

# More Models

Dominguez Center for Data Science Workshop

2025-04-16

## Load packages

```
# For data wrangling and plotting
library(tidyverse)

# For loading the meadowfoam dataset
library(Sleuth3)

# For fitting classification trees
library(rpart)

# For creating charts of trees
library(rpart.plot)
```

## Plan for today

- Cover lingering questions.
- Introduce a few more models:
  - Logistic regression
  - Classification trees
- Remember that you can either fill in the “more\_modeling.qmd” file or follow along with the “more\_modeling\_key.qmd” file.
  - Both documents have code related to topics we covered in previous weeks already filled in.

## Questions Round-Up

### But what does all of the output of `summary()` mean?

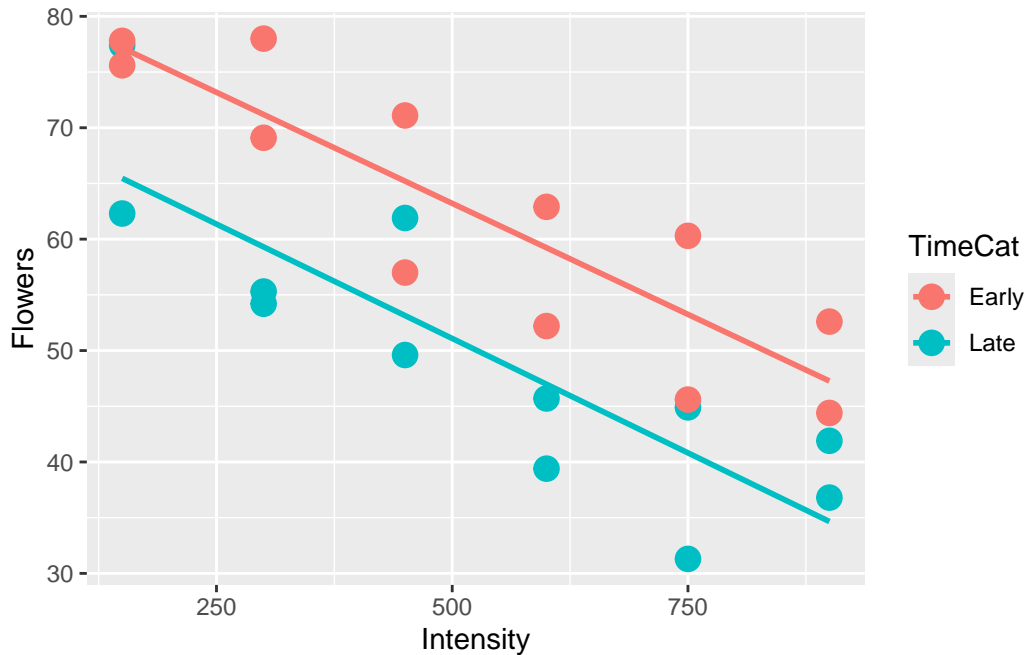
Let's return to the Meadowfoam flowers example from last time and build a model to predict the number of flowers on a plant based on the timing and intensity of the lighting.

In the code below, I fit two models: one with an interaction term and one without. Let's explore and interpret the `summary()` output.

```
# Load the data
data(case0901)

# Recode the timing variable
case0901 <- case0901 %>%
  mutate(TimeCat = case_match(Time,
                                1 ~ "Late",
                                2 ~ "Early"
  ))

# Visualize the data
ggplot(data = case0901,
       mapping = aes(x = Intensity,
                     y = Flowers,
                     color = TimeCat)) +
  geom_point(size = 4) +
  geom_smooth(method = lm, se = FALSE)
```



```
# Build a model without an interaction term
modFlowers <- lm(Flowers ~ Intensity + TimeCat, data = case0901)
summary(modFlowers)
```

Call:

```
lm(formula = Flowers ~ Intensity + TimeCat, data = case0901)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.652	-4.139	-1.558	5.632	12.165

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	83.464167	3.273772	25.495	< 2e-16 ***
Intensity	-0.040471	0.005132	-7.886	1.04e-07 ***
TimeCatLate	-12.158333	2.629557	-4.624	0.000146 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom

Multiple R-squared: 0.7992, Adjusted R-squared: 0.78

F-statistic: 41.78 on 2 and 21 DF, p-value: 4.786e-08

```
# Build a model with an interaction term
modFlowers_interact <- lm(Flowers ~ Intensity * TimeCat, data = case0901)
summary(modFlowers_interact)
```

Call:

```
lm(formula = Flowers ~ Intensity * TimeCat, data = case0901)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.516	-4.276	-1.422	5.473	11.938

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	83.146667	4.343305	19.144	2.49e-14	***
Intensity	-0.039867	0.007435	-5.362	3.01e-05	***
TimeCatLate	-11.523333	6.142360	-1.876	0.0753	.
Intensity:TimeCatLate	-0.001210	0.010515	-0.115	0.9096	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.598 on 20 degrees of freedom

Multiple R-squared: 0.7993, Adjusted R-squared: 0.7692

F-statistic: 26.55 on 3 and 20 DF, p-value: 3.549e-07

## And, how do I start using a local installation of RStudio?

- First download all of your projects from Posit as I will delete the workspace at the end of May.
- You can find directions [here](#) for downloading a local installation of RStudio.
- Bring any questions or issues with the transition to my R office hours on May 14th 1-3pm.

## Any other questions?

## Logistic Regression

Consider this model when:

- Response variable ( $y$ ): categorical (binary)

- Explanatory variable ( $x$ ): quantitative and/or categorical

## Come back to this data example:

“The social contract of Halloween is simple: Provide adequate treats to costumed masses, or be prepared for late-night tricks from those dissatisfied with your offer. To help you avoid that type of vengeance, and to help you make good decisions at the supermarket this weekend, we wanted to figure out what Halloween candy people most prefer. So we devised an experiment: Pit dozens of fun-sized candy varieties against one another, and let the wisdom of the crowd decide which one was best.” – Walt Hickey

“While we don’t know who exactly voted, we do know this: 8,371 different IP addresses voted on about 269,000 randomly generated matchups.<sup>2</sup> So, not a scientific survey or anything, but a good sample of what candy people like.”



```
# Load the data
candy <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-
  mutate(pricepercent = pricepercent*100)

# Look at the variables
glimpse(candy)
```

Rows: 85

Columns: 13

```
$ competitorname <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter~
$ chocolate      <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
$ fruity         <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1,~
```

```

$ caramel      <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ peanutyalmondy <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ nougat       <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ hard        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, ~
$ bar         <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ pluribus    <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, ~
$ sugarpercent <dbl> 0.732, 0.604, 0.011, 0.011, 0.906, 0.465, 0.604, 0.31~
$ pricepercent <dbl> 86.0, 51.1, 11.6, 51.1, 51.1, 76.7, 76.7, 51.1, 32.5, ~
$ winpercent  <dbl> 66.97173, 67.60294, 32.26109, 46.11650, 52.34146, 50.~

```

**Question of interest:** Does this candy have chocolate in it??

**Follow-up question:** Which variables in our dataset do you think would be good predictors of whether or not a candy will have chocolate in it?

## Visualize the relationships between the variables

### Build a logistic regression model

```

# Linear model
mod_lin <- lm(chocolate ~ as.factor(fruity) + winpercent, data = candy)
summary(mod_lin)

```

Call:

```
lm(formula = chocolate ~ as.factor(fruity) + winpercent, data = candy)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.63922	-0.14170	0.01374	0.17507	0.90544

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.010511	0.129843	-0.081	0.936
as.factor(fruity)1	-0.582356	0.065592	-8.879	1.26e-13 ***
winpercent	0.014034	0.002229	6.295	1.44e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.278 on 82 degrees of freedom  
Multiple R-squared: 0.6967, Adjusted R-squared: 0.6893  
F-statistic: 94.18 on 2 and 82 DF, p-value: < 2.2e-16

```
# Logistic model
```

## Prediction

You can use the `predict()` function in the same way as we did for linear regression.

```
# Start with a dataset of new observations  
new_cases <- data.frame(fruity = 0, winpercent = 50)
```

## Classification Trees

Consider this model when:

- Response variable ( $y$ ): categorical
- Explanatory variable ( $x$ ): quantitative and/or categorical

This is a more flexible model fit than logistic regression.

```
# Fit the model
```

```
# Visualize the tree
```

Which model is better: the logistic regression model or the classification tree?

Three metrics:

- Accuracy: Proportion of the guesses were correct
- Sensitivity: Among the chocolate candies, proportion of times that the model identified the candy has chocolate
- Specificity: Among the non-chocolate candies, proportion of times that the model identified the candy doesn't have chocolate

Let's draw a confusion matrix on the board.

```
# How do we access the predictions?

# Add predictions

# Confusion matrix

# Accuracy

# Sensitivity: Prob can identify that a candy has chocolate among the chocolate candies

# Specificity: Prob can identify that a candy doesn't have chocolate among candies without chocolate
```

## Resources related to modeling

We have just scratched the surface of building and interpreting models in R. Here are some additional resources on modeling:

- Chapters 5, 6, and 10 of [ModernDive](#)
- For a more machine learning approach, check out [Tidy Modeling with R](#) or [Introduction to Statistical Learning in R](#)
- For more advanced modeling, check out [Beyond Multiple Linear Regression](#)
- For bayesian model, check out [Bayes Rules!](#)

## Homework

- Download materials from Posit Cloud.
- Go back over all the materials we have covered and try to apply to your own work.
- Bring questions and ideas to my R office hours on May 14th 1-3pm.