# HANDOUT: Stat 230 - Comparing 2 regression lines

P.B. Matheson adapted from A.S. Wagaman

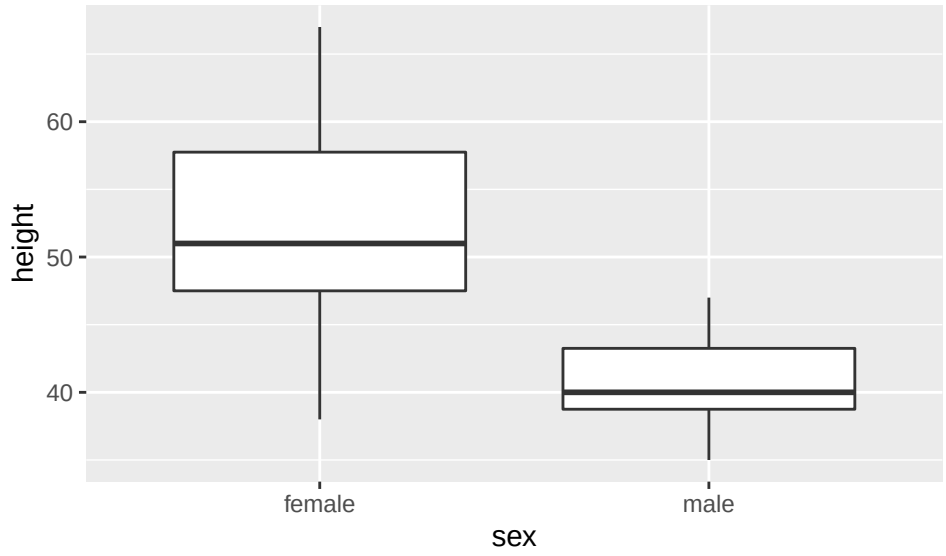
This example uses the data set on painted turtles, where we know their sex (defined in a binary way) and their height/length/width measurements.

```
turtle <- read.table("https://pmatheson.people.amherst.edu/stat230/paintedturtle.txt", header = TRUE)
```

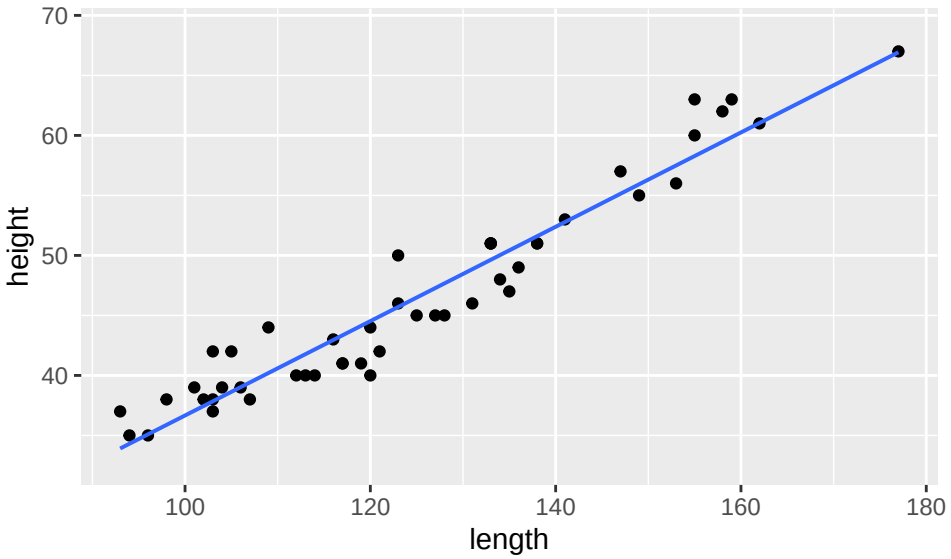
Suppose we want to try to predict the height of the turtles, using length and sex. Sex is currently coded in as female vs. male, but R will turn it into an indicator variable.

Let's look at how height (the outcome/response) relates to each possible predictor - length (quantitative var) and sex (categorical var).

```
gf_boxplot(height ~ sex, data = turtle) #for categorical predictor variable
```




```
gf_point(height ~ length, data = turtle) %>%
  gf_lm() #for quantitative predictor variable
```




```
cor(height ~ length, data = turtle)
```

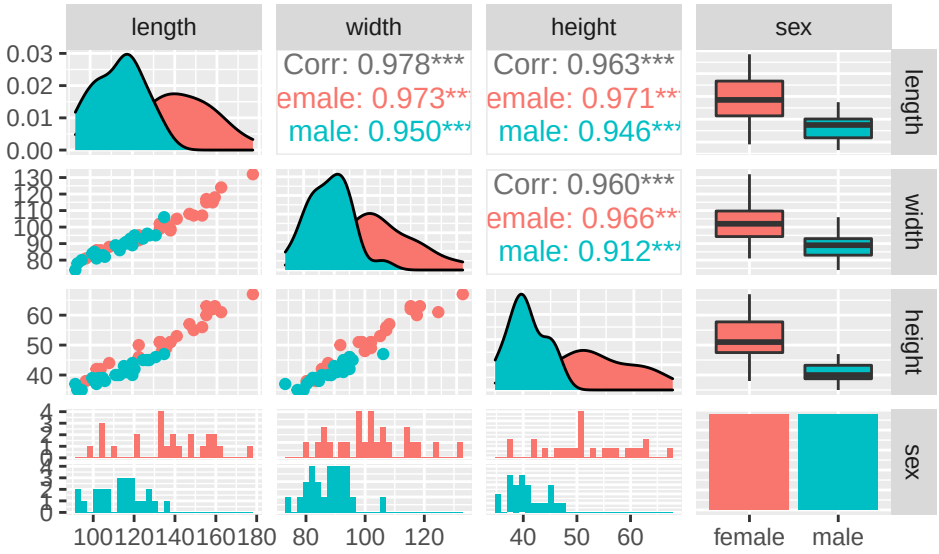
```
## [1] 0.9628899
```

In our previous exercise, we saw the linear relationship between height and length, but now we are adding the new categorical variable sex to the equation.

You can get the categorical variable included in the ggpairs plot too, carefully. Here, we don't really need this since we are just using one of the quantitative variables, but you might want something similar later!

```
#you actually don't want to use select here to drop the sex variable
ggpairs(turtle, columns=1:4, mapping=ggplot2::aes(color = sex))
```

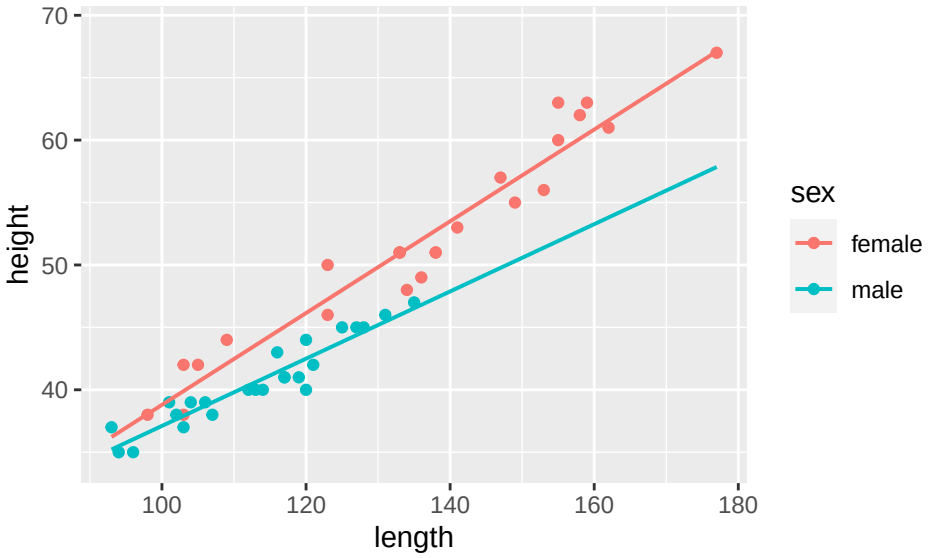
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#if you specified columns = 1:3, sex would only show in the color, no boxplots, etc.
```

There are two ways to fit a model that includes length and sex - a method that assumes parallel slopes for the two groups and one that doesn't (the interaction model). Look at the scatterplot with groups denoted by the sexes; it fits lines that don't need to have equal slopes. (We'll still fit each method so you can see the difference). Remember that by default, R fits the interaction model when you ask for lines by group, so this is showing what the model with interaction would do.

```
gf_point(height ~ length, data = turtle, color = ~ sex) %>%
  gf_lm()
```




**Parallel Slopes Model**

R is smart enough to treat the categorical variable sex as an indicator, even without us expressly framing it as one. It works in *alphabetical* order by default, setting female as the reference level getting a value of "0", and male as the "1". This means in our output, we won't see a line for sexfemale, but we will see values associated with sexmale, which is the newly created indicator variable R made to work with sex. Let's take a look at the fitted model.

```
fm <- lm(height ~ length + sex, data = turtle)
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.18082    2.06826   2.505   0.0159 *
## length       0.34446    0.01498  23.001  < 2e-16 ***
## sexmale     -3.52558    0.60743  -5.804  6.1e-07 ***
##
## Residual standard error: 1.745 on 45 degrees of freedom
## Multiple R-squared:  0.9583, Adjusted R-squared:  0.9565
## F-statistic: 517.6 on 2 and 45 DF,  p-value: < 2.2e-16
```

From the summary output, we can see that the fitted model is: predicted height = 5.18082 + 0.344(length) -3.526(sexmale)

Again, sexmale is what the sex variable has become when converted into an indicator variable. That's why we've written sexmale explicitly instead of sex in the equation so that is clear.

The 5.18082 is the intercept for females (as the reference level), and the -3.526 shifts that down for the males. This suggests that for turtles of the same length, males are on average 3.526 units shorter than females. Because sexmale is an indicator and it is not interacted with length, it only shifts the intercept.

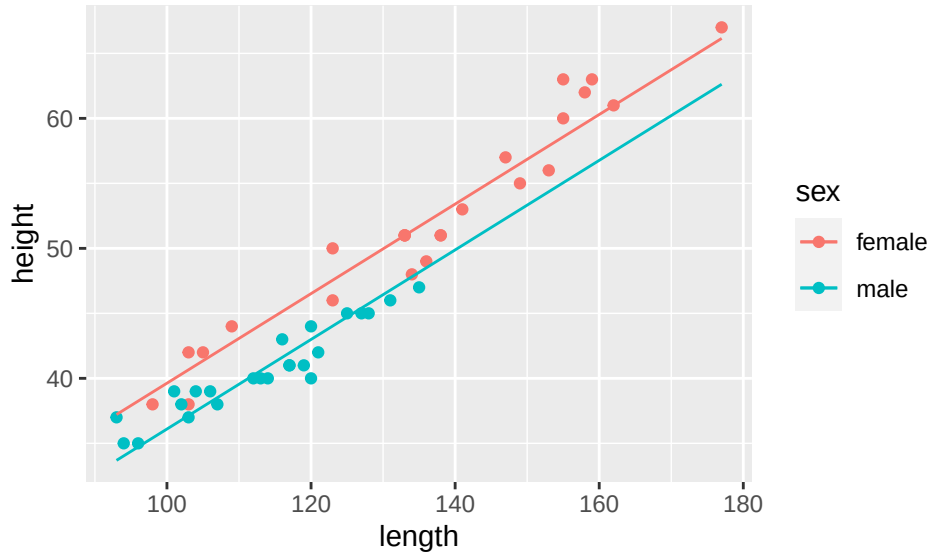
The slope of the regression line for length for both groups is 0.344. So, we'd say that for an increase of 1 unit in length, we expect average height to increase by 0.344 units for both male and female turtles.

The slope for the quantitative variable can't differ between the groups since the categorical variable was not entered as an interaction term with the quantitative predictor e.g., sex*length.

To plot these parallel lines, we use a function to do predictions:

```
myFun <- makeFun(fm)

gf_point(height ~ length, data = turtle, color = ~ sex) %>%
  gf_fun(myFun(length, sex = "female") ~ length, color = ~ "female") %>%
  gf_fun(myFun(length, sex = "male") ~ length, color = ~ "male")
```

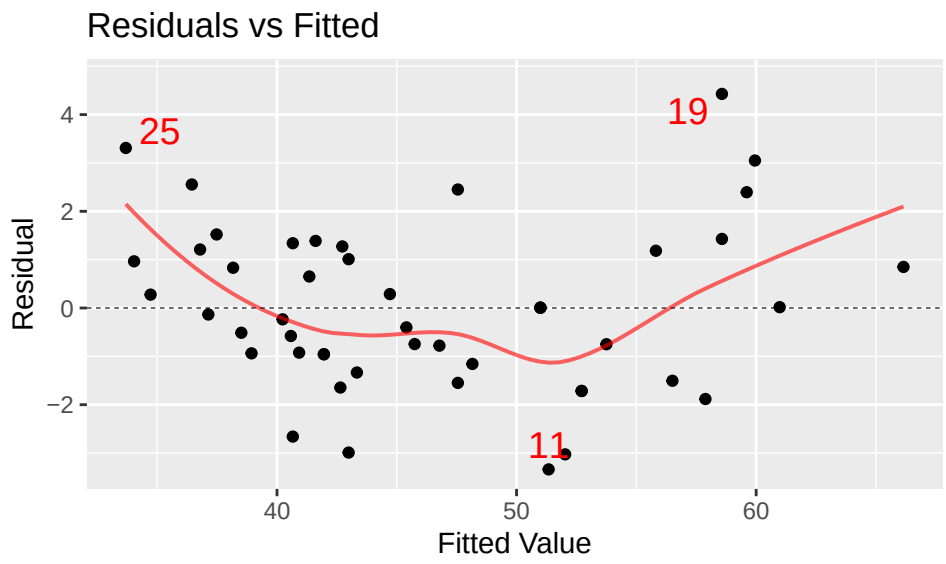



In this format, the function obtains the predictions for each group which are plotted along with the points (created as the first layer). The slopes are forced to be equal due to our model, and you can see that in the picture.

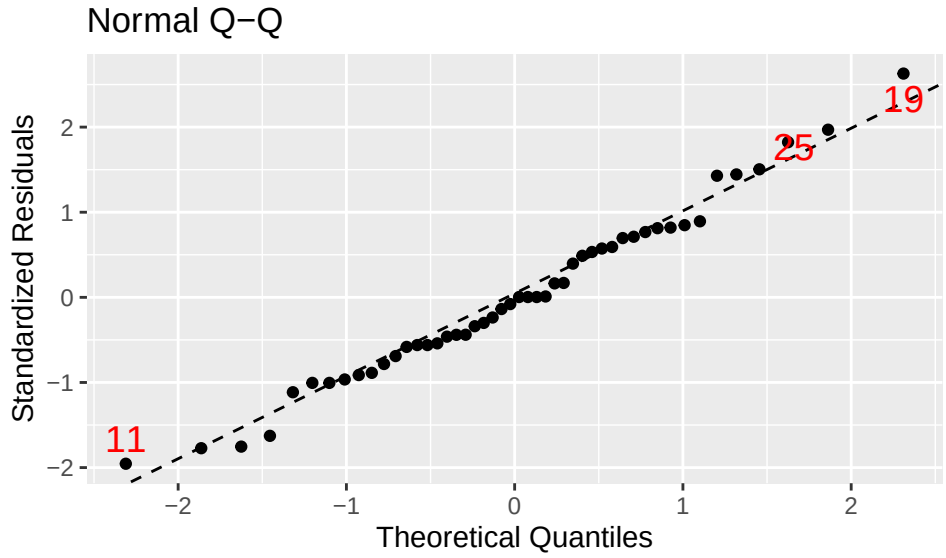
Assessing this fit, we can see the R^squared is 0.9583, and the ANOVA F-test indicates the overall model is significant, so at least one of the predictors is useful for predicting height. We can see both predictors are significant individually as well with their t-tests. But these inference procedures only work if the conditions are met, which we can check with:

```
mplot(fm, which = 1)#generates the residuals vs. fitted plot
```

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




```
mplot(fm, which = 2)#generates the QQ plot of residuals
```




The QQ-plot looks great but there may be some concern about the residuals vs. fitted plot. What is the issue? How might you go about fixing that? We may try that later.

We still have to assume that the turtles were independently selected and the data was collected through a random process (or that it is a representative sample) in order to do inference.

### Non-parallel Slopes Model

A slight tweak to the model adding the interaction term (we don't have to add any actual predictors to the data set) will mean that both the intercepts and slopes can vary for the groups conveyed by our categorical variable. There are two ways of fitting this model.

You can either do:

```
fm2 <- lm(height ~ length + sex + length:sex, data = turtle) #longhand
```

OR

```
fm2 <- lm(height ~ length * sex, data = turtle) #shorthand
msummary(fm2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.03999    2.16877   0.941  0.35204
## length          0.36755    0.01576  23.323  < 2e-16 ***
## sexmale         8.13218    3.89842   2.086  0.04281 *
## length:sexmale -0.09821    0.03250  -3.022  0.00418 **
##
## Residual standard error: 1.606 on 44 degrees of freedom
## Multiple R-squared:  0.9655, Adjusted R-squared:  0.9631
## F-statistic: 410.5 on 3 and 44 DF,  p-value: < 2.2e-16
```

The format variable1:variable2 adds their interaction term to the model. The format variable1*variable2 adds both main effects (individual variables) and their interaction to the model. The first format gives you more control if you have lots of interactions to add, but the other is faster for some investigations to start.

We'd report the overall model as:

$$\hat{height} = 2.04 + 0.37(length) + 8.13(sexmale) - 0.098(length : sexmale)$$

The last term could also be written (length*sexmale), indicating it is an interaction term - the use of the asterisk here is different from how R uses it in the shorthand for model specification. Assessing this model, we see that the R-squared is 0.9655, and the F-test still says the overall model is useful. We also see all three terms are significant with their individual t-tests for slopes, though the indicator (main effect for sexmale) itself is approaching borderline. Note that because the interaction is significant, we would keep the indicator (sexmale's main effect) in the model even if it was NOT significant by convention (If you want to keep a "higher" order term, you have to keep the associated lower order terms! Well, based on statistical convention.)

**IMPORTANT INTERPRETATION:

How would we come up with the regression lines for the two groups here? Well, female is the base or reference level, with sexmale = 0, so we can plug that into our equation and find that for females:
predicted height = 2.04 + 0.37(length)

For males, we have to combine the intercept and slope estimates because sexmale = 1, so we get:
predicted height = 10.17+0.27(length)
where 10.17 = 2.04+8.13, and 0.27=0.37-0.10

Again, writing out the full model, you'd still say:
predicted height = 2.04 + 0.37(length) +8.13(sexmale)-0.098(length*sexmale)

We should also practice interpreting these coefficients. When there is a categorical variable interacted with a quantitative one, some of the coefficients are telling us about one group relative to the other.

The intercept of 2.04 tells us that we expect a female turtle with 0 length to be on average 2.04 units tall (doesn't really make sense).

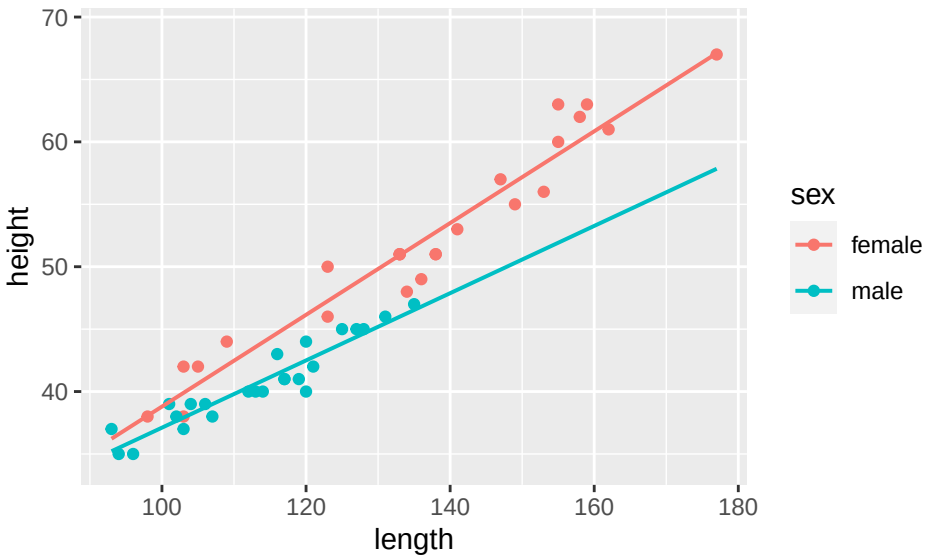
The length slope of 0.37 tells us that for female turtles, we expect average height to increase by 0.37 units per unit increase in length.

The sexmale slope of 8.13 tells us that once length is accounted for, having a male turtle instead of a female turtle adds 8.13 units to the average height. Note that this may seem counterintuitive, as male turtles are not taller than females on average... that's why talking about length already being accounted for is *very* important.

The interaction slope of -0.10 tells us that we would expect a decrease of 0.10 units in average height for each one unit increase in length for male turtles *relative* to female turtles.

Plotting these regression lines is what we did at the beginning of the example, so for review, that was:

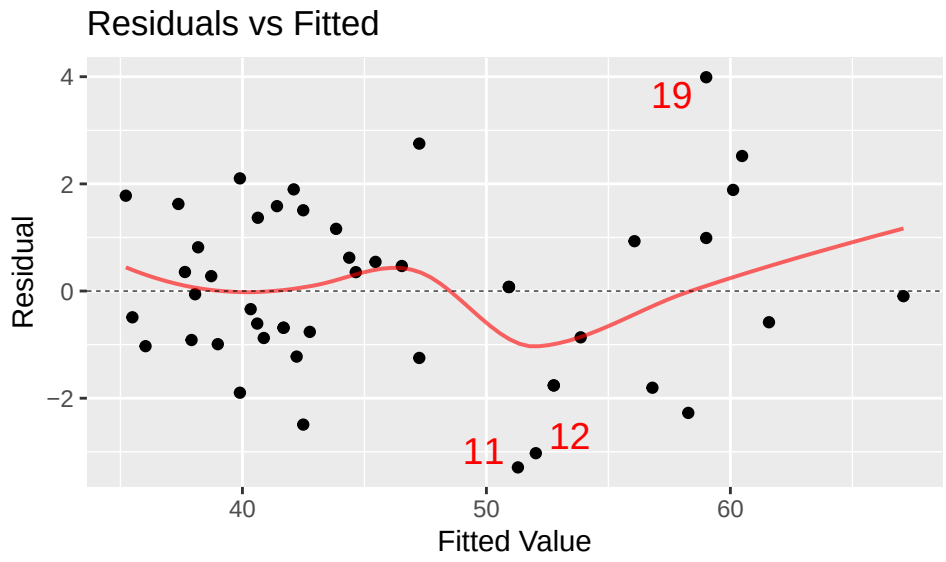
```
gf_point(height ~ length, data = turtle, color = ~ sex) %>%
  gf_lm()
```



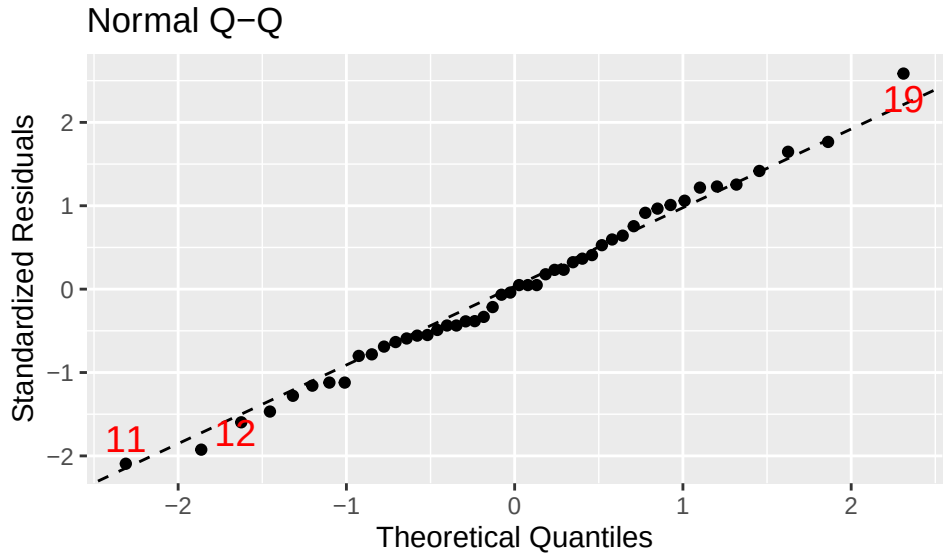

We should check conditions here to determine if using this model fixed the problem noted in the parallel slopes model. Do you think this model might have fixed it?

```
mplot(fm2, which = 1)#generates the residuals vs. fitted plot
```

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




```
mplot(fm2, which = 2)#generates the QQ plot of residuals
```




What do you conclude about the conditions?

Which model do you prefer for predicting height using length and sex? Parallel slopes or not?