Lab 11. Multiple Linear Regression

# Introduction

Wasps: what are their weaknesses? What are their strengths? What do they want and how do we keep it from them? It is time to strike with maximum force, so help Meaghan create a battle plan to hunt down and eradicate her nemesis. Nemeses? Nemsisises.

Whatever, build a multiple regression model to kill wasps, you get it.

# Learning Outcomes

By the end of today’s class you should be able to do the following in R:

* Use the lm() and summary() functions to perform a linear regression
* Use the melt() and facet\_wrap() functions to examine multiple data sets at once
* Use the geom\_smooth() function to create visualizations of trends
* Interpret multiple regression output
* Use predict() to predict new x output from given y variables
* Use math to defeat your enemies

# Part 1: Living Wasps

Earlier this season Meaghan had to change her office hours around to avoid the peak of wasp activity, which made her think there was a combination of temperature and time factors to their patterns. To start off with, we will be testing this theory using data collected sporadically and with great anger by Meaghan as she stalked through her house like a big game hunter armed with a bottle of soapy water, a flyswatter, and a clipboard.

## Part 1.1 Load Data and Packages

For this lab you will need the dplyr and ggplot2 packages, and the wasp dataset available on D2L.

library(dplyr)  
library(ggplot2)  
wasps <- read.csv("wasps.csv")

Our data isn’t complete however, as you need to do a little bit of column math. Specifically we need to make a total wasp column. Meaghan, in her scientific fervor, counted wasps in two separate rooms: the side room (where she films lectures) and the main room (where she eats, when not ducking to avoid being dive bombed). For our calculations we will need to add these together to get the total living wasp number (**Total.Wasps**).

wasps1 <- wasps %>%  
 mutate(Total.Wasps = Side.Room...Live+Main.Room...Live)

## Part 1.3 Wasp Time

While the temperature inside the house seemed to be useful, I also suspect that wasps have some sort of internal clock. That means that time of day was important.

Time of day is actually a pretty difficult number to parse in R. If you look at the **Time** column in this dataset, R does not initially understand it at all even though Excel had no problems with it. Use the class() function to see what R thought that column was.

class(wasps1$Time)

R thinks this is a character vector, much like a series of words. So you have to convince R that this is a number. But you also have to put the number into 24 hour time, because Meaghan recorded it in 12 hour time. So your steps are:

1. Make a number column called **TimeNumber**
2. Make 24 hour number column called **Hour24**

wasps2 <- wasps1 %>%  
 mutate(TimeNumber = as.numeric(gsub(":00 |PM|AM", "", Time))) %>%   
 mutate(Hour24 = ifelse(grepl("P", Time) & TimeNumber != 12, TimeNumber + 12, TimeNumber))

That code uses some functions and some twists that may be new to you. If you aren’t sure exactly what they do, a good way to find out is delete pieces and then re-run the code to see how it changes. Try the following changes to figure out what each piece does, then answer the questions below. **HINT**: it works a lot better if you do a change, run the code, then fix the code before doing another change…

* remove as.numeric() from the **TimeNumber** code
* change :00 |PM|AM to be :00 or PM|AM
* change P to be A or M
* remove & TimeNumber != 12,
* remove TimeNumber + 12

### QUESTION 1: What does gsub() do in the above code?

1. Makes the data numeric
2. Removes non-number data from the column
3. Makes a new column called TimeNumber

### QUESTION 2: What does grepl("P|p", Time) do in the code above?

1. It identifies columns that aren’t numeric
2. It identifies data that has a P or a p in it
3. It removes non-number data from the column

### QUESTION 3: Why can’t you just run as.numeric() on the Time column, but have to use gsub() as well?

1. Because R won’t interpret :00 or PM as a number
2. Because the data is too variable and can’t be identified
3. Because you have to change it to 24 hours

### QUESTION 4: Why does the ifelse() code have & TimeNumber != 12, as part of its criteria?

1. To make sure you don’t make noon become midnight
2. To make sure you don’t make noon become 1 AM
3. Because you’ve removed all noon data

# Part 2: Graphics

It is important to always visualize your data before you proceed, so you can catch weird results and errors as you go. Because we have raw data, scatterplots are the most effective visualization.

## Part 2.1 Individual Plots

In ggplot(), it is pretty simple to make a scatterplot. Use the geom\_point() function to do so.

ggplot(wasps2, aes( x = Temperature.Inside, y = Total.Wasps))+  
 geom\_point()

You can also add on a quick prediction line, using the geom\_smooth() function. If you want a prediction line that loosely fits the data (known as a **loess** curve), just leave the function as it is. If you want a linear model, specify method = "lm".

ggplot(wasps2, aes( x = Temperature.Inside, y = Total.Wasps))+  
 geom\_point()+  
 geom\_smooth()+  
 geom\_smooth(method = "lm", col = "purple", fill = "pink")

You can go through and do this for each variable (outside temperature, time, etc.) but it is kind of a pain. Instead, it makes sense to use the melt() function to set things up so that the facet\_wrap() function can run them all simultaneously.

## Part 2.2 Melting and Faceting

Load the reshape2 package; use install.packages("reshape2") if this doesn’t work.

library(reshape2)

Now, we’ll remove extra data columns by using the select() function from dplyr to focus on our columns. Then, we’ll make this into long format data using melt() and tell R that the **Date, Time,** and **Total.Wasps** columns are going to be kept still.

waspmelt<- wasps2 %>%  
 select(Date,Time, Hour24, Total.Wasps, Temperature.Inside, Temperature.Outside) %>%  
 melt(id.var = c("Date", "Time", "Total.Wasps"))

Finally, use the facet\_wrap() function to split out the data, so that the x axis is changing while the y axis is always the wasps.

ggplot(waspmelt, aes (x = value, y = Total.Wasps))+  
 geom\_point() +  
 facet\_wrap(~variable, scales = "free") +  
 geom\_smooth()

You might notice that the time of day does not seem to have a linear relationship with total wasp count. Wasps are most active around the middle of the day, rather than later in the evening or earlier in the morning. That means if we add this variable to a linear model, we won’t get quite the right answer - it’s not an *increase* or *decrease* in time that causes more wasps so much as it is the proximity to their most-active time of day.

## Part 2.3 Make a Better Time Column

To help adjust for the Wasping Hour problem, let’s make an **HourAdjusted** column. We’ll just standardize everything to 12 o’clock, by subtracting 12 hours from it. We don’t care about the direction, so we want to take the absolute value - 1:00 and 11:00 should have the same value.

wasps2 <- wasps2 %>%  
 mutate(HourAdjusted = abs(Hour24 - 12))

Then, to double check that this variable has the relationship we expected, let’s look to see if it is now linear:

ggplot(wasps2, aes(x = HourAdjusted, y = Total.Wasps))+  
 geom\_point()+  
 geom\_smooth()

I like to sometimes add gradient colors to see things a little more readily, and I want to double check that my new time variable is accurate. The following code adds a gradient, and a shape variable so I can make sure that mornings AND evenings are both further towards the right.

ggplot(wasps2, aes(x = HourAdjusted, y = Total.Wasps, col = Hour24))+  
 geom\_point(shape = ifelse(wasps2$Hour24 > 12, 17, 16), size = 3)+  
 geom\_smooth()+  
 scale\_color\_gradient(low = "red", high = "blue")

This tells me that both morning values and evening values are showing up in the same place, so our standardization seems to work fine.

# Part 3 Linear Model for Total Wasps

Now that you have looked at the data, it’s pretty easy to run a simple linear model.

## Part 3.1 Model Code

Use the + function to build out your x variables. With lm(), as long as you specify your data at the end you don’t need to use the $ sign for every column name.

model1 <- lm(Total.Wasps ~ Temperature.Outside + Temperature.Inside + HourAdjusted, data = wasps2)

Use the summary() function to evaluate your model and then go to the next section for help interpreting these results.

summary(model1)

## Part 3.2 Values Interpretation

R2 VALUE: According to these results, this model is explaining 41% of the variation in wasp number throughout the day (You can assume that “Meaghan’s super scary fake wasp nests” are explaining the rest). An R2 this high is a good sign that we’ve at least figured out some of the trend - in Biology and field sciences this is a decent R2 value.

P VALUES:

Both inside temperature and the time variable are significant (p < 0.05), meaning there is a significant relationship between them and total wasp number. This also means that our R2 value is valid, and our slopes are significant. But just because time is very significant doesn’t mean that it has the biggest impact - just the relationship least likely to be 0.

ESTIMATES: To find the actual impact, you have to look at the slope estimate. The bigger the number, the bigger the impact it has on your y variable. Think of the slope as being the amount of increase in wasp number according to one unit of that x variable.

## Part 3.3 Summary

The guiding question here was whether there was a relationship between living wasp number and the variables we measured - inside temperature, outside temperature, and time. If I was going to summarize this and explain it to someone else, here’s what I would write:

This model’s R2 value indicated these variables explained about 41% of the variance in wasp numbers. Inside temperature and time were the only statistically significant variables (p <0.05), but outside temperature was also close to statistical significance (p = .08).

Warm inside and outside temperatures increased wasp activity (positive slope estimates), and wasps were less active the further we got from noon (negative slope estimate). The strongest indicator of wasp number was time. For every hour away from noon, we had almost 2 fewer wasps (slope = - 1.978). For every one degree F increase in internal temperature, we had one half more wasp - or about one wasp every 2 degrees F increase (slope of .53).

## Part 3.4 Predict

Let’s say Meaghan is planning on giving an important lecture at her house on a day at about 1:00 in the afternoon. The average temperature at that time is about 45 degrees outside, and she’s hoping that by keeping her house a little cooler, she might be able to avoid having too many wasps. The easiest way to determine the number of wasps she mirth get at that point is to use the predict() function in R.

First, build a data frame that has all the x variables needed by the model:

temp1 <- data.frame(Temperature.Inside = 66, Temperature.Outside= 45, HourAdjusted = 1)

Then use the predict() function to see the number of wasps predicted.

predict(model1, newdata = temp1)

The predict function also will give you confidence intervals if you tell it to:

predict(model1, newdata = temp1, interval = "confidence")

So at 66 degrees inside at 1:00, she should predict about 6 to 18 wasps.

### QUESTION 5: Modify the above code to find out how many wasps she should expect if she has the internal temperature at about 76 (aka if she wants to feel her fingers)?

1. 12 - 18
2. 11 - 22
3. 9 - 25
4. 17 - 22

# Part 4: Linear Model for Dead Wasps

In addition to thinking there were more wasps on warm afternoons than there were in warm evenings, Meaghan also thought there was a link between how easy wasps were to kill when they were cold or sleepy.

Your job for the second part of this lab is to test that assumption and determine if there is a significant relationship between wasp mortality, temperature and time.

You can solve this question in a few different ways but you you need to make a few choices - what is your y variable, and what are your x variables? For a valid and accurate model, you have to make sure your choices

1. are likely to predict the data
2. contain the variables that are going to be important
3. do not have any duplication between y and x axes

Your y axis choices are (you have to make both using column math from this data): Total dead wasp number, percentage of dead wasps

Your x axis choices are: **Total.Wasps**, **HourAdjusted**, **Temperature.Inside** and **Temperature.Outside**. You may use some or all of these.

### QUESTION 6: Copy-paste the code you used to make your y axis.

### QUESTION 7: Copy-paste your model-building code (the lm function code).

### QUESTION 8: According to your results, is there a relationship between the temperature and time of day and the relative ease I had in killing wasps? Use your R2 value and slope p values to support your answer.

### QUESTION 9: According to your results, what was the strongest predictor for wasp mortality? Use your slope estimates to support your answer.

### QUESTION 10: One thing that was suggested to me for an “organic” method for killing wasps was to heat up my house and cool it rapidly every day, killing the wasps as they fell. According to your results, would this have worked? Why or why not?