```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
library(dplyr)
```

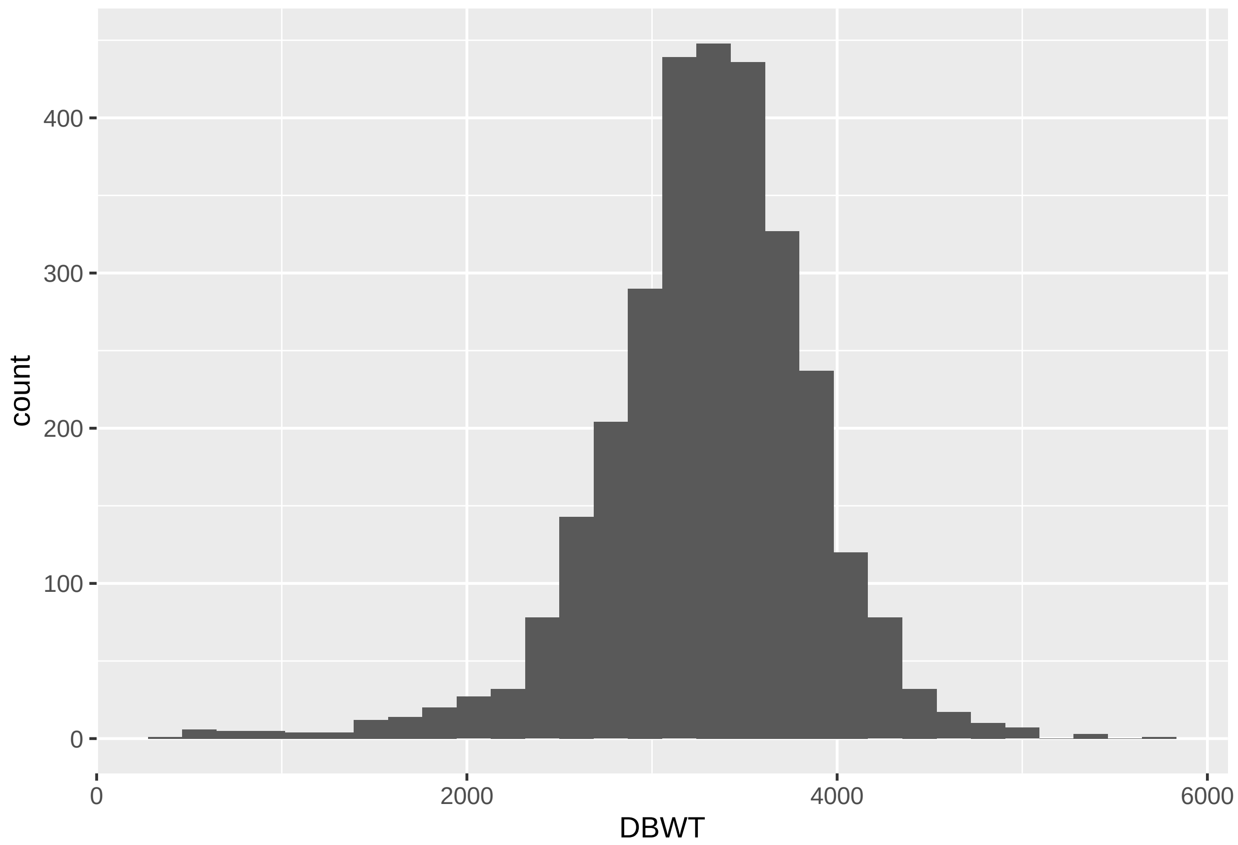
```
##
## Attaching package: ’dplyr’

## The following object is masked from ’package:car’:
##
##     recode

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
EDA_df <- read.csv("data/EDA.csv")
EDA_df$CIG_0_BIN <- factor(EDA_df$CIG_0_BIN)
EDA_df$PRECARE <- factor(EDA_df$PRECARE)
EDA_df$SEX <- factor(EDA_df$SEX)
EDA_df$RESTATUS <- factor(EDA_df$RESTATUS)
EDA_df$PAY <- factor(EDA_df$PAY)
EDA_df$NO_RISKS <- factor(EDA_df$NO_RISKS)
EDA_df$MRAVE6 <- factor(EDA_df$MRAVE6)
EDA_df$FRACE6  <- factor(EDA_df$FRACE6)
EDA_df$MEDUC <- factor(EDA_df$MEDUC)
EDA_df$FEDUC <- factor(EDA_df$FEDUC)
```

Response variable:

```
# response variable
ggplot(EDA_df, aes(x = DBWT)) +
  geom_histogram()
```

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

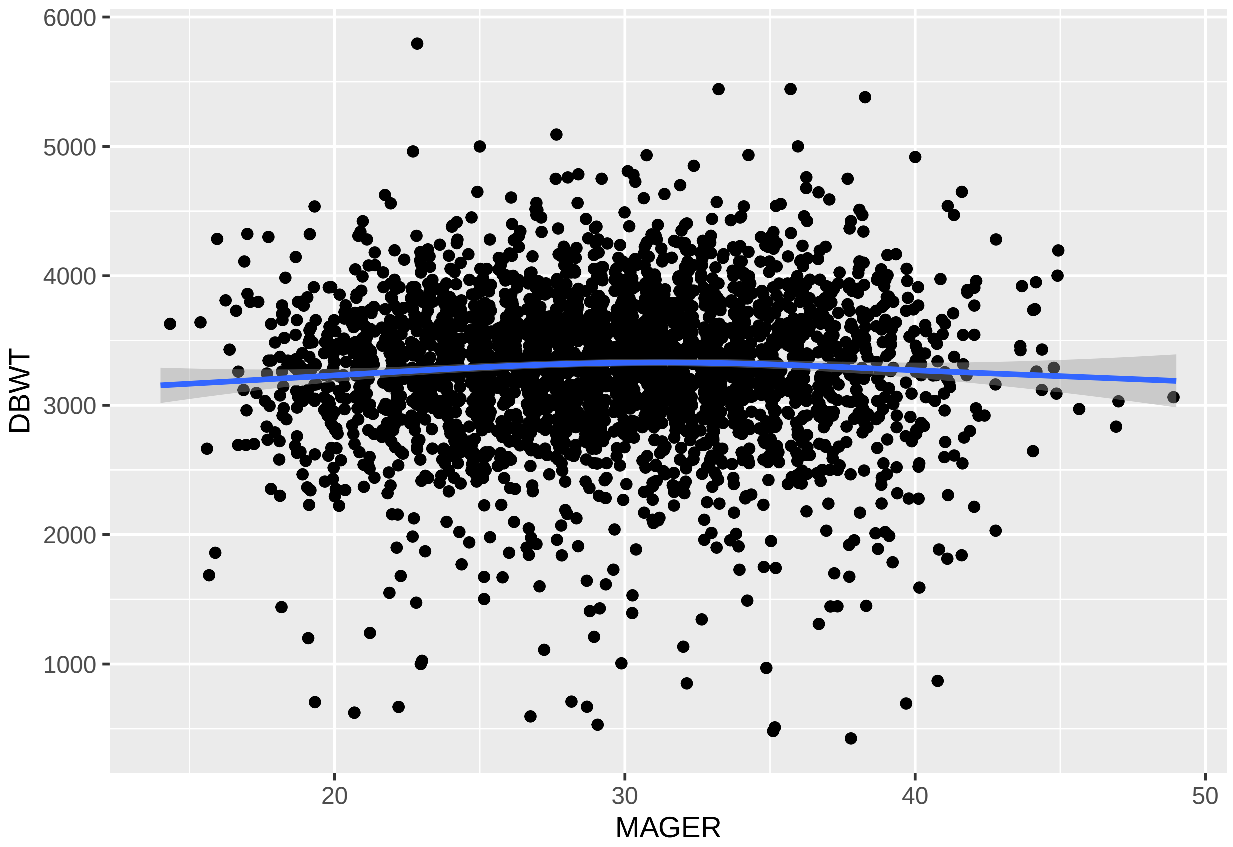


```r
# Measure of symmetry
DBWT_sym = (quantile(EDA_df$DBWT, 0.75) - median(EDA_df$DBWT)) /
  (median(EDA_df$DBWT) - quantile(EDA_df$DBWT, 0.25))
```

MAGER

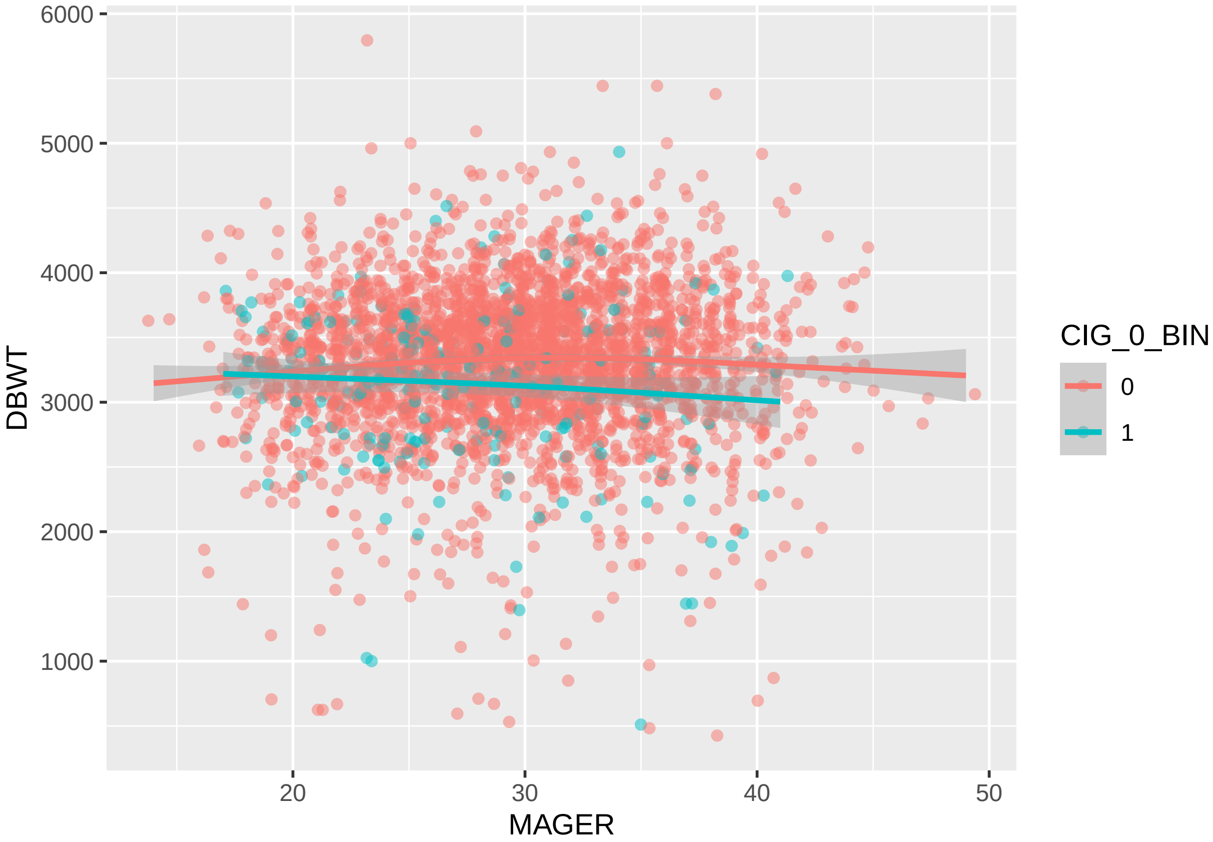
```r
ggplot(EDA_df, aes(x = MAGER, y = DBWT)) +
  geom_point(position = "jitter") +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```r
ggplot(EDA_df, aes(x = MAGER, y = DBWT)) +
  geom_point(position = "jitter", aes(colour = CIG_0_BIN), alpha = 0.5) +
  geom_smooth(aes(colour = CIG_0_BIN))
```

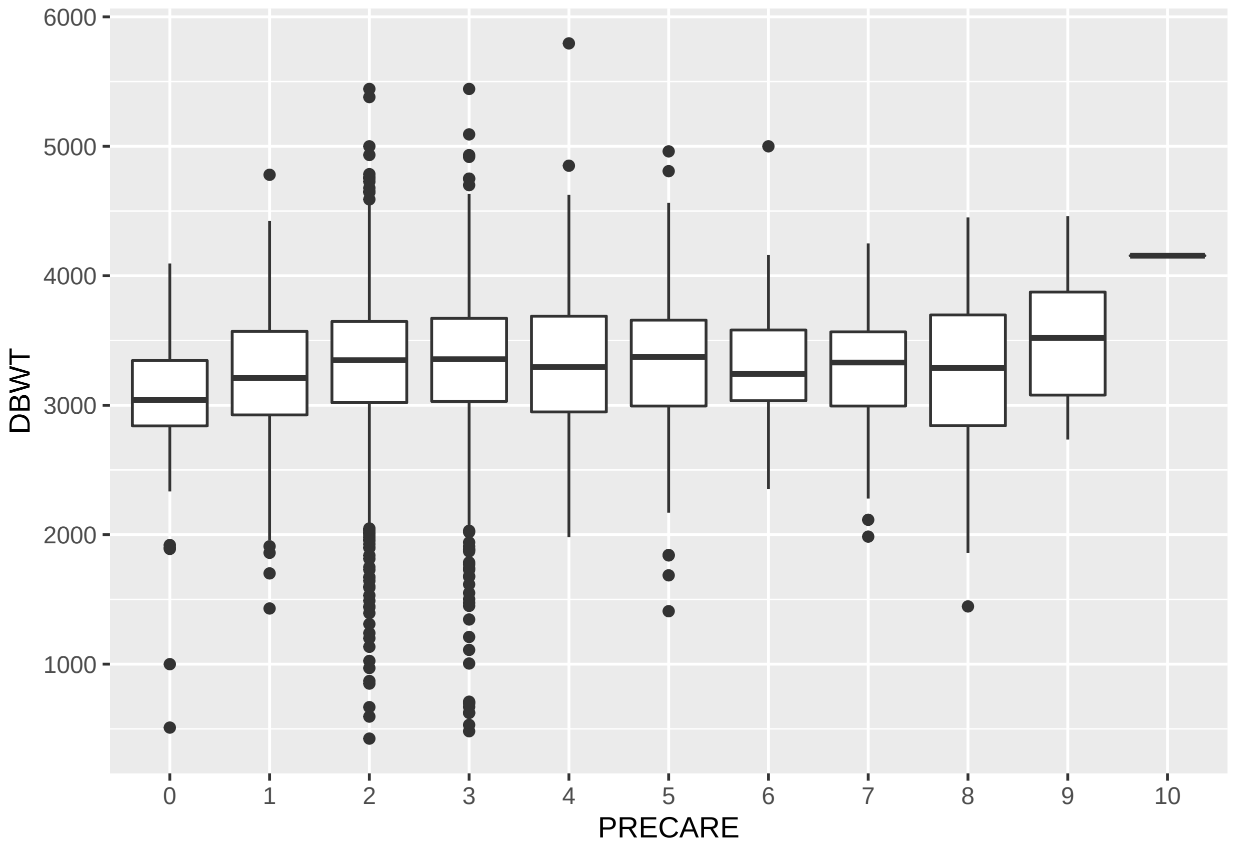
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Slightly negative slope for smoking mothers between `MAGER` and `DBWT`.

PRECARE

```
ggplot(EDA_df, aes(x = PRECARE, y = DBWT)) +
  geom_boxplot()
```

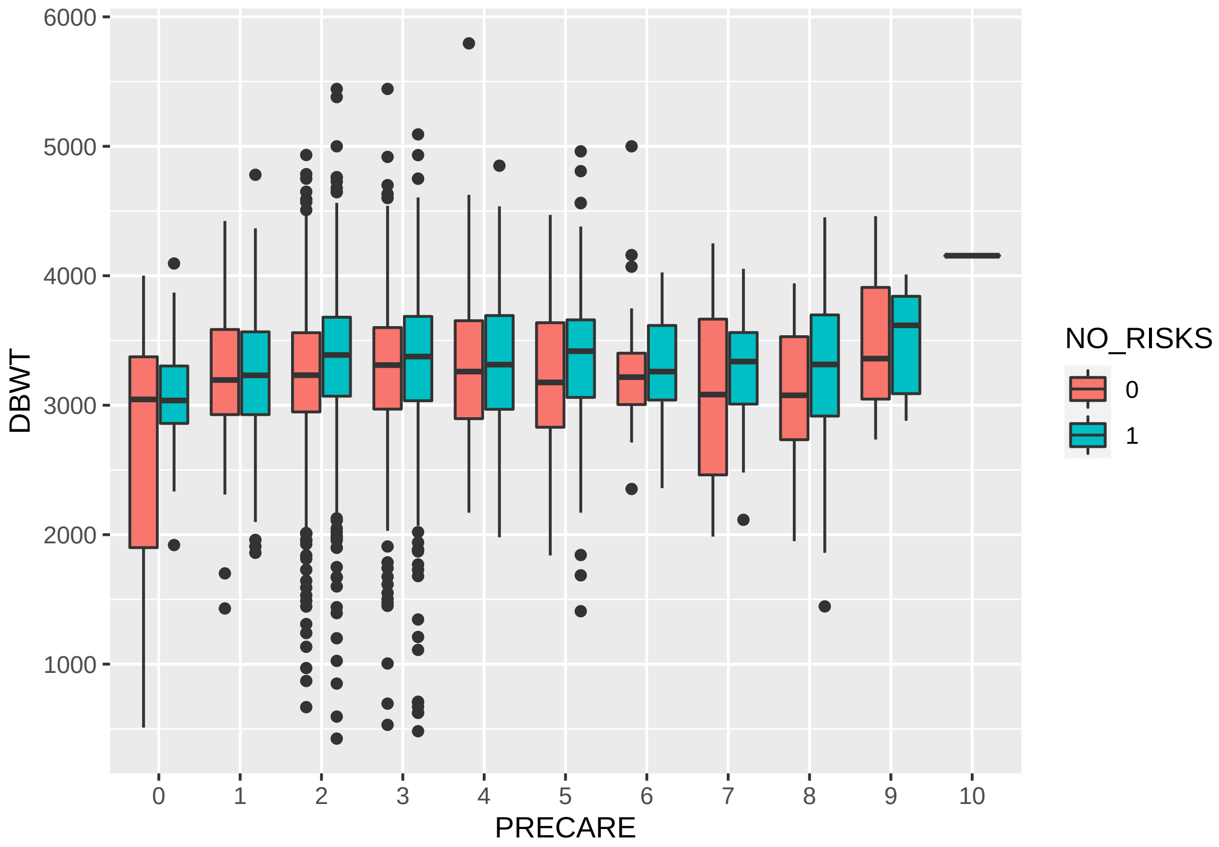


Higher `PRECARE` has higher `DBWT`.

```
sum(EDA_df$PRECARE == 10)
```

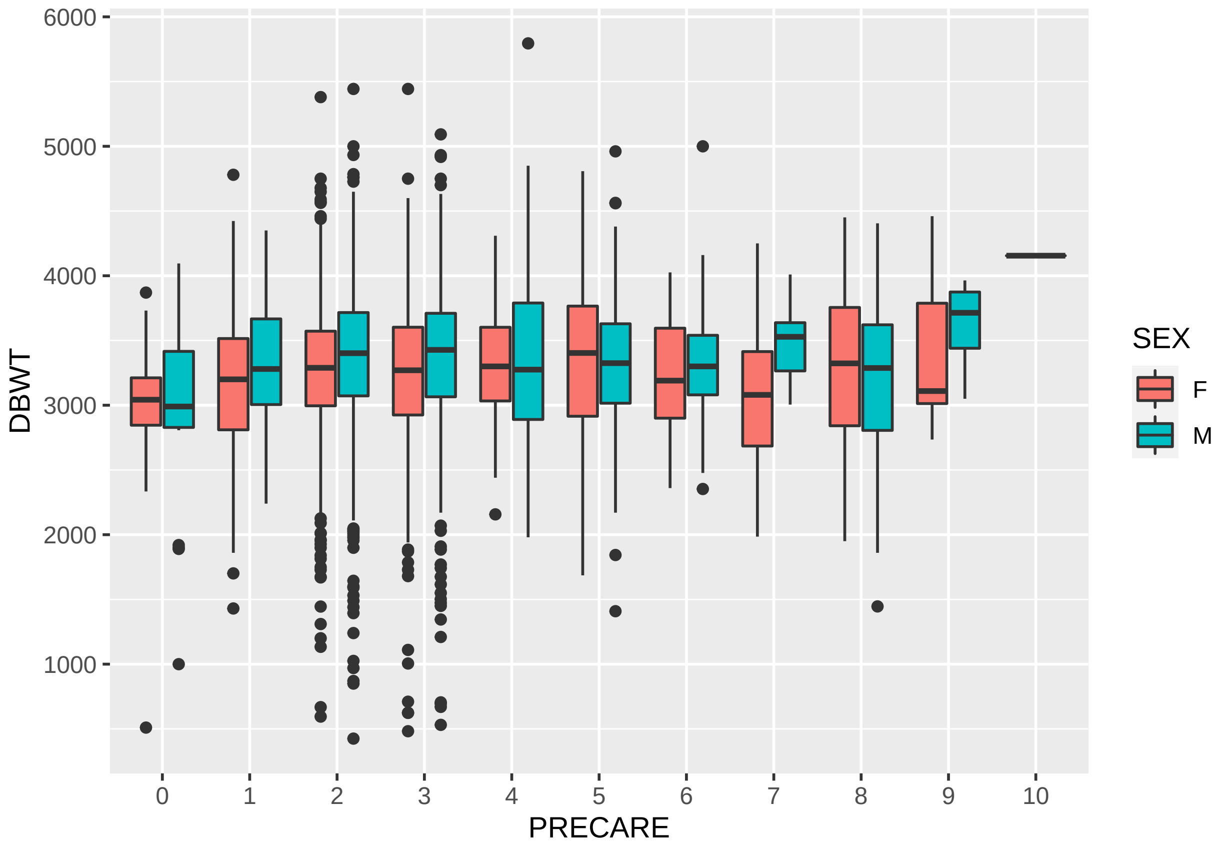
```
## [1] 1
```

```
ggplot(EDA_df, aes(x = PRECARE, y = DBWT)) +
  geom_boxplot(aes(fill = NO_RISKS))
```



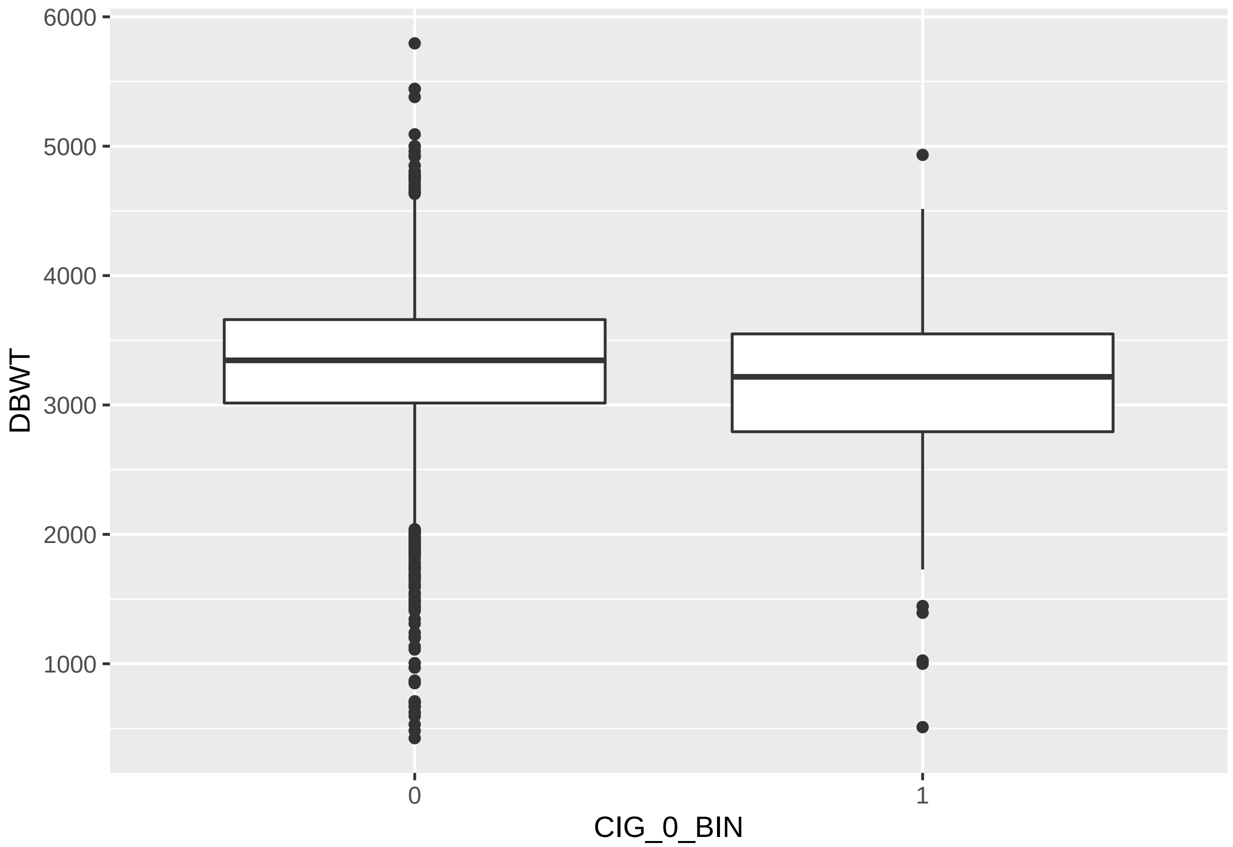
`NO_RISKS` matters more for higher `PRECARE`.

```
ggplot(EDA_df, aes(x = PRECARE, y = DBWT)) +
  geom_boxplot(aes(fill = SEX))
```



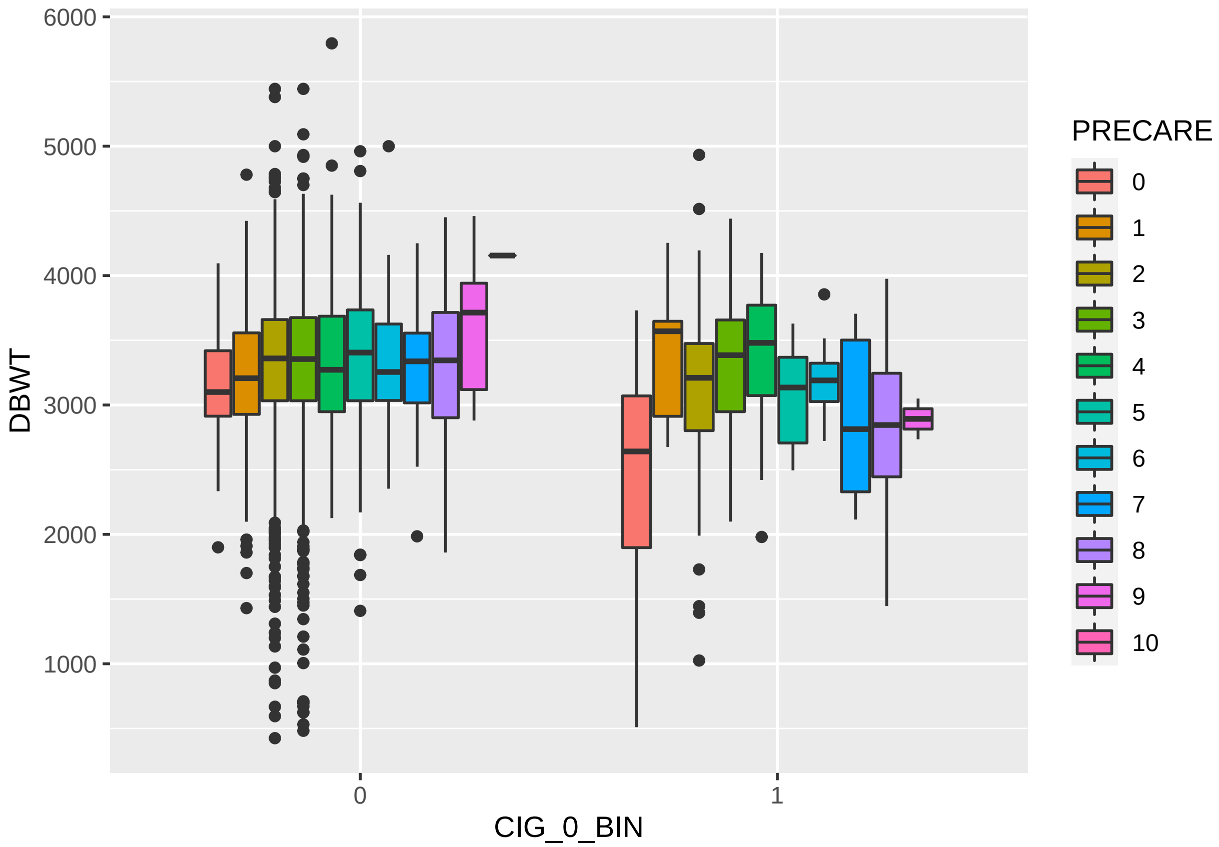
CIG_0

```
ggplot(EDA_df, aes(x = CIG_0_BIN, y = DBWT)) +
  geom_boxplot()
```



No smoking leads to higher `DBWT`.

```
ggplot(EDA_df, aes(x = CIG_0_BIN, y = DBWT)) +
  geom_boxplot(aes(fill = PRECARE))
```



`PRECARE` difference is more obvious in smoking mothers. But it might due to the relative smaller number of smoking mothers.

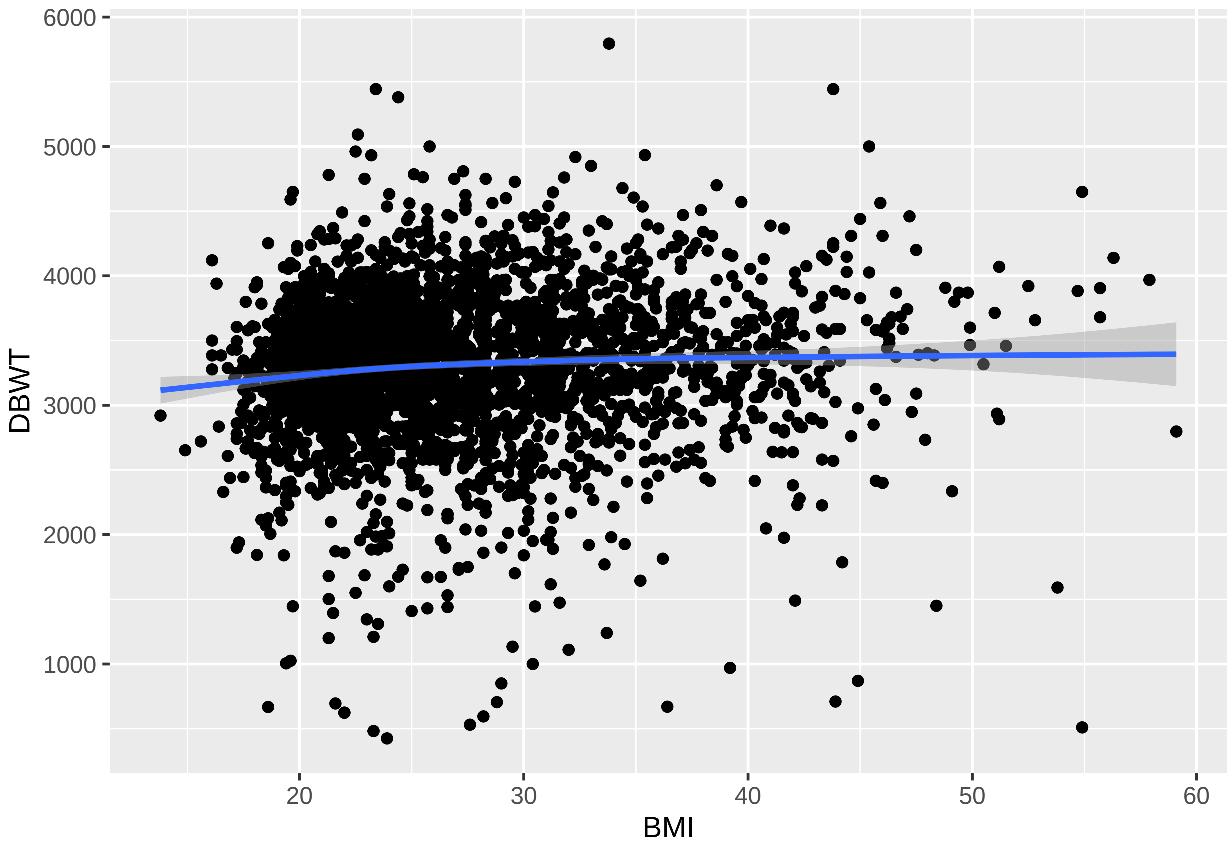
```
EDA_df %>% count(CIG_0_BIN)
```

```
##   CIG_0_BIN    n
## 1         0 2768
## 2         1  232
```

BMI:

```
ggplot(EDA_df, aes(x = BMI, y = DBWT)) +
  geom_point() +
  geom_smooth()
```

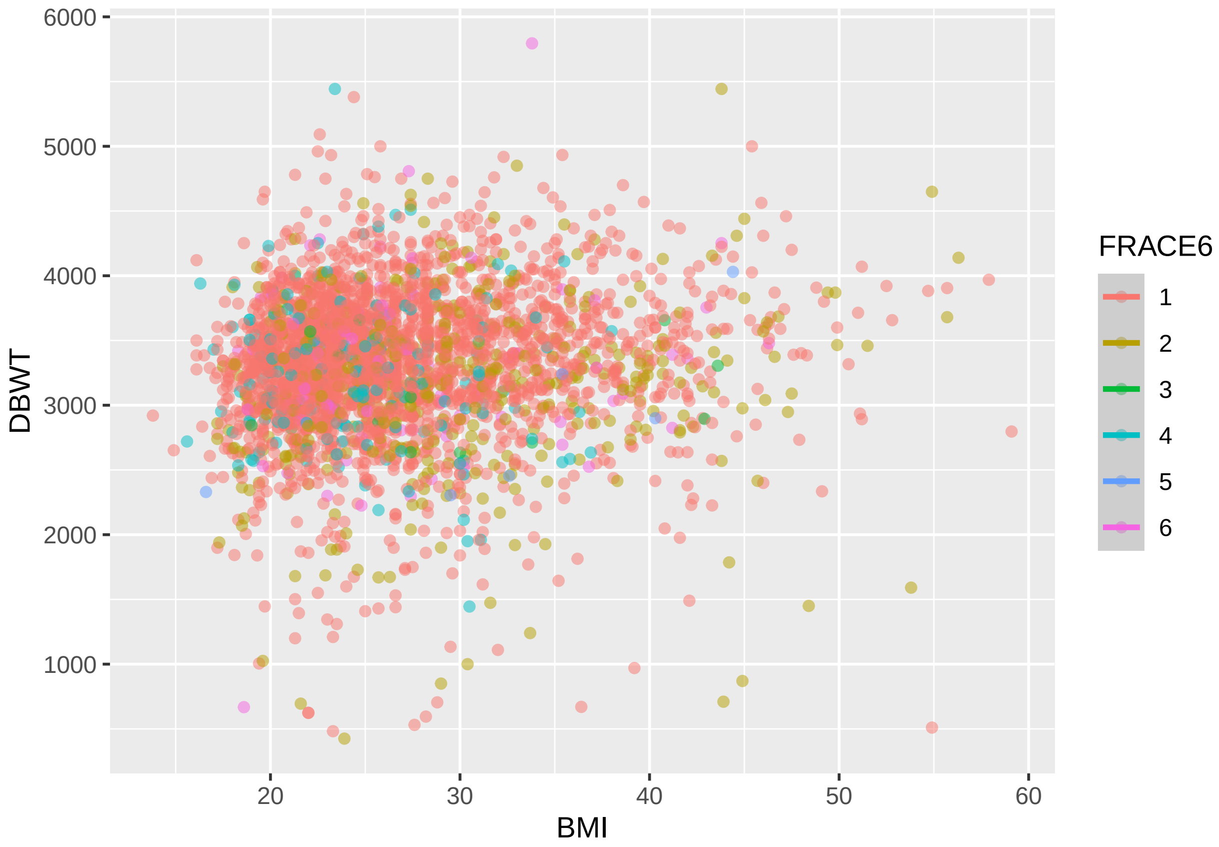
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
ggplot(EDA_df, aes(x = BMI, y = DBWT)) +
  geom_point(aes(colour = FRACE6), alpha = 0.5) +
  geom_smooth(aes(colour = FRACE6))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

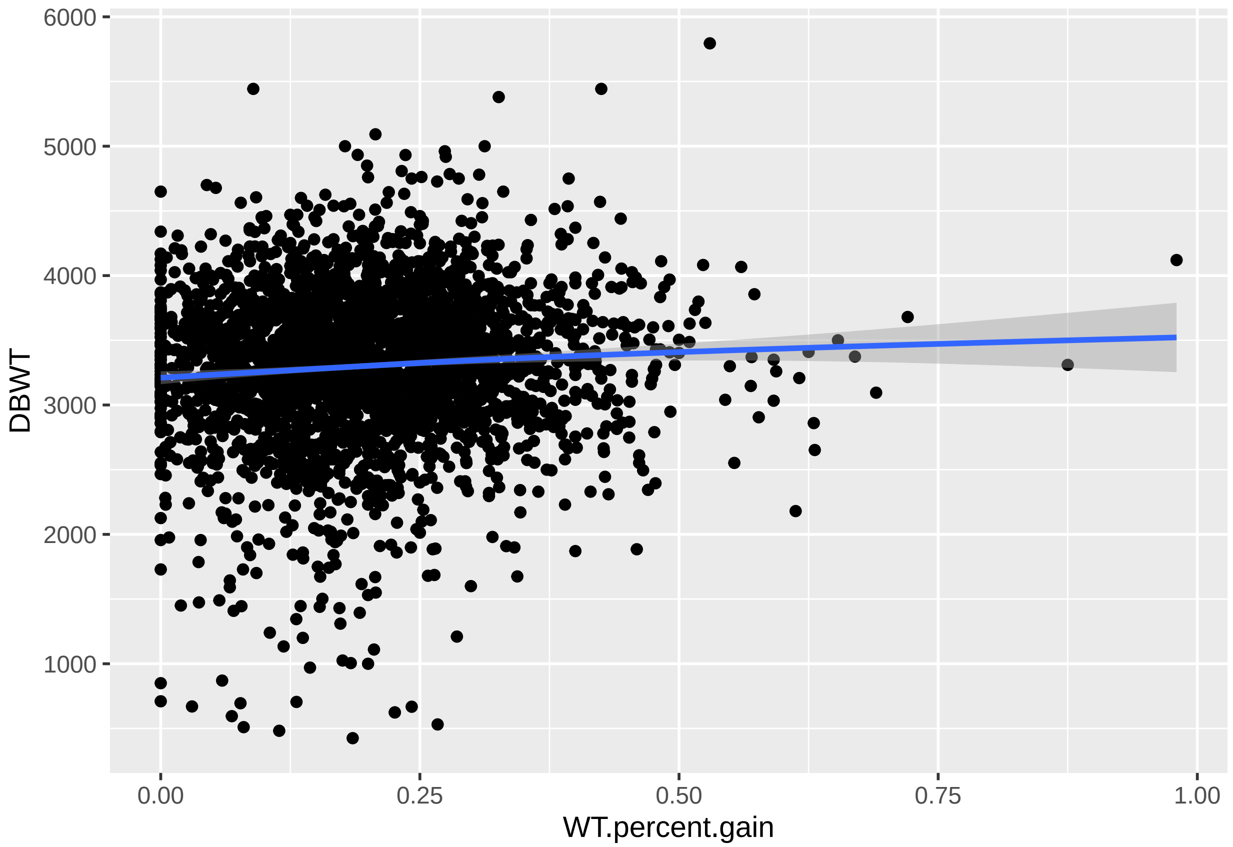
```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```



WTGAIN.percentage:

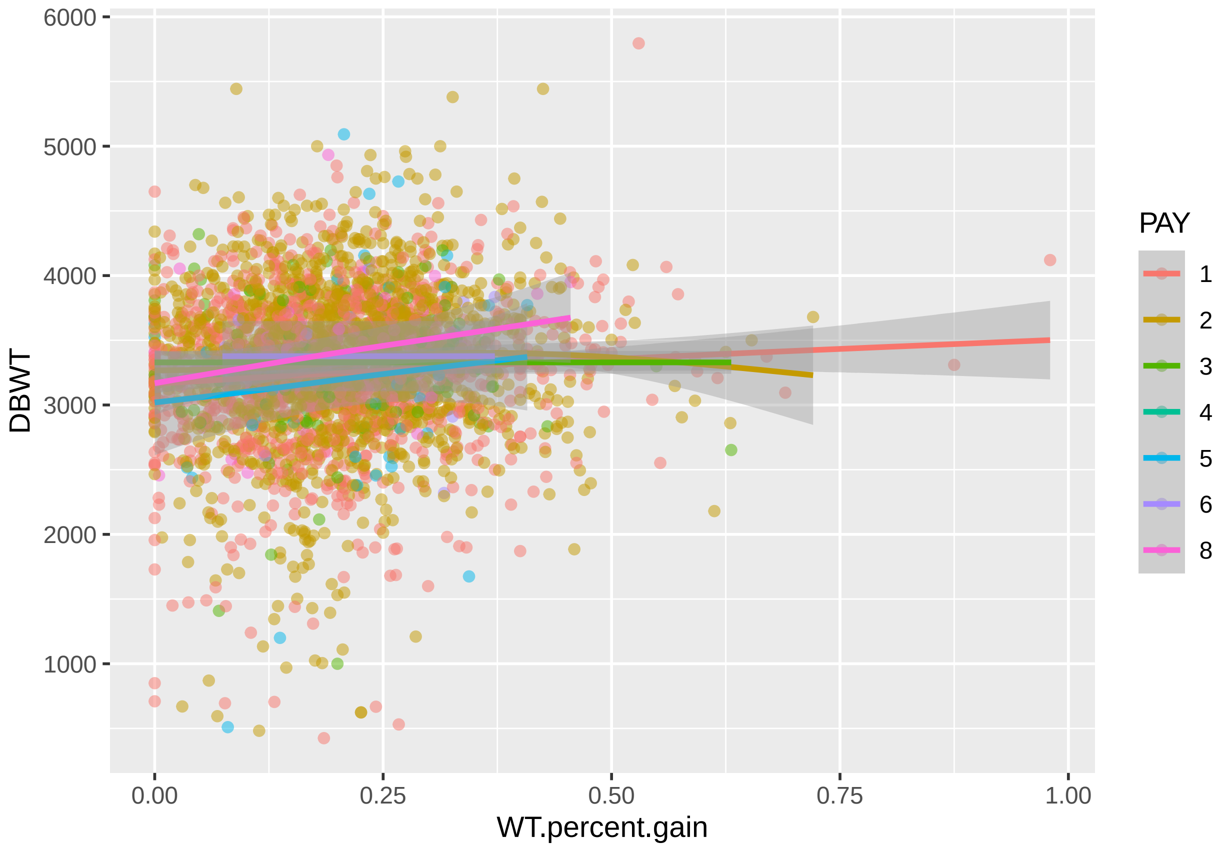
```
ggplot(EDA_df, aes(x = WT.percent.gain, y = DBWT)) +
  geom_point() +
  geom_smooth()
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



```
ggplot(EDA_df, aes(x = WT.percent.gain, y = DBWT)) +
  geom_point(aes(colour = PAY), alpha = 0.5) +
  geom_smooth(aes(colour = PAY))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


6000
5000
4000
3000
2000
1000
DBWT
0.00
0.25
0.50
0.75
1.00
WT.percent.gain
PAY
1
2
3
4
5
6
8