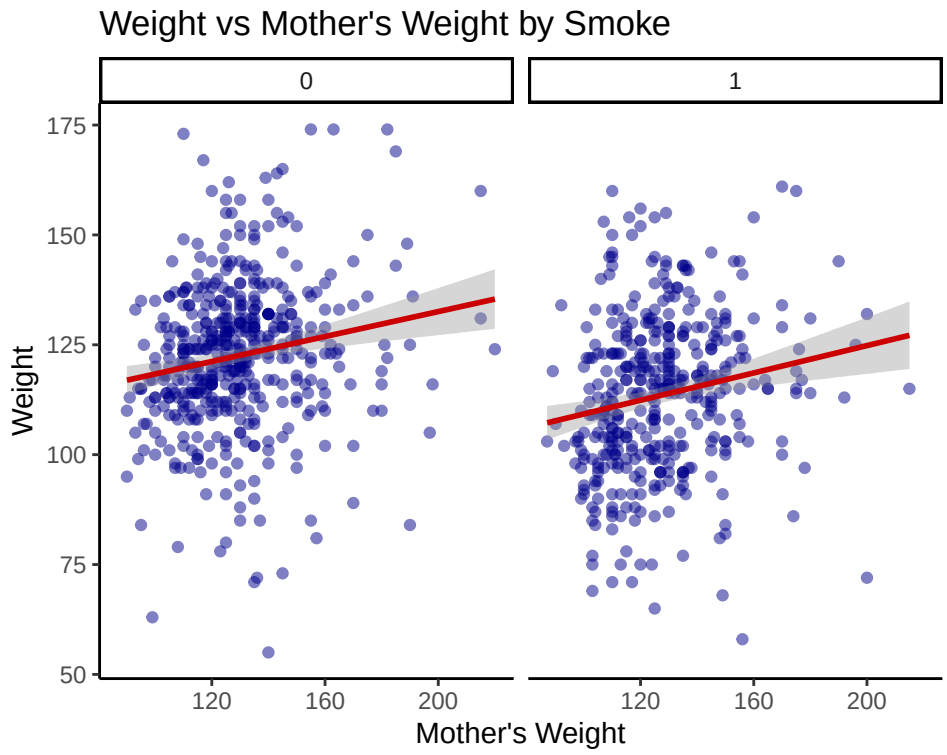

```
#no interaction

#Weight with mht by smoke
ggplot(smoking,aes(x=mht, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Height by Smoke",x="Mother's Height",y="Weight") +
  facet_wrap( ~ smoke,ncol=4)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

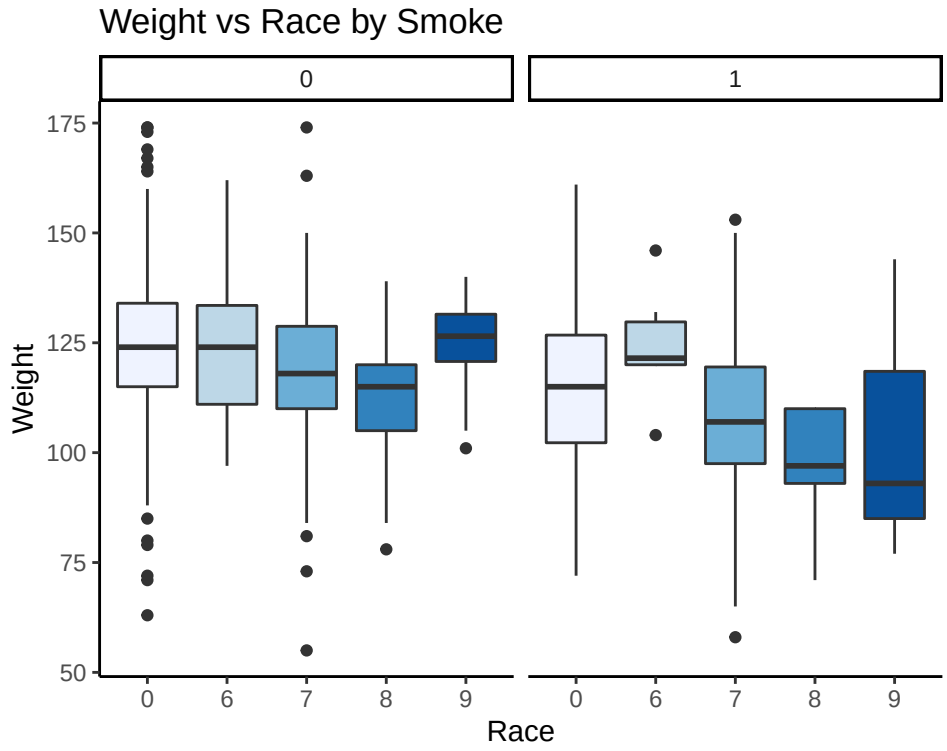



```
#no interaction

#Weight with mpregwt by smoke
ggplot(smoking,aes(x=mpregwt, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") + theme_classic() +
  labs(title="Weight vs Mother's Weight by Smoke",x="Mother's Weight",y="Weight") +
  facet_wrap( ~ smoke,ncol=4)
```

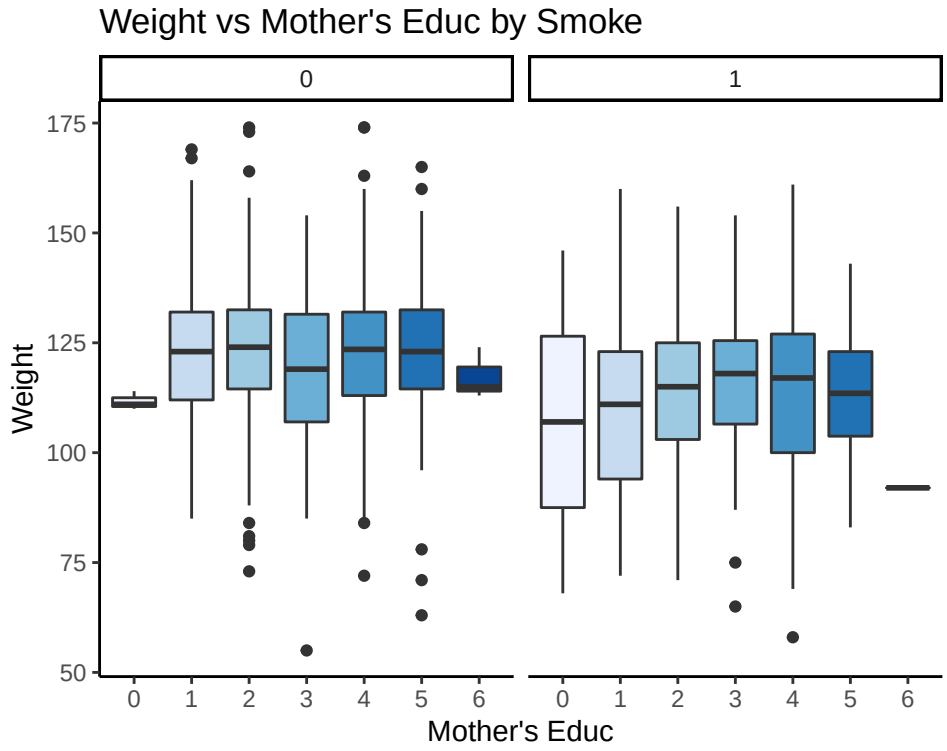
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#no interaction

####### Now interaction b/w categorical variables
#Weight with mrace by smoke
ggplot(smoking,aes(x=mrace, y=bwt.oz, fill=mrace)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Weight vs Race by Smoke",x="Race",y="Weight") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ smoke,ncol=4)
```

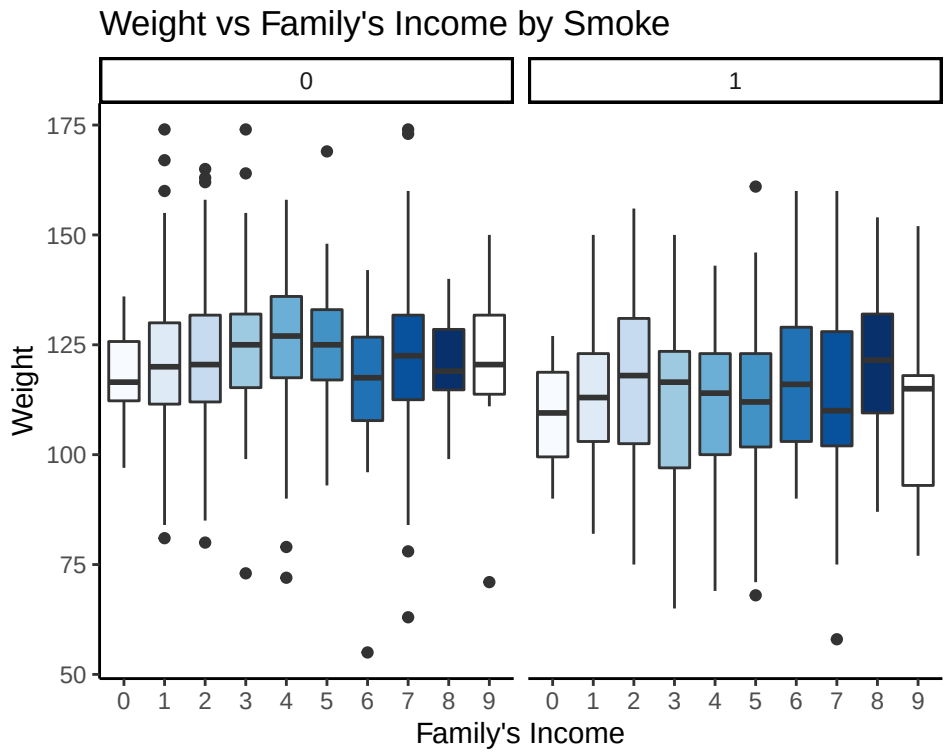



```
#there might be some interaction

#Weight with med by smoke
ggplot(smoking,aes(x=med, y=bwt.oz, fill=med)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Weight vs Mother's Educ by Smoke",x="Mother's Educ",y="Weight") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ smoke,ncol=4)
```




```
#trend is mostly the same with some differences

#Weight with inc by smoke
ggplot(smoking,aes(x=inc, y=bwt.oz, fill=inc)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Weight vs Family's Income by Smoke",x="Family's Income",y="Weight") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ smoke,ncol=4)
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Blues is 9
## Returning the palette you asked for with that many colors
```

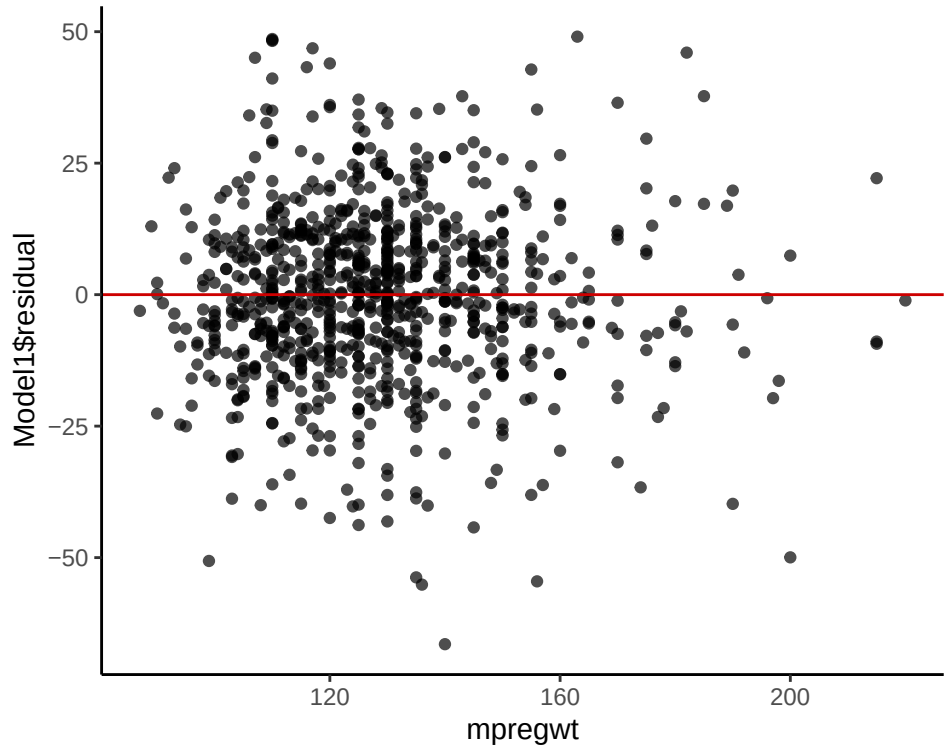



```
#trend is mostly the same with some differences

#TAKEAWAYS:
#We see no evidence of non-normality
#We also notice that linearity is a problem with all the continuous vars as they behave like categorica
#We see that the median for race and smoke is different suggesting they might be significant variables
#We might also consider interactions between `Race` and `Smoke`

###### Modeling and Model Assessment
#Based on our EDA we will include all the variables and the interaction b/w mrace and smoke

#First, a MLR model on weight with only main effects
Model1 <- lm(bwt.oz~parity+mrace+mage+med+mht+mpregwt+inc+smoke,data=smoking)
summary(Model1)
```

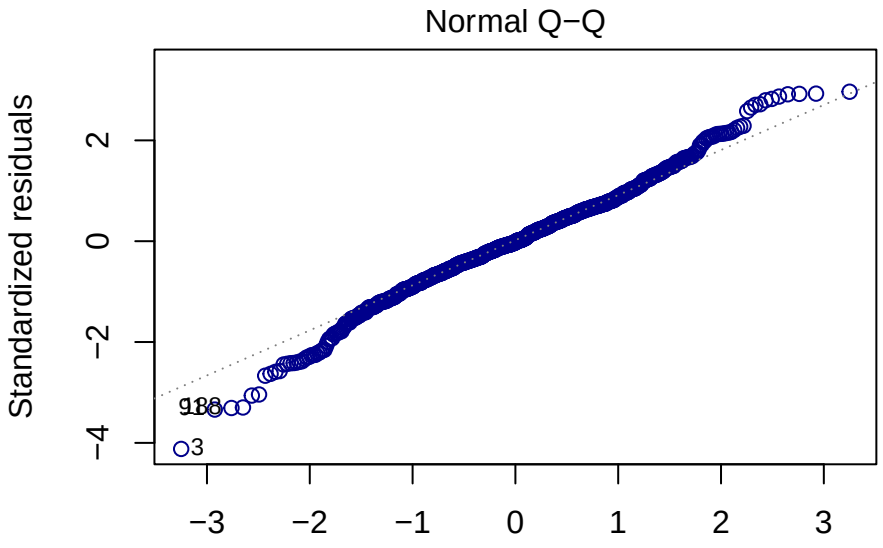
```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + mage + med + mht + mpregwt +
##     inc + smoke, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.464  -9.676  -0.334  10.373  49.055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.92805   17.88791   2.176 0.029815 *
## parity       0.78187    0.40086   1.950 0.051453 .
## mrace6       4.19494    3.53092   1.188 0.235144
```

```
## mrace7      -9.14155    1.59240  -5.741 1.31e-08 ***
## mrace8      -7.65440    3.12337  -2.451 0.014460 *
## mrace9      -2.68980    4.44035  -0.606 0.544836
## mage        -0.02705    0.13430  -0.201 0.840418
## med1         5.96206    7.80927   0.763 0.445403
## med2         8.06224    7.70613   1.046 0.295763
## med3         6.06152    8.00269   0.757 0.449000
## med4         8.51416    7.75189   1.098 0.272372
## med5         7.55992    7.78403   0.971 0.331723
## med6        -4.37963   11.32919  -0.387 0.699165
## mht          0.94625    0.26971   3.508 0.000475 ***
## mpregwt      0.10879    0.03297   3.300 0.001007 **
## inc1         3.43894    3.59253   0.957 0.338717
## inc2         5.22012    3.60482   1.448 0.147961
## inc3         2.13366    3.64001   0.586 0.557918
## inc4         2.76108    3.72187   0.742 0.458384
## inc5         2.17449    3.76549   0.577 0.563770
## inc6         1.22521    4.03056   0.304 0.761217
## inc7         2.14669    3.73493   0.575 0.565607
## inc8         3.30223    5.43759   0.607 0.543817
## inc9        -1.40629    5.07369  -0.277 0.781715
## smoke1      -9.23517    1.18005  -7.826 1.50e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 844 degrees of freedom
## Multiple R-squared:  0.1624, Adjusted R-squared:  0.1386
## F-statistic: 6.818 on 24 and 844 DF,  p-value: < 2.2e-16
```

```
#Let's do some model assessment with this before adding the interaction term
#Linearity
ggplot(smoking,aes(x=mpregwt, y=Model1$residual)) +
  geom_point(alpha = .7) +  geom_hline(yintercept=0,col="red3") + theme_classic()
```

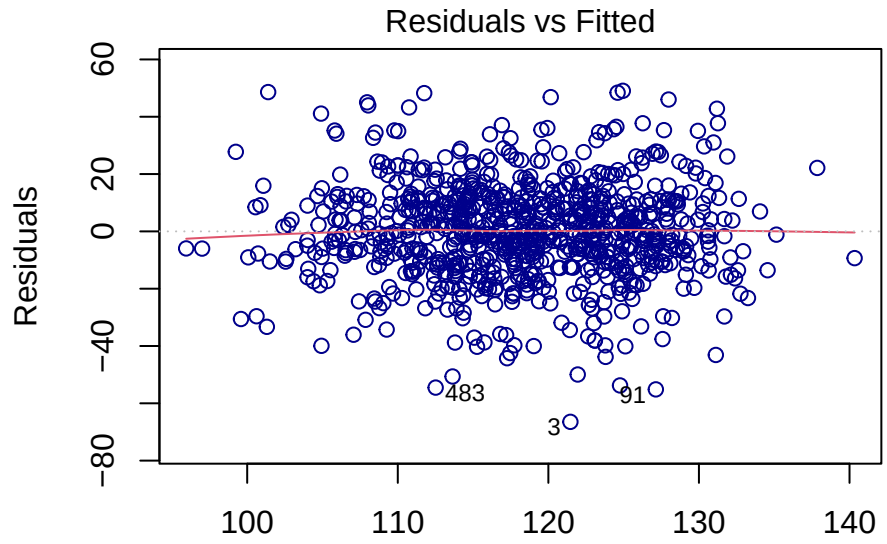



```
#Linearity is clearly not being satisfied for parity, mage, mht because they're behaving like categoric
#mpregwt is somewhat satisfying the linearity assumption

#Normality
plot(Model1,which=2,col=c("blue4"))
```



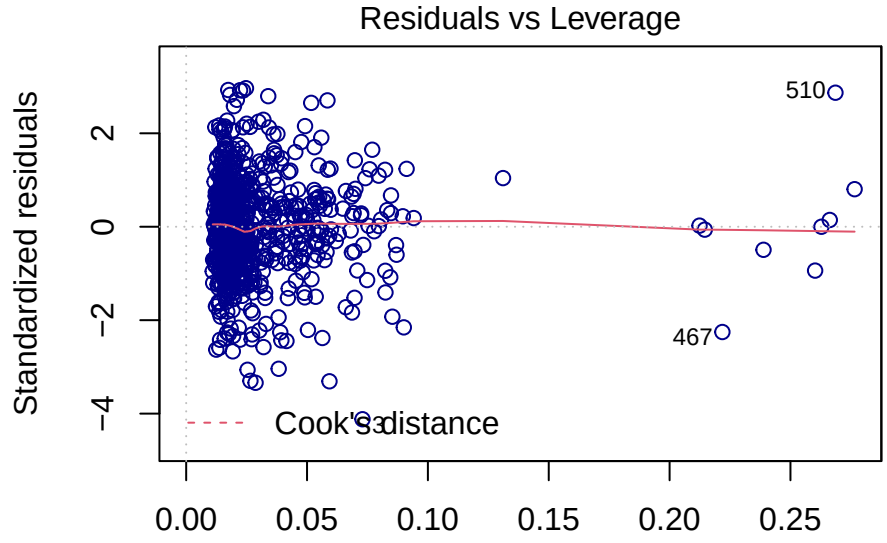
m(bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc + s

```
#normality is being reasonably satisfied aligned with what we saw during EDA

#variance and independence
plot(Model1,which=1,col=c("blue4"))
```



m(bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc + s

```
#both assumptions are being reasonably satisfied

#cooks distance, leverage, standardized residuals
plot(Model1,which=5,col=c("blue4"))
```



m(bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc + s

```r
#no outliers, leverage or influential points

#Lets fit the model to all the vars + interaction b/w smoke and race
Model1_inter1 <- lm(bwt.oz~parity+mage+med+mht+mpregwt+inc+smoke*mrace,data=smoking)
summary(Model1_inter1)
```

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mage + med + mht + mpregwt + inc +
##     smoke * mrace, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.892  -9.727  -0.478  10.277  49.779
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.04855   18.05668   1.830 0.067564 .
## parity          0.78586    0.40219   1.954 0.051039 .
## mage           -0.02935    0.13421  -0.219 0.826941
## med1            7.91925    7.87760   1.005 0.315048
## med2           10.41515    7.80018   1.335 0.182158
## med3            8.54184    8.10401   1.054 0.292174
## med4           10.72528    7.83389   1.369 0.171338
## med5            9.75695    7.86738   1.240 0.215256
## med6           -1.29493   11.44194  -0.113 0.909920
## mht             1.00614    0.27081   3.715 0.000216 ***
## mpregwt         0.10885    0.03294   3.304 0.000992 ***
## inc1            3.40576    3.58838   0.949 0.342839
## inc2            5.37770    3.60111   1.493 0.135722
## inc3            2.07149    3.63972   0.569 0.569417
## inc4            2.76973    3.72560   0.743 0.457428
## inc5            2.25508    3.76255   0.599 0.549103
## inc6            1.19257    4.02392   0.296 0.767020
## inc7            2.10175    3.73244   0.563 0.573514
## inc8            3.31021    5.43253   0.609 0.542469
## inc9           -1.09287    5.07721  -0.215 0.829624
## smoke1         -9.62129    1.36970  -7.024 4.44e-12 ***
## mrace6          0.84157    4.01796   0.209 0.834146
## mrace7        -10.14706    2.11689  -4.793 1.94e-06 ***
## mrace8         -5.57213    3.64341  -1.529 0.126548
## mrace9         -0.02785    4.97336  -0.006 0.995534
## smoke1:mrace6  14.53007    8.18930   1.774 0.076380 .
## smoke1:mrace7   2.40754    2.96941   0.811 0.417721
## smoke1:mrace8  -7.47122    6.70934  -1.114 0.265788
## smoke1:mrace9 -13.67585   10.96409  -1.247 0.212623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.73 on 840 degrees of freedom
## Multiple R-squared:  0.1692, Adjusted R-squared:  0.1415
## F-statistic: 6.109 on 28 and 840 DF,  p-value: < 2.2e-16
```

```r
#F-test to understand if our interaction is significant
anova(Model1,Model1_inter1)
```

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc +
##     smoke
## Model 2: bwt.oz ~ parity + mage + med + mht + mpregwt + inc + smoke *
##     mrace
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    844 236891
## 2    840 234972  4    1919.6 1.7156 0.1444
```

```r
#interaction is not significant

#check multicollinearity with vif
vif(Model1)
```

```
##    parity    mrace6    mrace7    mrace8    mrace9      mage      med1      med2
##  1.759380  1.078532  1.229877  1.135494  1.035518  1.817370 24.020507 42.828416
##      med3      med4      med5      med6       mht   mpregwt      inc1      inc2
## 10.144138 33.308799 28.043925  1.820738  1.444099  1.451002  5.796681  5.623822
##      inc3      inc4      inc5      inc6      inc7      inc8      inc9    smoke1
##  5.415298  4.555957  4.392344  3.082723  4.812058  1.654457  1.879475  1.072168
```

```r
#med has very high vif which is problematic

#Let's remove med and see if it improves our model
Model2 <- lm(bwt.oz~parity+mrace+mage+mht+mpregwt+inc+smoke,data=smoking)
summary(Model2)
```

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + mage + mht + mpregwt +
##     inc + smoke, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.170  -9.595  -0.209  10.568  50.357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.32918   16.18734   2.739 0.006301 **
## parity       0.67507    0.37851   1.783 0.074867 .
## mrace6       3.25494    3.50166   0.930 0.352872
## mrace7      -9.11563    1.57046  -5.804 9.11e-09 ***
## mrace8      -7.57216    3.08498  -2.455 0.014307 *
## mrace9      -2.43751    4.41789  -0.552 0.581274
## mage        -0.03304    0.12786  -0.258 0.796156
## mht          0.98545    0.26477   3.722 0.000211 ***
## mpregwt      0.10611    0.03272   3.243 0.001230 **
## inc1         3.64192    3.57120   1.020 0.308112
```

```
## inc2         5.57834    3.58968   1.554 0.120558
## inc3         2.71324    3.61862   0.750 0.453583
## inc4         3.36644    3.70019   0.910 0.363185
## inc5         2.72551    3.74232   0.728 0.466633
## inc6         1.77708    4.00750   0.443 0.657561
## inc7         2.76569    3.70144   0.747 0.455154
## inc8         4.09821    5.39855   0.759 0.447985
## inc9        -0.59985    5.03275  -0.119 0.905153
## smoke1      -9.45824    1.16252  -8.136 1.44e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 850 degrees of freedom
## Multiple R-squared:  0.1573, Adjusted R-squared:  0.1395
## F-statistic: 8.814 on 18 and 850 DF,  p-value: < 2.2e-16
```

```
#doesn't improve the model

#Let's check for multicollinearity between med and inc as they intuitively seem to be correlated
Model1_inter2 <- lm(bwt.oz~parity+mage+med*inc+mht+mpregwt+smoke+mrace,data=smoking)
summary(Model1_inter2)
```

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mage + med * inc + mht + mpregwt +
##     smoke + mrace, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.973  -9.451   0.000   9.616  50.883
##
## Coefficients: (16 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.20999   24.36913   0.993 0.320779
## parity        0.70278    0.41530   1.692 0.090990 .
## mage         -0.04472    0.13682  -0.327 0.743883
## med1         11.50184   18.08623   0.636 0.524993
## med2         10.43042   18.03713   0.578 0.563240
## med3         25.51563   23.58399   1.082 0.279619
## med4         12.97202   17.81854   0.728 0.466821
## med5         24.69607   18.29287   1.350 0.177383
## med6        -46.85048   24.13909  -1.941 0.052624 .
## inc1         52.96336   32.01014   1.655 0.098399 .
## inc2         12.14805   20.34901   0.597 0.550685
## inc3         54.10766   24.14333   2.241 0.025291 *
## inc4         -6.20718    8.46978  -0.733 0.463856
## inc5        -24.58382   23.54247  -1.044 0.296691
## inc6        -15.21785    8.64572  -1.760 0.078760 .
## inc7         -7.60390    8.00578  -0.950 0.342498
## inc8         -9.91333    9.79228  -1.012 0.311669
## inc9        -20.26363    9.04527  -2.240 0.025347 *
## mht           1.07347    0.27740   3.870 0.000118 ***
## mpregwt       0.10802    0.03359   3.216 0.001350 **
## smoke1       -9.68104    1.19873  -8.076 2.43e-15 ***
```

```
## mrace6        3.30290    3.60865   0.915 0.360322
## mrace7       -8.16173    1.62579  -5.020 6.36e-07 ***
## mrace8       -7.64759    3.16139  -2.419 0.015781 *
## mrace9       -2.32944    4.50277  -0.517 0.605064
## med1:inc1   -46.76493   32.74743  -1.428 0.153665
## med2:inc1   -44.37692   32.71606  -1.356 0.175343
## med3:inc1   -65.13351   36.73234  -1.773 0.076574 .
## med4:inc1   -48.81608   32.79586  -1.488 0.137015
## med5:inc1   -58.09992   33.15474  -1.752 0.080088 .
## med6:inc1          NA         NA      NA       NA
## med1:inc2   -10.67125   21.69484  -0.492 0.622938
## med2:inc2    -1.10422   21.56015  -0.051 0.959166
## med3:inc2   -13.52611   27.21497  -0.497 0.619318
## med4:inc2    -4.26278   21.51980  -0.198 0.843028
## med5:inc2   -12.01453   22.08831  -0.544 0.586639
## med6:inc2    43.24555   31.79172   1.360 0.174122
## med1:inc3   -52.96044   25.45813  -2.080 0.037814 *
## med2:inc3   -47.52500   25.17082  -1.888 0.059372 .
## med3:inc3   -74.09245   30.22310  -2.452 0.014437 *
## med4:inc3   -47.52955   25.04560  -1.898 0.058090 .
## med5:inc3   -61.35101   25.57599  -2.399 0.016676 *
## med6:inc3          NA         NA      NA       NA
## med1:inc4    10.68279   11.62777   0.919 0.358511
## med2:inc4    16.13354   11.18647   1.442 0.149624
## med3:inc4    -2.88103   19.88318  -0.145 0.884828
## med4:inc4     8.95752   10.97047   0.817 0.414449
## med5:inc4          NA         NA      NA       NA
## med6:inc4          NA         NA      NA       NA
## med1:inc5    33.64251   25.17960   1.336 0.181893
## med2:inc5    29.55494   24.67328   1.198 0.231327
## med3:inc5    19.20520   29.63494   0.648 0.517131
## med4:inc5    29.40577   24.68464   1.191 0.233903
## med5:inc5    20.35363   25.05207   0.812 0.416772
## med6:inc5          NA         NA      NA       NA
## med1:inc6    22.91791   12.93745   1.771 0.076865 .
## med2:inc6    22.00631   11.81565   1.862 0.062900 .
## med3:inc6     4.56252   20.21240   0.226 0.821469
## med4:inc6    22.92984   11.46733   2.000 0.045881 *
## med5:inc6          NA         NA      NA       NA
## med6:inc6          NA         NA      NA       NA
## med1:inc7    38.06045   12.84429   2.963 0.003134 **
## med2:inc7    12.69364   10.81518   1.174 0.240867
## med3:inc7    -2.62488   19.18417  -0.137 0.891203
## med4:inc7    10.73761   10.79948   0.994 0.320389
## med5:inc7          NA         NA      NA       NA
## med6:inc7          NA         NA      NA       NA
## med1:inc8   -18.30953   20.56166  -0.890 0.373480
## med2:inc8    27.88089   14.03940   1.986 0.047382 *
## med3:inc8          NA         NA      NA       NA
## med4:inc8    19.85665   15.15947   1.310 0.190619
## med5:inc8          NA         NA      NA       NA
## med6:inc8          NA         NA      NA       NA
## med1:inc9          NA         NA      NA       NA
## med2:inc9    33.17773   13.14133   2.525 0.011771 *
```

```
## med3:inc9          NA         NA      NA       NA
## med4:inc9    33.55620   13.77917   2.435 0.015096 *
## med5:inc9          NA         NA      NA       NA
## med6:inc9          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.56 on 806 degrees of freedom
## Multiple R-squared:  0.2187, Adjusted R-squared:  0.1586
## F-statistic: 3.639 on 62 and 806 DF,  p-value: < 2.2e-16
```

```
#F-test to understand if our interaction is significant
anova(Model1,Model1_inter2)
```

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc +
##     smoke
## Model 2: bwt.oz ~ parity + mage + med * inc + mht + mpregwt + smoke +
##     mrace
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    844 236891
## 2    806 220962 38     15929 1.5291 0.02278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#interaction is significant and we might want to remove one of these variables

#stepwise for all other interactions
NullModel <- lm(bwt.oz~smoke+mrace,data=smoking)
FullModel <- lm(bwt.oz~smoke*mrace + smoke*med + smoke*inc +
                  med*inc + parity + mage + mht + mpregwt,
                data=smoking)
n <- nrow(smoking)
Model_step <- step(NullModel, scope = formula(FullModel),direction="both",trace=0, k = log(n))
Model_step$call
```

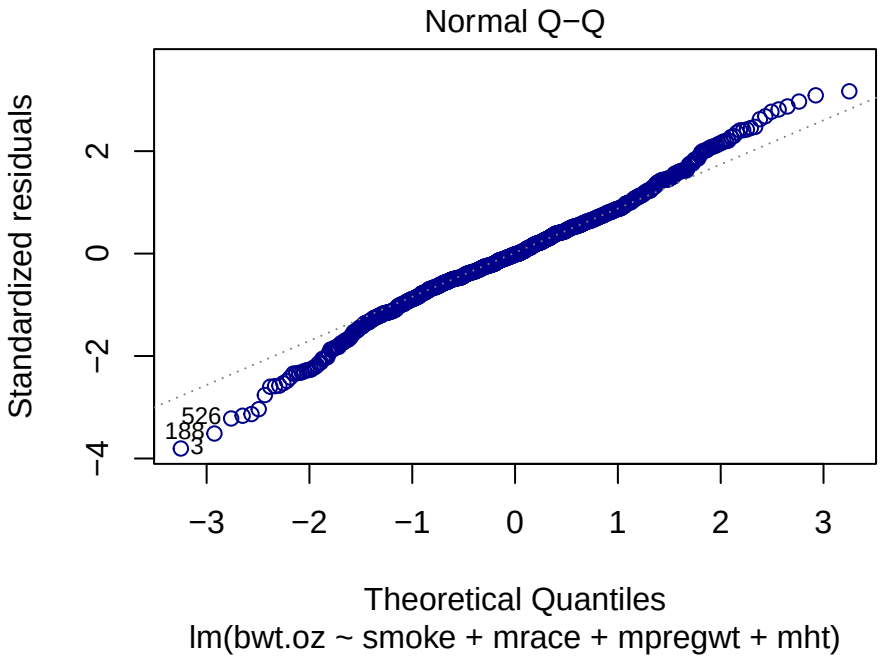
```
## lm(formula = bwt.oz ~ smoke + mrace + mpregwt + mht, data = smoking)
```

```
#lm(formula = bwt.oz ~ smoke + mrace + mpregwt + mht, data = smoking)

#making the final model
final_model <- lm(bwt.oz~smoke + mrace + mpregwt + mht, data = smoking)
summary(final_model)
```

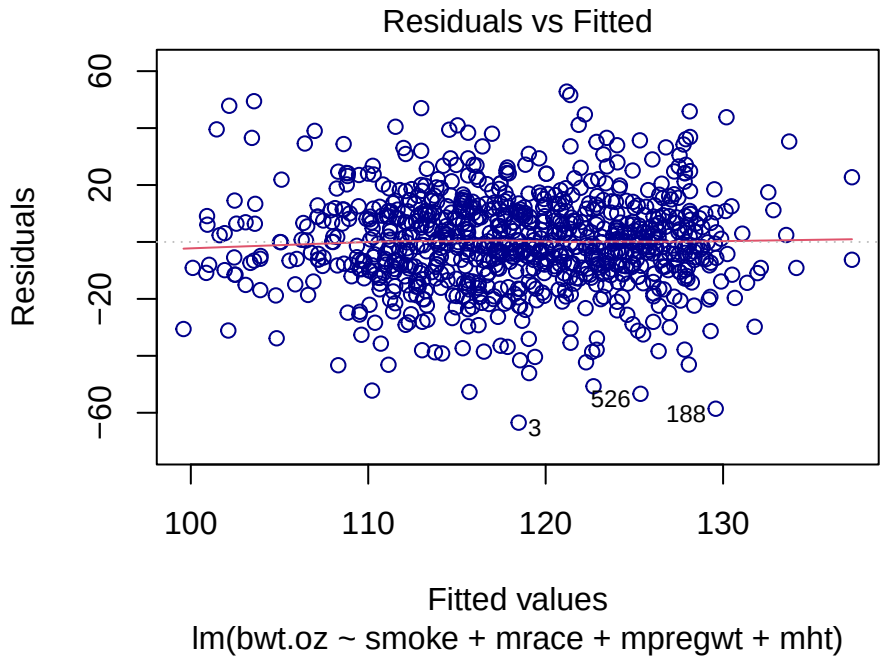
```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mrace + mpregwt + mht, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.478  -9.329  -0.221  10.015  52.811
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.24651   15.32152   3.475 0.000536 ***
## smoke1      -9.27445    1.15392  -8.037 3.02e-15 ***
## mrace6       3.62858    3.47071   1.045 0.296092
## mrace7      -8.19323    1.48938  -5.501 4.98e-08 ***
## mrace8      -8.14676    3.03959  -2.680 0.007498 **
## mrace9      -1.66704    4.39268  -0.380 0.704407
## mpregwt      0.11788    0.03189   3.696 0.000233 ***
## mht          0.87571    0.25977   3.371 0.000782 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 861 degrees of freedom
## Multiple R-squared:  0.1469, Adjusted R-squared:   0.14
## F-statistic: 21.19 on 7 and 861 DF,  p-value: < 2.2e-16
```

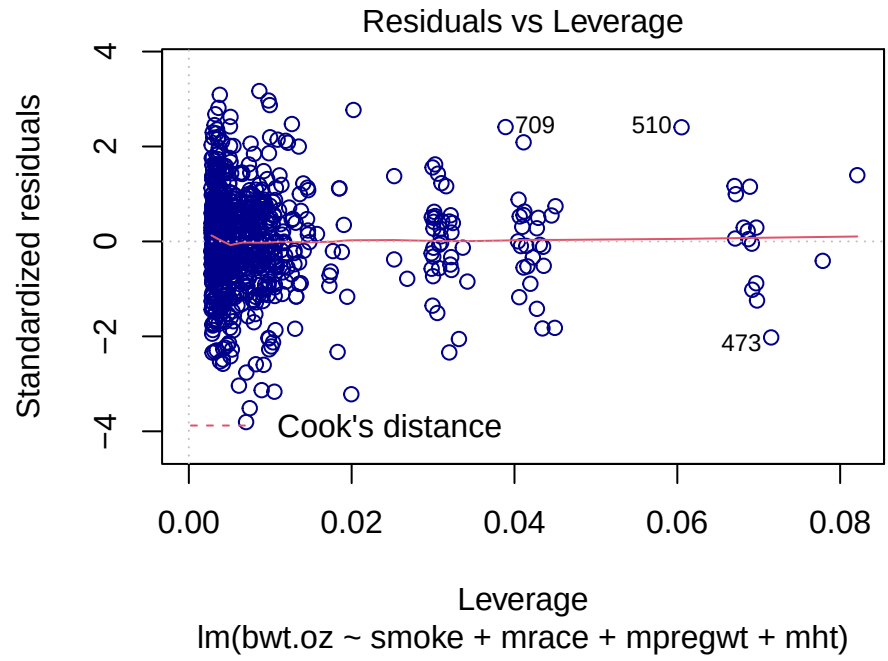
```
#let's check model assumptions of the final model
#Normality
plot(final_model,which=2,col=c("blue4"))
```




```
#normality is being reasonably satisfied aligned with what we saw during EDA

#variance and independence
plot(final_model,which=1,col=c("blue4"))
```




```
#both assumptions are being reasonably satisfied

#cooks distance, leverage, standardized residuals
plot(final_model,which=5,col=c("blue4"))
```




```
#few outliers - how do we extract outliers from this graph?

#confidence intervals
confint(final_model,level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept)  23.17461918 83.3184049
## smoke1      -11.53927673 -7.0096196
## mrace6       -3.18345964 10.4406137
## mrace7      -11.11646052 -5.2699973
## mrace8      -14.11263531 -2.1808879
## mrace9      -10.28865816  6.9545760
## mpregwt       0.05528321  0.1804827
## mht           0.36585416  1.3855620
```

```
#centering the vars to interpret the intercept
summ(final_model, center = TRUE)
```

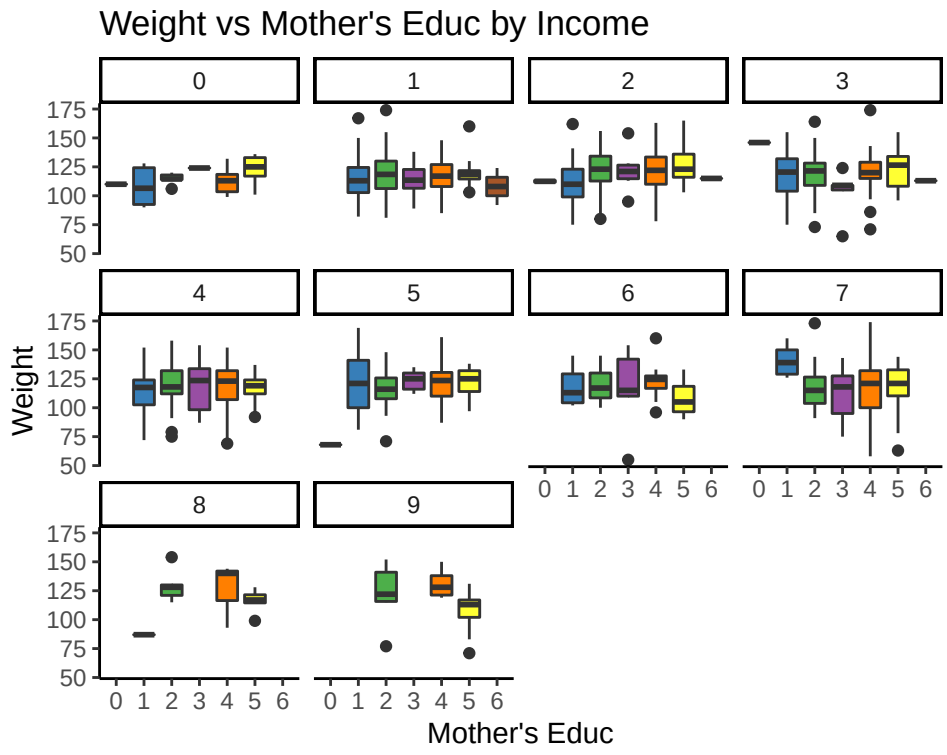
Observations	869
Dependent variable	bwt.oz
Type	OLS linear regression

F(7,861)	21.19
R^2	0.15
Adj. R^2	0.14

	Est.	S.E.	t val.	p
(Intercept)	124.50	0.88	141.45	0.00
smoke	-9.27	1.15	-8.04	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
mpregwt	0.12	0.03	3.70	0.00
mht	0.88	0.26	3.37	0.00

Standard errors: OLS; Continuous predictors are mean-centered.

```
#plot for interaction between income and education
ggplot(smoking,aes(x=med, y=bwt.oz, fill=med)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Set1") +
  labs(title="Weight vs Mother's Educ by Income",x="Mother's Educ",y="Weight") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ inc,ncol=4)
```




```
#image 1 (histogram of mrace)
counts <- table(smoking$smoke, smoking$mrace)
barplot(counts, main="Distribution of Mothers' Race by Smoking Status",
        xlab="Race", col=c("lightblue","yellow"),
        legend = rownames(counts))
```