Problem 2.6

Stephen Blatti

October 12, 2017

2.6 Access the data from the url http://www.stat.berkeley.edu/users/statlabs/data/babies.data and store the info in an object named BABIES using the fctn read.table(). A description of the var's can be found at http://www.stat.berkeley.edu/users/statlabs/labs.html

These data are a subset from a much larger sstudy dealing with child health and developement.

```r
data_url <- "http://www.stat.berkeley.edu/users/statlabs/data/babies.data"
BABIES <- read.table(file = url(data_url), header = TRUE)
head(BABIES)
```

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

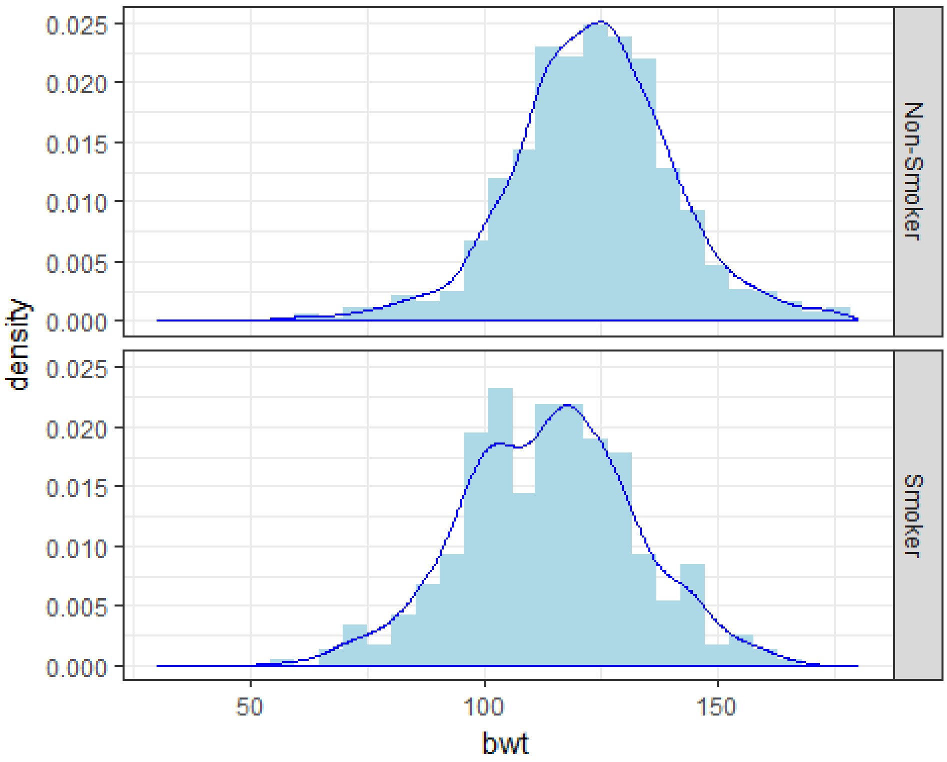
a) Create a "clean" data set that removes subjests if any observations on the subject are unknown. Note that bwt, gestation, parity, height, weight and smoke use values of 999, 999, 9, 99, 999, and 9 to denote unknown. Store the modified data set in an object named CLEAN.

```r
CLEAN <- with(data = BABIES, BABIES[bwt != 999 & gestation != 999 & parity !=
9 & age != 99 & height != 99 & weight != 999 & smoke != 9, ])
```

b) Use the info in CLEAN to create a density histogram of the birth weights of babies whose mothers have never smoked (smoke = 0) and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke (smoke=1). Make the range of the x-axis 30 to 180 (ounces) for both histograms. Superimpose a density curve over each histogram.

```r
library(ggplot2)
CLEAN$smoke <- factor(CLEAN$smoke, levels = 0:1, labels =c("Non-Smoker",
"Smoker"))
ggplot(data = CLEAN, aes(x = bwt, y = ..density..)) + geom_histogram(fill =
"lightblue") + geom_density(color = "blue") + facet_grid(smoke ~.) + xlim(30,
180) + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing missing values (geom_bar).
```



c) Based on the histograms in (b), characterize the distribution of baby birth weight for both non-smoking and smoking mothers.

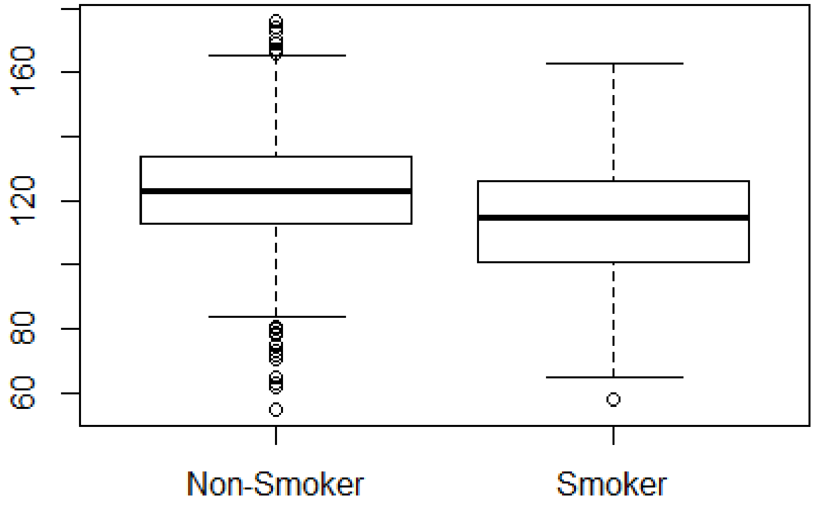
```
# Based on the density histograms in (b), the distributions of birth weights for
# both smoking and non-smoking mothers are unimodal and symmetric.
```

d) What is the mean weight difference between babies of smokers and non-smokers? Can you think of any reasons not to use the mean as a measure of center to compare birth weights in this problem?

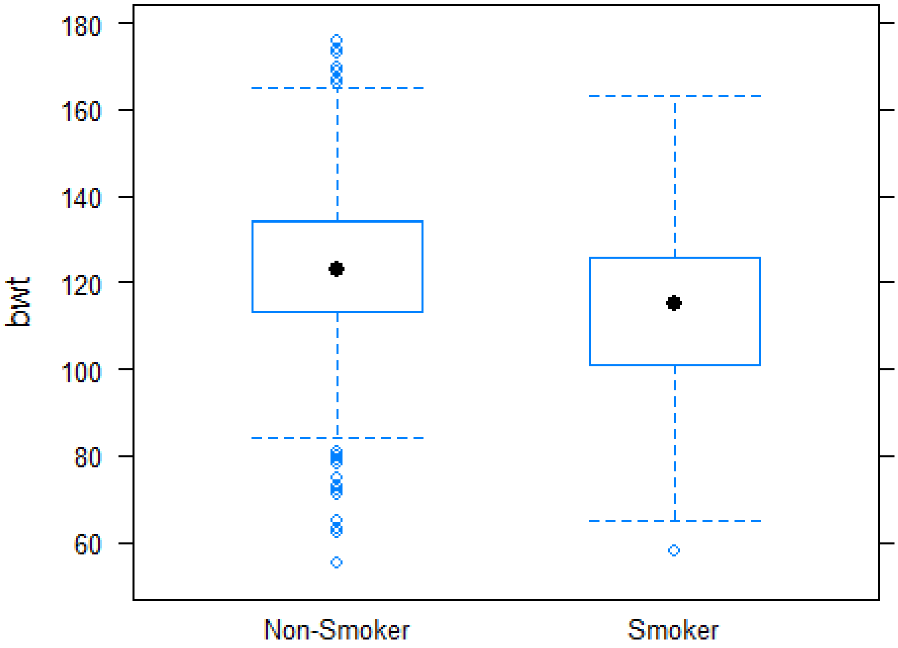
```
(result <- tapply(CLEAN$bwt, CLEAN$smoke, mean))

## Non-Smoker     Smoker
##   123.0853   113.8192

(DIFF <- result[1] - result[2])

## Non-Smoker
##   9.266143
```

e) Create side—by-side boxplots to compare the birth weights of babies whose mother's smoked and those who currently smoke. Use traditional graphics (boxplot()) as well as Trellis/lattice graphs to create the boxplots (bwplot()).

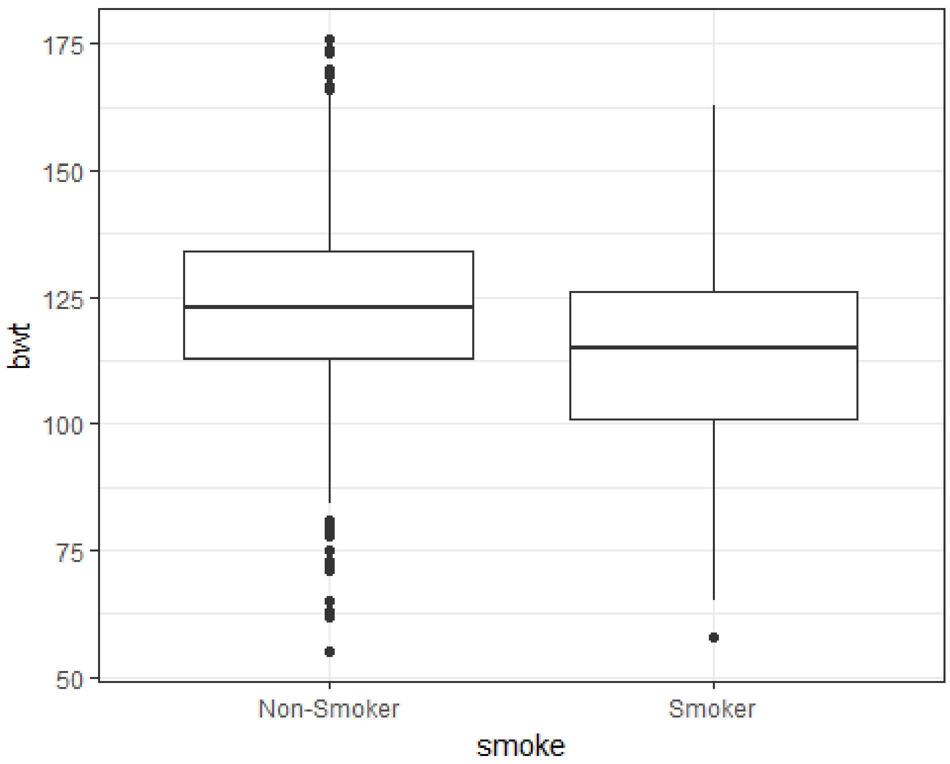
```
library(lattice)
boxplot(bwt ~ smoke, data = CLEAN)
```



```
bwplot(bwt ~ smoke, data = CLEAN)
```



```
ggplot(data = CLEAN, aes(x = smoke, y = bwt)) + geom_boxplot() + theme_bw()
```

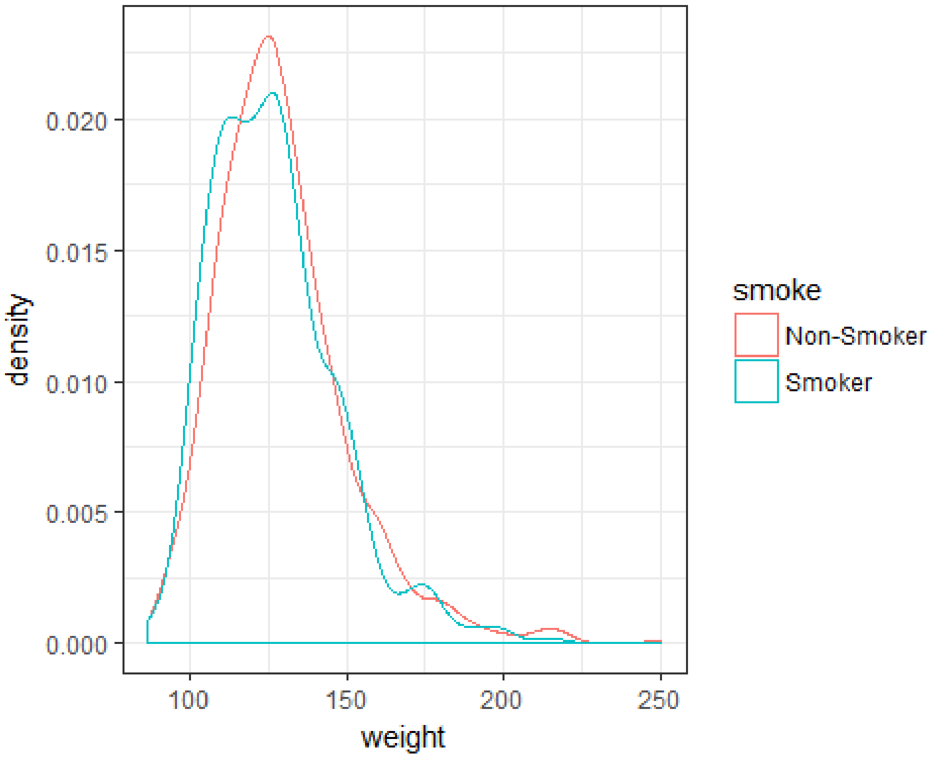


f) What is the median weight difference between babies who are firstborn and who are not?

```
(result <- tapply(CLEAN$bwt, CLEAN$parity, median))

##   0   1
## 120 118

(dif <- result[1] - result[2])

## 0
## 2
```

g) Create a single graph of the densities for pre-pregnancy weight for mothers who have never smoked and for mothers who currently smoke. Make sure both densities appear on the same graphics device and place a color coded legend in the top right corner of the graph.

```
ggplot(data = CLEAN, aes(x = weight, color = smoke)) + geom_density() +
theme_bw()
```



h) Characterize the pre-pregnancy distribution of weight for mothers who have never smoked and for mothers who currently smoke.

```
median(CLEAN$weight[CLEAN$smoke == "Smoker"])

## [1] 125

IQR(CLEAN$weight[CLEAN$smoke == "Smoker"])

## [1] 24.5

median(CLEAN$weight[CLEAN$smoke == "Non-Smoker"])

## [1] 126

IQR(CLEAN$weight[CLEAN$smoke == "Non-Smoker"])

## [1] 25

# The distribution of pre-pregnancy weight for mothers who are non-smokers
and mothers who currently smoke is unimodal and skewed to the right.
```

i) What is the mean pre-pregnancy weight difference between mothers who do not smoke and those who do? Can you think of any reasons not to use the mean as a measure of center to compare pre-pregnancy weights in this problem?

```
(result <- tapply(CLEAN$weight, CLEAN$smoke, mean))

## Non-Smoker     Smoker
##   129.4797   126.9194
```

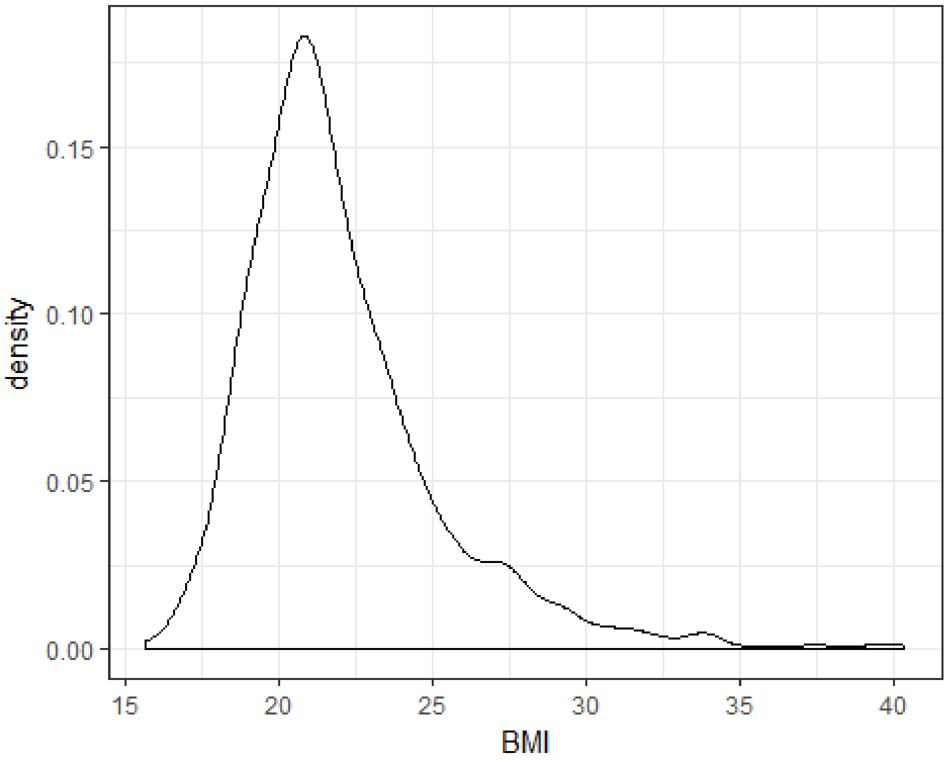
```
(dif <- result[1] - result[2])

## Non-Smoker
##    2.56033
```

j) Compute the body weight index (BMI) for each mother in CLEAN. Recall that BMI is defined as kg/m2 (0.0254 m = 1 in., and 0.45359 kg = 1 lb.). Add the variables weight in kg, height in m, and BMI to CLEAN and store the result in CLEANP.

```
CLEANP <- within(data = CLEAN, expr = {
  weight_SU = 0.45359 * weight
  height_SU = 0.0254 * height
  BMI = weight_SU / height_SU^2
})
head(CLEANP)

##   bwt gestation parity age height weight      smoke      BMI height_SU
## 1 120       284      0  27     62    100 Non-Smoker 18.28996    1.5748
## 2 113       282      0  33     64    135 Non-Smoker 23.17234    1.6256
## 3 128       279      0  28     64    115     Smoker 19.73940    1.6256
## 5 108       282      0  23     67    125     Smoker 19.57746    1.7018
## 6 136       286      0  25     62     93 Non-Smoker 17.00966    1.5748
## 7 138       244      0  33     62    178 Non-Smoker 32.55612    1.5748
##   weight_SU
## 1  45.35900
## 2  61.23465
## 3  52.16285
## 5  56.69875
## 6  42.18387
## 7  80.73902
```

k) Characterize the distribution of BMI.

```
median(CLEANP$BMI)

## [1] 21.28422

IQR(CLEANP$BMI)

## [1] 3.408279

ggplot(data = CLEANP, aes(x = BMI)) + geom_density() + theme_bw()
```



```
# The distribution of BMI is unimodal skewed to the right
```

l) Group pregnant mothers according to their BWI quartile. Find the mean and standard deviation for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Find the median and IQR for baby birth weights in each quartile for mothers who have never smoked and those who currently smoke. Based on your answers, would you characterize birth weight in each group as relatively symmetric or skewed? Create histograms and densities of bwt conditioned on BWI quartiles and whether the mother smokes to verify your previous assertions about the shape.

```
values <- quantile(CLEANP$BMI)
CLEANP <- within(data = CLEANP, expr = {
  Quartiles <- cut(BMI, values, include.lowest = TRUE)
})
tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke), mean)

##             Non-Smoker   Smoker
## [15.7,19.9]   121.8125 110.5662
## (19.9,21.3]   123.4696 114.6754
## (21.3,23.3]   122.3552 117.4393
## (23.3,40.4]   124.4869 113.4020

tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke), sd)

##             Non-Smoker   Smoker
## [15.7,19.9]   15.34434 18.44459
## (19.9,21.3]   18.07778 17.66076
## (21.3,23.3]   16.70788 17.74355
## (23.3,40.4]   19.04755 18.82923
```

```
tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke), median)
```

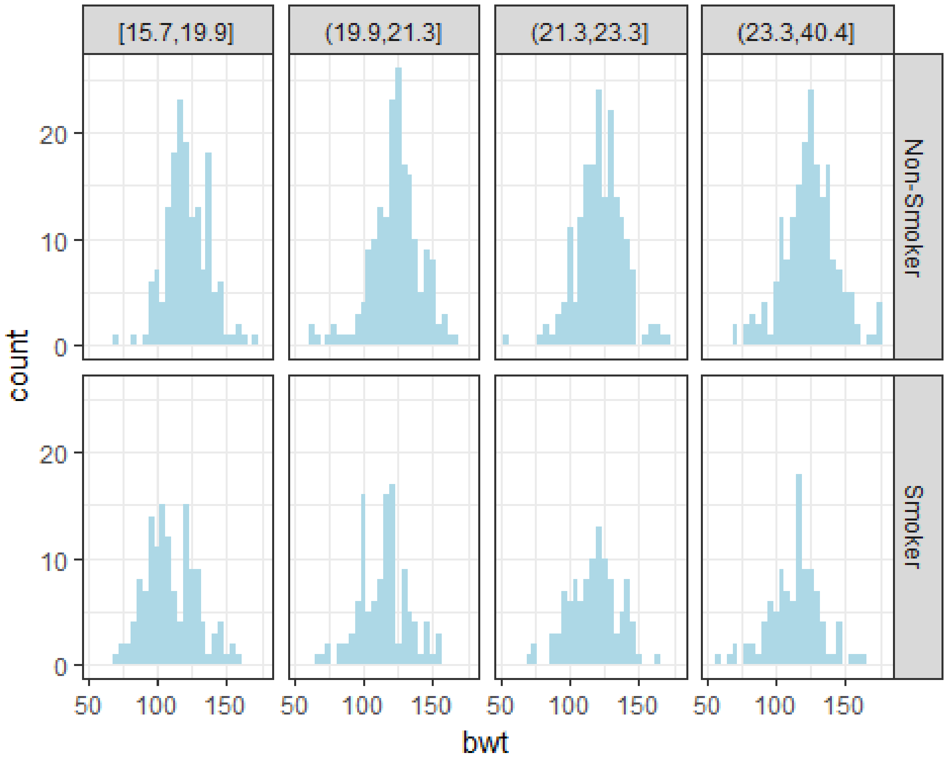
```
##             Non-Smoker Smoker
## [15.7,19.9]      120.5  108.5
## (19.9,21.3]      125.0  116.0
## (21.3,23.3]      122.0  118.0
## (23.3,40.4]      125.0  115.0
```

```
tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke), IQR)
```

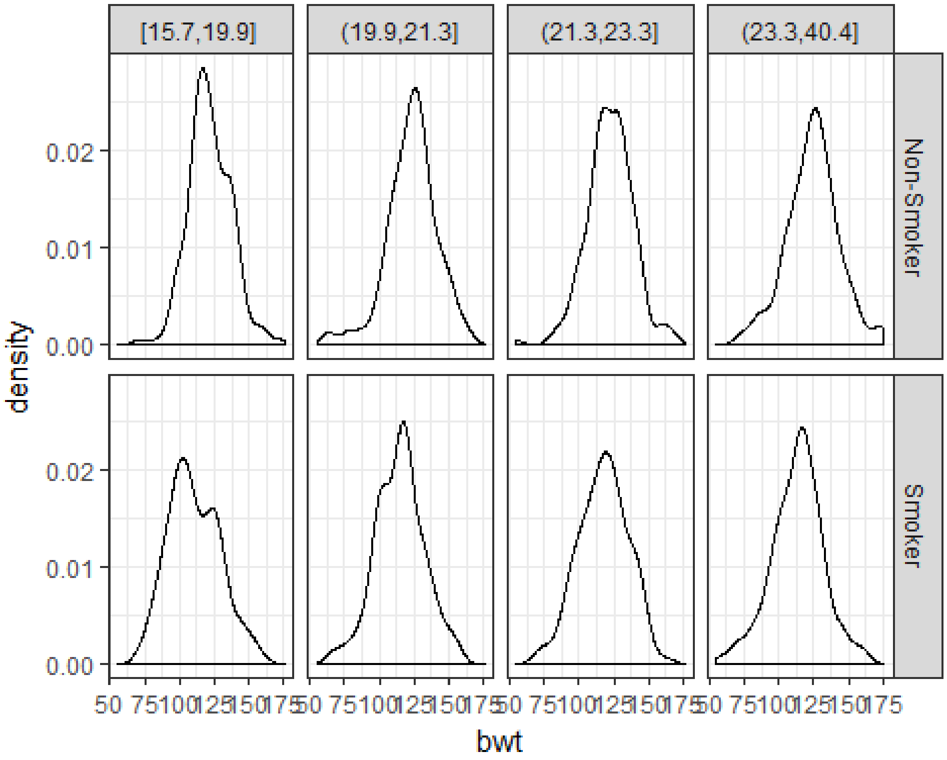
```
##             Non-Smoker Smoker
## [15.7,19.9]       20.0  25.25
## (19.9,21.3]       21.0  21.75
## (21.3,23.3]       19.5  24.00
## (23.3,40.4]       23.0  22.00
```

```
ggplot(data = CLEANP, aes(x = bwt)) + geom_histogram(fill="lightblue") +
theme_bw() + facet_grid(smoke ~ Quartiles)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



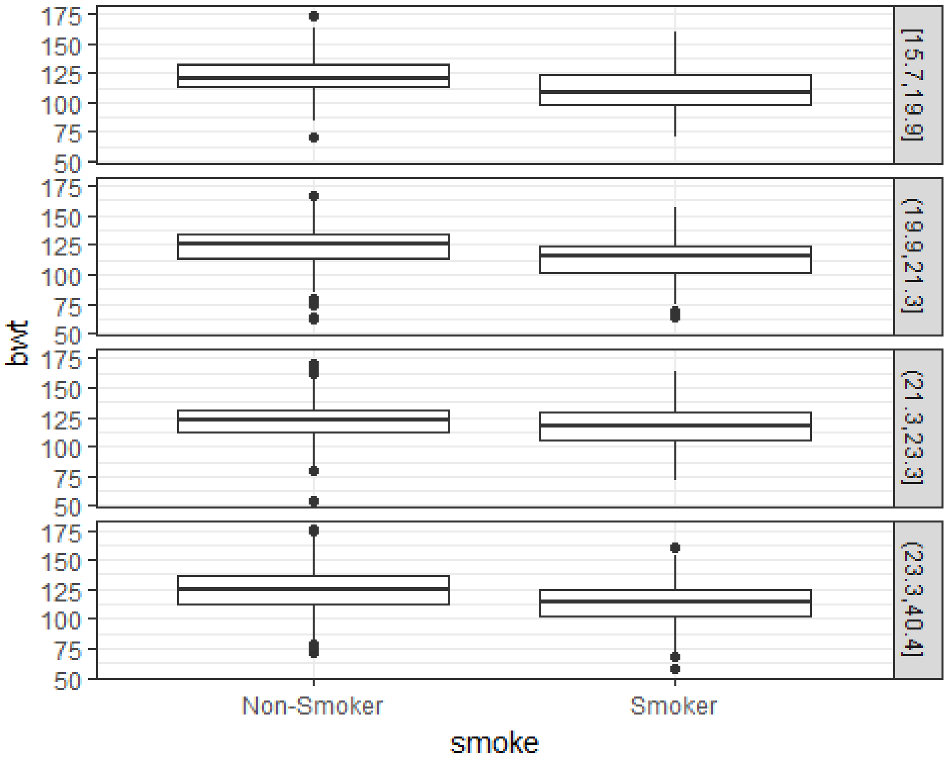
```
ggplot(data = CLEANP, aes(x = bwt)) + geom_density() + theme_bw() +
facet_grid(smoke ~ Quartiles)
```



```
# Birth weight in each quartile appears to be symmetric regardless of the
mother's smoking status.
```

m) Create side-by-side boxplots of bwt based on whether the mother smokes conditioned on BWI quartiles. Does this graph verify your findings in (I)?

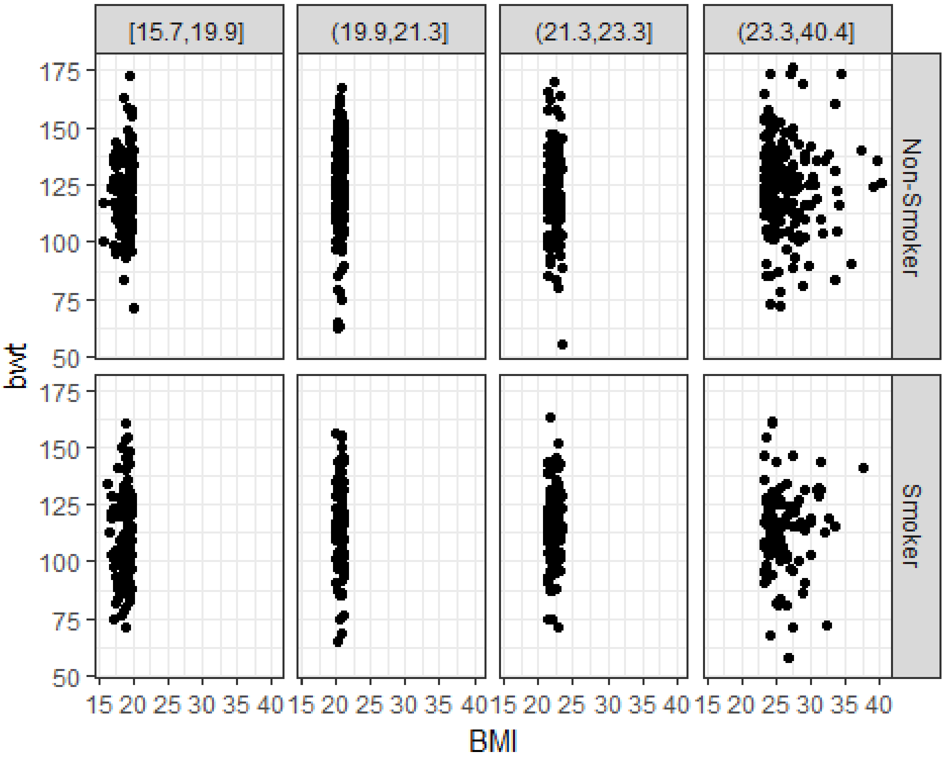
```
ggplot(data = CLEANP, aes(x = smoke, y = bwt)) + geom_boxplot() +
facet_grid(Quartiles~.) + theme_bw()
```



```
# The boxplots also suggest the distribution of bwt is symmetric for both smokers
# and non-smokers in each quartile.
```

n) Does it appear that BWI is related to the birth weight of a baby? Create a scatterplot Of birth weight (bwt) versus BMI while conditioning on BWI quartiles and whether the mother smokes to help answer the question.

```
ggplot(data = CLEANP, aes(x = BMI, y = bwt)) + geom_point() +
facet_grid(smoke ~ Quartiles) + theme_bw()
```



```
# There appears to be no association between birth weight and BMI.
```

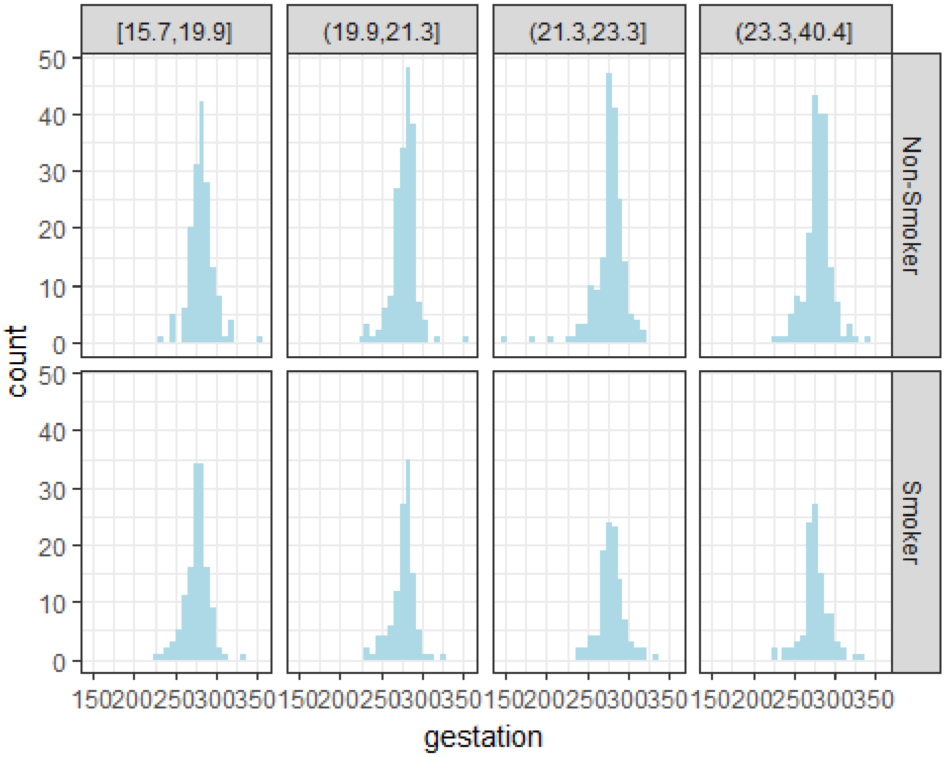
o) Replace baby birth weight (bwt) with gestation length (gestation) and answer questions (l), (m), and (n).

```
tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke), mean)

##               Non-Smoker   Smoker
## [15.7,19.9]     282.8938 277.2132
## (19.9,21.3]     279.0331 277.4649
## (21.3,23.3]     277.4372 279.6636
## (23.3,40.4]     280.4764 277.4412

tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke), sd)

##               Non-Smoker   Smoker
## [15.7,19.9]     14.57214 14.55330
## (19.9,21.3]     14.57810 14.40082
## (21.3,23.3]     20.08376 15.33890
## (23.3,40.4]     15.48780 16.77727

tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke), median)

##               Non-Smoker Smoker
## [15.7,19.9]          283    279
## (19.9,21.3]          281    280
## (21.3,23.3]          279    279
## (23.3,40.4]          281    277
tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke), IQR)
```

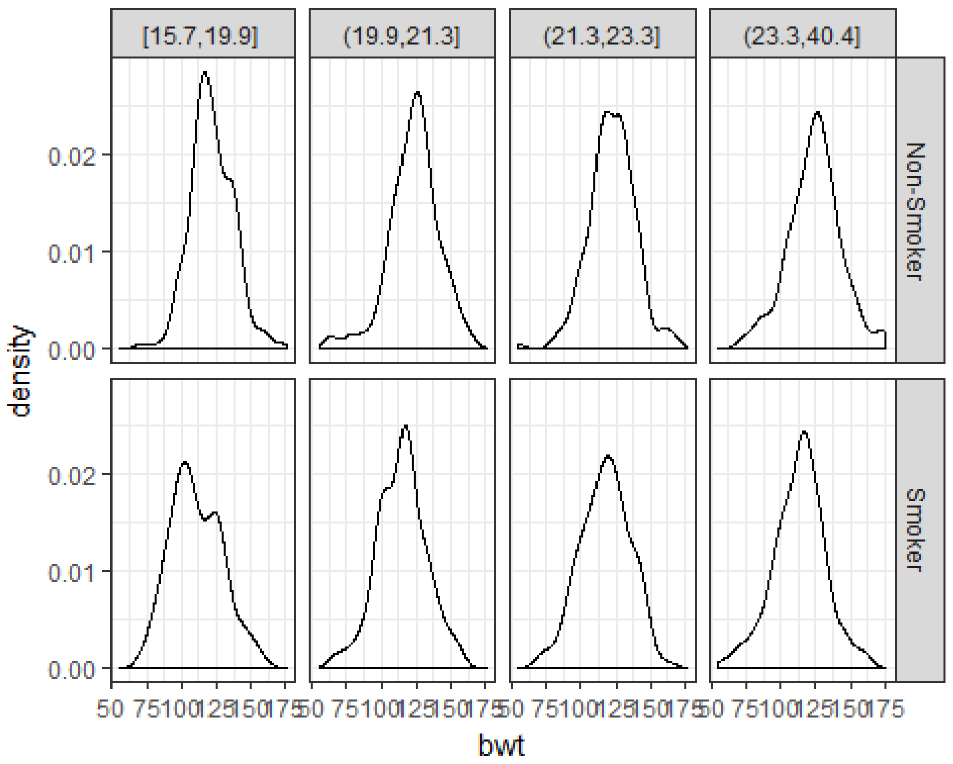
```
##             Non-Smoker Smoker
## [15.7,19.9]      14.25   16.0
## (19.9,21.3]      16.00   14.5
## (21.3,23.3]      15.00   17.0
## (23.3,40.4]      16.50   14.0
```

```
ggplot(data = CLEANP, aes(x = gestation)) + geom_histogram(fill =
"lightblue") +theme_bw() + facet_grid(smoke ~ Quartiles)
```

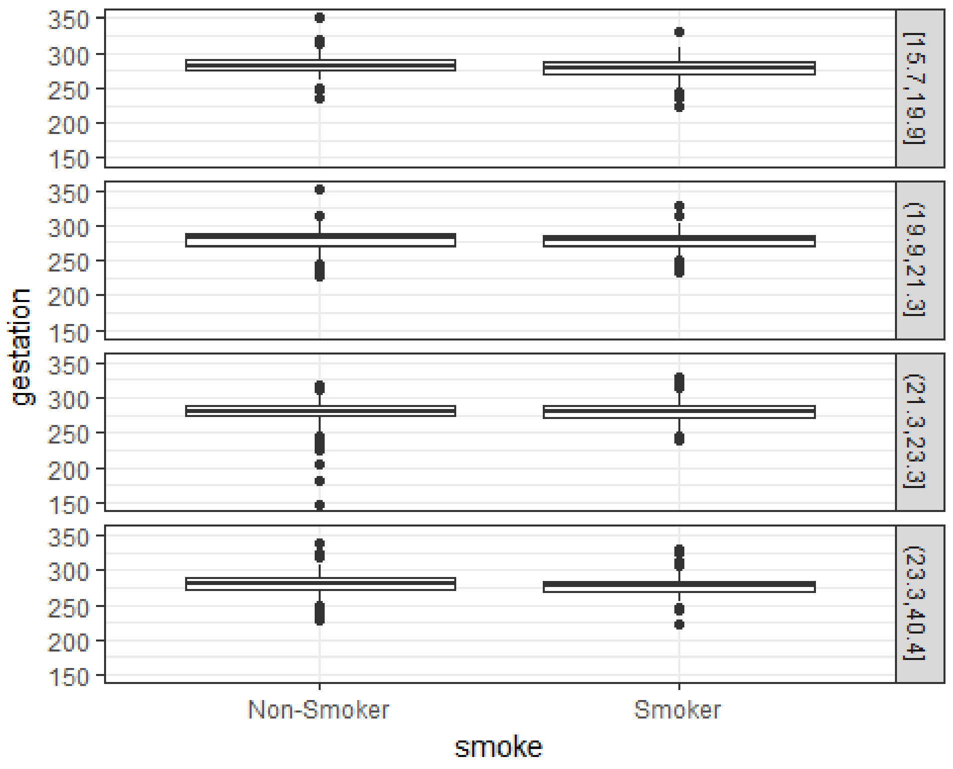
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



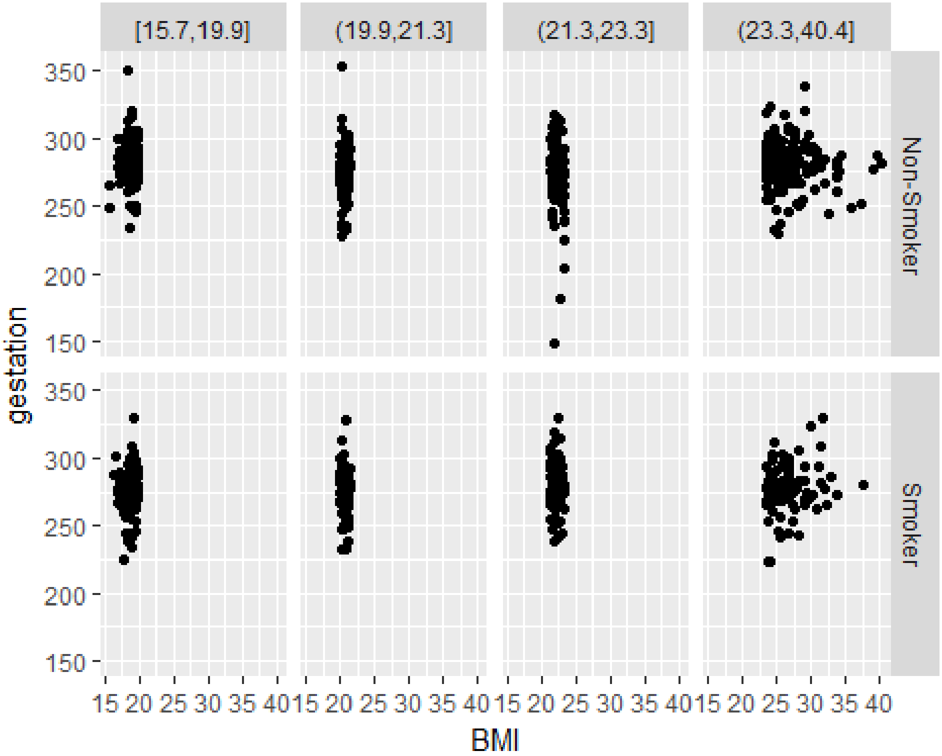
```
ggplot(data = CLEANP, aes(x = bwt)) + geom_density() + theme_bw() +
facet_grid(smoke ~ Quartiles)
```



```
ggplot(data = CLEANP, aes(x = smoke, y = gestation)) + geom_boxplot() +
facet_grid(Quartiles~.) + theme_bw()
```



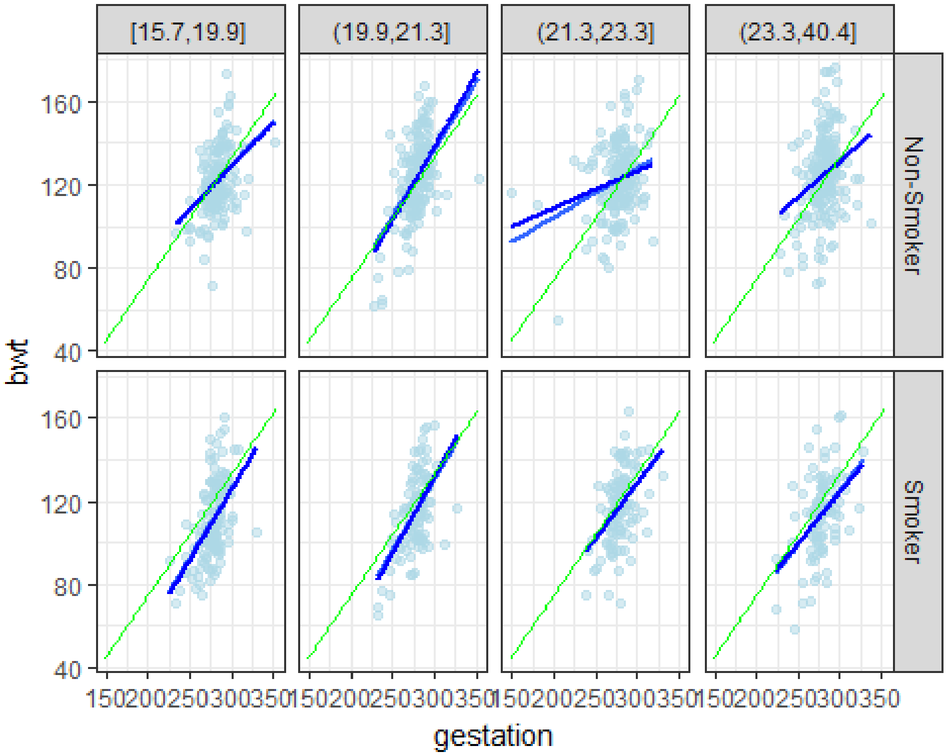
```
# Gestation in each quartile appears to be symmetric regardless of the
mother’s smoking status.
# The histograms, density and box plots confirm this
ggplot(data = CLEANP, aes(x = BMI, y = gestation)) + geom_point() +
facet_grid(smoke ~ Quartiles)
```



```
  #+ theme_bw()
# There doesn't appear to be any association between gestation and BMI
```

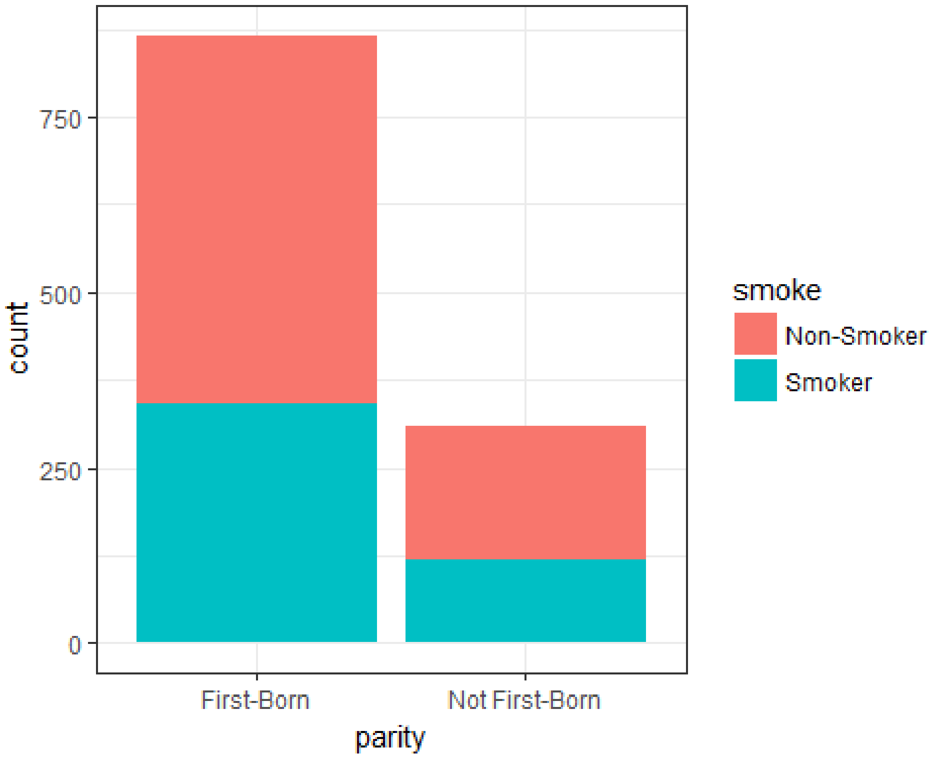
p) Create a scatterplot of bwt versus gestation conditioned on BWI quartiles and Whether the mother smokes. Fit straight lines to the data using lm(), lqs(), and r1m(); and display the lines in the scatterplots. What do you find interesting about the resulting graphs?

```
library(MASS)
fit_lqs <- lqs(bwt ~ gestation, data = CLEANP)
x_vals <- seq(min(CLEANP$gestation), max(CLEANP$gestation),length.out = 100)
df <- data.frame(gestation = x_vals)
df$bwt <- predict(fit_lqs, newdata = df)
ggplot(data = CLEANP, aes(x = gestation, y = bwt)) + geom_point(alpha = 0.5,
color ="lightblue") +
  facet_grid(smoke ~ Quartiles) +
  theme_bw() +
  stat_smooth(method = "lm", se = FALSE) +
  stat_smooth(method = "rlm", se = FALSE, color = "blue") +
  geom_line(data = df, color = "green")
```



q) Create a table of smoke by parity. Display the numerical results in a graph. What percent of mothers did not smoke during the pregnancy of their first child?

```
CLEANP$parity <- factor(CLEAN$parity, levels = 0:1, labels =c("First-Born",
"Not First-Born"))
table1 <- xtabs(~smoke + parity, data = CLEANP)
prop.table(table1, 2)

##               parity
## smoke          First-Born Not First-Born
##   Non-Smoker  0.6062356      0.6168831
##   Smoker      0.3937644      0.3831169

ggplot(data = CLEANP, aes(x = parity, fill = smoke)) + geom_bar() +
theme_bw()
```


count
750
500
250
0
First-Born
Not First-Born
parity
smoke
Non-Smoker
Smoker