Data Analysis #1

Dustin Leatherman

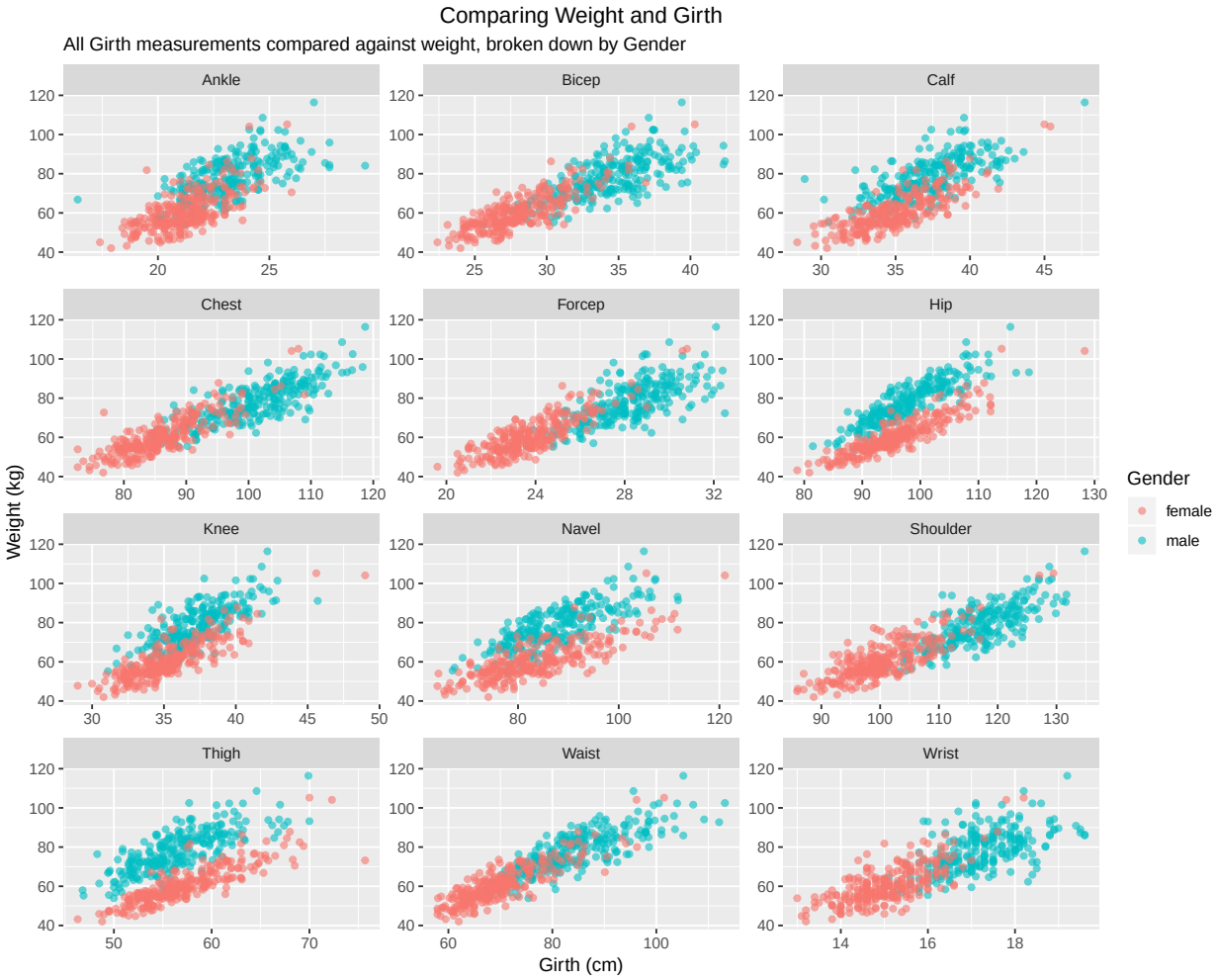
February 9, 2019

Body Girth and Physical Activity

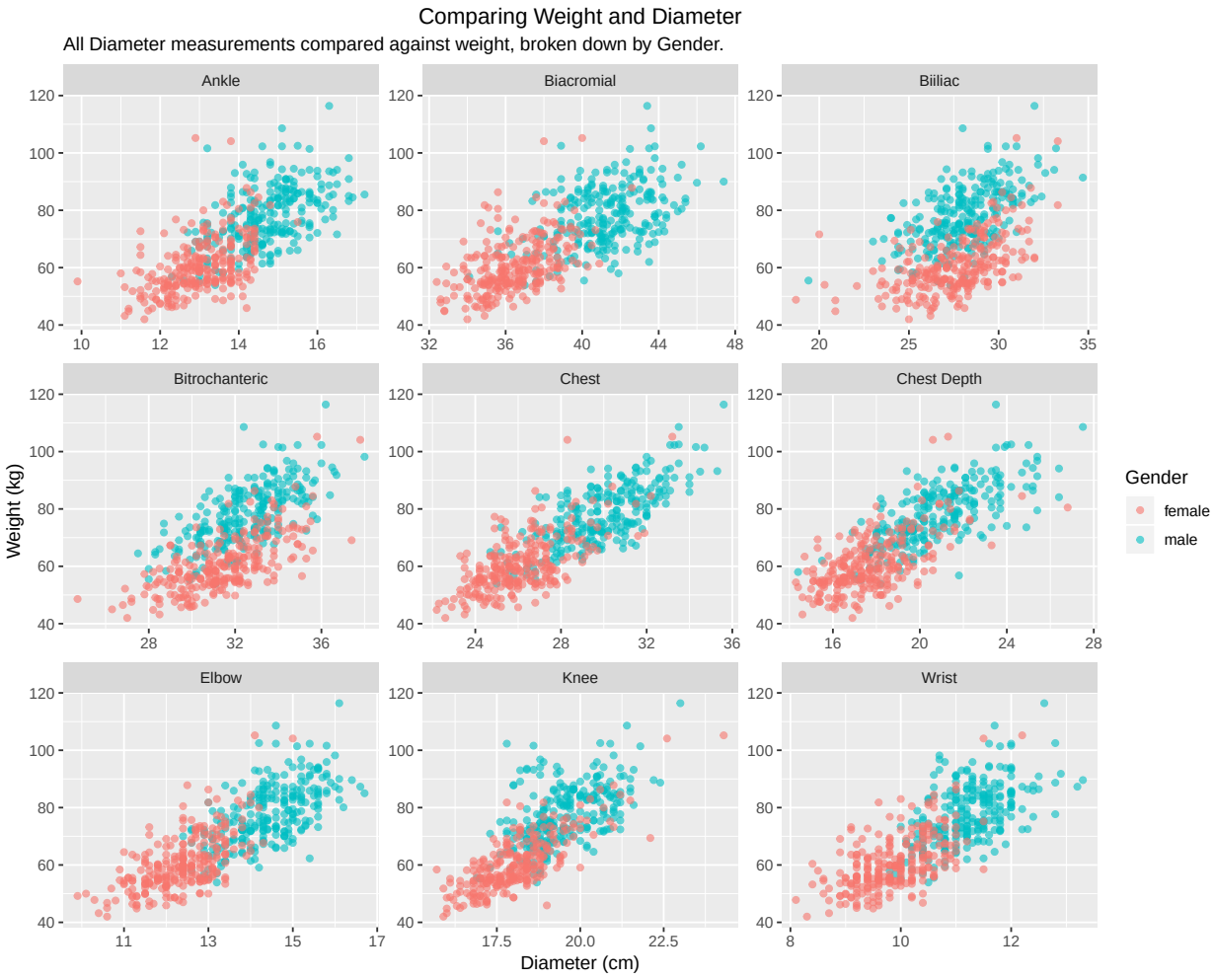
Introduction

Body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, were measured for 507 physically active individuals - 247 men and 260 women. This dataset was initially published in the article, "Exploring Relationships in Body Dimensions" in the *Journal of Statistics Education*.

Many measurements were taken but the researchers are particularly interested in predicting an individual's weight in kilograms from just their waist girth in centimeters. Even so, it is beneficial to compare weight with all other measurements to see why the researchers made their decision.



```r
# Show all diameter-related scatterplots with relation to weight. Color-coded by gender.
bdimsWithGender %>%
  select(
    Weight = wgt,
    Gender = gender,
    Biacromial = bia.di,
    Biiliac = bii.di,
    Bitrochanteric = bit.di,
    `Chest Depth` = che.de,
    Chest = che.di,
    Elbow = elb.di,
    Wrist = wri.di,
    Knee = kne.di,
    Ankle = ank.di) %>%
  # remove wgt and gender from gathered data. We don't want to include these as explanatory variables i
  gather(c(-Weight,-Gender), key = "variable", value = "value") %>%
  ggplot(aes(x = value, y = Weight, color = Gender)) +
    geom_point(alpha = 0.6) +
    facet_wrap(~ variable , ncol = 3, scales = "free") +
      labs(color = "Gender", subtitle = "All Diameter measurements compared against weight, broken down
      ylab("Weight (kg)") +
      xlab("Diameter (cm)") +
      ggtitle("Comparing Weight and Diameter") +
        theme(plot.title = element_text(hjust = 0.5))
```




```
# Show the rest of the variables of interest. In this case, it is only age and height
plot.height <- bdimsWithGender %>%
  select(Weight = wgt, Gender = gender, Height = hgt) %>%
  ggplot(aes(x = Height, y = Weight, color = Gender)) +
    geom_point(alpha = 0.6) +
      ylab("Weight (kg)") +
      xlab("Height (cm)") +
      ggtitle("Comparing Weight and Height") +
        theme(plot.title = element_text(hjust = 0.5))

plot.age <- bdimsWithGender %>%
  select(Weight = wgt, Gender = gender, Age = age) %>%
  ggplot(aes(x = Age, y = Weight, color = Gender)) +
    geom_point(alpha = 0.6, show.legend = FALSE) +
      ylab("Weight (kg)") +
      xlab("Age (years)") +
      ggtitle("Comparing Weight and Age") +
        theme(plot.title = element_text(hjust = 0.5))

grid.arrange(plot.age, plot.height,
               widths = c(1,1),
```

```
                ncol = 2)
```

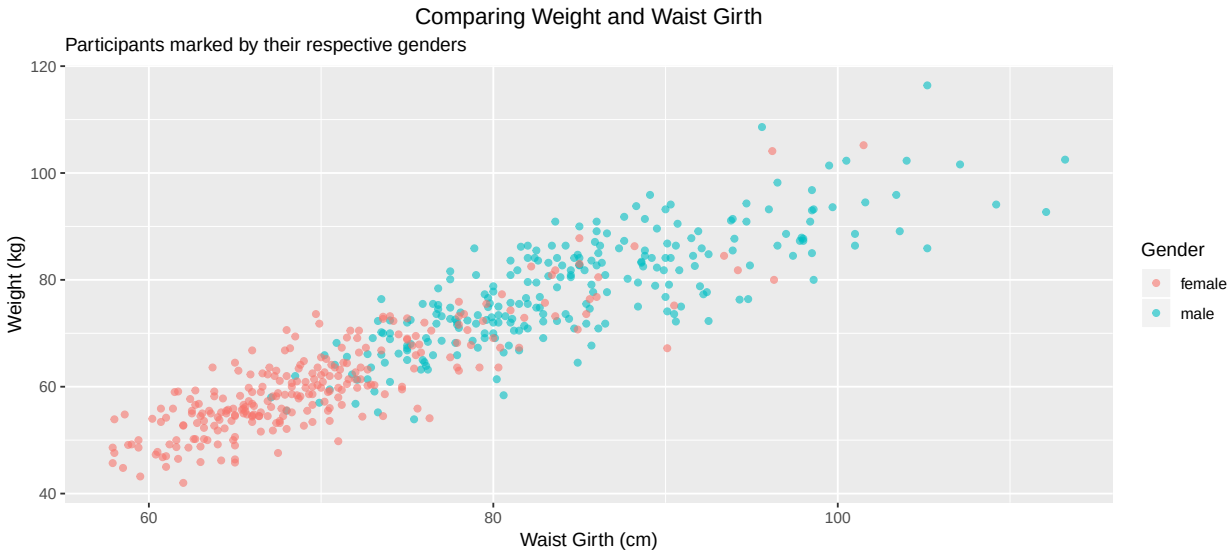
All girth and diameter explanatory variables have a positive correlation with weight. It appears that girth of body parts in general may have a stronger correlation to weight than diameter. This can be determined from the more focused trendline in the girth plots than the diameter plots. In all cases, there appears to be a clear distinction between males and females in terms of weight. Males appear to be heavier than females for all measurements barring age. There are two visible patterns within the graphs. The first being those which females and males have little overlapping variables. Examples of this include Forcep Girth vs Weight, Chest Girth vs Weight, and Wrist Girth vs Weight. The second pattern is that for a male and female of the same measurement, the male is generally heavier. This may indicate that this measurement is not as impactful as one where there are fewer overlapping values. Examples of this include Thigh Girth vs Weight, Calf Girth vs Weight, and Bicep Diameter vs Weight. Waist Girth vs Weight has a fairly clear and focused trendline. There appears to be a decent amount of intersection between males and females in the middle so it is understandable why this is the main variable of interest.

The simple linear regression model they have used is as follows:

μ {Weight|Waist girth} = $\beta_0 + \beta_1$Waist girth

This linear model does not take gender into account. As seen in the scatterplots, gender appears to be associated with weight so the researchers are interested in whether they should be using different lines for men and women.

```
bdimsWithGender %>%
  select("wgt", "wai.gi", "gender") %>%
  ggplot(aes(x = wai.gi, y = wgt, color = gender)) +
    geom_point(alpha = 0.6) +
      labs(color = "Gender", subtitle = "Participants marked by their respective genders") +
      ylab("Weight (kg)") +
      xlab("Waist Girth (cm)") +
      ggtitle("Comparing Weight and Waist Girth") +
        theme(plot.title = element_text(hjust = 0.5))
```



These are our questions of interest:

- After taking into account Waist Girth, is the deviation in weight between men and women variable or fixed?
- Based on what is known, what is the most appropriate model for estimating average weight?

Methods

For convenience, the single parameter model is as follows:

μ {Weight|Waist girth} = $\beta_0 + \beta_1$Waist girth

The most common method to create a separate lines model is to add an interaction term. Since the researchers are interested in the deviation between men and women, gender should be added as an indicator variable as well as an interaction term with waist girth.

The separate lines model is as follows:

μ {Weight|Waist girth, MALE} = $\beta_0 + \beta_1$Waist girth + β_2MALE + β_3(MALE $\times$ Waist girth)

```
# canned waist girth values per gender
new_data <- expand.grid(wai.gi = c(60,80,100), gender = unique(bdimsWithGender$gender))

# I tried for longer than I care to admit to make models work within the tidyverse framework
# I was able to get it to run just fine in my console but would run into strange issues when
# generating the report here. It's a shame because I really like Tidyverse and the principles
# behind it. Alas, there must be some error (user or otherwise) that makes it not play nicely with RMar
model.initial <- lm(wgt ~ wai.gi, data = bdimsWithGender)
model.initial.conf <- confint_tidy(model.initial)
model.initial.conf.pred <- predict(model.initial, newdata = new_data, interval = "confidence")

model.separate_lines <- lm(wgt ~ wai.gi * gender, data = bdimsWithGender)
model.separate_lines.conf <- confint_tidy(model.separate_lines)
model.separate_lines.conf.pred <- predict(model.separate_lines, newdata = new_data, interval = "confide
```

```
kable(
  new_data %>%
    bind_cols(
      model.initial.conf.pred %>%
        tidy,
      model.separate_lines.conf.pred %>%
        tidy
      ) %>%
    arrange(wai.gi),
  align = c('l'),
  digits = 4,
  col.names = c("Waist Girth (cm)","Gender","Fit", "Lower", "Upper", "Fit", "Lower", "Upper"),
  caption = "Fit + Confidence Interval for specified values of Waist Girth"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  add_header_above(c(" " = 2, "Single Parameter (kg)" = 3, "Separate Lines (kg)" = 3)) %>%
  row_spec(0, bold = T)
```

Table 1: Fit + Confidence Interval for specified values of Waist Girth

		Single Parameter (kg)			Separate Lines (kg)		
Waist Girth (cm)	**Gender**	**Fit**	**Lower**	**Upper**	**Fit**	**Lower**	**Upper**
60	male	50.5464	49.6299	51.4629	54.4994	52.4107	56.5882
60	female	50.5464	49.6299	51.4629	49.9445	48.8241	51.0648
80	male	72.4565	71.9397	72.9733	73.7754	72.9843	74.5666
80	female	72.4565	71.9397	72.9733	71.6835	70.5349	72.8322
100	male	94.3666	93.2108	95.5224	93.0515	91.6261	94.4768
100	female	94.3666	93.2108	95.5224	93.4226	90.6071	96.2381

The confidence intervals for the mean weight after taking into account Waist Girth is the same between genders for the Single Parameter model. This is expected since gender is not a parameter in the Single Parameter model. The confidence intervals for the Separate Lines model varies between gender after taking into account Waist Girth. These confidence intervals appear to be wider than the Single Parameter model. From a non-technical perspective, this means that by including Gender and its interaction with Waist Girth, there is more variability in expected results for a given waist girth and gender.

The range appears to be wider at Waist Girth values of 60 and 100 than 80 cm which indicates that there is more variability in the tails of the Separate line model . The Waist Girth tails (60 and 100cm) of the Single Parameter model are wider than the center value (80cm) but not as wide as the Separate Lines model.

Results

Linear Terms

There is moderate evidence in the Separate Lines model that Waist Girth reduces weight for males compared to females.

```
kable(
  tidy(model.initial),
```

```
  align = c('l'),
  digits = 4,
  col.names = c("Explanatory Variable","Fit (kg)","Standard Error (kg)", "T-Statistic", "p-value"),
  caption = "Single Parameter Model"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1:1, bold = T, border_right = T) %>%
  row_spec(0, bold = T)
```

Table 2: Single Parameter Model

Explanatory Variable	Fit (kg)	Standard Error (kg)	T-Statistic	p-value
(Intercept)	-15.1838	1.7929	-8.4688	0
wai.gi	1.0955	0.0231	47.5139	0

```
kable(
  tidy(model.separate_lines),
  align = c('l'),
  digits = 4,
  col.names = c("Explanatory Variable","Fit (kg)","Standard Error (kg)", "T-Statistic", "p-value"),
  caption = "Separate Lines Model"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1:1, bold = T, border_right = T) %>%
  row_spec(0, bold = T)
```

Table 3: Separate Lines Model

Explanatory Variable	Fit (kg)	Standard Error (kg)	T-Statistic	p-value
(Intercept)	-15.2727	3.2321	-4.7254	0.0000
wai.gi	1.0870	0.0460	23.6129	0.0000
gendermale	11.9441	4.7407	2.5195	0.0121
wai.gi:gendermale	-0.1232	0.0615	-2.0019	0.0458

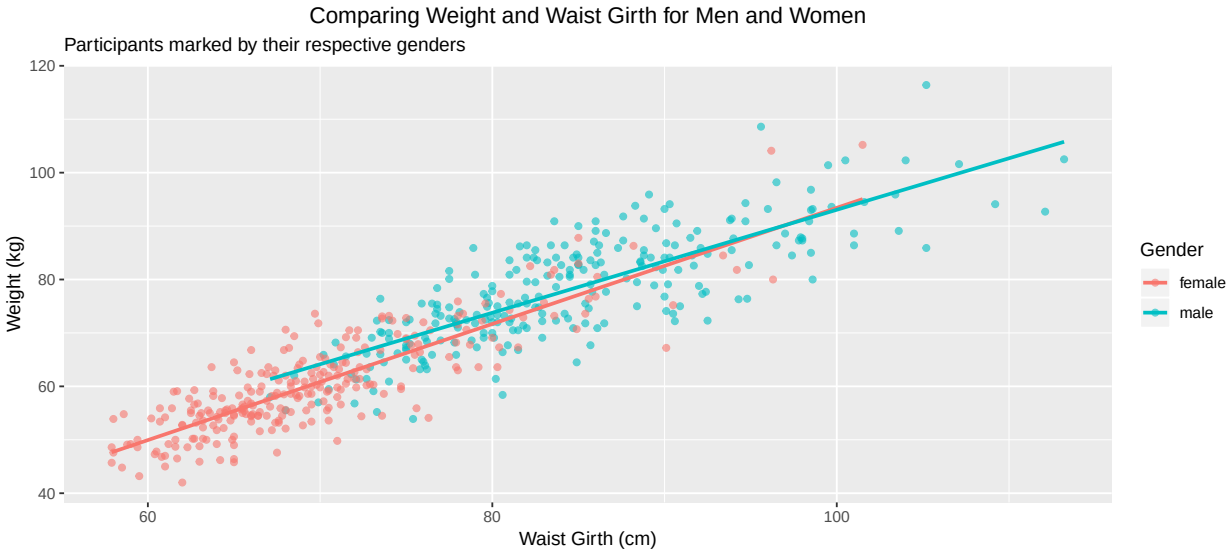
```
kable (
  anova(model.initial, model.separate_lines) %>% tidy,
  align = c('l'),
  digits = 4,
  #col.names = c("Explanatory Variable","Fit (kg)","Standard Error (kg)", "T-Statistic", "p-value"),
  caption = "Sum of Squares F-Test to compare Single Parameter and Separate Lines Model"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  row_spec(0, bold = T)
```

There is convincing evidence that the Single Parameter model does not have a better fit than the Separate Lines model (p-value = 0.0001).

Table 4: Sum of Squares F-Test to compare Single Parameter and Separate Lines Model

res.df	rss	df	sumsq	statistic	p.value
505	16474.61	NA	NA	NA	NA
503	15893.42	2	581.1905	9.1969	1e-04

```
model.separate_lines %>%
  ggplot(aes(x = wai.gi, y = wgt, color = gender)) +
    geom_point(alpha = 0.6) +
    geom_line(aes(wai.gi, .fitted), size = 1) +
      labs(color = "Gender", subtitle = "Participants marked by their respective genders") +
      ylab("Weight (kg)") +
      xlab("Waist Girth (cm)") +
      ggtitle("Comparing Weight and Waist Girth for Men and Women") +
        theme(plot.title = element_text(hjust = 0.5))
```



The Separate Lines model is the better for estimating weight with Waist Girth than the Single Parameter model initially proposed by the researchers. More analysis is needed as there appears to be some outliers and influencers that affect the slopes.

Quadratic Terms

The male weight appears to curve downwards for the high Waist Girth values. The Separate Lines model indicates that this is the case since the interaction term produced a negative value. It is worth seeing whether or not a quadratic term for Waist Girth provides a better fit.

```
model.separate_lines.poly <- lm(wgt ~ poly(wai.gi, 2) * gender, data = bdimsWithGender)

# Display Fit and Summary Information
kable(
  model.separate_lines.poly %>% tidy,
  align = c('l'),
  digits = 4,
  col.names = c("Explanatory Variable","Fit (kg)","Standard Error (kg)", "T-Statistic", "p-value"),
```

```
  caption = "Separate Lines with Quadratic Waist Girth"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1:1, bold = T, border_right = T) %>%
  row_spec(0, bold = T)
```

Table 5: Separate Lines with Quadratic Waist Girth

Explanatory Variable	Fit (kg)	Standard Error (kg)	T-Statistic	p-value
(Intercept)	68.4941	0.5106	134.1550	0.0000
poly(wai.gi, 2)1	274.5599	15.3865	17.8442	0.0000
poly(wai.gi, 2)2	6.5141	12.8452	0.5071	0.6123
gendermale	0.7365	0.8518	0.8646	0.3877
poly(wai.gi, 2)1:gendermale	5.7818	22.2824	0.2595	0.7954
poly(wai.gi, 2)2:gendermale	-44.4178	17.2457	-2.5756	0.0103

```
# Display SS F-Test Table
kable (
  anova(model.separate_lines, model.separate_lines.poly) %>% tidy,
  align = c('l'),
  digits = 4,
  caption = "Sum of Squares F-Test to compare Separate Lines and Separate Lines Quadratic Model"
) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  row_spec(0, bold = T)
```

Table 6: Sum of Squares F-Test to compare Separate Lines and Separate Lines Quadratic Model

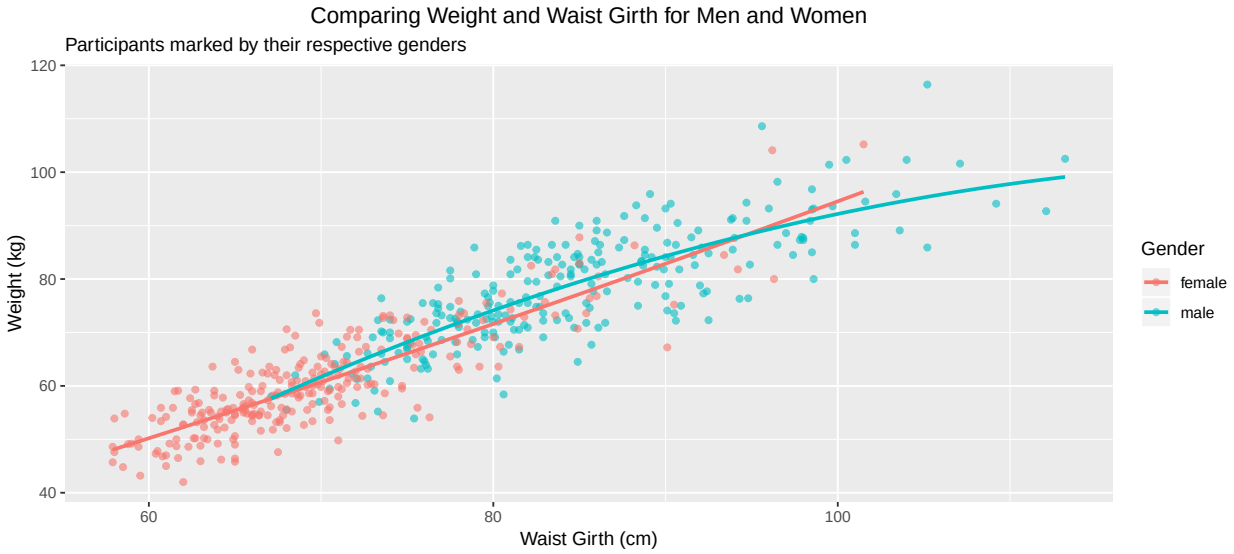
res.df	rss	df	sumsq	statistic	p.value
503	15893.42	NA	NA	NA	NA
501	15548.70	2	344.713	5.5536	0.0041

```
# For some reason, this renders fine in my console but not through here
# This function is preferred since it reuses the model but I couldn't figure it out.
# The plot below this works as expected.

# ggplot(model.separate_lines.poly, aes(x = wai.gi, y = wgt, color = gender)) +
#   geom_point(alpha = 0.6) +
#   geom_line(aes(y = .fitted)) +
#     labs(color = "Gender", subtitle = "Participants marked by their respective genders") +
#     ylab("Weight (kg)") +
#     xlab("Waist Girth (cm)") +
#     ggtitle("Comparing Weight and Waist Girth for Men and Women") +
#       theme(plot.title = element_text(hjust = 0.5))

# Graph the polynomial model against the scatterplot
ggplot(bdimsWithGender) +
  geom_point(aes(x = wai.gi, y = wgt, color = gender), alpha = 0.6) +
    # gender isnt included but comparing the graphs, it doesnt affect the line that is drawn so it shou
```

```
  stat_smooth(aes(x = wai.gi, y = wgt, color = gender), method = "lm", formula = y ~ poly(x, 2), se =
  labs(color = "Gender", subtitle = "Participants marked by their respective genders") +
  ylab("Weight (kg)") +
  xlab("Waist Girth (cm)") +
  ggtitle("Comparing Weight and Waist Girth for Men and Women") +
    theme(plot.title = element_text(hjust = 0.5))
```



There is convincing evidence that the Separate Lines Model is not as good of a fit as the Separate Lines Quadratic model (Sum of Squares F-Test. p-value = 0.0041). Having a quadratic term in this model makes sense since the right tail does not appear to be as linear as the rest of the terms. Statistically, the model is a better fit since it closer maps to the data. Thhe introduction of a quadratic term and its interactions show that the linear term for wai.gi is still significant but now the the interaction between wai.gi and MALE is no longer significant. Instead, the interaction between quadratic wai.gi and MALE is now significant. This indicates that the Waist Girth of males reduces weight in a quadratic fashion. The negative coefficient for the quadratic Waist Girth, MALE interaction mirrors that of the Separate Lines model meaning that the overall results are still the same but they can be predicted better using the quadratic model.