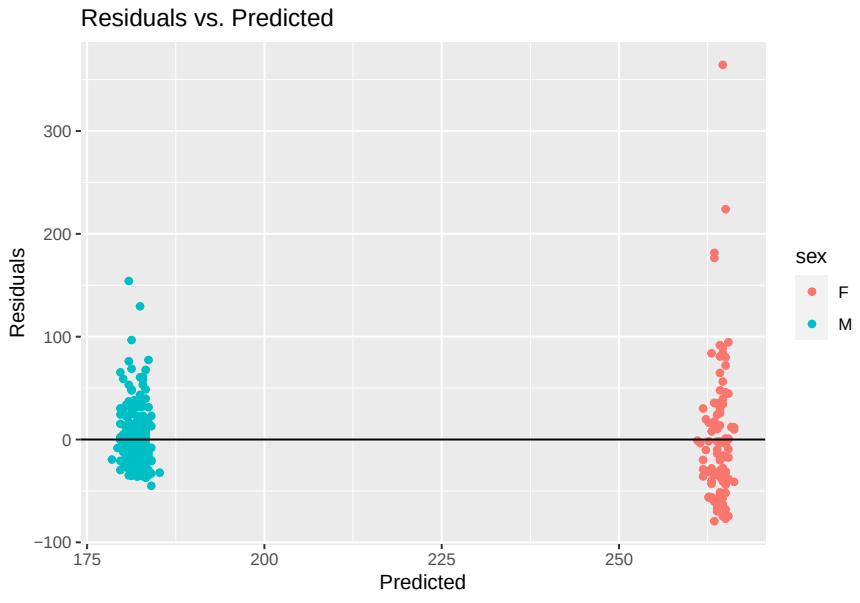
Lesson 12 Companion

Differing effects:

Can you remember all the way back to example I used in **Lesson 11**? Do you remember the issue in the diagnostic plot I said I would discuss in this lesson? Let's take a closer look at that situation. First let me recreate the model and diagnostic plot and add some color.

```
library(tidyverse)

IOCT = read.csv("IOCT_tab_data.csv")

model = lm(IOCT$IOCT_Time ~ IOCT$height + IOCT$sex)

#There is probably a more elegant way of doing this but
# this is what I've got right now.
IOCT_with_model = IOCT %>%
  select(IOCT_Time, height, sex) %>%
  cbind(pred = model$fitted.values) %>%
  cbind(resid = model$residuals)

IOCT_with_model %>%
  ggplot(aes(x = pred, y = resid, color = sex)) +
  geom_point() +
  labs(x = "Predicted", y = "Residuals",
       title = "Residuals vs. Predicted") +
  geom_hline(yintercept = 0)
```

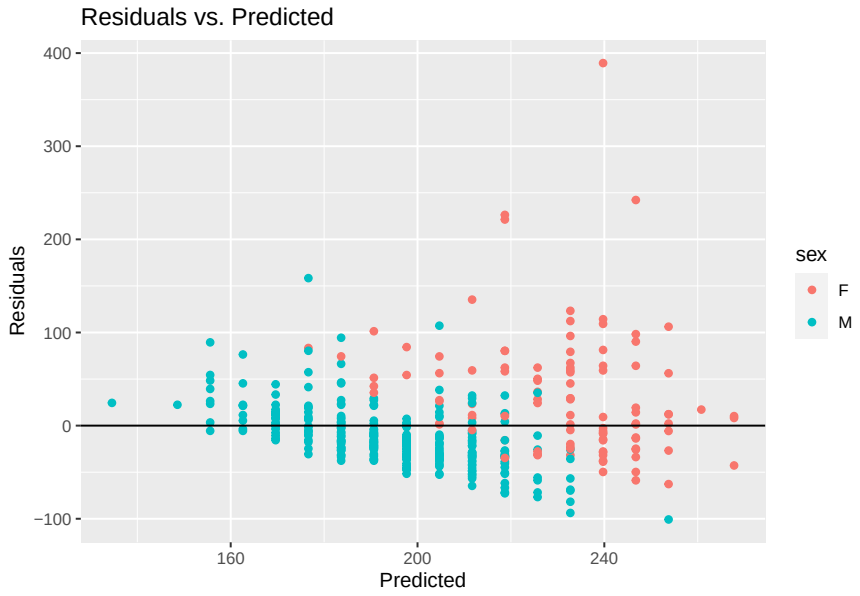


You should, hopefully, quickly identify that there appear to be two clumps of points on this diagnostic plots. Furthermore, like teenagers at a high-school dance, it appears as if the points that are clumped together

all belong to the same sex. This seperation itself doesn't really concern me but it does lead me to want to investigate some other things concerning my model.

For one, if you see seperation by category in your diagnostic plot and you haven't included that categorical variable as a term in the model, you really should. You generally don't see horizontal seperation in these cases, it is generally more of a vertical seperation as in the example below where I remove *sex* from this model.

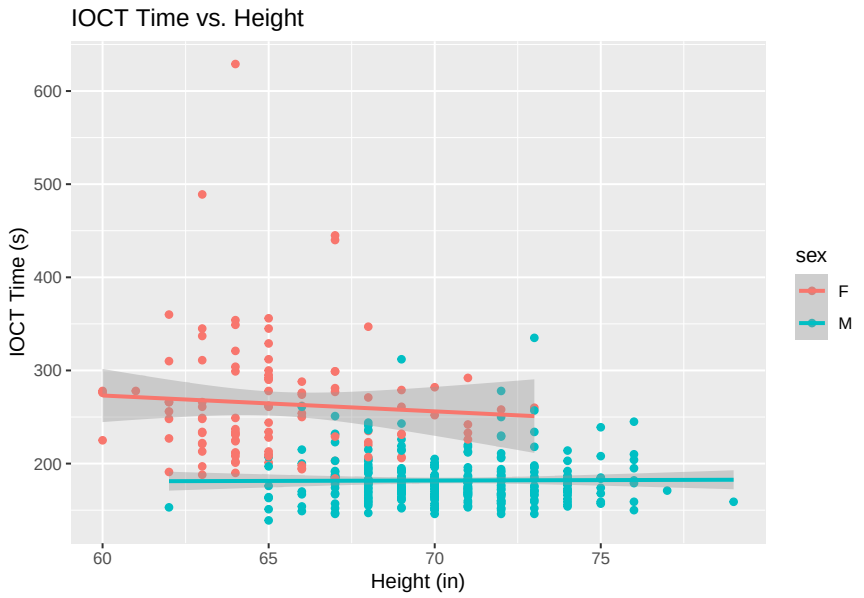
```
model2 = lm(IOCT$IOCT_Time ~ IOCT$height)

#There is probably a more elegant way of doing this but
# this is what I've got right now.
IOCT_with_model2 = IOCT %>%
  select(IOCT_Time, height, sex) %>%
  cbind(pred = model2$fitted.values) %>%
  cbind(resid = model2$residuals)

IOCT_with_model2 %>%
  ggplot(aes(x = pred, y = resid, color = sex)) +
  geom_point() +
  labs(x = "Predicted", y = "Residuals",
       title = "Residuals vs. Predicted") +
  geom_hline(yintercept = 0)
```



This tells me that we are generally overpredicting male times and underpredicting female times and therefore we should include *sex* as a term in the model.

The second thing that I think of when I see this sort of by-category seperation is an **interaction term**. Let me make another plot to better illustrate this:

```
IOCT %>%
  ggplot(aes(x = height, y = IOCT_Time, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Height (in)", y = "IOCT Time (s)",
       title = "IOCT Time vs. Height")
```



Hopefully, me sneaking in **geom_smooth()** doesn't blow your mind too much. It's available for quick visualizations but you have to be careful using it because it produces vastly different results depending on the ordering of you arguments in **ggplot**. Here it produced two seperate lines (one for each sex) because I first colored the points by *sex*.

My point in producing this plot is that you can see that the slopes are slightly different between the genders. This suggests that the association between *height* and *IOCT Time* is different for males and females. This tells me that there may be a significant interaction between these two variables. To say again: if one variable changes the association between an explanatory variable and the response variable, we say that there is an interaction between these two explanatory variables.

Adding an interaction term:

In order to properly account for interactions, we need to add an interaction term to our model. Pay close attention to the terms in the model below.

```
int_model = lm(IOCT$IOCT_Time ~ IOCT$height + IOCT$sex + IOCT$height*IOCT$sex)

summary(int_model)
```

```
##
## Call:
## lm(formula = IOCT$IOCT_Time ~ IOCT$height + IOCT$sex + IOCT$height *
##     IOCT$sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.00 -20.90  -4.42  11.26 362.70
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              375.267     98.099   3.825 0.000153 ***
## IOCT$height               -1.703      1.502  -1.134 0.257588
## IOCT$sexM               -199.660    117.451  -1.700 0.089959 .
## IOCT$height:IOCT$sexM      1.792      1.759   1.019 0.309028
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 380 degrees of freedom
## Multiple R-squared:  0.4481, Adjusted R-squared:  0.4438
## F-statistic: 102.9 on 3 and 380 DF,  p-value: < 2.2e-16
```

You see that we have three terms in the model: the two "main effects" of *height* and *sex* and the interaction term of *height*×*sex*. I'm not going to rehash everything about interpreting the model but I do want to make a few points.

Interpreting coefficients in the presence of an interaction term:

Let's start off by writing out the equation of our model with an interaction term:

$\hat{y} = -1.703x_1 - 199.660x_2 + 1.792x_1x_2 + 375.267$

where x_1 is height in inches and x_2 is sex (0 is female, 1 is male)

To interpret our coefficients with the interaction term, there are a couple extra steps. First let's consider intepreting the effect of height on IOCT time in this model.

Height

Without the interaction term we would say: if sex is held constant, for every one inch increase in height, we expect a 1.703 second decrease in IOCT time. The issue with this is that the first term is no longer the only x_1 term in the model. With the introduction of the interaction term, this value only applies when $x_2 = 0$ (sex is female) and therefore removes the other x_1 term. Coincidentally, this change in the slope depending on the value of *sex* is exactly what we saw in our two-line scatter plot above.

Sex

Without the interaction term we would say: if height is fixed, males have a 199.660 second faster IOCT time on average. Taking into account our interaction term, we see that this interpretation only applies when height is equal to zero... which doesn't make a lot of sense. This basically invalidates the B_2 term for our model. Instead, if we wish to get an idea of the association between sex and IOCT time with the interaction term, we would need to substitute some height values, calculate the model prediction, and compare those.

Interpreting the p-values

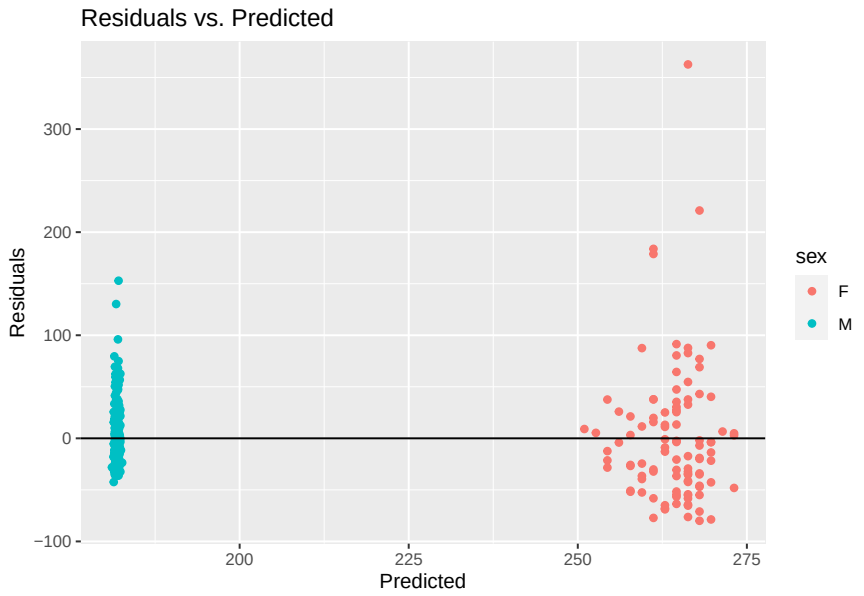
Don't hate me for pointing this out at this point but it is important to recognize that, even though we saw differing slopes on the scatter plot, the interaction term is not actually statistically significant at any reasonable level (p-value of 0.309). If I built this interaction-term model and saw that p-value, I would remove the interaction term and keep on rocking with the main effects-only model. This is the only time you'll hear me talk about removing a term from your causal model.

All for naught?:

Let me rebuild the original diagnostic plot using the interaction term model:

```
IOCT_with_int_model = IOCT %>%
  select(IOCT_Time, height, sex) %>%
  cbind(pred = int_model$fitted.values) %>%
  cbind(resid = int_model$residuals)

IOCT_with_int_model %>%
  ggplot(aes(x = pred, y = resid, color = sex)) +
  geom_point() +
  labs(x = "Predicted", y = "Residuals",
```

```
      title = "Residuals vs. Predicted") +
  geom_hline(yintercept = 0)
```



By golly the two groups are still there. Well I told you it probably wasn't going to fix the seperation between these groups. **Diagnostic** plots help your **diagnose** issues with your model. Just like a bad MRI: sometimes the tendon is torn and sometimes you just have a funny looking tendon.