Final: Modules 1 - 6

Alex Shapovalov

2023-12-08

Note to students: Make sure to put your name in the “author” section above and do not change what is in the “date” section.

Final Exam Instructions

- The exam should be submitted as a pdf file. The easiest way is to knit as an html file and convert to a pdf. There is a video in module 1 about how you can do that!
- Partial credit is awarded based off of the work shown.
- If a question asks for a graph, table, or calculation (like an average), make sure that it appears in your knitted document.
- Your exam should be your own work. While you can use the internet for help, any major deviations to methods seen in this course will be marked incorrect, even if it gives the correct answer.
- The code should be readable and commented. If I’m unsure what your code did, I can’t award partial credit!

Setup

```
knitr::opts_chunk$set(echo = TRUE)

# Load the tidyverse, skimr, gt, and ggplot2 packages
pacman::p_load(tidyverse, skimr, gt, ggplot2, moderndive, GGally)

### Setting the default themes for ggplot ###
theme_set(theme_bw())
theme_update(plot.title = element_text(hjust = 0.5))

# Throughout this assignment we will be looking at a a study on pregnancies between 1960 and 1967.
# The data was collected from women in the Kaiser Foundation Health Plan in the San Francisco East Bay Area.
# This data was gathered from https://www.openintro.org/data/index.php?data=babies
# For a summary of the variables, please see the attached text file.
baby <- read_csv("babies.csv")
```

1) Understanding the Data

1.1) Cases and Variables **(1 pt)**

How many cases and how many variables are in this dataset? Classify each variable as either categorical or quantitative (how they are being treated in R, not how the researchers wanted to treat them). Display any code used to answer these questions below.

```
tail(baby, 6)
```

```
## # A tibble: 6 × 8
##    case   bwt gestation parity   age height weight smoke
##   <dbl> <dbl>     <dbl>  <dbl> <dbl>  <dbl>  <dbl> <dbl>
## 1  1231   132       270      0    27     65    126     0
## 2  1232   113       275      1    27     60    100     0
## 3  1233   128       265      0    24     67    120     0
## 4  1234   130       291      0    30     65    150     1
## 5  1235   125       281      1    21     65    110     0
## 6  1236   117       297      0    38     65    129     0
```

- [number of cases here] 1,236 cases (rows)
- [Number of variables here] 8 variables
- [Classify your variables here] There are 8 variables, all are numeric except for parity and smoke which are categorical

1.2) Data Cleaning **(1 pt)**

The variables of parity and smoke are not being treated correctly by R given what they represent. Additionally, having a response of “0” or “1” is not very helpful.

Create a new object called baby2 that transforms them into the correct type of variable and changes the responses in the following way: For parity: - **If the original response was a 0, instead have the response be recorded as “first”** - **If the original response was a 1, instead have the response be recorded as “not first” For smoke:** - **If the original response was a 0, instead have the response be recorded as “no”** - **If the original response was a 1, instead have the response be recorded as “yes”**

We are also missing some values for the gestation, smoke, and height variables which may cause an issue later, so let’s remove those missing values. Then show the skimmed version of the data to make sure the missing values were removed. Additionally, show the first 5 rows in the new baby2 datset.

```
baby2 <- baby

baby2$parity <- as.character(baby2$parity)
baby2$smoke <- as.character(baby2$smoke)

baby2 <- mutate(baby2, smoke = ifelse(smoke == "1", "yes", "no"))
baby2 <- mutate(baby2, parity = ifelse(parity == "1", "first", "not first"))

baby2 <- baby2 %>% filter(!is.na(gestation)) %>% filter(!is.na(smoke)) %>% filter(!is.na(height))

skim(baby2)
```

Data summary

Name	baby2
Number of rows	1192
Number of columns	8
_______________________	
Column type frequency:	
character	2
numeric	6
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
parity	0	1	5	9	0	2	0
smoke	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
case	0	1.00	621.70	357.27	1	312.75	621.5	932.25	1236	▇▇▇▇▇
bwt	0	1.00	119.48	18.26	55	108.00	120.0	131.00	176	▁▂▇▃▁
gestation	0	1.00	279.19	16.02	148	272.00	280.0	288.00	353	▁▁▁▇▁
age	1	1.00	27.23	5.80	15	23.00	26.0	31.00	45	▂▇▅▂▁
height	0	1.00	64.04	2.53	53	62.00	64.0	66.00	72	▁▁▇▇▁
weight	17	0.99	128.49	20.73	87	114.50	125.0	139.00	250	▃▇▁▁▁

```
head(baby2, 5)
```

```
## # A tibble: 5 × 8
##    case   bwt gestation parity      age height weight smoke
##   <dbl> <dbl>     <dbl> <chr>     <dbl>  <dbl>  <dbl> <chr>
## 1     1   120       284 not first    27     62    100 no
## 2     2   113       282 not first    33     64    135 no
## 3     3   128       279 not first    28     64    115 yes
## 4     5   108       282 not first    23     67    125 yes
## 5     6   136       286 not first    25     62     93 no
```

1.3) Data exploration **(1 pt)**

Create two separate tables, each displaying the mean, standard deviation, median, and IQR for the following: - **birth weight based on whether the mother smoked during the pregnancy** - **birth weight based on whether this was the mother’s first pregnancy**

```
# Table when grouping by smoke
smoke <- baby2 %>% group_by(smoke) %>% summarize(mean = mean(bwt), sd = sd(bwt), med = median(bwt), iqr = IQR(bw
t))
smoke
```

```
## # A tibble: 2 × 5
##   smoke  mean    sd   med   iqr
##   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 no     123.  17.4  124.    21
## 2 yes    114.  18.2  115     25
```

```
# Table when grouping by mother's first pregnancy
pregnancy <- baby2 %>% group_by(parity) %>% summarize(mean = mean(bwt), sd = sd(bwt), med = median(bwt), iqr = IQ
R(bwt))
pregnancy
```

```
## # A tibble: 2 × 5
##   parity     mean    sd   med   iqr
##   <chr>     <dbl> <dbl> <dbl> <dbl>
## 1 first      118.  17.3  118   21.8
## 2 not first  120.  18.6  120.  23
```

*Use the above two tables to answer the following question: Which variable (smoking of mother or whether this is the mother’s first child) seems to have have more of an effect on the birth weight of a baby?**

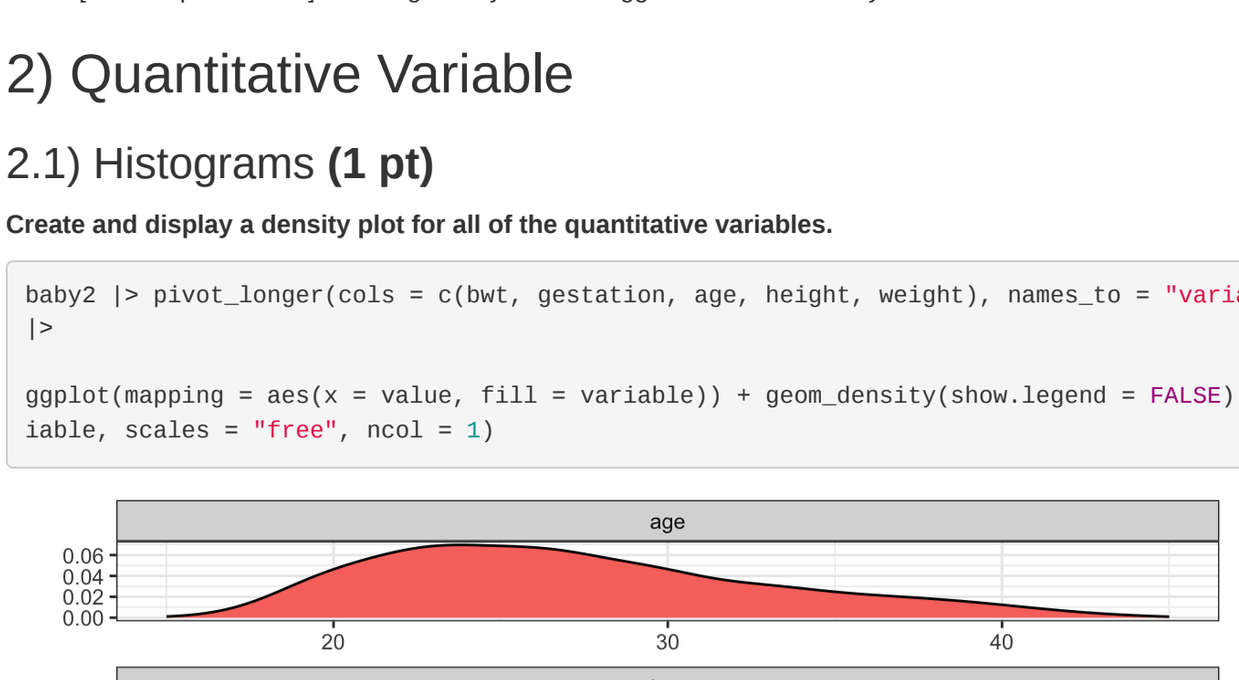
- [Your response here] Smoking clearly has the biggest effect with nearly a 10 ounce decrease in mean birthweight

2) Quantitative Variable

2.1) Histograms **(1 pt)**

Create and display a density plot for all of the quantitative variables.

```
baby2 |> pivot_longer(cols = c(bwt, gestation, age, height, weight), names_to = "variable", values_to = "value")
|>

ggplot(mapping = aes(x = value, fill = variable)) + geom_density(show.legend = FALSE) + facet_wrap(facets = ~ var
iable, scales = "free", ncol = 1)
```



*Describe the shape of the data using the plots above.**

- [Your answer here] Age is centered around 26 Bwt is centered around 123 Gestation is heavily centered around 275 Height is centered around 63 Weight is centered around 125 - 130

2.2) Mean vs Median **(1 pt)**

Should we use the mean or median of birth weight to describe what a typical birth weight would be? Briefly explain why.

- [Your answer here] Median is fine to use because the distribution is not left or right tailed. The distribution of weight is normal so we would expect median to equal mean

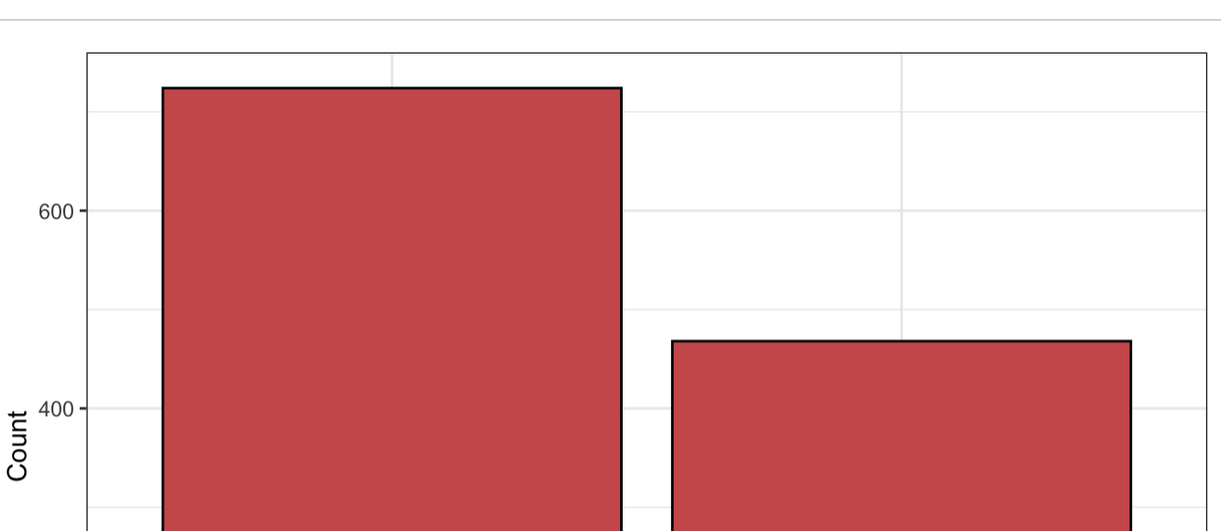
3) Categorical variables

3.1) Frequency Tables and Graphs **(1 pt)**

Make and display a bar graph for each of the categorical variables in this dataset. Create and display a relative frequency table for the categorical variables as well.

```
# Create the bar graphs
ggplot(data = baby2) + geom_bar(aes(x = smoke, y = ..count.., group = 1), fill = "indianred", color = "black") +
labs(y = "Count", x = "Smoke")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
ggplot(data = baby2) + geom_bar(aes(x = parity, y = ..count.., group = 1), fill = "lightblue", color = "black") +
labs(y = "Count", x = "Parity")
```


```
# Frequency Table for smoking:
table(baby2$smoke)/length(baby2$smoke)
```

```
##
##        no       yes
## 0.6073826 0.3926174
```

```
# Frequency Table for parity:
table(baby2$parity)/length(baby2$parity)
```

```
##
##     first not first
## 0.2600671 0.7399329
```

3.2 Two-Way Table **(1 pt)**

Create and display a two different two-way tables for the two variables of smoke and parity: - **One showing the frequencies for each combination** - *One showing the proportions for each combination**

```
# Make the two-way freq table
baby2freq <- table(baby2$smoke, baby2$parity)
baby2freq
```

```
##
##       first not first
##   no    190       534
##   yes   120       348
```

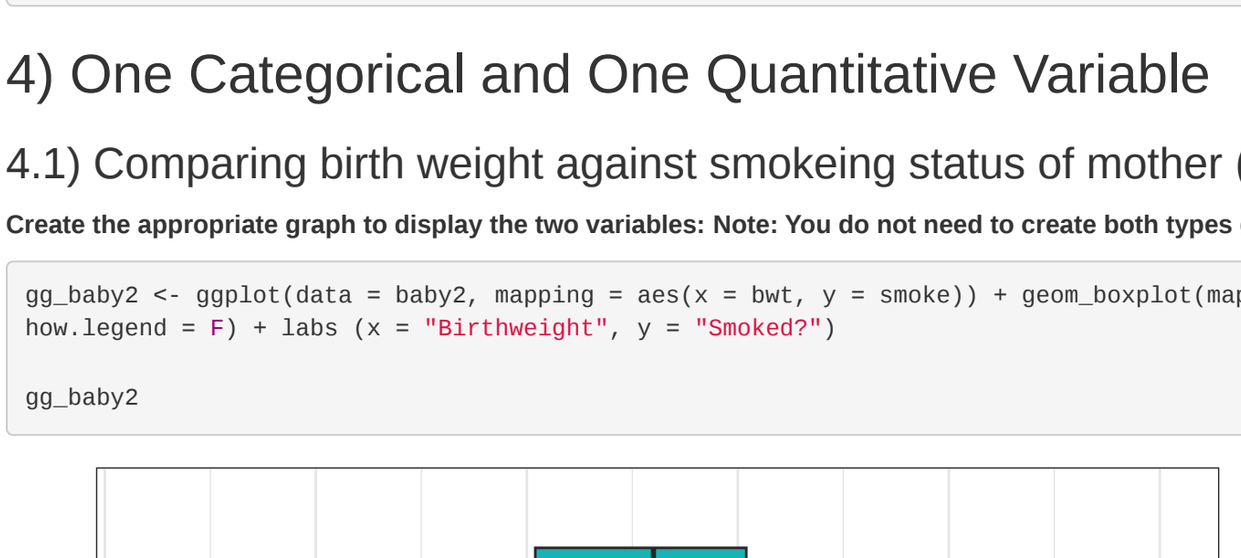
```
# Make the two-way proportion table
```

4) One Categorical and One Quantitative Variable

4.1) Comparing birth weight against smokeing status of mother **(1 pt)**

Create the appropriate graph to display the two variables: Note: You do not need to create both types graphs, only one of them.

```
gg_baby2 <- ggplot(data = baby2, mapping = aes(x = bwt, y = smoke)) + geom_boxplot(mapping = aes(fill = smoke), s
how.legend = F) + labs (x = "Birthweight", y = "Smoked?")

gg_baby2
```



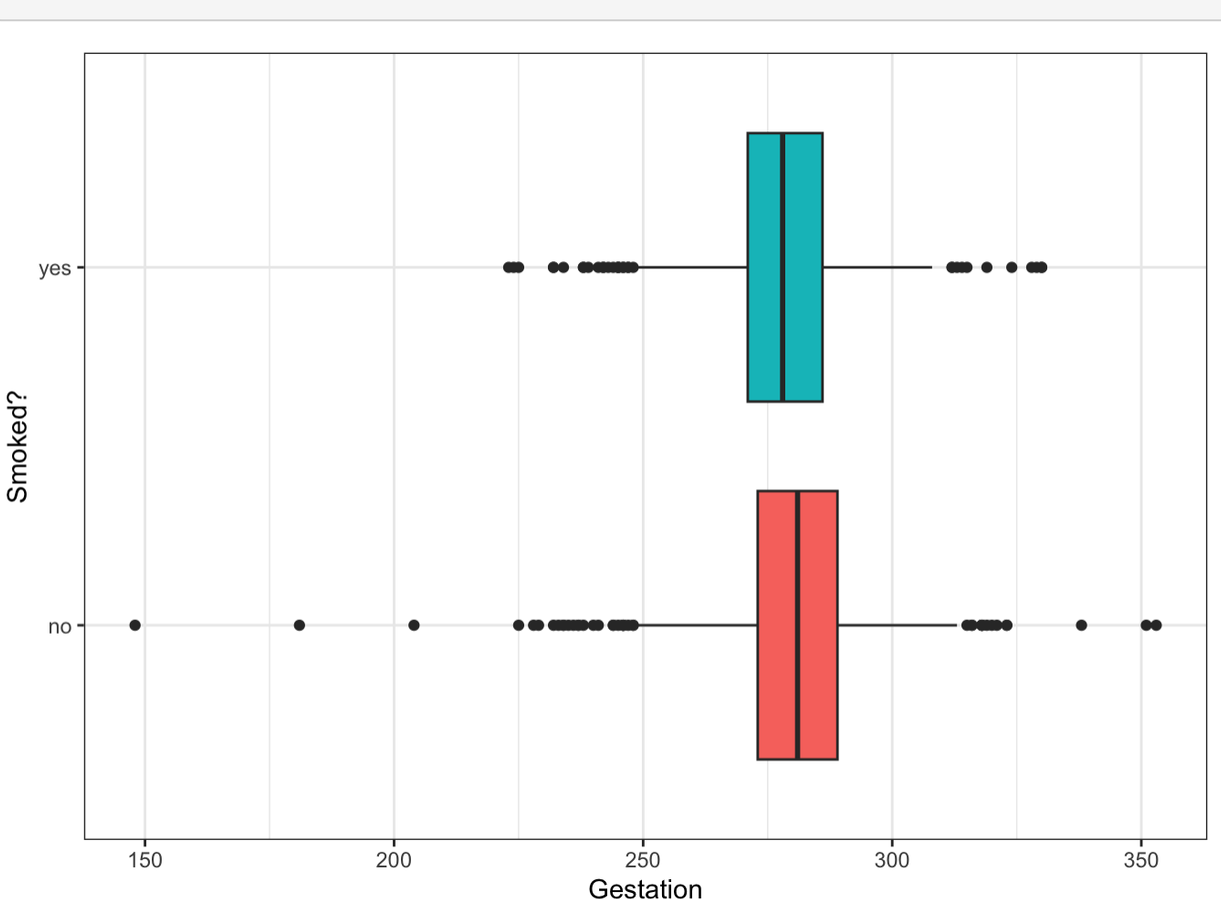
*Compare the distributions:**

- [Your answer here] The mothers that did smoke have a much lower birthweight on average than the mothers who didn’t smoke, but with little outliers.

4.2) Comparing gestation against smokeing status of mother **(1 pt)**

Create the appropriate graph to display the two variables: Note: You do not need to create both types graphs, only one of them.

```
gg_baby2 <- ggplot(data = baby2, mapping = aes(x = gestation, y = smoke)) + geom_boxplot(mapping = aes(fill = smo
ke), show.legend = F) + labs (x = "Gestation", y = "Smoked?")

gg_baby2
```



*Compare the distributions:**

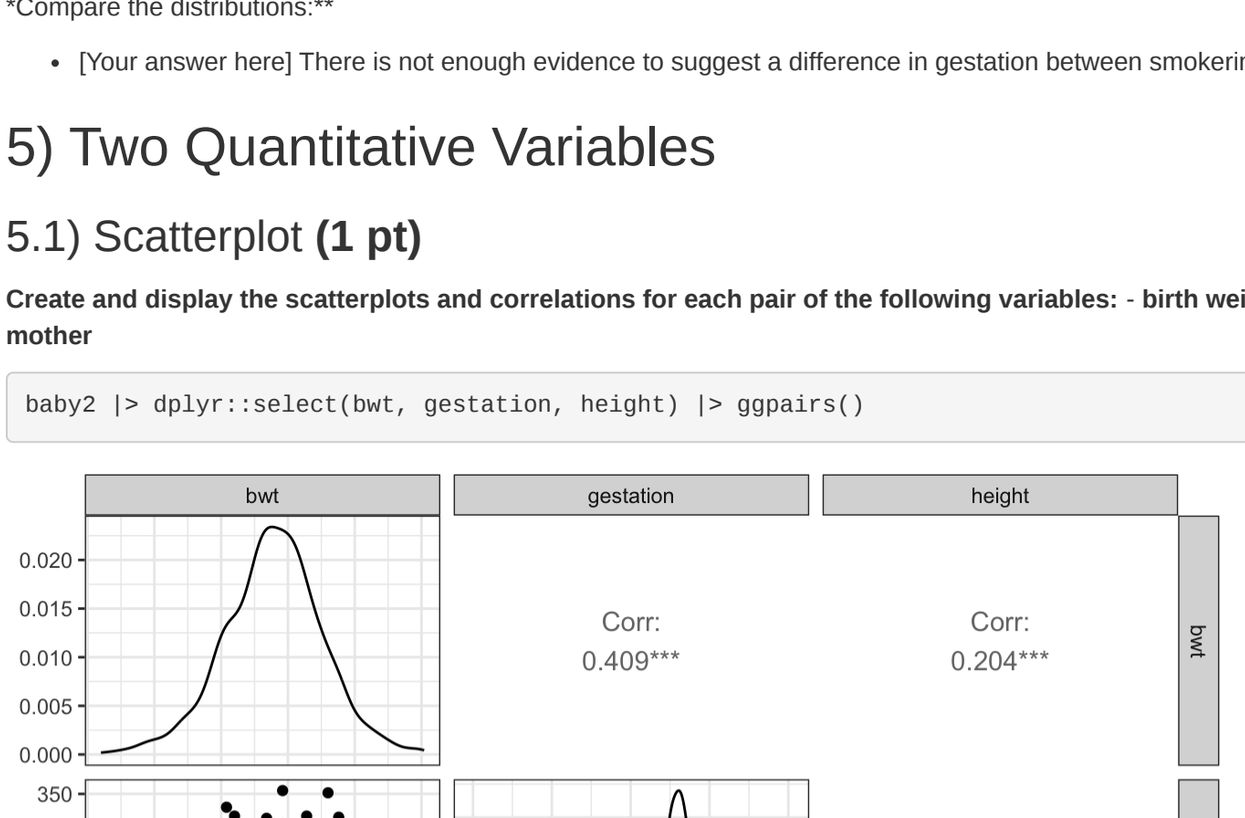
- [Your answer here] There is not enough evidence to suggest a difference in gestation between smokering and non-smoking mothers

5) Two Quantitative Variables

5.1) Scatterplot **(1 pt)**

Create and display the scatterplots and correlations for each pair of the following variables: - **birth weight** - **gestation** - **height of the mother**

```
baby2 |> dplyr::select(bwt, gestation, height) |> ggpairs()
```



5.2) Describing Scatterplots **(1 pt)**

Given the scatterplots and correlation, which variable would better predict the birth weight of a child on it’s own: gestation or height of the mother?

- [Your answer here] Height because gestation is always very similar no matter what, height can influence it more because that influences the weight of the mother

If you wanted to use gestation and height of the mother to predict the birth weight of the baby, would multicollinearity be an issue? Briefly justify your response.

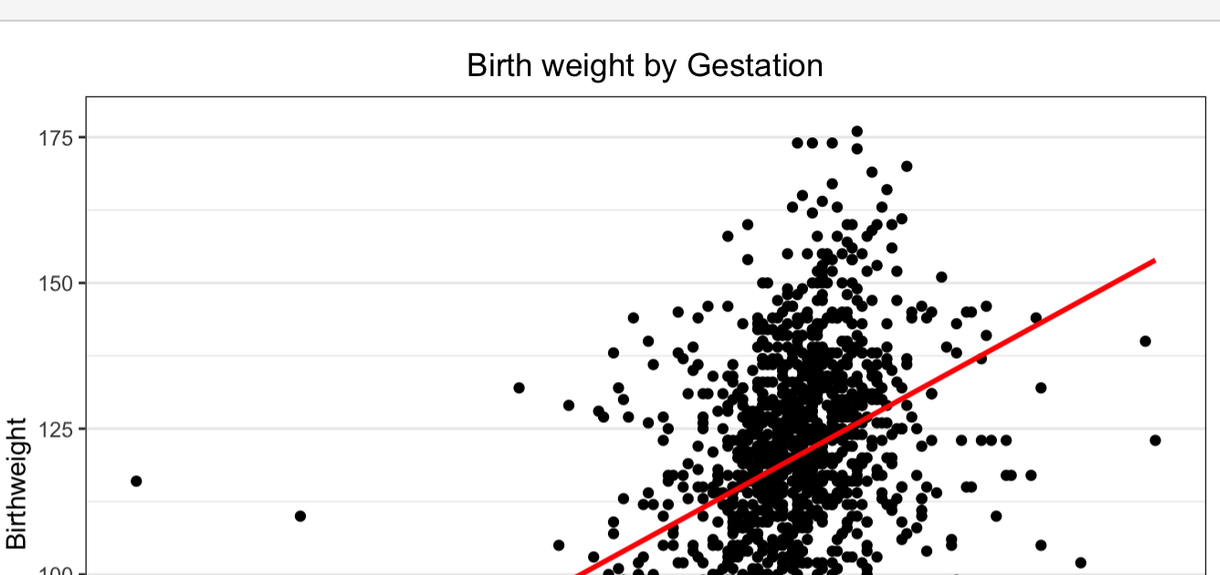
- [Your answer here] I don’t believe so because gestation is such an invariable factor, that it probably doesn’t correlate with much else

6) Models

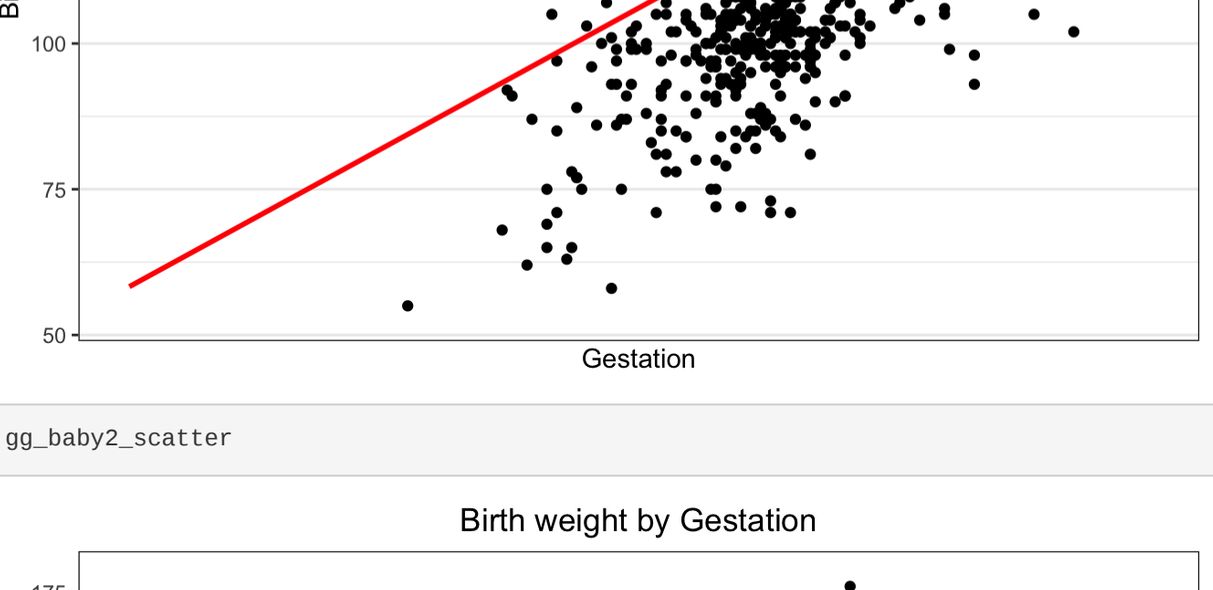
6.1) Simple Linear **(1.5 pts)**

Create the simple linear model that uses only gestation to predict the birth weight of the baby. Then display the model estimates as well as the r-squared, r-squared adjusted, an residual standard error. (Note: these do not have to all appear in the same output)

```
gg_baby2_scatter <- baby2 |>

ggplot(mapping = aes(x = gestation, y = bwt)) + geom_point() + labs(x = "Gestation", y = "Birthweight", title =
"Birth weight by Gestation") + scale_x_continuous(breaks = seq(2, 13, by = 2))

gg_baby2_scatter + geom_smooth(method = "lm", se = F, color = "red", formula = y ~ x)
```




```
gg_baby2_scatter
```

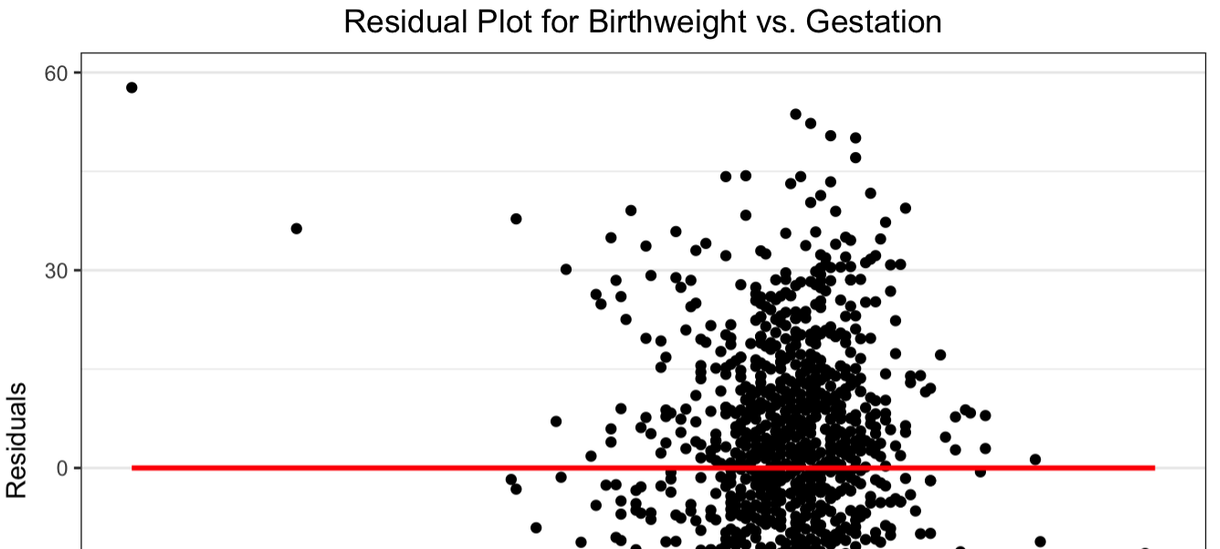


*Display the appropriate residual plot for this model.**

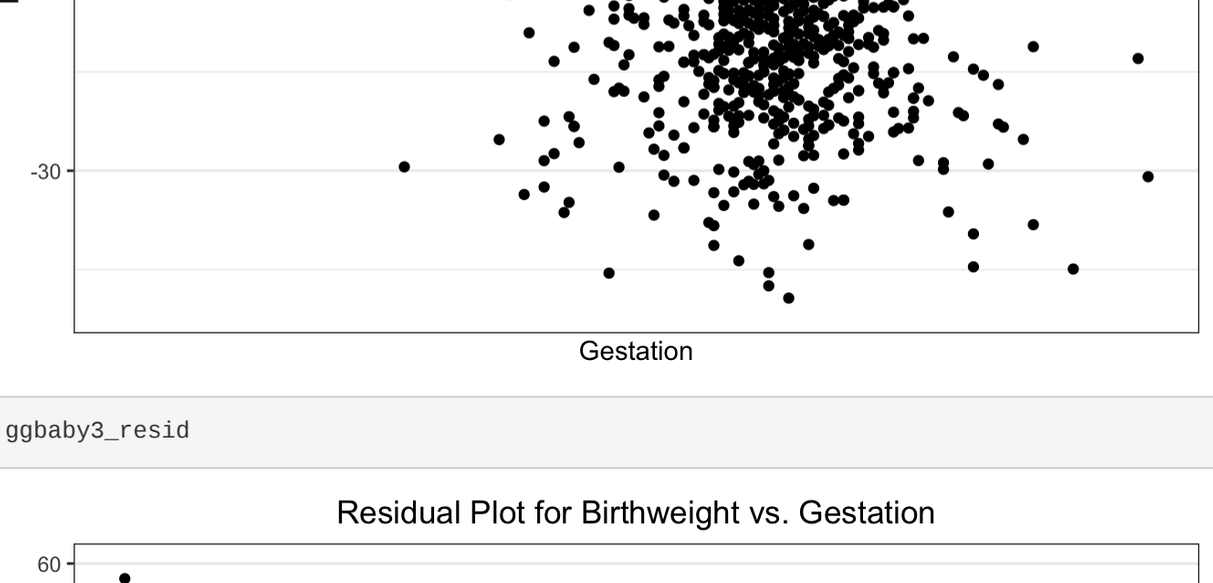
```
baby2_lm <- lm(formula = bwt ~ gestation, data = baby2)

baby2_lm
```

```
##
## Call:
## lm(formula = bwt ~ gestation, data = baby2)
##
## Coefficients:
## (Intercept)    gestation
##    -10.7375       0.4664
```

```
baby3 <- baby2 |> mutate(bwt_hat = baby2_lm$fitted, residual = baby2_lm$residuals)

ggbaby3_resid <- baby3 |>

ggplot(mapping = aes(x = gestation, y = residual)) + geom_point() + labs(title = "Residual Plot for Birthweight v
s. Gestation", x = "Gestation", y = "Residuals") + scale_x_continuous(breaks = seq(2, 13, by = 2))

ggbaby3_resid + geom_smooth(method = "lm", se = F, color = "red",formula = y ~ x)
```




```
ggbaby3_resid
```



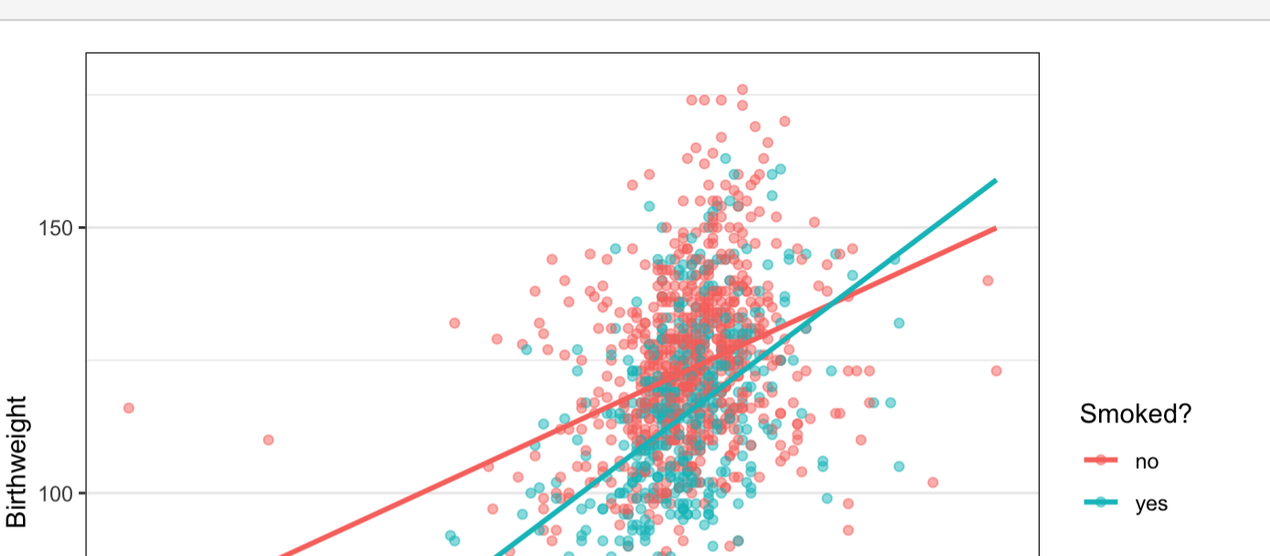
6.2) Interaction - gestation and smoke **(1.5 pts)**

Create the interaction model that uses gestation and whether the mother smoked or not to predict the birth weight of the baby. Then display the model estimates as well as the r-squared, r-squared adjusted, an residual standard error. (Note: these do not have to all appear in the same output)

```
gg_baby2_1 <- ggplot(data = baby2, mapping = aes(x = gestation, y = bwt, color = smoke)) + geom_point(alpha = 0.
5) + labs(x = "Gestation", y = "Birthweight", color = "Smoked?") + scale_x_continuous(breaks = seq(2, 12, by =
2))

baby2_interact <- lm(formula = bwt ~ smoke * gestation, data = baby2)

get_regression_table(baby2_interact) |> dplyr::select(term, estimate)
```

```
## # A tibble: 4 × 2
##   term                 estimate
##   <chr>                   <dbl>
## 1 intercept              20.3
## 2 smoke: yes            -73.5
## 3 gestation               0.367
## 4 smoke: yes:gestation    0.234
```

```
gg_baby2_1 + geom_smooth(method = "lm", formula = y ~ x, se = F, fullrange = T)
```



*Display the appropriate residual plot for this model.**

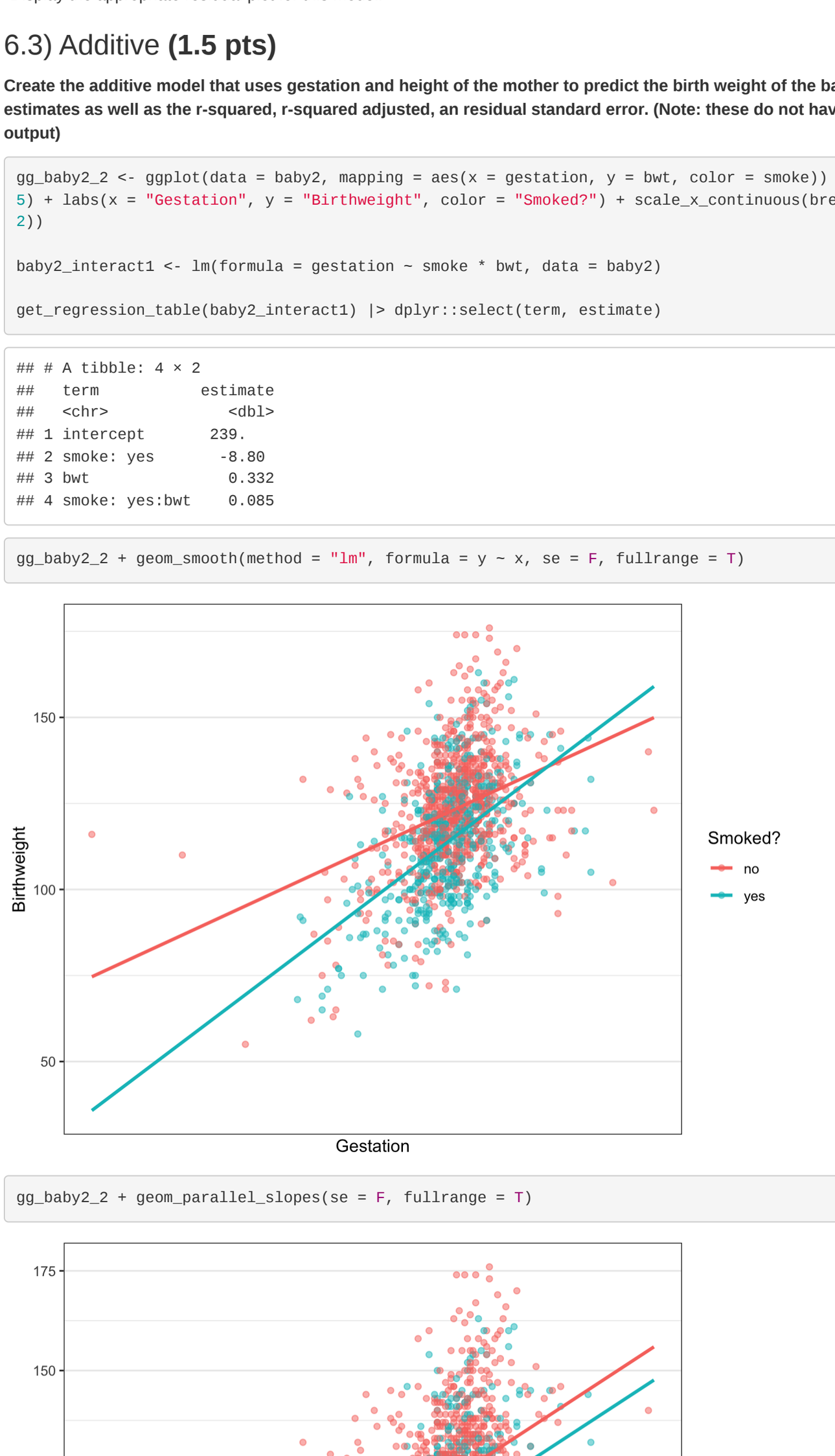
6.3) Additive **(1.5 pts)**

Create the additive model that uses gestation and height of the mother to predict the birth weight of the baby. Then display the model estimates as well as the r-squared, r-squared adjusted, an residual standard error. (Note: these do not have to all appear in the same output)

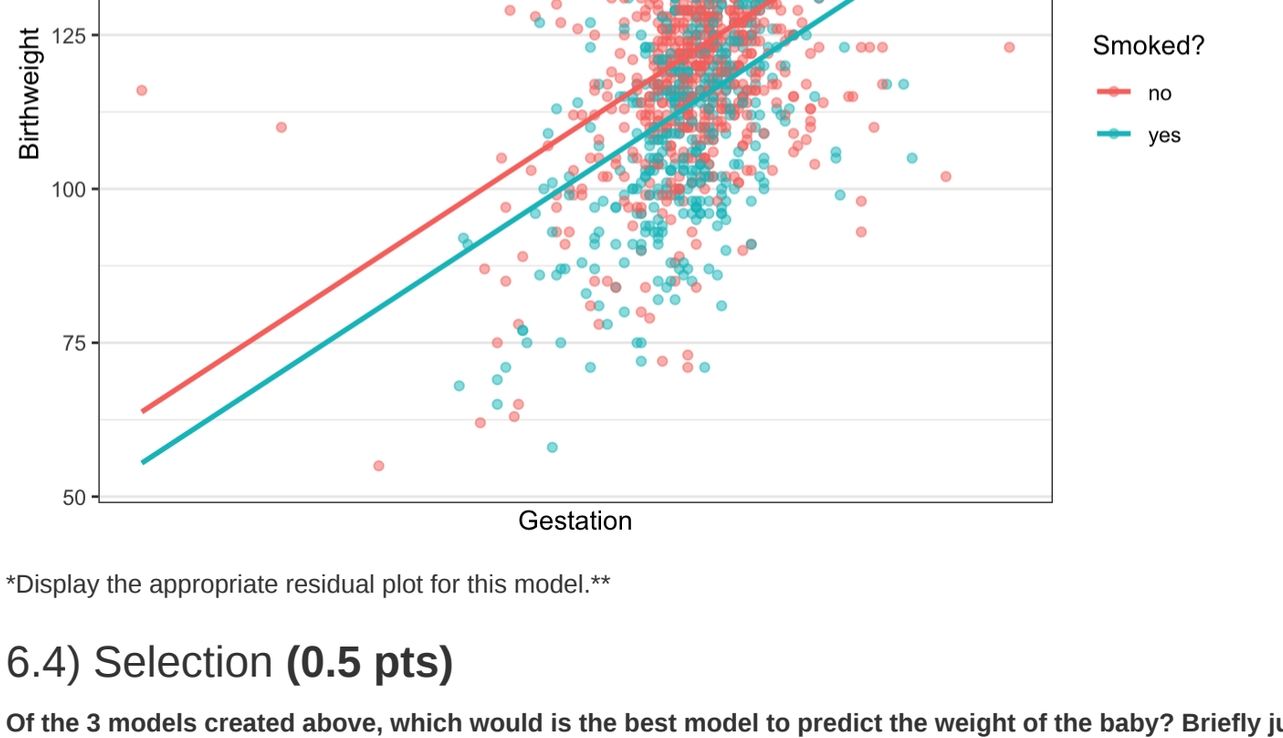
```
gg_baby2_2 <- ggplot(data = baby2, mapping = aes(x = gestation, y = bwt, color = smoke)) + geom_point(alpha = 0.
5) + labs(x = "Gestation", y = "Birthweight", color = "Smoked?") + scale_x_continuous(breaks = seq(2, 12, by =
2))

baby2_interact1 <- lm(formula = gestation ~ smoke * bwt, data = baby2)

get_regression_table(baby2_interact1) |> dplyr::select(term, estimate)
```

```
## # A tibble: 4 × 2
##   term           estimate
##   <chr>             <dbl>
## 1 intercept       239.
## 2 smoke: yes       -8.80
## 3 bwt               0.332
## 4 smoke: yes:bwt    0.085
```

```
gg_baby2_2 + geom_smooth(method = "lm", formula = y ~ x, se = F, fullrange = T)
```




```
gg_baby2_2 + geom_parallel_slopes(se = F, fullrange = T)
```



*Display the appropriate residual plot for this model.**

6.4) Selection **(0.5 pts)**

Of the 3 models created above, which would is the best model to predict the weight of the baby? Briefly justify your response.

- [Your answer here] The parallel lines additives model because it clearly shows that if a mother smokes, the birthweight line is much lower