

Georgia Birth Weight EDA

Alejandro Hernandez

2024-06-03

```
birthwt <- read.csv("data/cdc birthwt.csv")

names(birthwt) <- c("mother.id", "birth.order", "birth.weight", "maternal.age", "child.id")

# define breaks in maternal age factor
breaks <- c(min(birthwt$maternal.age), 20, 30, 40,
           max(birthwt$maternal.age))

birthwt <- birthwt %>%
  mutate(
    # create maternal age factor
    maternal.age.factor = cut(maternal.age, breaks=breaks, include.lowest = T),
    # create birth weight factor
    birth.weight.factor = ifelse(birth.weight < 2500,
                                 ifelse(birth.weight < 1500,
                                       "Very low (less than 1500g)",
                                       "Low (less than 2500g)"),
                                 "Not low"),
    # create birth weight binary factor
    birth.weight.binary = ifelse(birth.weight < 2500,
                                 "Low (less than 2500g)",
                                 "Not low")) %>%
    # define birth order (and above) as factors
    mutate_at(vars(birth.order, birth.weight.factor, birth.weight.binary),
              as.factor)

# create a interpregnancy variable
birthwt <- birthwt %>%
  arrange(mother.id, birth.order) %>%
  group_by(mother.id) %>%
  mutate(interval = maternal.age - lag(maternal.age)) %>%
  ungroup()

birthwt %>% filter(interval < 0)

## # A tibble: 2 x 9
##   mother.id birth.order birth.weight maternal.age child.id maternal.age.factor
##       <int>      <fct>        <int>      <int>      <int>      <fct>
## 1      57939 3            3430        19      1063 [12,20]
## 2     212988 3            2470        20      3648 [12,20]
## # i 3 more variables: birth.weight.factor <fct>, birth.weight.binary <fct>,
## #   interval <int>
```

```
# mothers 57939 and 212988 both have negative intervals for their third births
# we will remove all their data and do not expect measurable changes in future
# analysis, their data is not special among the sample
birthwt <- birthwt %>% filter(!mother.id %in% c(57939, 212988))
```

Validation

```
# validate each mother has 5 children
birthwt %>%
  group_by(mother.id) %>%
  summarize(Count = n()) %>%
  all(.Count == 5)

## [1] TRUE

# summarize numeric variables
birthwt %>%
  select(birth.weight, maternal.age) %>%
  summary

##    birth.weight   maternal.age
##    Min. :312   Min. :12.00
##    1st Qu.:2850  1st Qu.:18.00
##    Median :3175   Median :21.00
##    Mean   :3156   Mean   :21.65
##    3rd Qu.:3515  3rd Qu.:24.00
##    Max.   :5528   Max.   :42.00
```

Exploratory data analysis

Birth weight

Birth weight is an infant's weight that is optimally measured in the hours following birth.

The World Health Organization (WHO) defines low birth weight as below 2500 g (5.5 lbs / 5 lbs 8 oz) and very low birth weight as below 1500 g (3.3 lbs / 3 lbs 5 oz). An infant with a low birth weight is often also premature, and at greater risk of health complications (e.g., underdeveloped lungs, inability to maintain body temperature, difficulty gaining weight, intestinal disease, bleeding in the brain, and sudden death).

Factors that are believed to cause low birth weight include maternal health and age, and multiple-baby pregnancies. Effects from race are also believed to impact birth weight: among Americans, Black mothers are twice as likely as white mothers to have low birth weights.

In some cases, the validity of birth weight data is of concern, as measures may be taken days after birth, after which significant weight loss may have occurred. We assume the source of our data to be valid.

1. What are the quartiles and average birth weight among all mothers? What quantile defines the threshold of low birth weights? How many births fall into each quartile and each category of (low / not low)?
2. Describe the distribution of all birth weights.

3. Describe the distribution of each mother's average birth weight. How does this distribution compare to that of all birth weights?

```
#####
# Birth weight
#####

# numeric summary of birth weight
round(c(mean = mean(birthwt$birth.weight),
       sd = sd(birthwt$birth.weight),
       IQR = IQR(birthwt$birth.weight),
       range = diff(range(birthwt$birth.weight)),
       quantile(birthwt$birth.weight)), 2)

##      mean      sd      IQR      range      0%      25%      50%      75%      100%
## 3156.13  570.25  665.00  5216.00  312.00  2850.00  3175.00  3515.00  5528.00

# numeric summary of individual average birth weight
birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.weight = mean(birth.weight)) %>%
  reframe(c(mean = mean(avg.weight),
            sd = sd(avg.weight),
            IQR = IQR(avg.weight),
            quantile(avg.weight))) %>%
  round(2) %>% as.list

## $`c(...)`'
##      mean      sd      IQR      0%      25%      50%      75%      100%
## 3156.13  416.70  532.35 1690.40 2891.80 3166.80 3424.15 4745.80

# size of each quartile
birthwt %>%
  mutate(quartile.range = cut(birth.weight,
                               breaks = quantile(birthwt$birth.weight),
                               include.lowest = TRUE,
                               dig.lab = 5)) %>%
  group_by(quartile.range) %>%
  summarize(n = length(quartile.range)) %>%
  mutate(quartile = 1:4) %>%
  select(Quartile=quartile, Range=quartile.range, Size=n) %>%
  kable(caption = "Quartiles of birth weight")
```

Table 1: Quartiles of birth weight

Quartile	Range	Size
1	[312,2850]	1100
2	(2850,3175]	1097
3	(3175,3515]	1109
4	(3515,5528]	1074

```

# number of low births
birthwt %>%
  group_by(birth.weight.factor) %>%
  summarize(n = length(birth.weight.factor),
            prop = round(n / nrow(.), 2)) %>%
  arrange(desc(n)) %>%
  select(birth.weight.factor, n, prop) %>%
  kable(caption = "Sizes of birth weight groups",
        col.names = c("Weight group", "N", "Proportion"))

```

Table 2: Sizes of birth weight groups

Weight group	N	Proportion
Not low	3933	0.90
Low (less than 2500g)	389	0.09
Very low (less than 1500g)	58	0.01

```

# distribution of all birth weights
gg_bw_all <- birthwt %>%
  ggplot(aes(x = birth.weight, y = after_stat(density))) +
  geom_histogram(bins = 30) +
  xlab("Birth weight") + ylab("") +
  geom_vline(xintercept = c(1500, 2500), color = c("red", "blue")) +
  geom_vline(xintercept = quantile(birthwt$birth.weight), color = "grey") +
  theme_bw()

# distribution of mothers' average birth weight
# must compute quartiles separately
quartiles <- birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.weight = mean(birth.weight)) %>%
  reframe(quartile(avg.weight)) %>% as.list

gg_bw_indv <- birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.weight = mean(birth.weight)) %>%
  ggplot(aes(x = avg.weight, y = after_stat(density))) +
  geom_histogram(bins = 30) +
  xlab("Average birth weight") + ylab("") +
  labs(caption = "Quartiles marked by vertical grey lines") +
  geom_vline(xintercept = c(1500, 2500), color = c("red", "blue")) +
  geom_vline(xintercept = quartiles[[1]], color = "grey") +
  theme_bw()

gridExtra::grid.arrange(gg_bw_all, gg_bw_indv)

```

Maternal age

Maternal age is the age at which a mother gives birth. Previous studies suggest that maternal age is associated with birth weight (younger and older mothers have greater rates of preterm births).

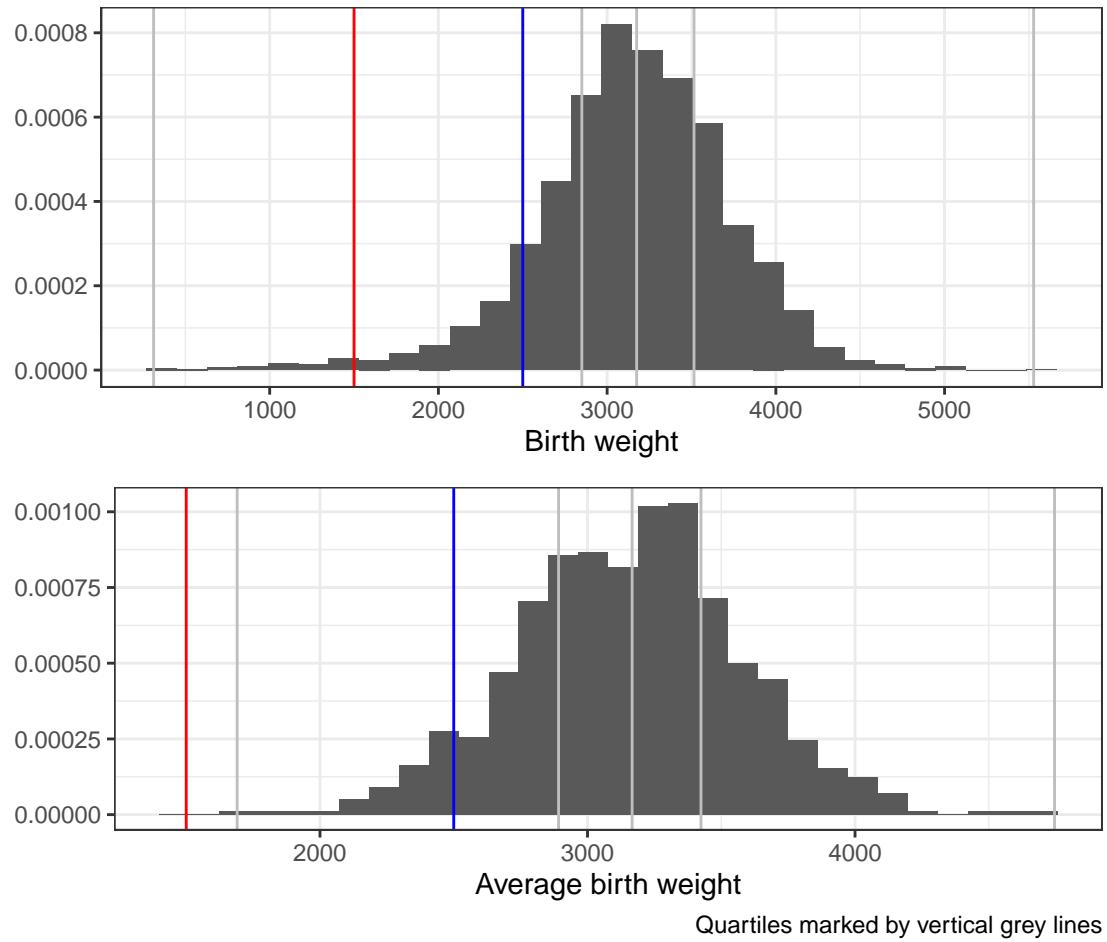


Figure 1: Distributions of birth weight with thresholds of low (blue) and very low (red) birth weight

- What are the quartiles and average maternal age among all mothers? What quantile defines the thresholds of middle aged? How many births fall into each quantile and each category of (young / middle age)?
- Describe the distribution of all maternal ages.
- Describe the distribution of each mother's average maternal age. How does this distribution compare to that of all maternal ages?

```
#####
# Maternal age

# numeric summary of maternal age
round(c(mean = mean(birthwt$maternal.age),
       sd = sd(birthwt$maternal.age),
       IQR = IQR(birthwt$maternal.age),
       range = diff(range(birthwt$maternal.age)),
       quantile(birthwt$maternal.age)), 2)

##   mean     sd    IQR range    0%   25%   50%   75%  100%
## 21.65  4.63  6.00 30.00 12.00 18.00 21.00 24.00 42.00

# numeric summary of individual average maternal age
birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.age = mean(maternal.age)) %>%
  reframe(c(mean = mean(avg.age),
            sd = sd(avg.age),
            IQR = IQR(avg.age),
            quantile(avg.age))) %>%
  round(2) %>% as.list

## $`c(...)`'
##   mean     sd    IQR    0%   25%   50%   75%  100%
## 21.65  3.70  4.25 15.40 19.00 20.80 23.25 38.20

# size of each quartile
birthwt %>%
  mutate(quartile.range = cut(maternal.age,
                               breaks = quantile(birthwt$maternal.age),
                               include.lowest = TRUE)) %>%
  group_by(quartile.range) %>%
  summarize(n = length(quartile.range)) %>%
  mutate(quartile = 1:4) %>%
  select(quartile, quartile.range, n) %>%
  kable(caption = "Quantiles of maternal age",
        col.names = c("Quartile", "Range", "Size"))
```

Table 3: Quantiles of maternal age

Quartile	Range	Size
1	[12,18]	1211
2	(18,21]	1188
3	(21,24]	955
4	(24,42]	1026

```
# condensed version of above code (this has worse readability):
# table(cut(birthwt$maternal.age,
#             breaks = quantile(birthwt$maternal.age),
#             include.lowest = T))

# size of each age group
birthwt <- birthwt %>%
  # create new age group variable
  mutate(age.group = cut(maternal.age,
                        breaks = c(0, 18, 25, 35, 45),
                        right = F))

birthwt %>%
  group_by(age.group) %>%
  summarize(n = length(age.group),
            prop = round(n / nrow(.), 2)) %>%
  kable(caption = "Sizes of maternal age groups",
        col.names = c("Group", "Size", "Proportion"))
```

Table 4: Sizes of maternal age groups

Group	Size	Proportion
[0,18)	789	0.18
[18,25)	2565	0.59
[25,35)	959	0.22
[35,45)	67	0.02

```
# distribution of all maternal ages
gg_ma_all <- birthwt %>%
  ggplot(aes(x = maternal.age, y = after_stat(density))) +
  geom_histogram(bins = 30) +
  xlab("Maternal age") + ylab("") +
  scale_x_continuous(breaks = seq(15, 40, by = 5)) +
  geom_vline(xintercept = quantile(birthwt$maternal.age), color = "grey") +
  theme_bw()

# distribution of mothers' average maternal age
# must compute quartiles separately
quartiles <- birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.age = mean(maternal.age)) %>%
  reframe(quartile(avg.age)) %>% as.list
```

```

gg_ma_indv <- birthwt %>%
  group_by(mother.id) %>%
  summarize(avg.age = mean(maternal.age)) %>%
  ggplot(aes(x = avg.age, y = after_stat(density))) +
  geom_histogram(bins = 30) +
  xlab("Average maternal age") + ylab("") +
  labs(caption = "Quartiles marked by vertical grey lines") +
  geom_vline(xintercept = quartiles[[1]], color = "grey") +
  theme_bw()

gridExtra::grid.arrange(gg_ma_all, gg_ma_indv)

```

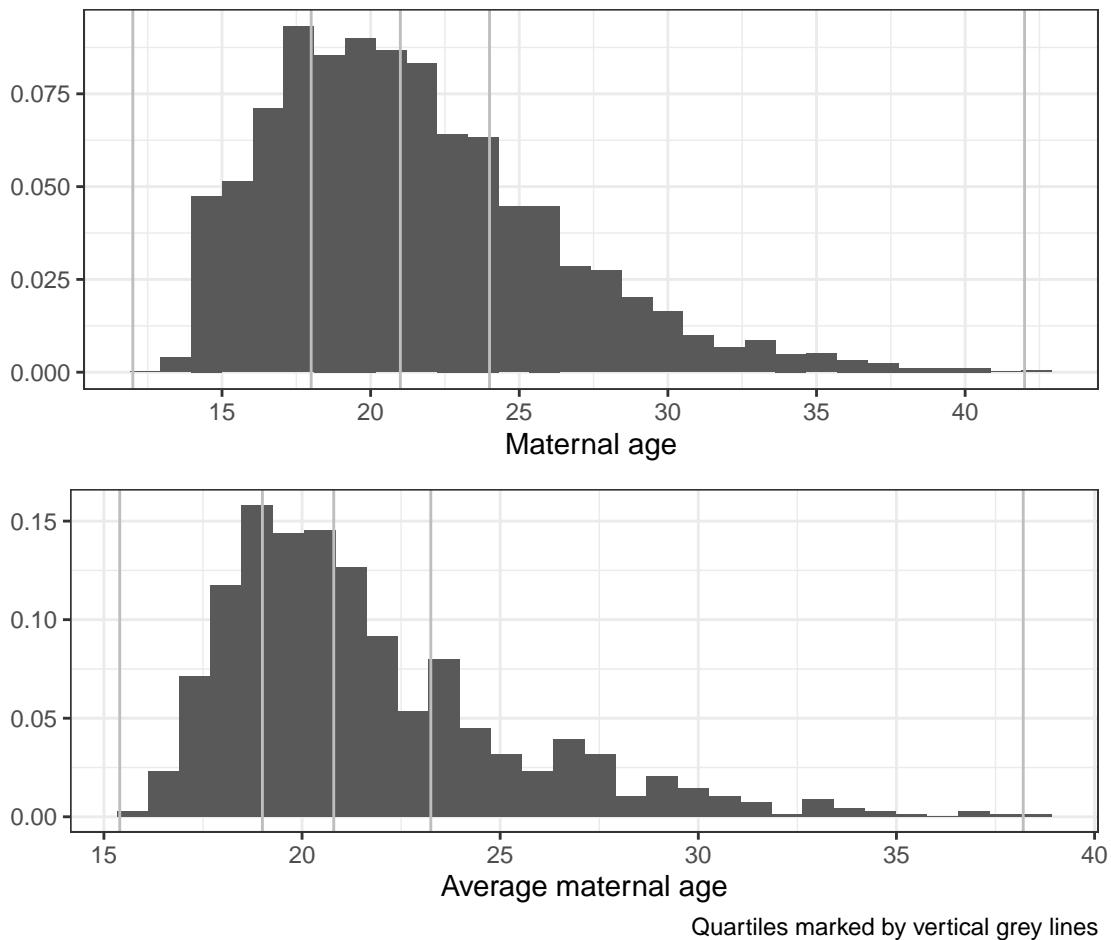


Figure 2: Distributions of maternal age and average maternal age. Average maternal age is average age of a mother over their five births- it is unique to each mother.

Interpregnancy interval

Interpregnancy interval is the time elapsed between one child's birth and the subsequent child's conception, often measured in months. This interval can be brief: mothers may become pregnant as early as four weeks after delivery.⁴ Similar to maternal age, extremely short and long intervals are associated with health

risks for the mother and second-born child. Interpregnancy intervals less than 18 months introduce moderate risk to children and significant risk is associated with intervals shorter than 6 months; intervals greater than 5-10 years is associated with increased risk of adverse health outcomes for both mother and child.⁴

With available data, our best estimate of this interval is the difference in maternal age between two subsequent births, measured in years, which in certain cases may be a slight underestimate.

1. What are the group sizes and mode for all interval values?
2. Describe the distribution of all intervals.
3. Describe the distribution of each mother's interval mode. How does this distribution compare to that of all intervals?

Interdelivery interval is the time elapsed between two subsequent births. May we more accurately estimate this interval using our data? Not necessarily- the accuracy of using the difference in maternal ages to estimate an interpregnancy versus an interdelivery interval varies by situation and relies on information unavailable to us. We have no workaround for this issue, other than to define short and long intervals according to respective thresholds (which differ by ?? 9 months?) and examine the extent to which results agree.

```
#####
# Interpregnancy interval
#####

# distribution of maternal age interval
birthwt %>%
  filter(!is.na(interval) & interval >= 0) %>%
  group_by(Interval = interval) %>%
  summarize(Count = n()) %>%
  mutate(Percentage = paste(round(100 * Count/sum(Count), 2), "%", sep = "")) %>%
  kable()
```

Interval	Count	Percentage
0	35	1%
1	1465	41.81%
2	1335	38.1%
3	398	11.36%
4	166	4.74%
5	73	2.08%
6	21	0.6%
7	9	0.26%
8	2	0.06%

Visualizing relationships with birth weight

Boxplots

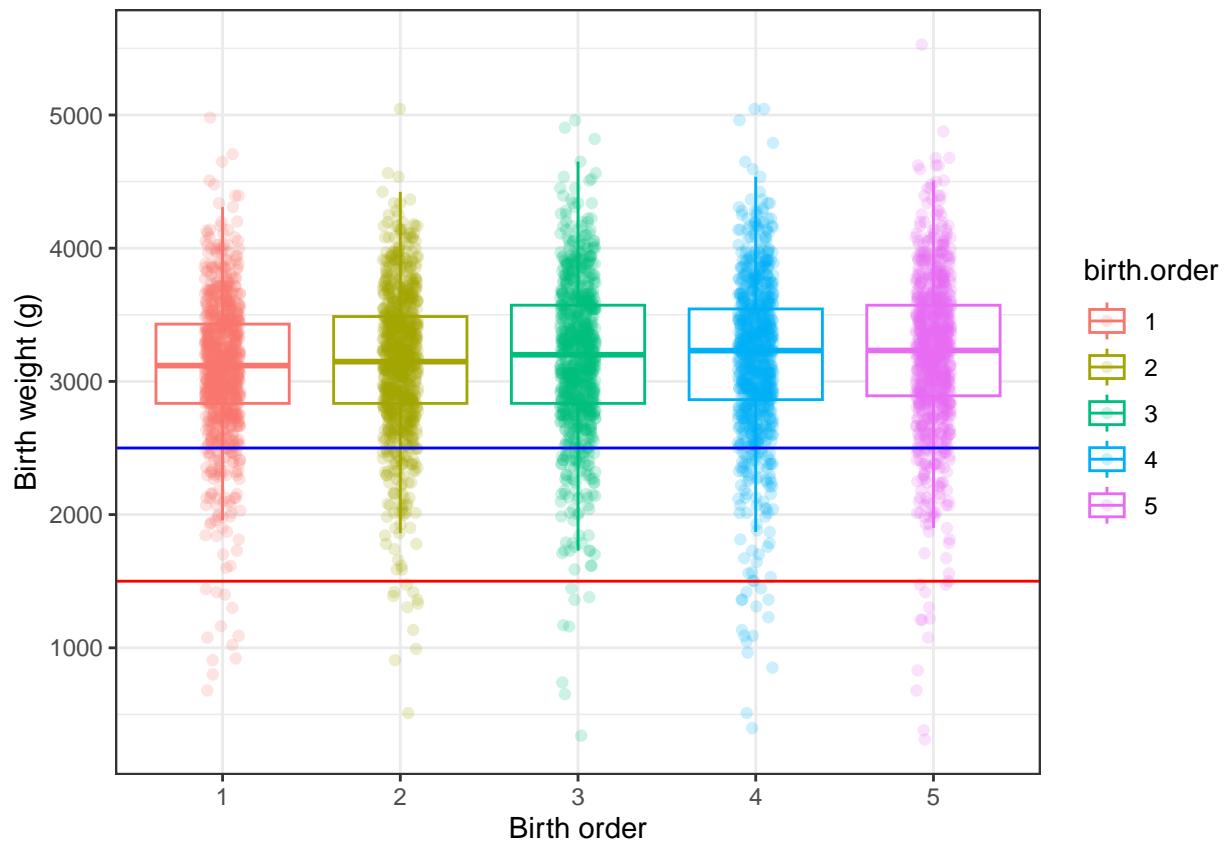
1. Produce boxplots of birth weights across birth order for all mothers, marking the threshold for low birth weight. Is there a visual trend?

```

gg1.weight.order <- birthwt %>%
  ggplot(aes(x = birth.order, y = birth.weight, group = birth.order,
             color = birth.order)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Birth order") + ylab("Birth weight (g)") +
  theme_bw()

gg1.weight.order

```

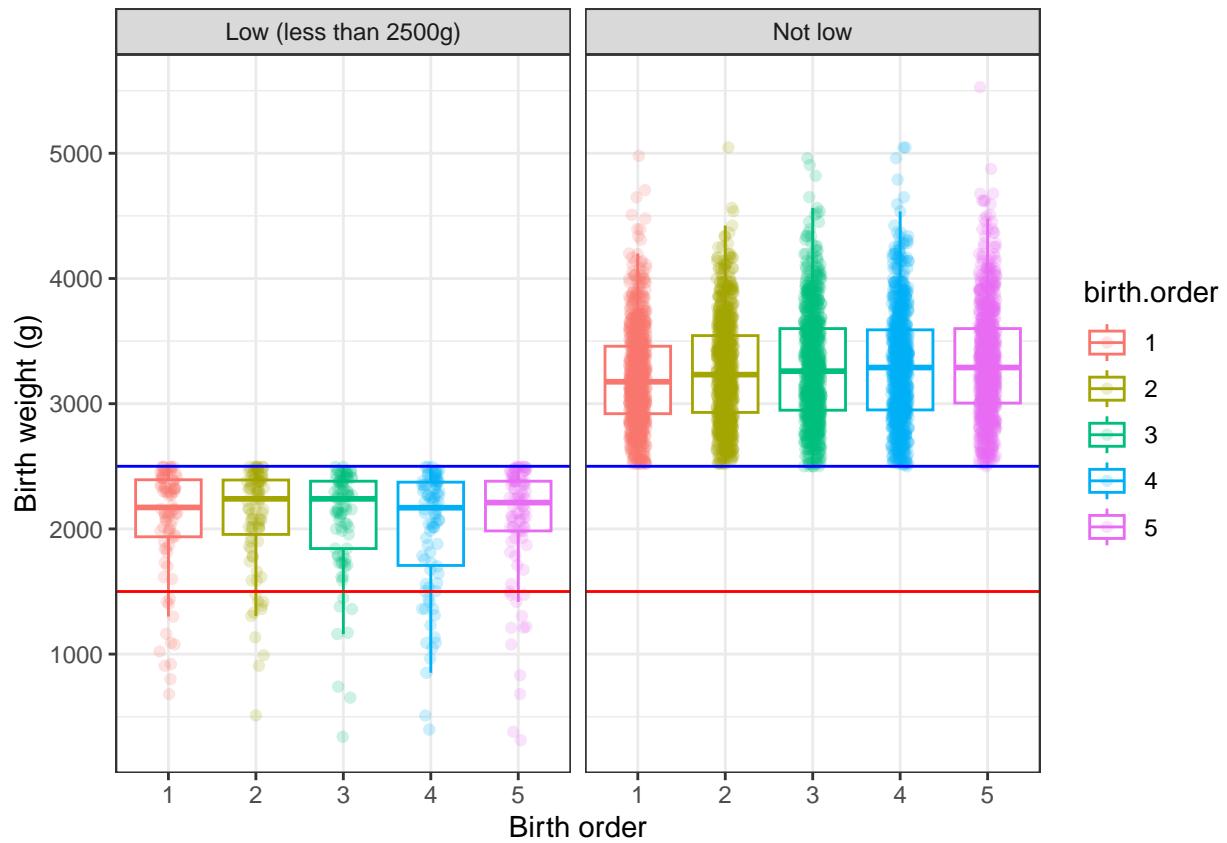


(a) Stratify the plot by low / not low birth weight

```

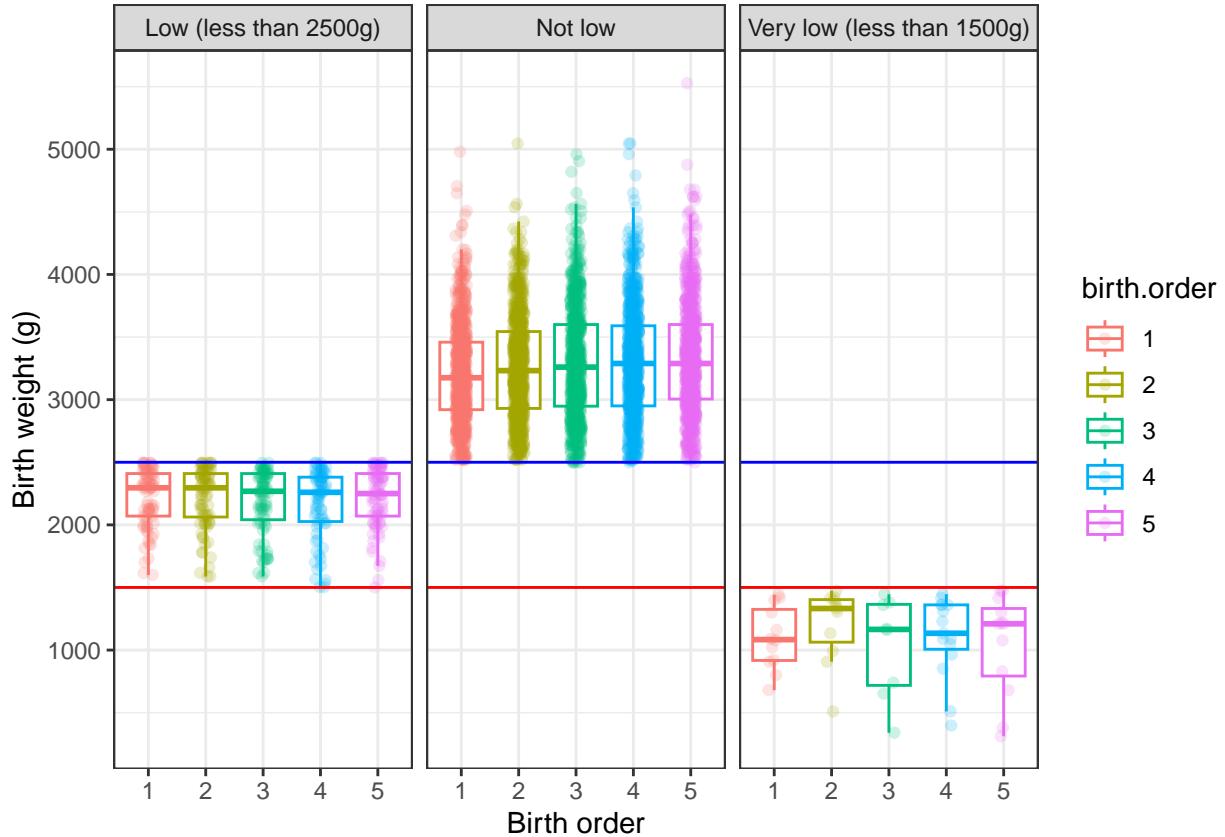
gg1.weight.order + facet_wrap(vars(birth.weight.binary))

```



(b) Stratify the plot by very low / low / not low birth weight

```
gg1.weight.order + facet_wrap(vars(birth.weight.factor))
```

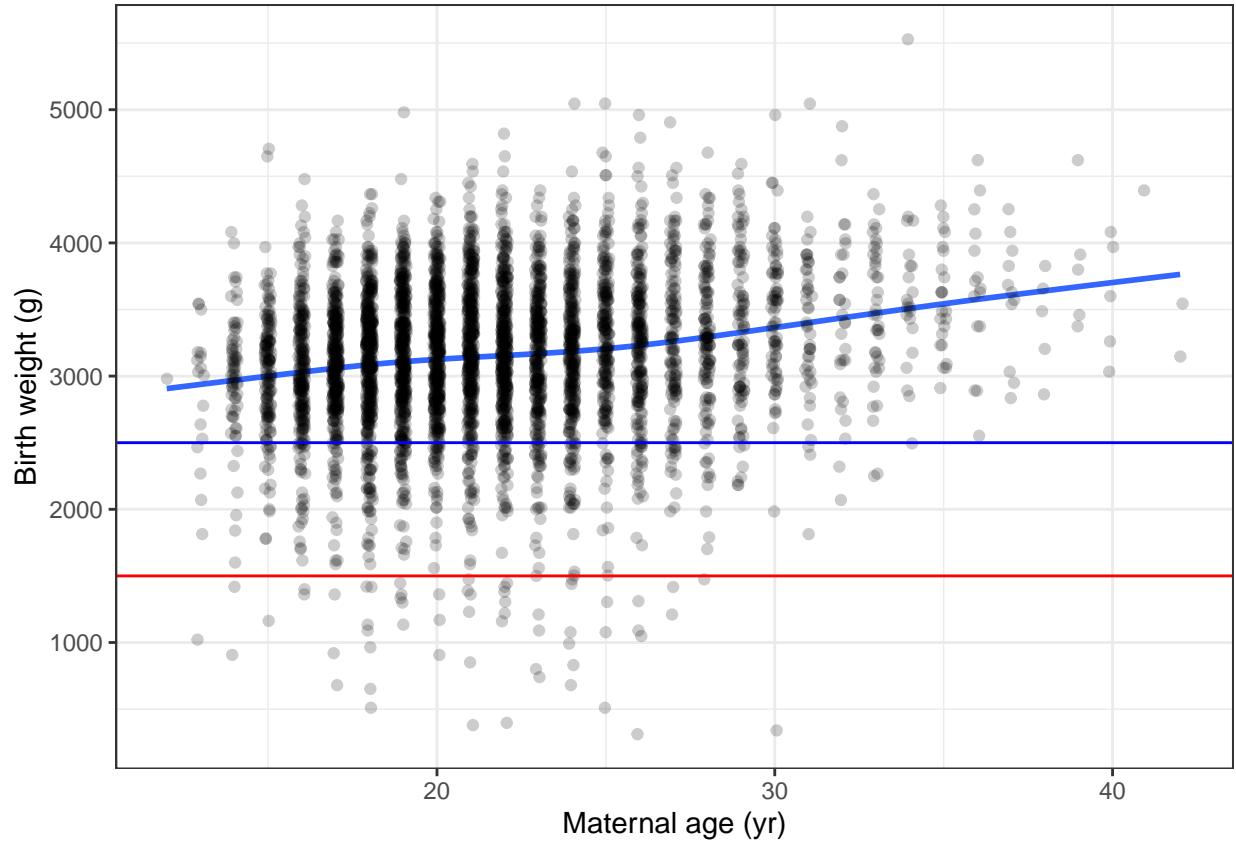


- Produce a boxplot of birth weight across maternal age for all mothers, marking the threshold for low birth weight. Is there a visual trend?

```
gg2.weight.age <- birthwt %>%
  ggplot(aes(x = maternal.age, y = birth.weight)) +
  geom_smooth(se = F) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Maternal age (yr)") + ylab("Birth weight (g)") +
  theme_bw()

gg2.weight.age

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

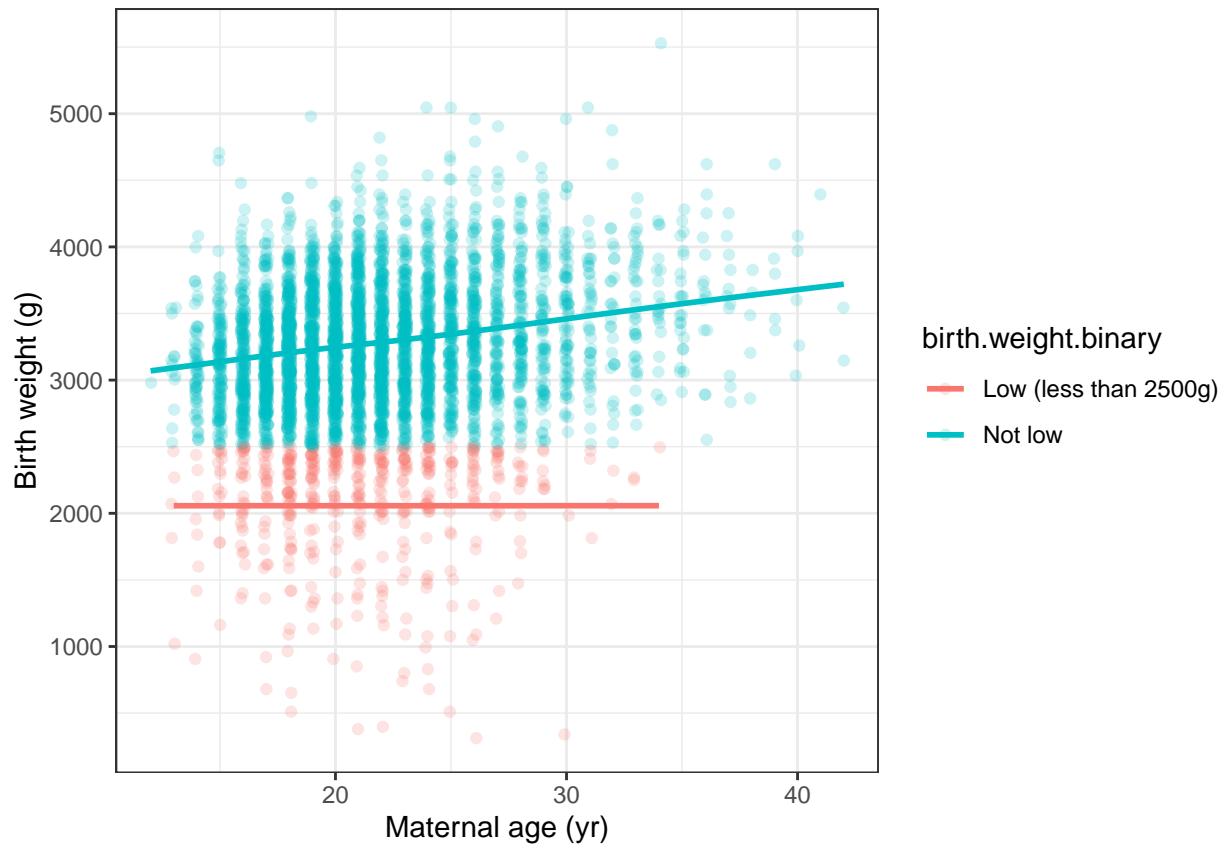


(a) Stratify the plot by low / not low birth weight

```
gg2a.weight.age <- birthwt %>%
  ggplot(aes(x = maternal.age, y = birth.weight,
             group = birth.weight.binary, color = birth.weight.binary)) +
  geom_smooth(se = F) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  # geom_hline(yintercept = 2500, color = "blue") +
  # geom_hline(yintercept = 1500, color = "red") +
  xlab("Maternal age (yr)") + ylab("Birth weight (g)") +
  theme_bw()

gg2a.weight.age
```

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

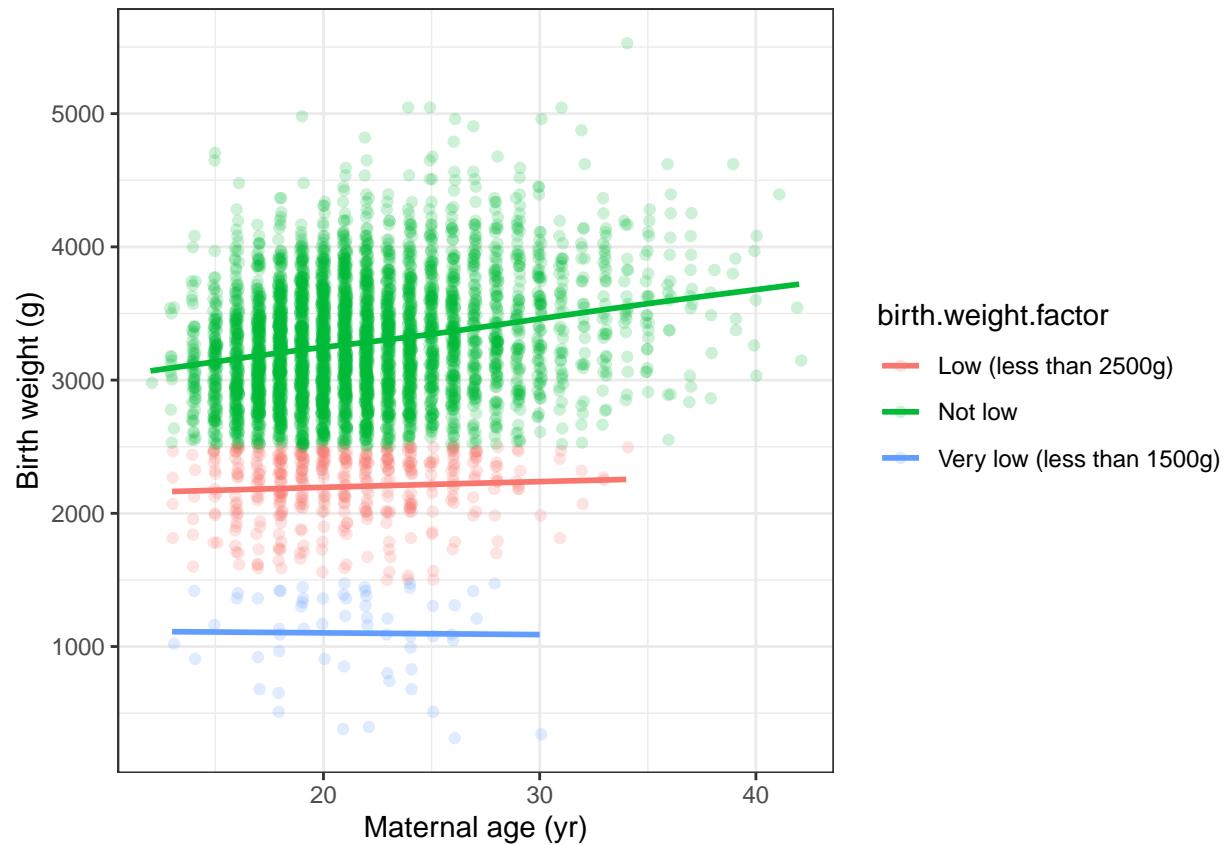


(b) Stratify the plot by very low / low / not low birth weight

```
gg2b.weight.age <- birthwt %>%
  ggplot(aes(x = maternal.age, y = birth.weight,
             group = birth.weight.factor, color = birth.weight.factor)) +
  geom_smooth(se = F) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  # geom_hline(yintercept = 2500, color = "blue") +
  # geom_hline(yintercept = 1500, color = "red") +
  xlab("Maternal age (yr)") + ylab("Birth weight (g)") +
  theme_bw()

gg2b.weight.age
```

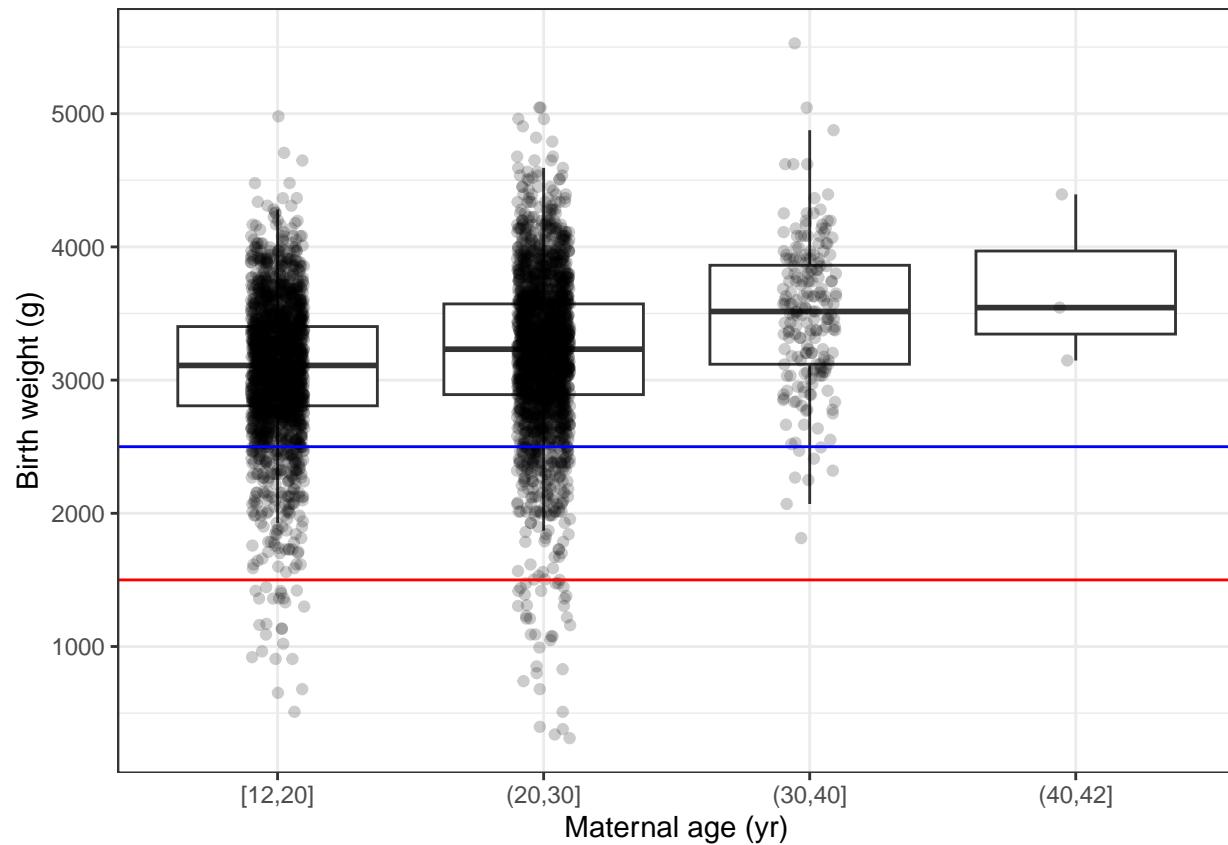
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



(c) Repeat the above for birth weight across maternal age factor

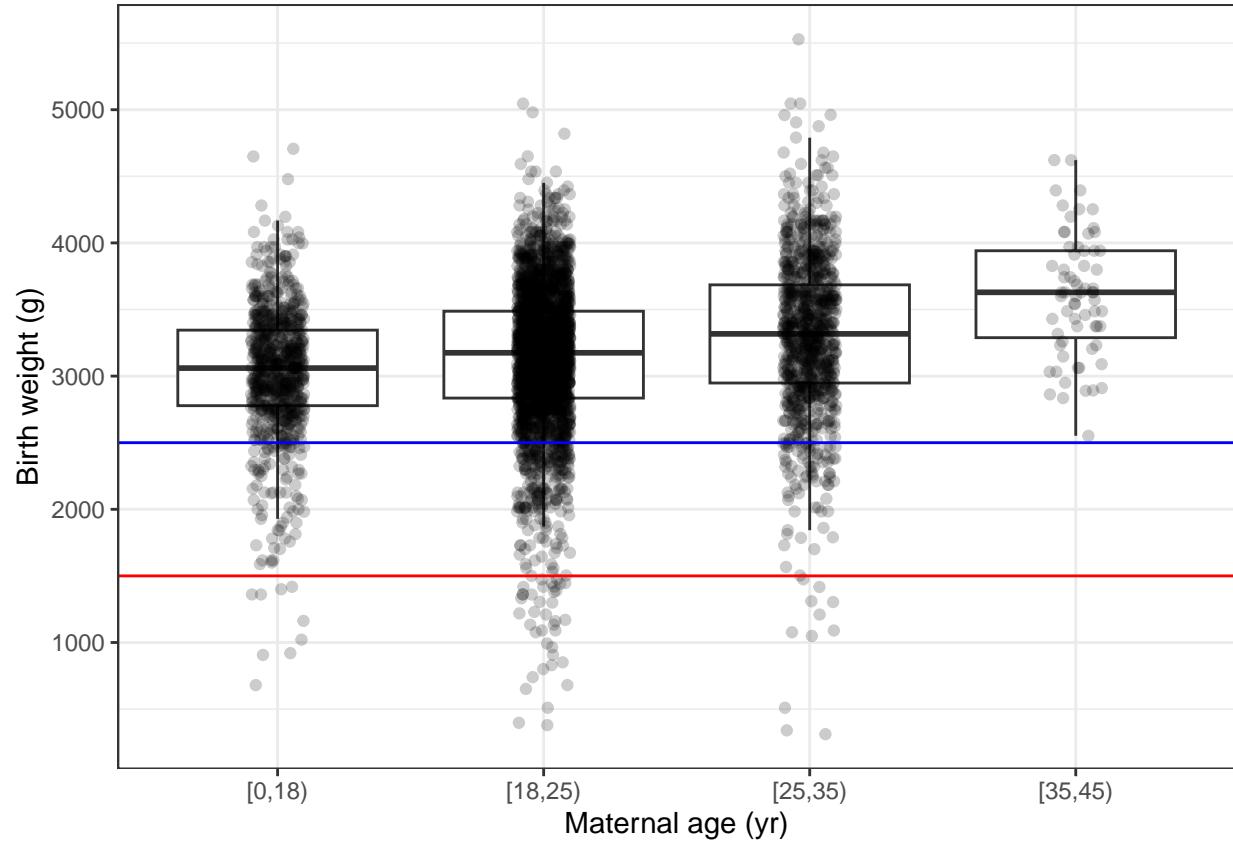
```
# using age grouped by decade
gg2c.weight.age <- birthwt %>%
  ggplot(aes(x = maternal.age.factor, y = birth.weight)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Maternal age (yr)") + ylab("Birth weight (g)") +
  theme_bw()

gg2c.weight.age
```



```
# using age grouped by popular age cutoffs
gg2c.weight.age <- birthwt %>%
  ggplot(aes(x = age.group, y = birth.weight)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1, alpha = 0.2) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Maternal age (yr)") + ylab("Birth weight (g)") +
  theme_bw()

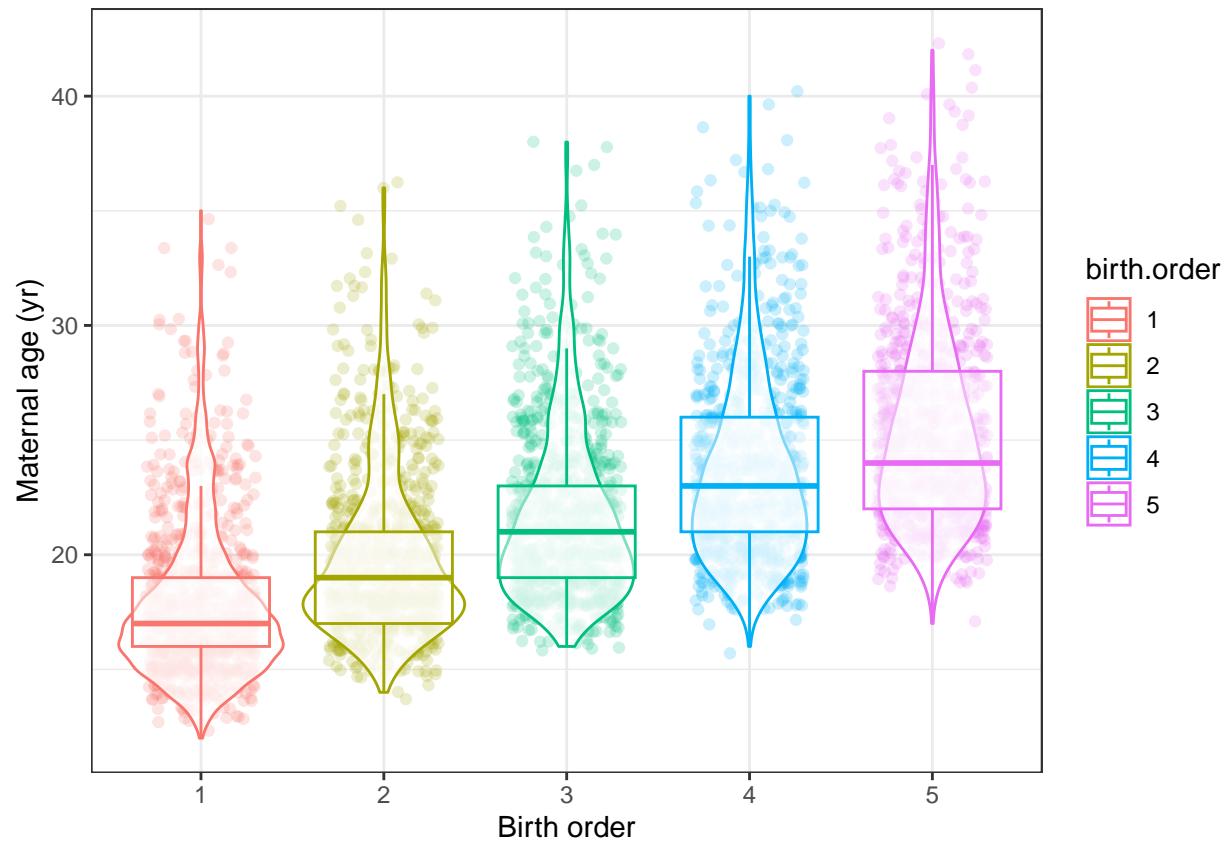
gg2c.weight.age
```



3. Produce boxplots of maternal age across birth order for all mothers. Is there a visual trend?

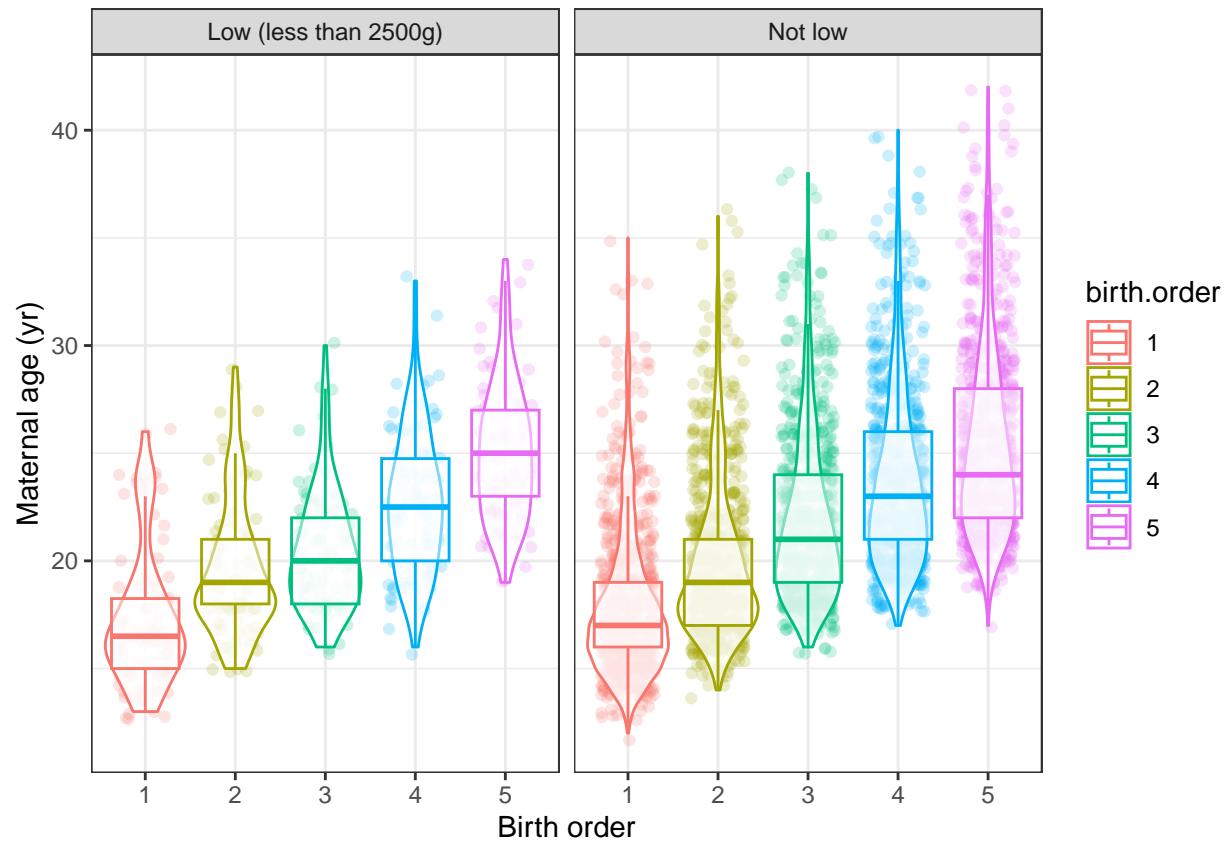
```
gg3.age.order <- birthwt %>%
  ggplot(aes(x = birth.order, y = maternal.age, group = birth.order,
             color = birth.order)) +
  geom_jitter(width = 0.3, alpha = 0.2) +
  geom_violin(alpha = 0.8) +
  geom_boxplot(outlier.shape = NA, alpha = 0.5) +
  xlab("Birth order") + ylab("Maternal age (yr)") +
  theme_bw()

gg3.age.order
```



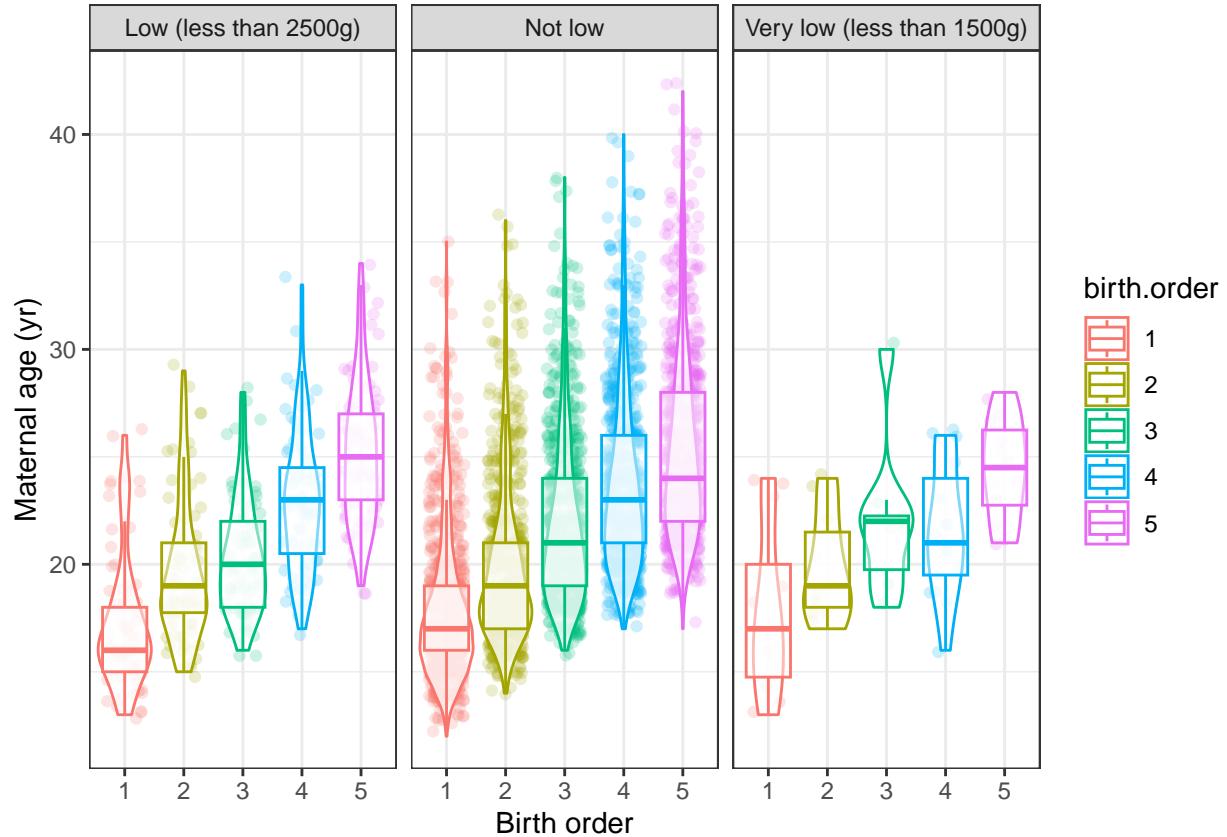
(a) Stratify the plot by low / not low birth weight

```
gg3.age.order + facet_wrap(vars(birth.weight.binary))
```



(b) Stratify the plot by very low / low / not low birth weight

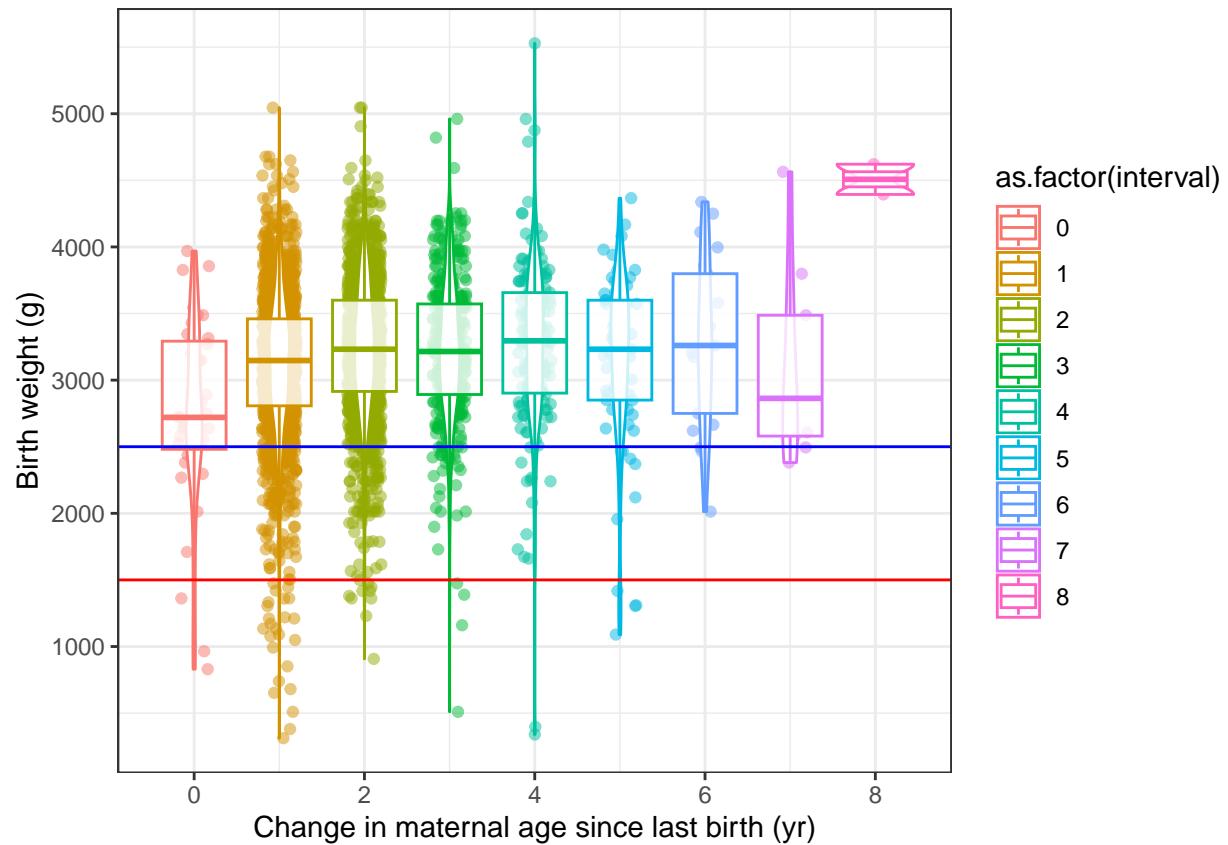
```
gg3.age.order + facet_wrap(vars(birth.weight.factor))
```



4. Produce a boxplot of birth weight across interval for non-firstborns (such that interval is defined), marking the threshold for low birth weight. Is there a visual trend?

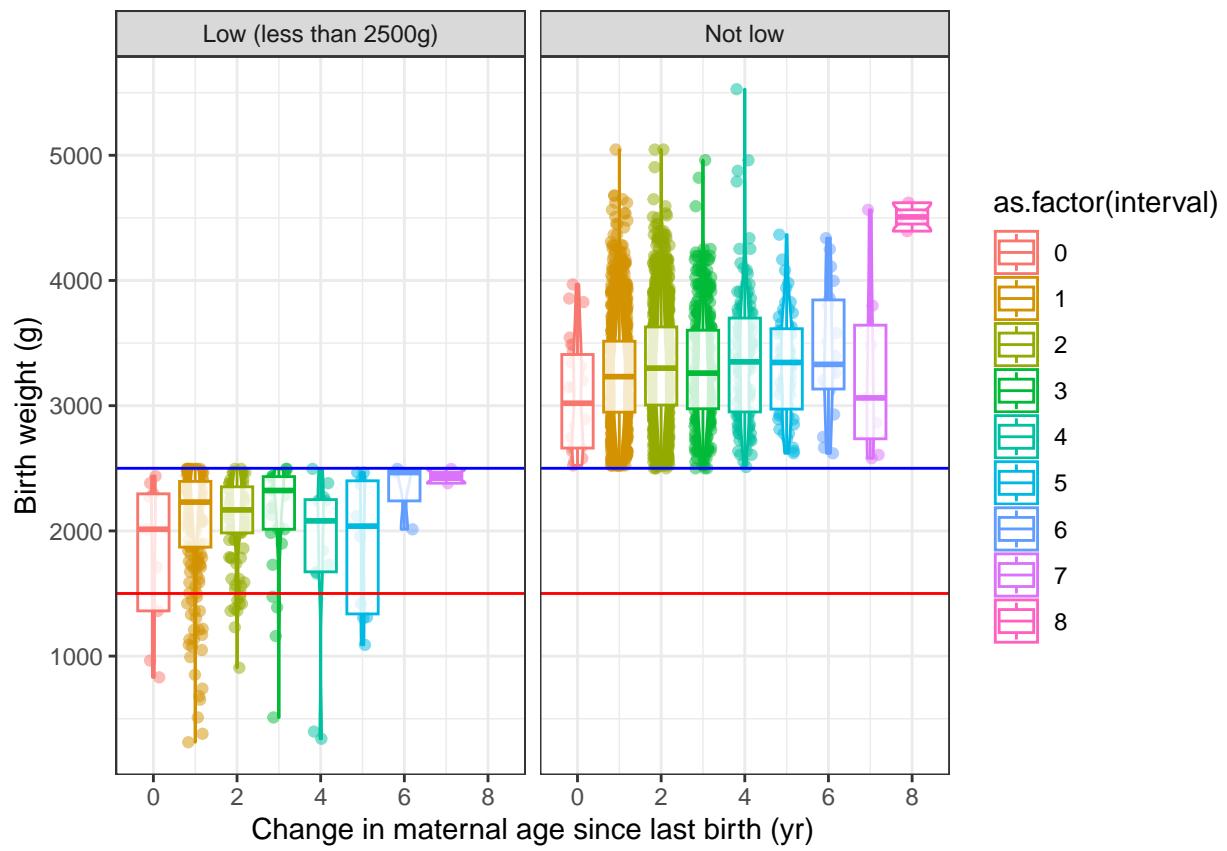
```
gg4.weight.interval <- birthwt %>%
  filter(!is.na(interval)) %>%
  ggplot(aes(x = interval, y = birth.weight, group = interval,
             color = as.factor(interval))) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  geom_violin() +
  geom_boxplot(outlier.shape = NA, alpha = 0.8) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Change in maternal age since last birth (yr)") +
  ylab("Birth weight (g)") +
  theme_bw()

gg4.weight.interval
```



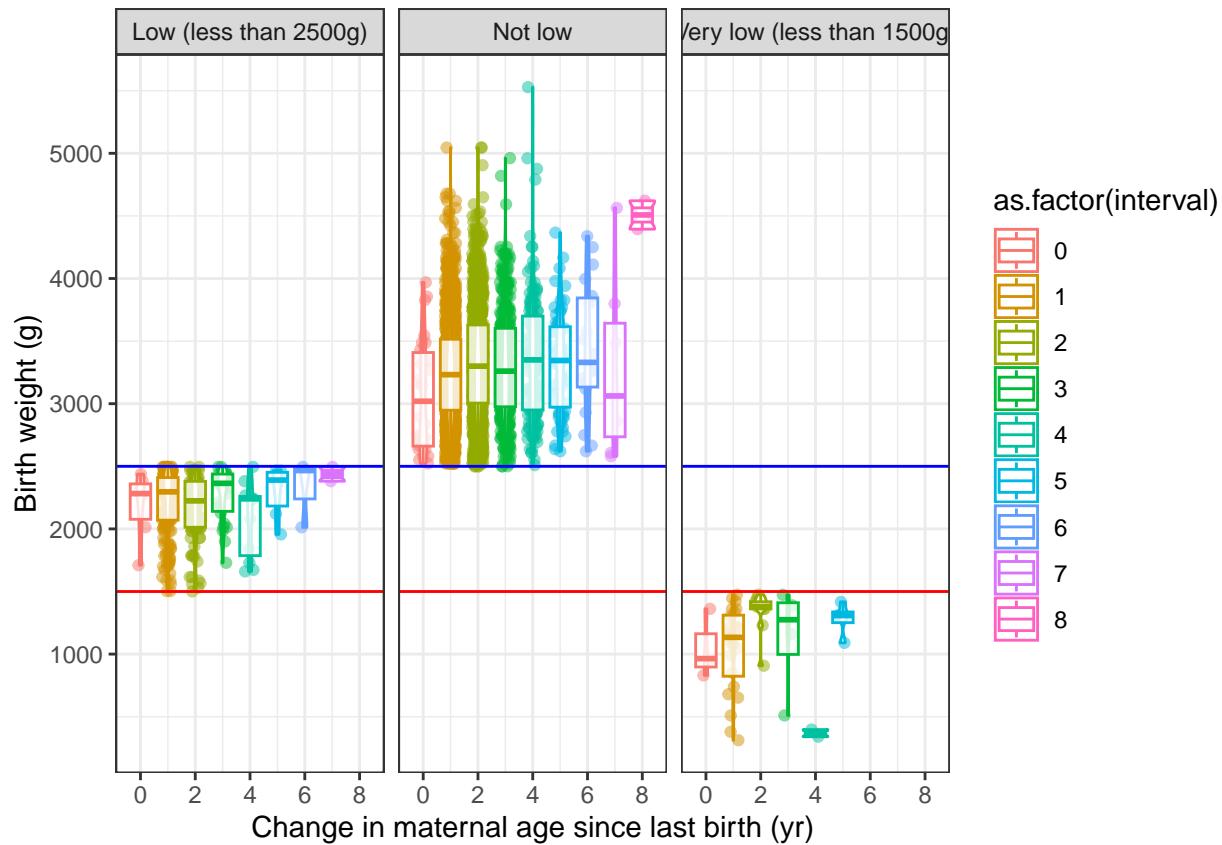
(a) Stratify the plot by low / not low birth weight

```
gg4.weight.interval + facet_wrap(vars(birth.weight.binary))
```



(b) Stratify the plot by very low / low / not low birth weight

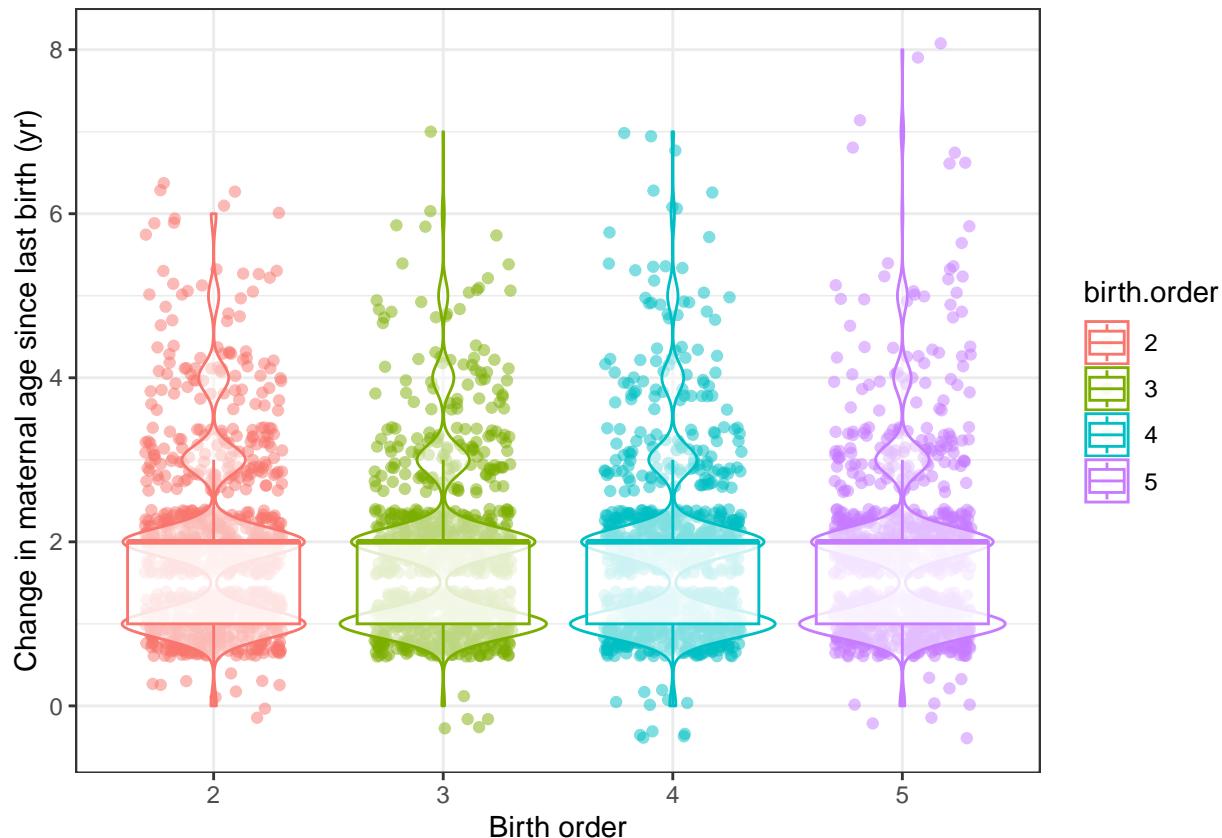
```
gg4.weight.interval + facet_wrap(vars(birth.weight.factor))
```



5. Produce a boxplot of interval across birth order for all non-firstborns (such that interval is defined). Is there a visual trend?

```
gg5.interval.order <- birthwt %>%
  filter(!is.na(interval)) %>%
  ggplot(aes(x = birth.order, y = interval, group = birth.order,
             color = birth.order)) +
  geom_jitter(width = 0.3, alpha = 0.5) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(outlier.shape = NA, alpha = 0.8) +
  xlab("Birth order") + ylab("Change in maternal age since last birth (yr)") +
  theme_bw()

gg5.interval.order
```



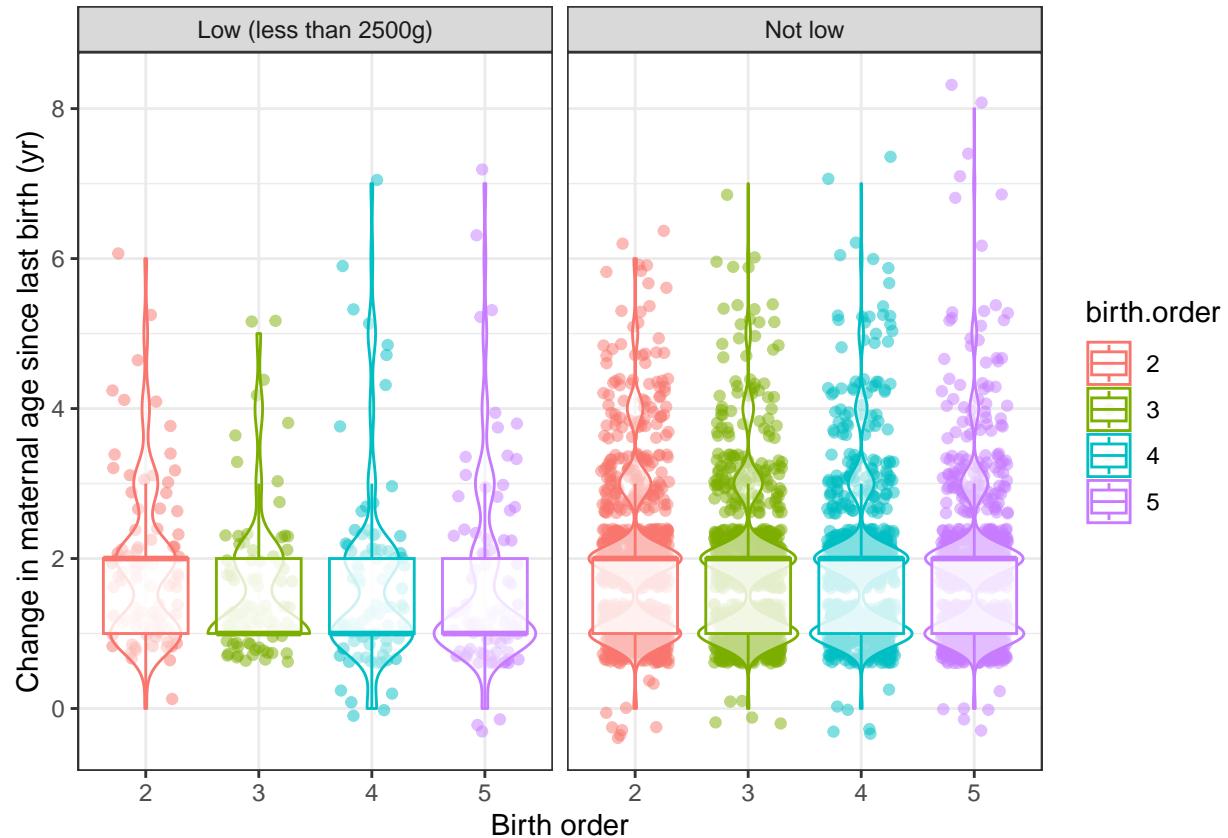
```
# table of counts
table_bw_intv <- birthwt %>%
  select(interval, birth.order) %>%
  filter(!is.na(interval)) %>%
  table()

# table of percents
round(table_bw_intv / nrow(table_bw_intv))
```

```
##           birth.order
## interval 1 2 3 4 5
##   0       0 0 1 1 1
##   1       0 37 42 42 42
##   2       0 37 37 37 37
##   3       0 13 11 10 11
##   4       0 6 4 4 4
##   5       0 2 2 2 2
##   6       0 1 0 1 0
##   7       0 0 0 0 1
##   8       0 0 0 0 0
```

(a) Stratify the plot by low / not low birth weight

```
gg5.interval.order + facet_wrap(vars(birth.weight.binary))
```



```
# table of counts
table_bw_intv_notlow <- birthwt %>%
  filter(!is.na(interval) & birth.weight.binary == "Not low") %>%
  select(interval, birth.order) %>%
  table()
```

```
table_bw_intv_low <- birthwt %>%
  filter(!is.na(interval) & birth.weight.binary != "Not low") %>%
  select(interval, birth.order) %>%
  table()
```

```
# table of percents
round(table_bw_intv_notlow / nrow(table_bw_intv))
```

```
##          birth.order
## interval 1 2 3 4 5
##      0 0 1 1 1 1
##      1 0 33 36 37 36
##      2 0 33 34 34 35
##      3 0 11 10 9 10
##      4 0 6 4 4 3
##      5 0 2 2 2 2
##      6 0 1 0 1 0
```

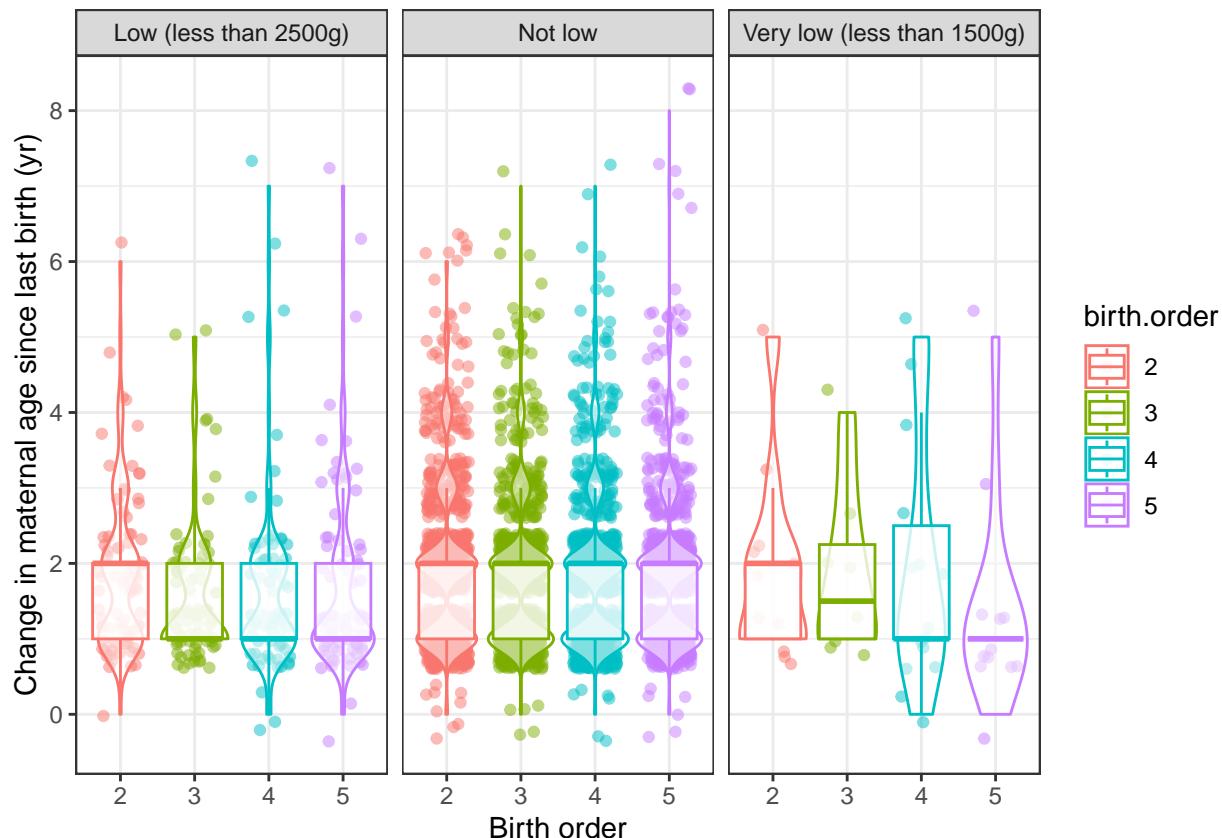
```
##      7 0 0 0 0 0
##      8 0 0 0 0 0 0
```

```
round(table_bw_intv_low / nrow(table_bw_intv))
```

```
##      birth.order
## interval 1 2 3 4 5
##      0 0 0 0 1 0
##      1 0 4 6 5 6
##      2 0 4 3 3 2
##      3 0 1 0 0 1
##      4 0 0 0 0 0
##      5 0 0 0 0 0
##      6 0 0 0 0 0
##      7 0 0 0 0 0
```

(b) Stratify the plot by very low / low / not low birth weight

```
gg5.interval.order + facet_wrap(vars(birth.weight.factor))
```



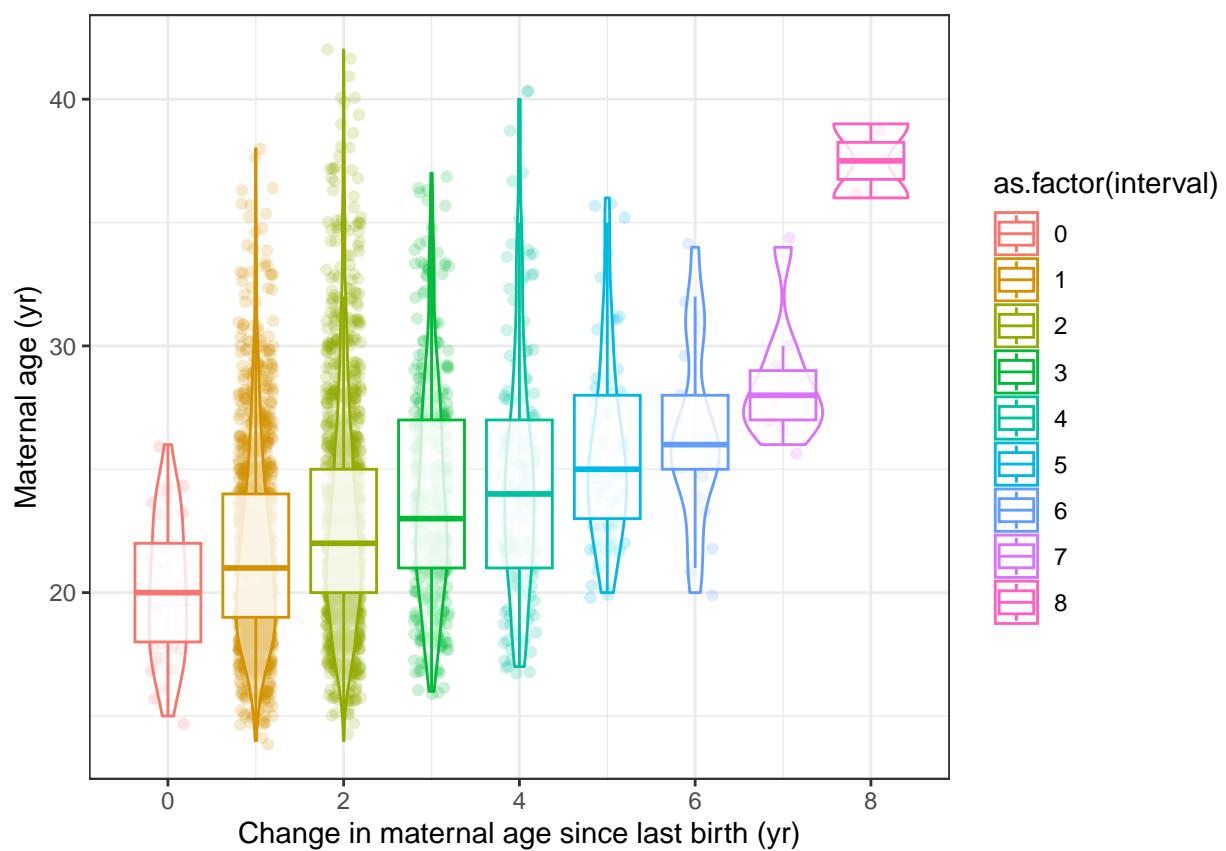
- Produce a boxplot of maternal age across interval for all non-firstborns (such that interval is defined). Is there a visual trend?

```

gg6.interval.age <- birthwt %>%
  filter(!is.na(interval)) %>%
  ggplot(aes(x = interval, y = maternal.age, group = interval,
             color = as.factor(interval))) +
  geom_jitter(width = 0.2, alpha = 0.2) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(outlier.shape = NA, alpha = 0.8) +
  xlab("Change in maternal age since last birth (yr)") +
  ylab("Maternal age (yr)") +
  theme_bw()

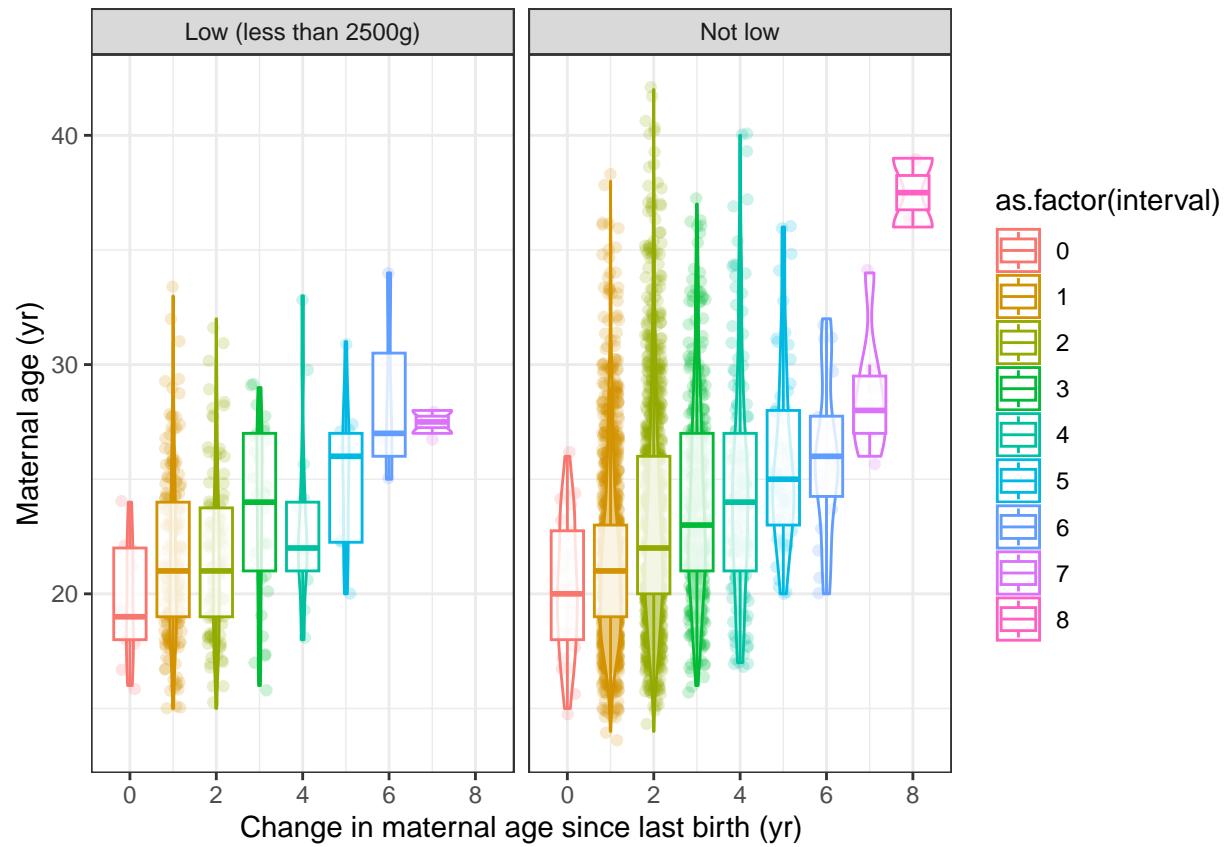
```

gg6.interval.age



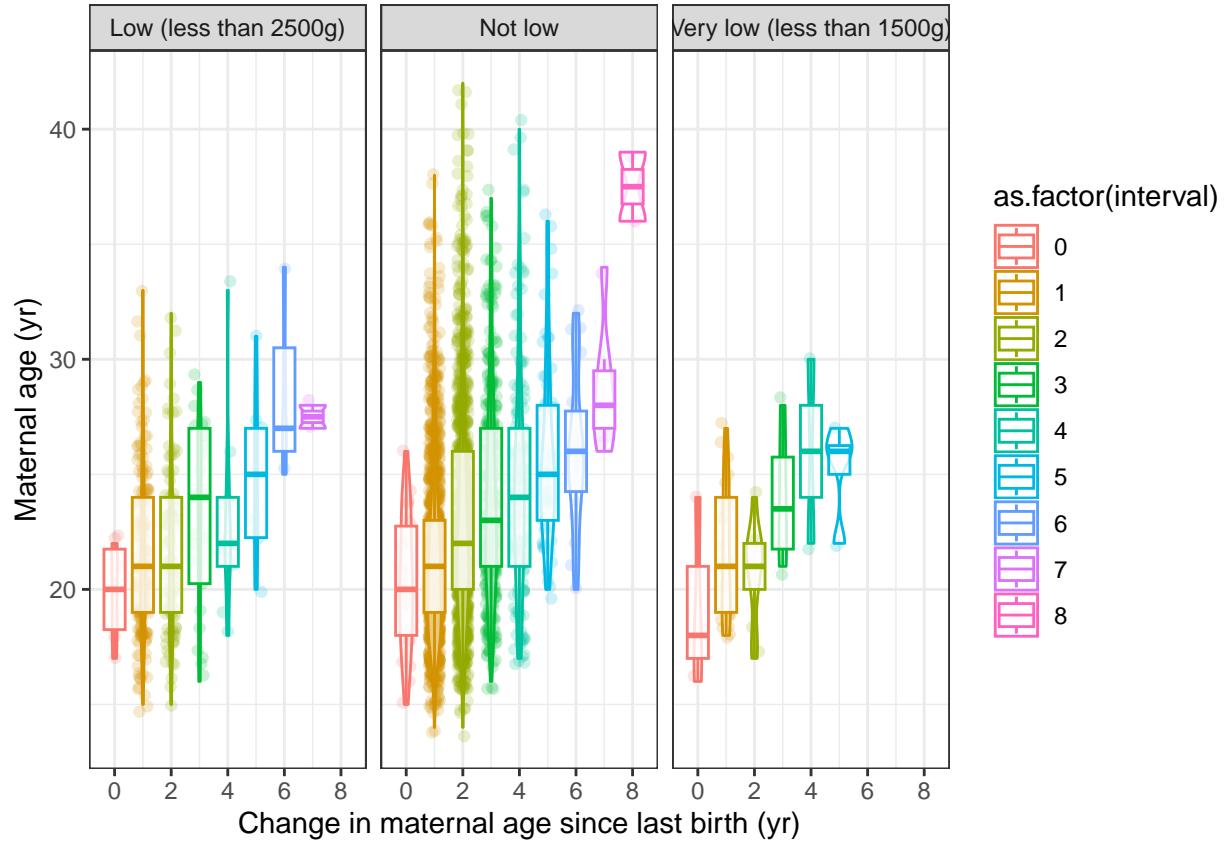
(a) Stratify the plot by low / not low birth weight

```
gg6.interval.age + facet_wrap(vars(birth.weight.binary))
```



(b) Stratify the plot by very low / low / not low birth weight

```
gg6.interval.age + facet_wrap(vars(birth.weight.factor))
```



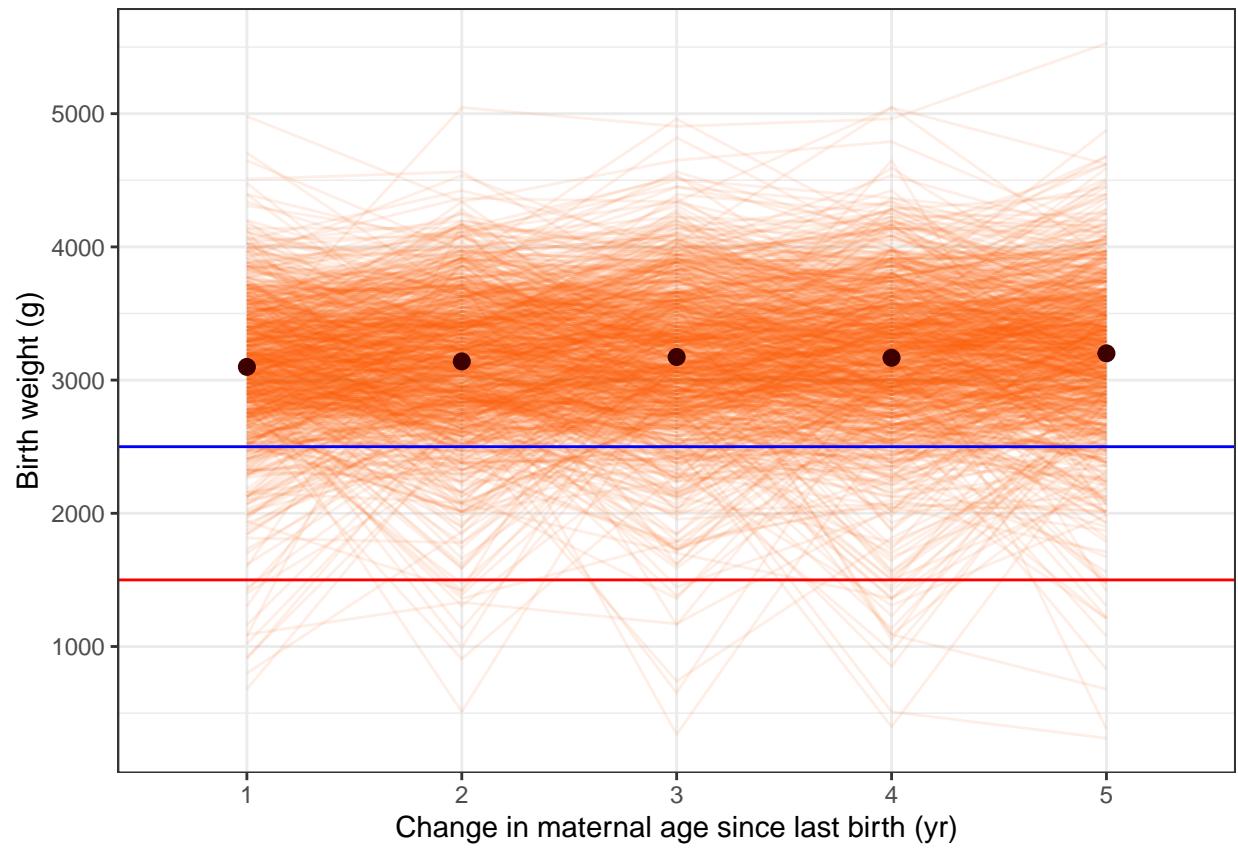
Spaghetti plots

- Produce a spaghetti plot of birth weights versus birth order, marking the threshold for low birth weight; include individual and overall averages. Does baseline birth weight seem to inform subsequent values?

```
# set.seed(540) # optional

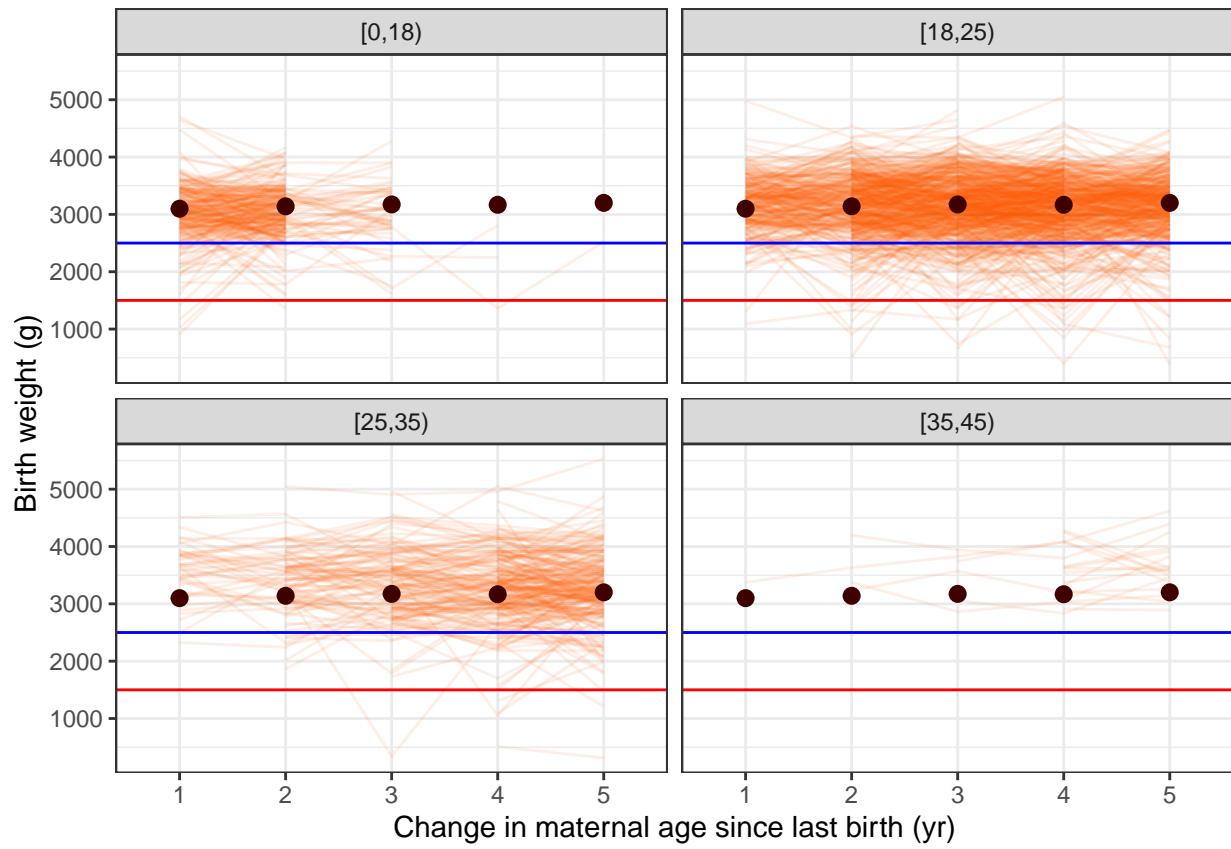
# produce a spaghetti plot of birth weight over time
gg7 <- birthwt %>%
  # plot repeated measures of birth weights
  ggplot(aes(y = birth.weight, x = birth.order)) +
  geom_line(aes(group = mother.id), alpha = 0.1, color = colors[1]) +
  geom_point(data = . %>%
    group_by(birth.order) %>%
    summarize(mean = mean(birth.weight)),
    aes(x = birth.order, y = mean), size = 2.5, color = colors[3]) +
  geom_hline(yintercept = 2500, color = "blue") +
  geom_hline(yintercept = 1500, color = "red") +
  xlab("Change in maternal age since last birth (yr)") +
  ylab("Birth weight (g)") +
  theme_bw()

gg7
```



(a) Color the plot by age group

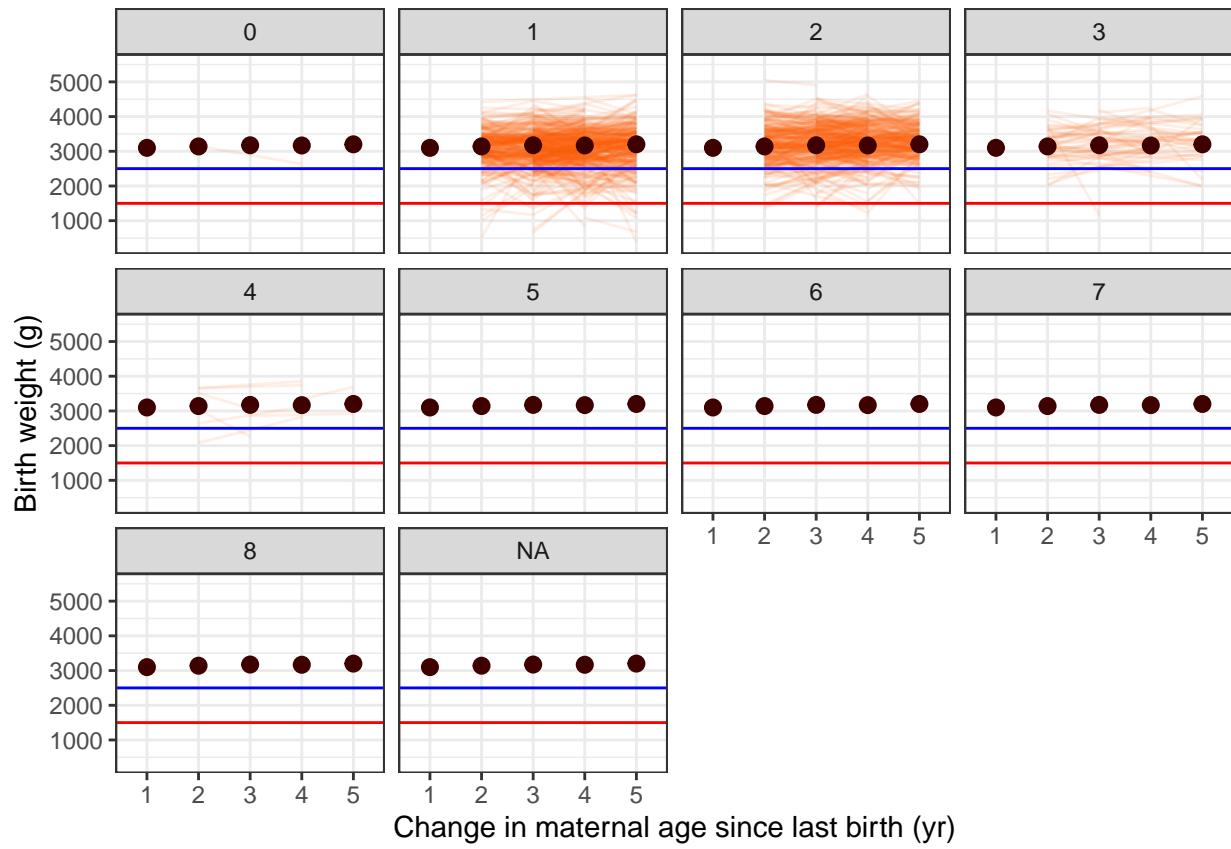
```
gg7 + facet_wrap(vars(age_group))
```



(b) Color the plot by interval

```
gg7 + facet_wrap(vars(interval))
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



8. Produce a spaghetti plot of birth weights versus maternal age, marking the threshold for low birth weight; include individual and overall averages. Does baseline birth weight seem to inform subsequent values?
- (a) Color the plot by (i) age group, (ii) interval
9. Produce a spaghetti plot of percent in birth weight group versus maternal age. Repeat for number in birth weight group versus maternal age. Compare the two. Does the most common birth weight group change as mothers age?

Modeling

Modeling birth weight

```

# Fixed intercept and random slope for each subject
lmm_model_fixedint <- lme4::lmer(
  birth.weight ~ as.numeric(birth.order) * maternal.age + (1 | mother.id),
  data = birthwt)

# Random intercept and random slope for each subject
lmm_model <- lme4::lmer(
  birth.weight ~ as.numeric(birth.order) * maternal.age + (as.numeric(birth.order) | mother.id),
  data = birthwt)

# Fit GEE model with independent working correlation structure
gee_independent <- geepack::geeglm(
  birth.weight ~ as.numeric(birth.order) * maternal.age, id = mother.id,
  data = birthwt, corstr = "independence")

# Fit GEE model with exchangeable working correlation structure
gee_exchangeable <- geepack::geeglm(
  birth.weight ~ as.numeric(birth.order) * maternal.age, id = mother.id,
  data = birthwt, corstr = "exchangeable")

# Fit GEE model with AR(1) working correlation structure
gee_ar1 <- geeglm(
  birth.weight ~ as.numeric(birth.order) * maternal.age, id = mother.id,
  data = birthwt, corstr = "ar1")

# Summary of the models
cbind(
  lmm_model_fixedint %>% summary %>% coef %>% data.frame %>%
    select(LMM.fixedint = Estimate, SE = Std..Error),
  lmm_model %>% summary %>% coef %>% data.frame %>%
    select(LMM = Estimate, SE = Std..Error),
  gee_independent %>% summary %>% coef %>%
    select(GEE.indp = Estimate, SE = Std.err),
  gee_exchangeable %>% summary %>% coef %>%
    select(GEE.exch = Estimate, SE = Std.err)) %>%
  # pretty print
  kable(digits = 3,
        caption = "Coefficient estimates and corresponding standard errors")

```

Birth weight ~ Birth order * Maternal age

Table 6: Coefficient estimates and corresponding standard errors

	LMM.fixedint	SE	LMM	SE	GEE.indp	SE	GEE.exch	SE
(Intercept)	2582.061	91.860	2565.706	88.355	2538.900	91.140	2581.975	88.773
as.numeric(birth.order)	4.803	23.511	11.972	25.267	-3.849	25.620	4.795	25.059

	LMM.fixedint	SE	LMM	SE	GEE.indp	SE	GEE.exch	SE
maternal.age	30.320	4.917	31.084	4.706	33.112	4.983	30.326	4.832
as.numeric(birth.order):maternal.age	-1.410	1.068	-1.725	1.109	-1.283	1.109	-1.410	1.078

```
# Fixed intercept and random slope for each subject
lmm_model_fixedint <- lme4::lmer(
  birth.weight ~ interval * maternal.age + (1 | mother.id),
  data = birthwt %>% na.omit)

# Random intercept and random slope for each subject
lmm_model <- lme4::lmer(
  birth.weight ~ interval * maternal.age + (as.numeric(birth.order) | mother.id),
  data = birthwt %>% na.omit)
```

Birth weight ~ Interbirth interval * Maternal age

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00259451 (tol = 0.002, component 1)
```

```
# Fit GEE model with independent working correlation structure
gee_independent <- geepack::geeglm(
  birth.weight ~ interval * maternal.age, id = mother.id,
  data = birthwt %>% na.omit, corstr = "independence")

# Fit GEE model with exchangeable working correlation structure
gee_exchangeable <- geeglm(
  birth.weight ~ interval * maternal.age, id = mother.id,
  data = birthwt %>% na.omit, corstr = "exchangeable")

# Fit GEE model with AR(1) working correlation structure
gee_ar1 <- geeglm(
  birth.weight ~ interval * maternal.age, id = mother.id,
  data = birthwt %>% na.omit, corstr = "ar1")

# Summary of the models
cbind(
  lmm_model_fixedint %>% summary %>% coef %>% data.frame %>%
    select(LMM.fixedint = Estimate, SE = Std..Error),
  lmm_model %>% summary %>% coef %>% data.frame %>%
    select(LMM = Estimate, SE = Std..Error),
  gee_independent %>% summary %>% coef %>%
    select(GEE.indp = Estimate, SE = Std.err),
  gee_exchangeable %>% summary %>% coef %>%
    select(GEE.exch = Estimate, SE = Std.err)) %>%
  # pretty print
  kable(digits = 3,
        caption = "Coefficient estimates and corresponding standard errors")
```

Table 7: Coefficient estimates and corresponding standard errors

	LMM.fixedint	SE	LMM	SE	GEE.indp	SE	GEE.exch	SE
(Intercept)	2741.680	96.951	2722.493	97.096	2641.496	120.574	2741.592	98.438
interval	41.947	39.763	48.155	39.813	7.689	51.965	41.939	40.964
maternal.age	16.999	4.167	18.024	4.210	21.566	5.268	17.003	4.263
interval:maternal.age	-0.781	1.610	-1.089	1.616	0.623	2.138	-0.780	1.635

```
# Fixed intercept and random slope for each subject
lmm_model_fixedint <- lme4::lmer(
  birth.weight ~ interval + maternal.age + as.numeric(birth.order) + (1 | mother.id),
  data = birthwt %>% na.omit)

# Random intercept and random slope for each subject
lmm_model <- lme4::lmer(
  birth.weight ~ interval + maternal.age + as.numeric(birth.order) + (as.numeric(birth.order) | mother.id),
  data = birthwt %>% na.omit)
```

Birth weight ~ Interbirth interval + Birth order + Maternal age

```
## Warning in checkConv(attr(opt, "derivs")), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00462825 (tol = 0.002, component 1)
```

```
# Fit GEE model with independent working correlation structure
gee_independent <- geepack::geeglm(
  birth.weight ~ interval + maternal.age + as.numeric(birth.order),
  id = mother.id, data = birthwt %>% na.omit, corstr = "independence")
```

```
# Fit GEE model with exchangeable working correlation structure
gee_exchangeable <- geeglm(
  birth.weight ~ interval + maternal.age + as.numeric(birth.order),
  id = mother.id, data = birthwt %>% na.omit, corstr = "exchangeable")
```

```
# Fit GEE model with AR(1) working correlation structure
gee_ar1 <- geeglm(
  birth.weight ~ interval + maternal.age + as.numeric(birth.order),
  id = mother.id, data = birthwt %>% na.omit, corstr = "ar1")
```

```
# Summary of the models
cbind(
  lmm_model_fixedint %>% summary %>% coef %>% data.frame %>%
    select(LMM.fixedint = Estimate, SE = Std..Error),
  lmm_model %>% summary %>% coef %>% data.frame %>%
    select(LMM = Estimate, SE = Std..Error),
  gee_independent %>% summary %>% coef %>%
    select(GEE.indp = Estimate, SE = Std.err),
  gee_exchangeable %>% summary %>% coef %>%
    select(GEE.exch = Estimate, SE = Std.err)) %>%
  # pretty print
```

```

kable(digits = 3,
      caption = "Coefficient estimates and corresponding standard errors")

```

Table 8: Coefficient estimates and corresponding standard errors

	LMM.fixedint	SE	LMM	SE	GEE.indp	SE	GEE.exch	SE
(Intercept)	2721.442	62.943	2716.925	63.266	2640.754	62.892	2721.303	63.030
interval	18.090	7.990	16.951	7.968	17.075	10.131	18.086	8.448
maternal.age	21.774	3.726	22.194	3.758	26.898	3.843	21.783	3.870
as.numeric(birth.order)	-21.993	9.500	-22.806	9.738	-31.480	9.941	-22.010	9.807

Modeling log odds of low birth weight

```
birthwt <- birthwt %>%
  mutate(birth.weight.binary <- ifelse(birth.weight.binary == "Not low", 0, 1))

# Fit GLMM to model log odds of low birth weight
glmm_model <- lme4::glmer(
  birth.weight.binary ~ maternal.age + interval + as.numeric(birth.order) +
  (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

round(glmm_model %>% summary %>% coef, 3)

##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.934     0.439   4.403   0.000
## maternal.age                0.047     0.024   1.911   0.056
## interval                     0.149     0.070   2.127   0.033
## as.numeric(birth.order)    -0.070     0.071  -0.988   0.323

glmm_model <- lme4::glmer(
  birth.weight.binary ~ maternal.age + interval +
  (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

round(glmm_model %>% summary %>% coef, 3)

##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.000     0.435   4.595   0.000
## maternal.age                0.032     0.019   1.674   0.094
## interval                     0.165     0.068   2.417   0.016
```

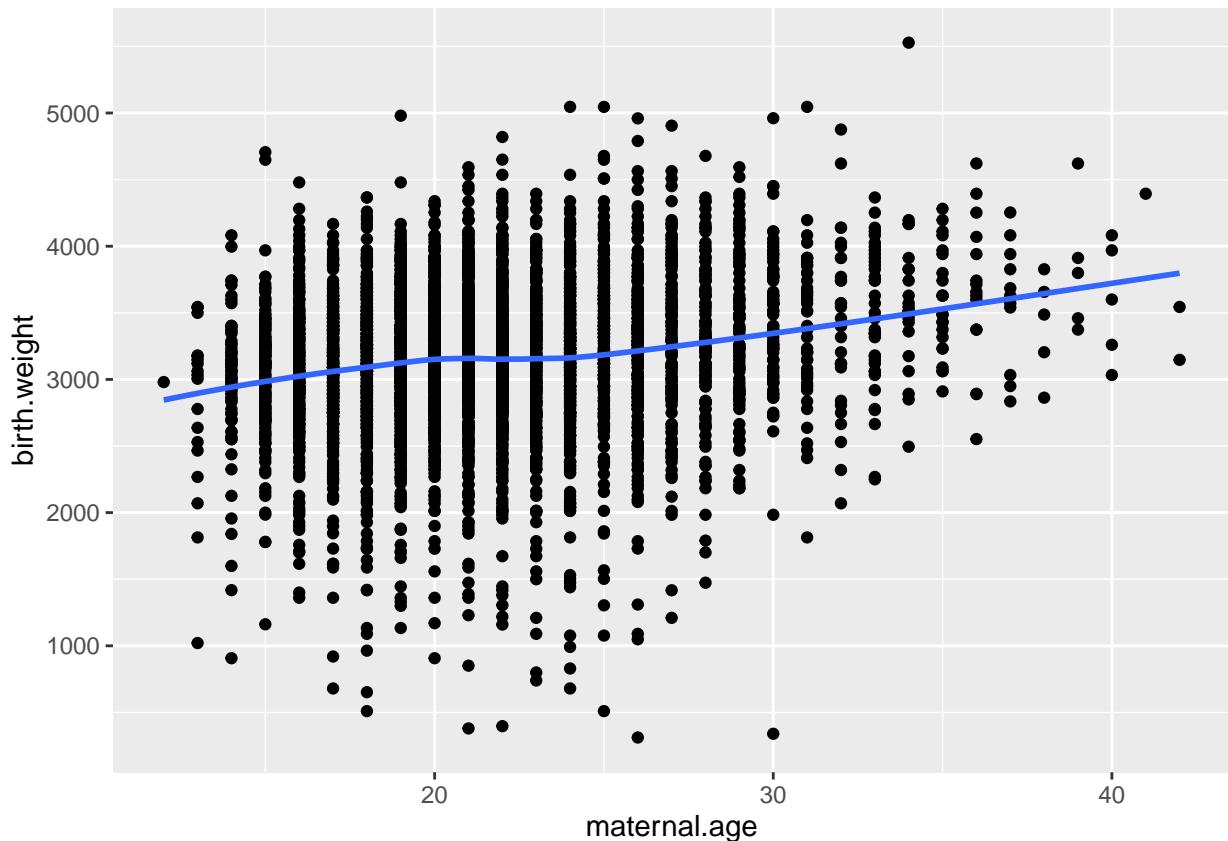
2. Model birth weight using a linear fixed effects model and interpret its coefficients.
3. Compare the estimates and standard errors of the two models. Do they suggest different conclusions about effects? Do you believe a nonlinear model is well-motivated?
4. Model low-weight birth using logistic regression and interpret its coefficients.

Birth weight vs covariates

Covariates: Maternal age, interpregnancy interval, birth order

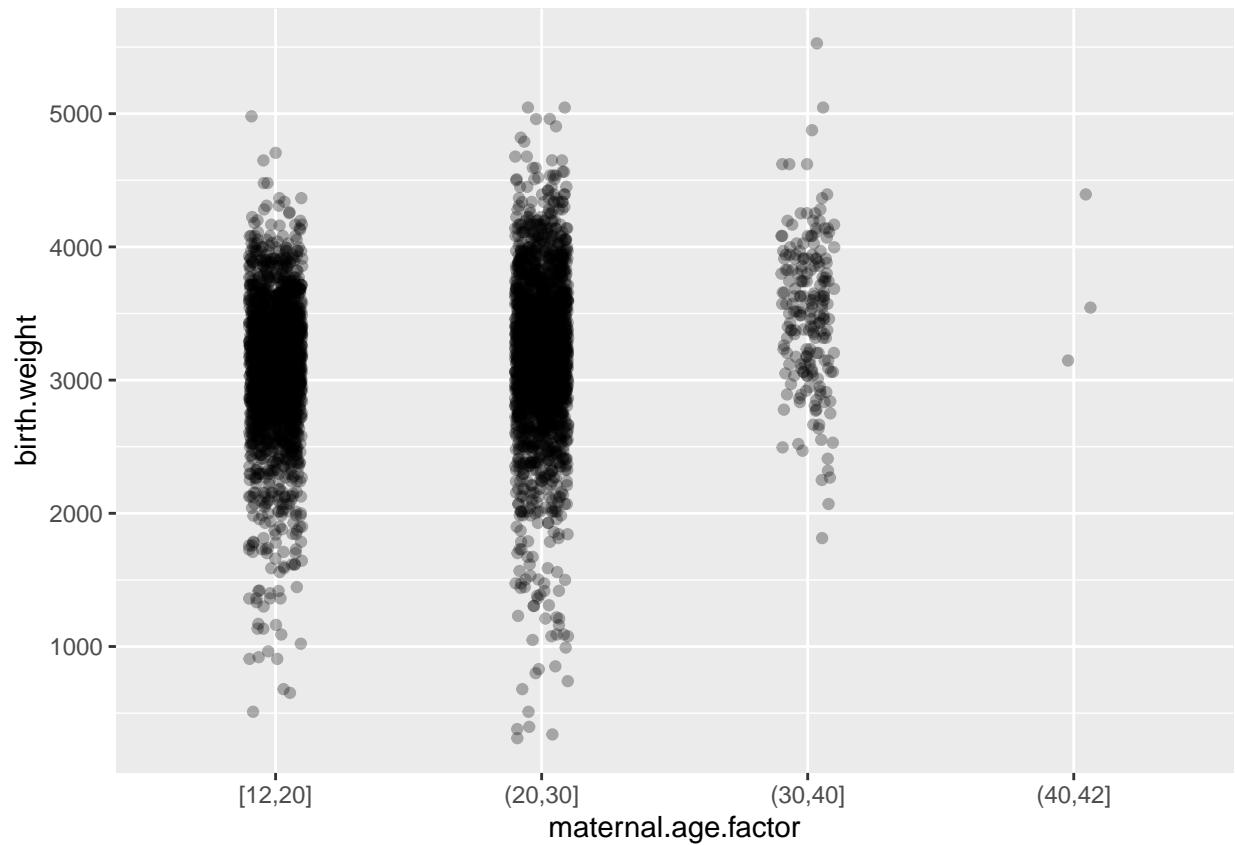
```
# plot birth weight vs maternal age
birthwt %>%
  ggplot(aes(x = maternal.age, y = birth.weight)) +
  geom_point() +
  geom_smooth(method = "loess", se = F)
```

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



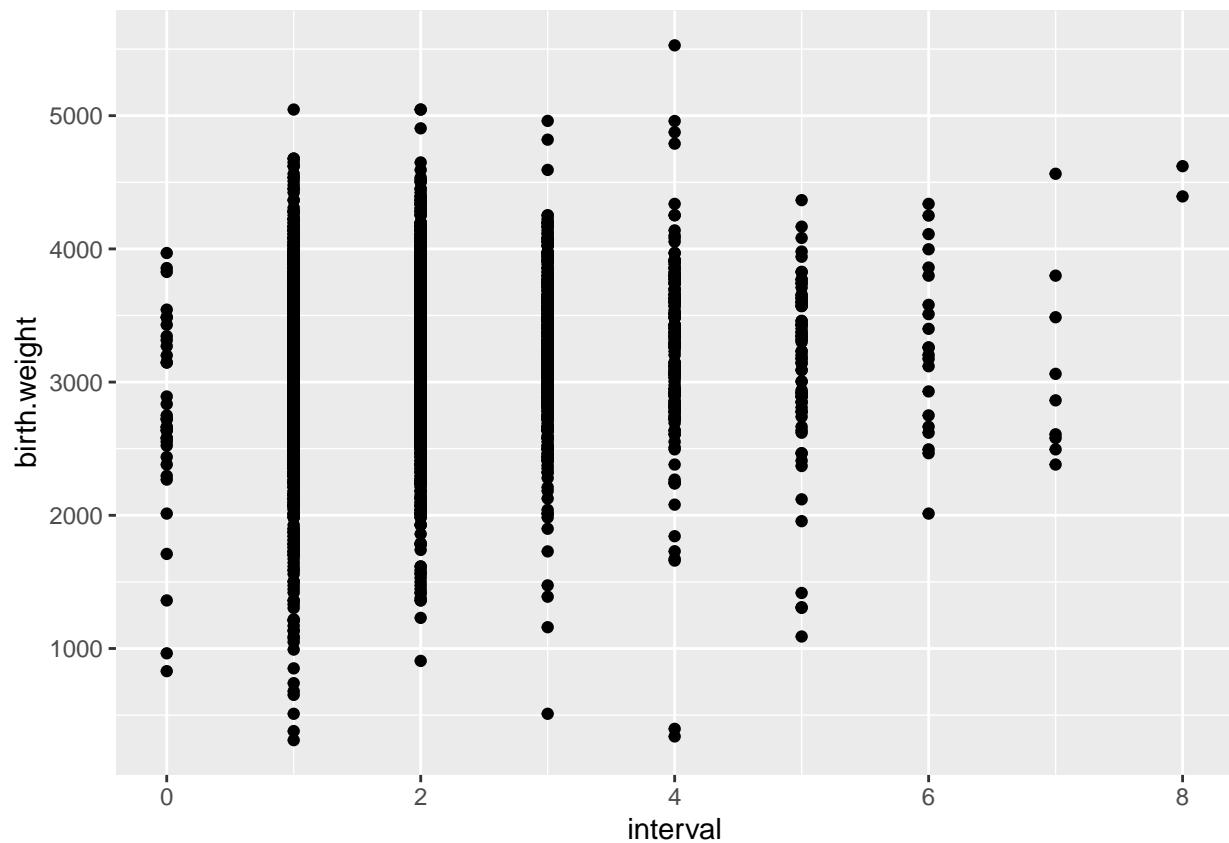
```
birthwt %>%
  ggplot(aes(x = maternal.age.factor, y = birth.weight)) +
  geom_jitter(width = 0.1, height = 0, alpha = 0.3) +
  geom_smooth(method = "loess", se = F)
```

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

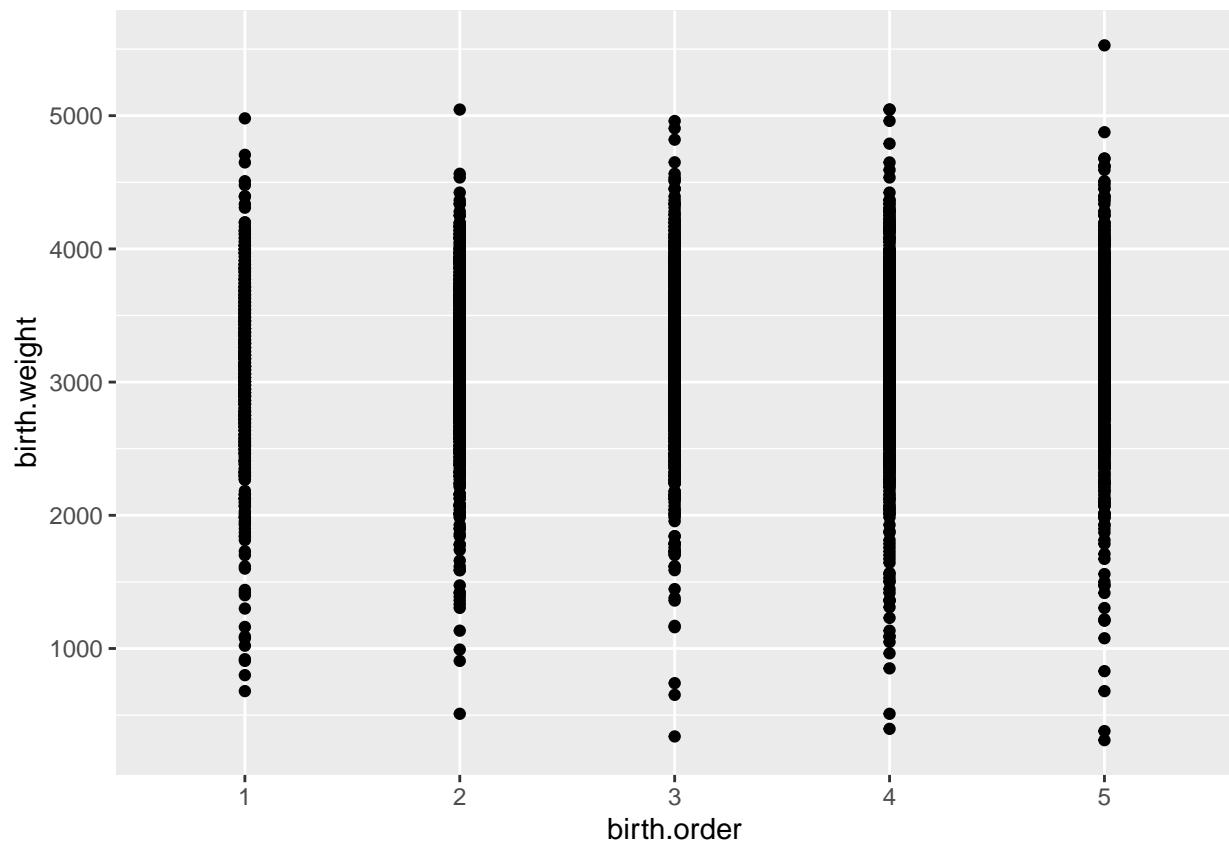


```
# plot birth weight vs interpregnancy interval
birthwt %>%
  ggplot(aes(y = birth.weight, x = interval)) +
  geom_point()

## Warning: Removed 876 rows containing missing values ('geom_point()').
```



```
# plot birth weight vs birth order
birthwt %>%
  ggplot(aes(y = birth.weight, x = birth.order)) +
  geom_point()
```



```
# plot birth weight vs birth order
birthwt %>%
  ggplot(aes(y = birth.weight, x = birth.order)) +
  geom_point()
```

