ADS 503 - Applied Predictive Modeling (M6)

Summer 2024 - Week 6

Dave Hurst

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Start Recording!

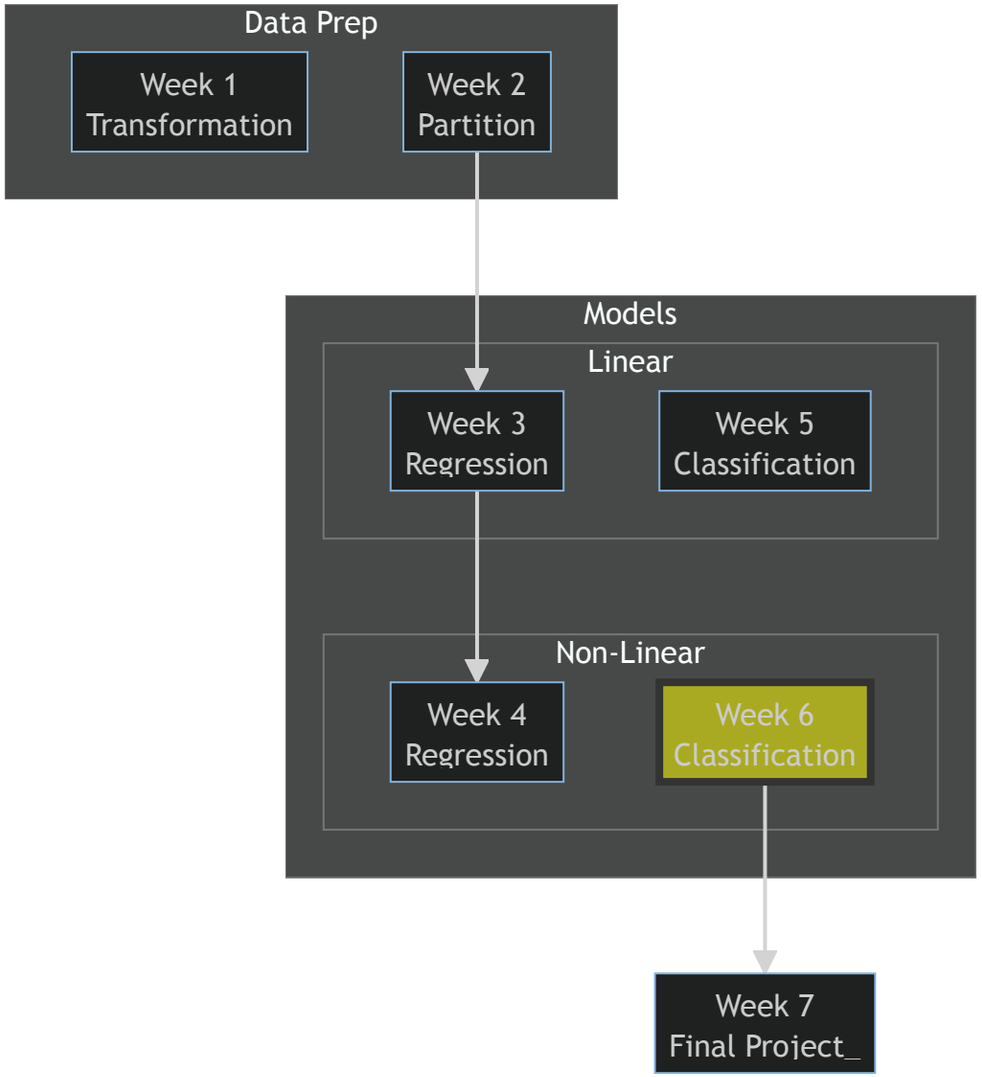
University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Agenda

- Posit.Cloud - reminder: cancel your plans
- Course Map
- Assignment 5 Review
- Assignment 6 Tips
- Shiny
- Final Project Discussion
- QA

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Course Map



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 5 Review

5.1.a Given the size of the dataset and the injury status distribution, describe if you would create a separate training and testing data set

281 observations ; 184/192 biological/chemical predictors.

A: Probably not

5.2.a Like the hepatic injury data, these data suffer from imbalance. Given this imbalance, should the data be split into training and test sets?

With only 96 observations it does not make practical sense to split the data into training and testing sets. It is more practical to use resampling methods to build and estimate the model performance.

B: Definitely not

Assignment 6 Tips

- Base R (optional `tidyverse`) code necessary for
  - Problem 1
  - Problem 2
  - Problem 3.b
- Show calculations when/if code is not used

```r
data61a <- tibble::tibble(
    X1 = 1:6,
    X2 = 2:7,
    X3 = seq(30, 80, by = 10),
)
class61a <- c("A", "B", "A", "B", "B", "A")
```

Given these observations:

X1	X2	X3	Class
1	2	30	A
2	3	40	B
3	4	50	A
4	5	60	B
5	6	70	B
6	7	80	A

Predict `Class` for these:

X1	X2	X3
9	4	15
7	8	45
8	7	75

Using only Base R and tidyverse:

```r
euclid_dist <- function(p1, p2) {
    # ... code
}
```

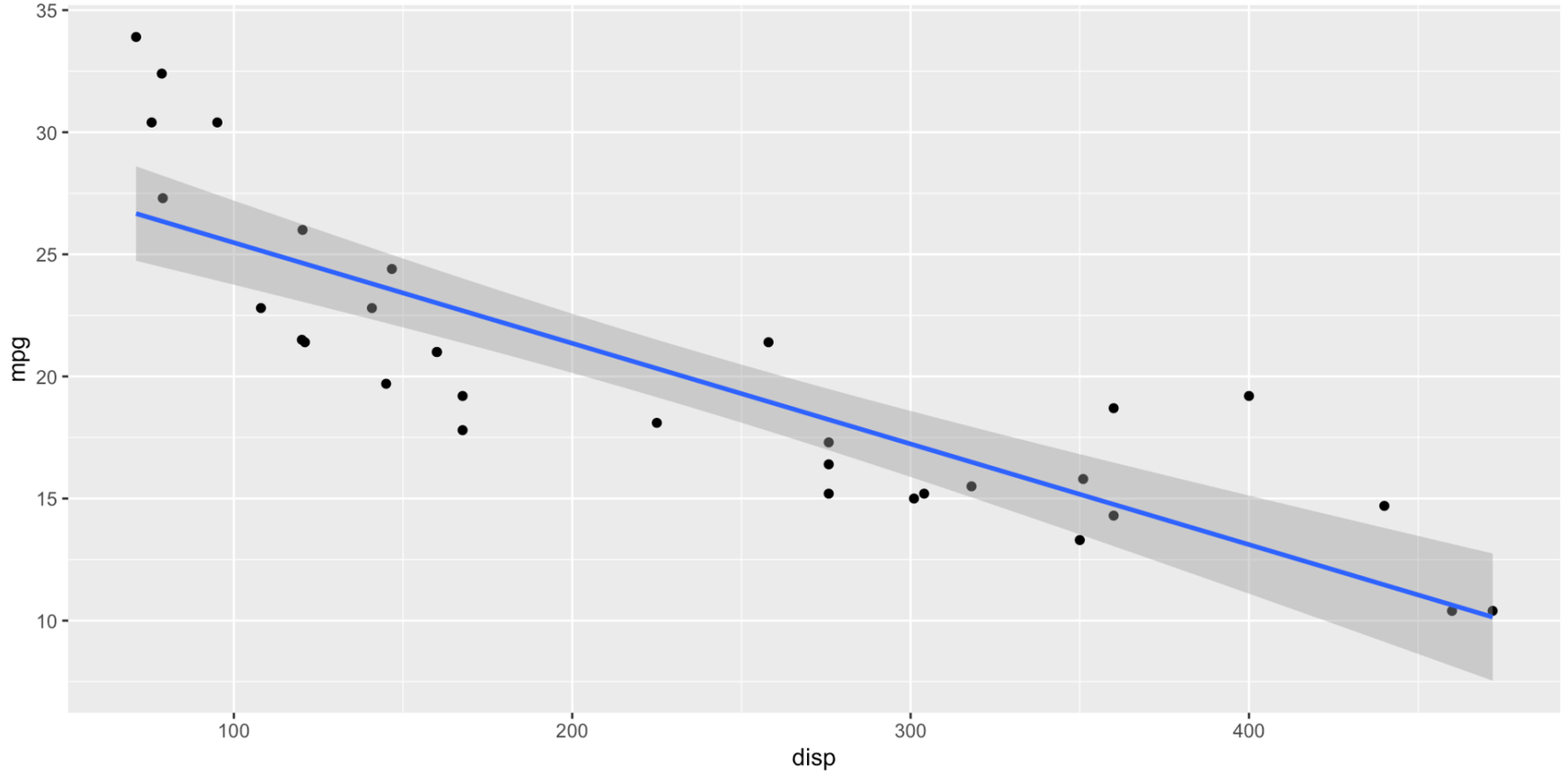
```r
#recast data as a list of vectors (rows)
data61b_list <- data61b |>
    split(seq(nrow(data61b))) |>
    map(\(row) as.numeric(unlist(row)))

target_point <- c(9, 14, 15)

# Calculate the Euclidean distances using purrr::map
distances <- map_dbl(data61b_list, ~ euclid_dist(.x, target_point))
min_index <- which.min(distances)
closest_class <- class61a[min_index]
closest_class
```

```
[1] "A"
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

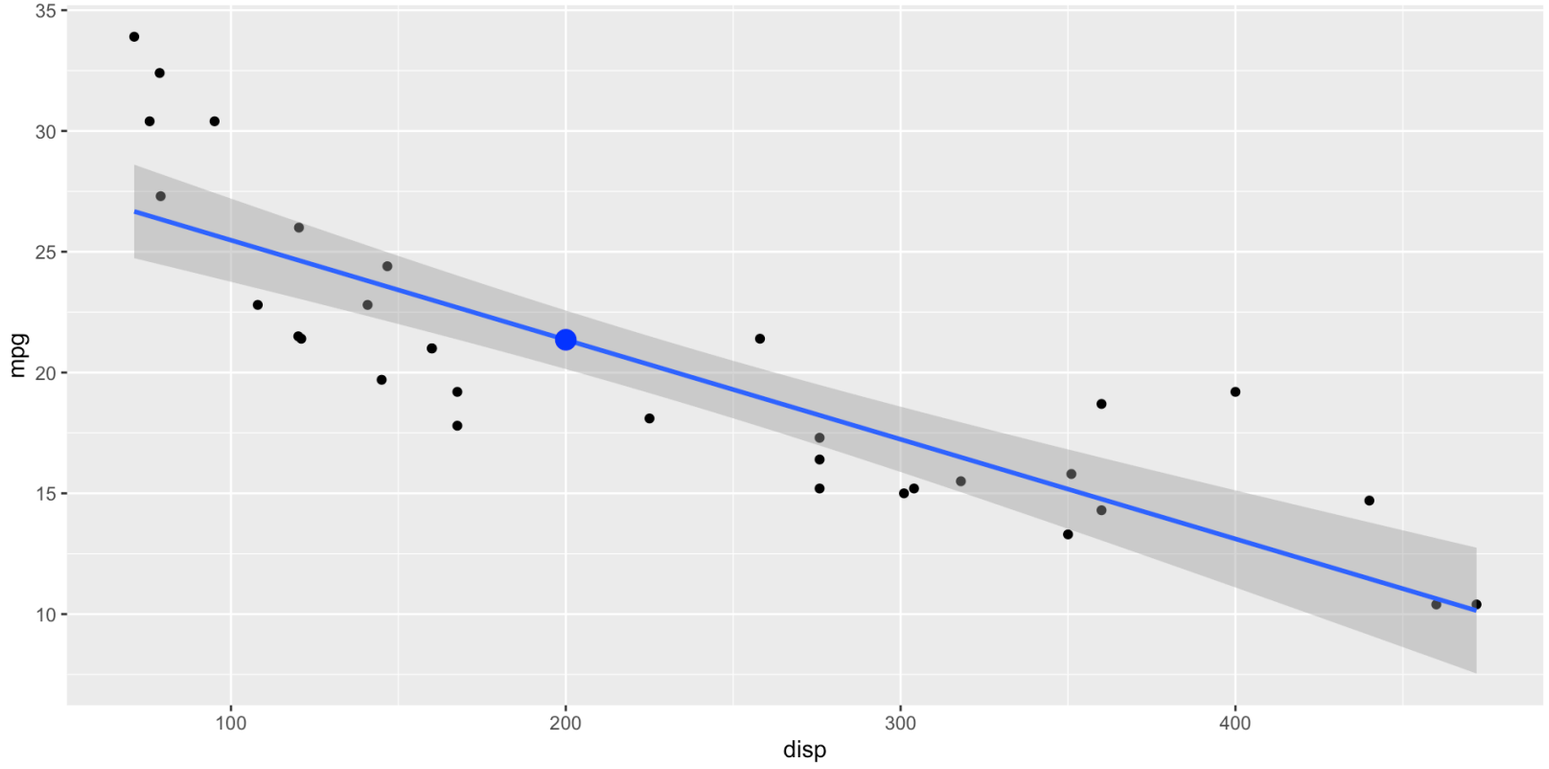
Shiny Example

```
1 data(mtcars)
2 mtcars |>
3     ggplot(aes(disp, mpg)) +
4     geom_point() +
5     geom_smooth(method = "lm")
```



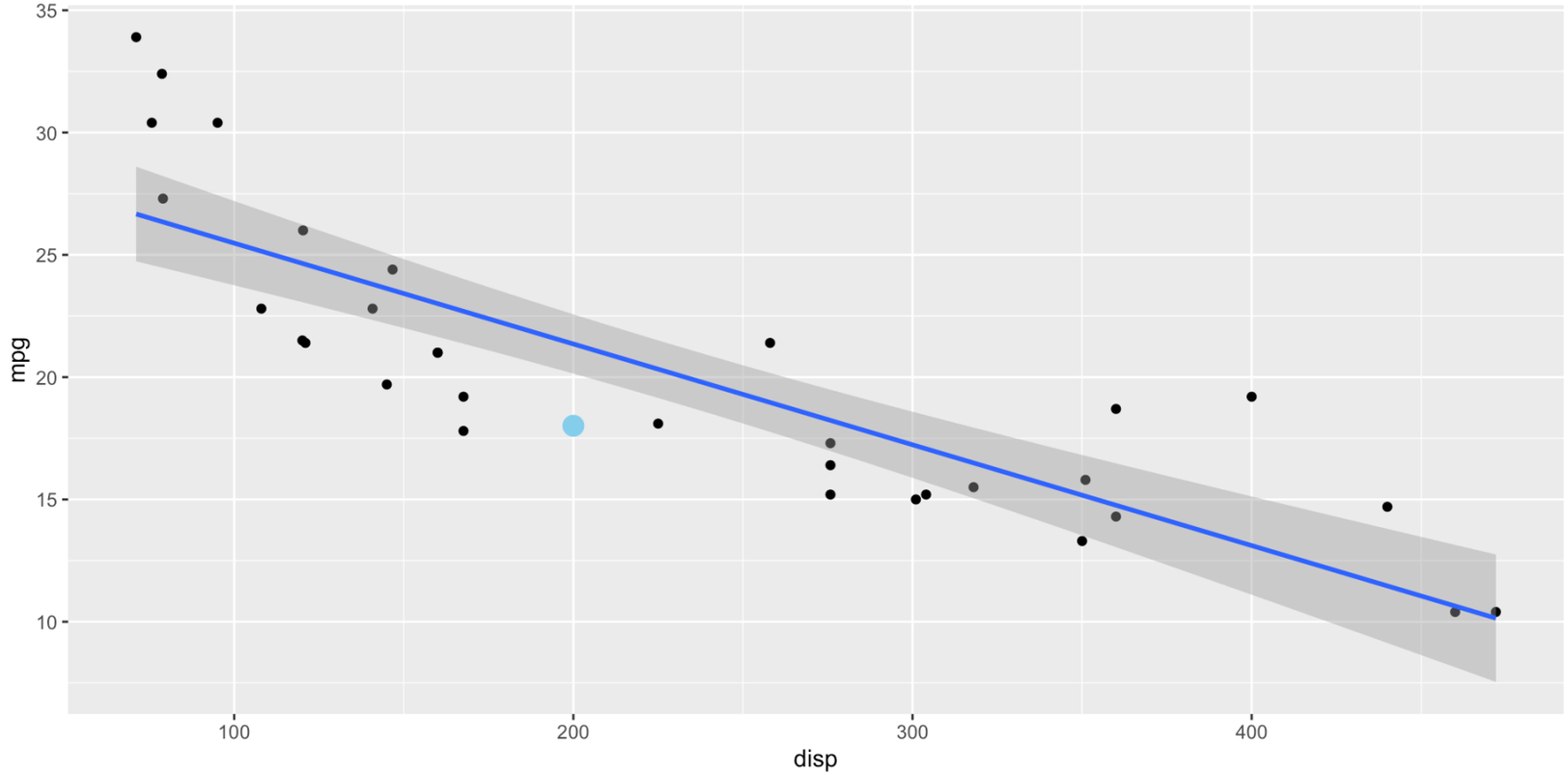
University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

```r
library(caret)
lm_disp <- train( mpg ~ disp, data = mtcars,
                  trControl = trainControl(method = "cv"),
                  method = "lm")
ex1 <- tibble(disp = 200)
yhat1 <- predict(lm_disp, newdata = ex1)
ex1$mpg <- yhat1

mtcars |>
    ggplot(aes(disp, mpg)) +
    geom_point() +
    geom_smooth(method = "lm") +
    geom_point(data = ex1, color = "blue", size = 4)
```



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

```r
lm_disp_wt <- train( mpg ~ disp + wt, data = mtcars,
                     trControl = trainControl(method = "cv"),
                     method = "lm")
ex2 <- tibble(disp = 200, wt = 4.0)
yhat2 <- predict(lm_disp_wt, newdata = ex2 )
ex2$mpg <- yhat2

mtcars |>
    ggplot(aes(disp, mpg)) +
    geom_point() +
    geom_smooth(method = "lm") +
    geom_point(data = ex2, color = "skyblue", size = 4)
```



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

APM w/ R

By David A Hurst

Applied Predictive Modeling with R

Upload a dataset to analyze

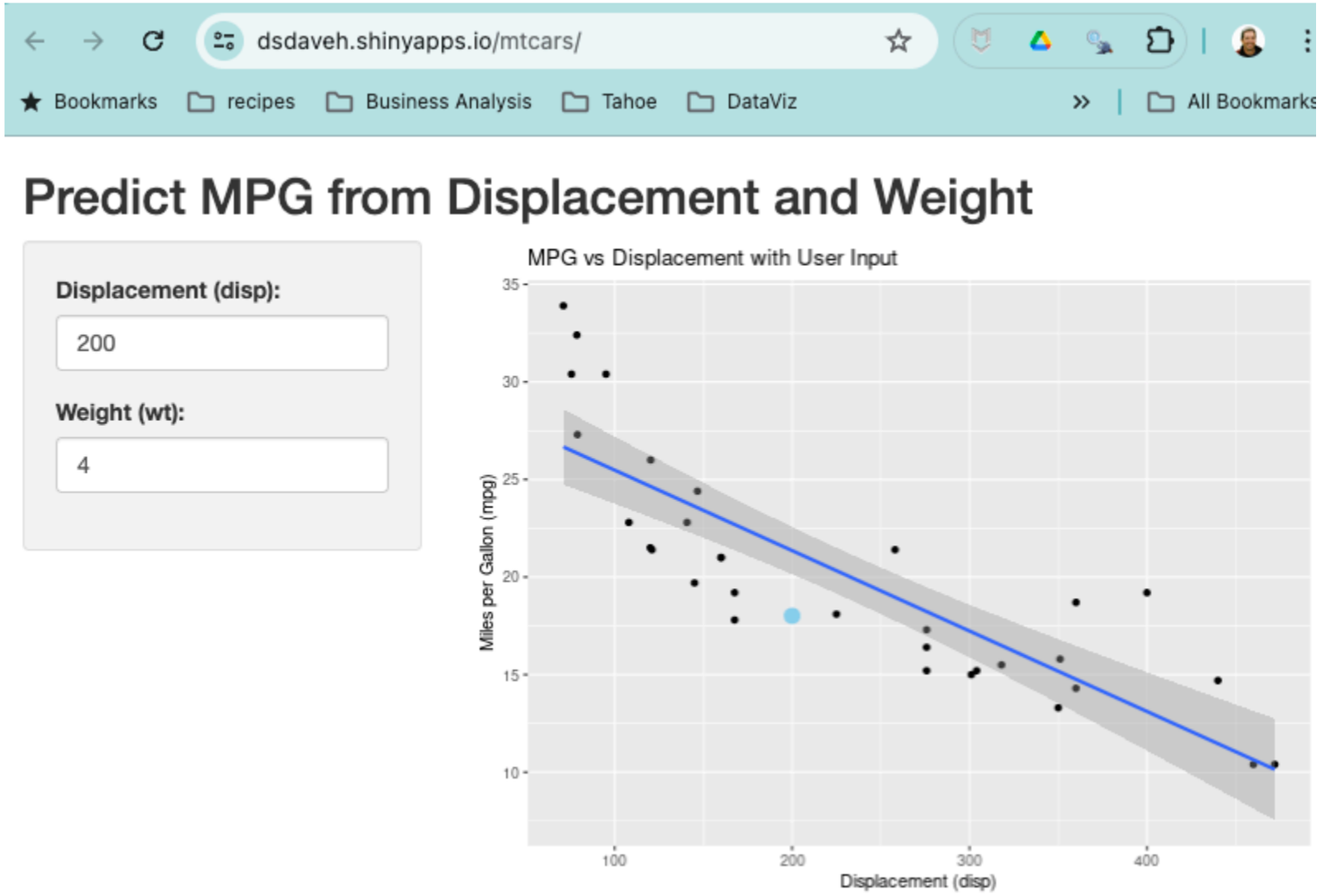
Ask me what a specific argument does

Give me a code starter template for a textbook exercise

Help me write a shiny app for the following model. I want the user to be able to input the displacement `disp` or weight `wt` of a vehicle and have it show up on the plot as I've shown in the example:

```
library(caret)
lm_disp_wt <- train( mpg ~ disp + wt, data = mtcars,
                     trControl = trainControl(method = "cv"),
                     method = "lm")
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING



https://dsdaveh.shinyapps.io/mtcars/

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Final Project

(see Canvas for full requirements)

- Video Presentation - 10-15 Mins
  - Model performance and hyperparameter tuning
  - Results and final model selection. *Use summary tables.*
  - AUDIENCE: your data science peers/technical audience
- Technical Report
  - Include a clearly defined problem statement.
  - You can include graphs and output tables only if you use them in your discussion. *This restriction includes code output.*
- Executive summary PowerPoint:
  - *should make us want to read your paper!*

Q&A

University of San Diego®
SHILEY-MARCOS SCHOOL OF ENGINEERING