Assignment 8

gzahn

December 30, 2018

Assignment 8

In this assignment, you will use R (within R-Studio) to load a data set and perform:

- point estimates (e.g., mean())
- interval estimates (e.g., confidence intervals)
- Hypothesis testing / Model fitting (with linear models and ANOVA)

All file paths should be relative, starting from the Assignment_8 directory!! (where you found this file)

This means that you need to create a new R-Project named "Assignment_8.Rproj" in your Assignment_8 directory, and work from scripts within that.

For credit...

1. Push a completed version of your Rproj and R-script (details at end of this assignment) to GitHub
2. Submit a plaintext file to Canvas answering the questions at the end of the assignment.

Your tasks:

- This has a lesson summary about modeling and testing
- At the end of the document are instructions for what you need to do
- Essentially, you will be loading a new data set, finding the best model, and then making predictions
- Your R code should be readible and well-documented

Models

A statistical model is a simple (hopefully) equation that *explains* trends in your data

"In what way does variable Y (The response) depend on other variables (X1 .... Xn) in the study?

The model attempts to approximate this relationship

Decisions, decisions...

- Which variable is your response?
- Which variables are explanatory?
- Are the explanatory variables continuous, categorical, or both?

```
1. All continuous: Regression
2. All categorial: ANOVA
3. Mix:            ANCoVA
```

Is the response variable continuous, a count, a proportion, a category????

- Continuous: Regression, ANOVA, ANCoVA
- Catergorical: ANOVA
- Proportion: Logistic regression
- Count: Log-Linear model
- Binary: Binary logistic

Here's a handy website: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

What models look like in R (the very basics)

Y ~ X This means "Y, is modeled as a function of X"

Y ~ X1 + X2 This means "Y, is modeled as a function of X1 AND X2" (two explanatory variables)

Y ~ X1:X2 This means "Y, is modeled as a function of THE INTERACTION BETWEEN X1 AND X2 (only the interaction term)

Y ~ X1*X2 This means "Y, is modeled as a function of X1 AND X2 AND THE INTERACTION BETWEEN X1 AND X2

R comes with a large variety of models you can choose from. Other packages provide hundreds more. Or you could write you own, if you are some sort of maniac!

The basic idea though, is to try to fit your data to a given model and then see how well it fits. If the fit is good, you can use that model to make meaningful (but not perfect) predictions.

Let's look at an example:

First, load some useful packages...

```
library(modelr)
library(broom)
library(dplyr)
library(fitdistrplus)
library(tidyr)
```

Next, load some data...

```
data("mtcars")
glimpse(mtcars)
```

```
## Observations: 32
## Variables: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8...
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8...
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 1...
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 18...
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92...
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3...
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 1...
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0...
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0...
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3...
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2...
```

Our dependent variable is mpg (it's the thing we want to know about)

Any or all of the other independent variables might help explain why it varies between different cars

Let's try a simple linear model with displacement and horsepower as explanatory variables...

```
mod1 = lm(mpg ~ disp, data = mtcars)
summary(mod1)
```

```
##
## Call:
## lm(formula = mpg ~ disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8922 -2.2022 -0.9631  1.6272  7.2305
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
## disp        -0.041215   0.004712  -8.747 9.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.251 on 30 degrees of freedom
## Multiple R-squared:  0.7183, Adjusted R-squared:  0.709
## F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

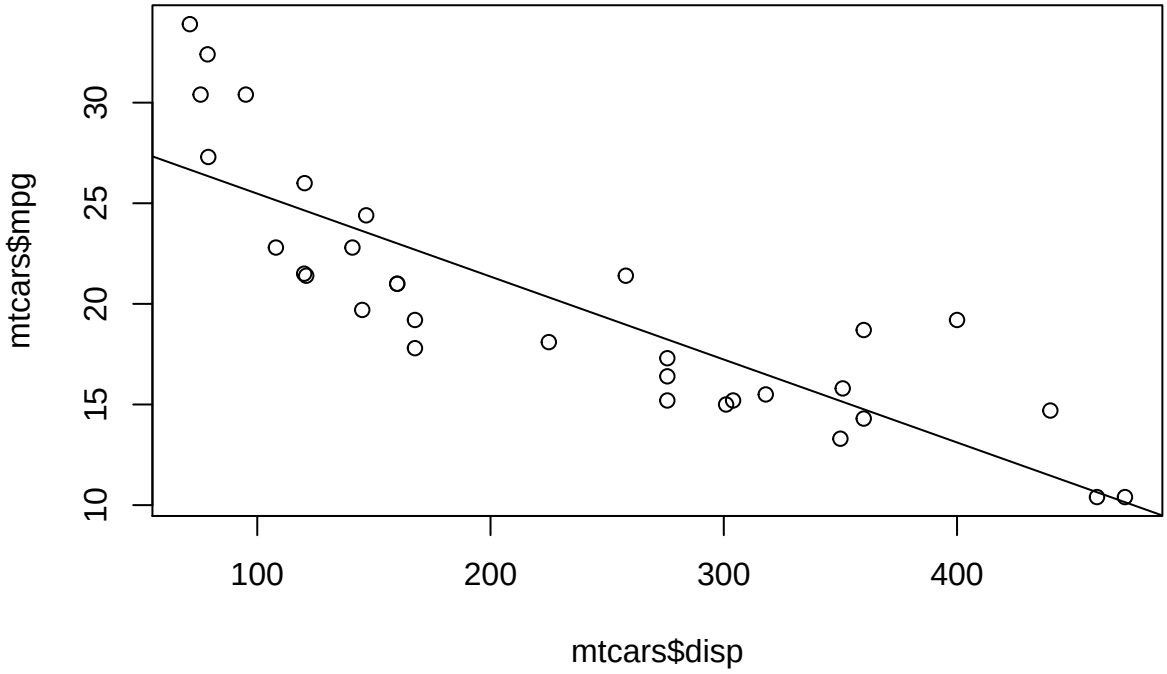
That's a lot of summary information about our model and it's kinda confusing.

- "Call" shows what model you ran
- "Residuals" are essentially measures of how well your actual data fit the model
- "Coefficients" can be thought of as the intercept and slope of the best-fit line

If you want a deeper explanation, see this website: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R

It may help to look at it visually...

```
plot(mtcars$mpg ~ mtcars$disp)
abline(mod1)
```



On this figure:

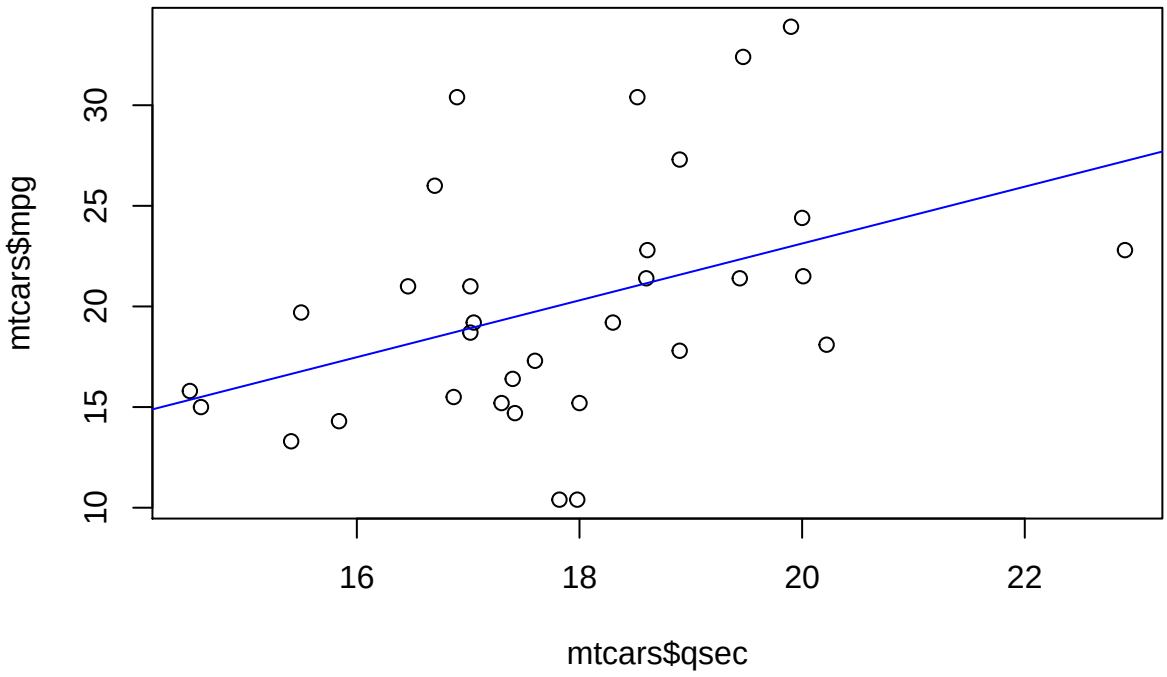
- The Residuals are the average distance of each point from the best-fit line
- The Coefficient for displacement is the slope

In other words, looking at our model summary tells us that this model predicts that for every increase in displacement of 1 cu.in. we can expect our fuel efficiency to drop by 0.04 mpg

Perhaps there's a better model in our data, however...

Let's look at one that incorporates speed instead

```
mod2 = lm(mpg ~ qsec, data = mtcars)
plot(mtcars$mpg ~ mtcars$qsec)
abline(mod2, col="Blue")
```



How to compare these two different models?

First thing is that you can visually look at the scatter around each line. By this measure, it seems clear our second model is a poorer fit than the first.

We can explicity measure that scatter. We call this the mean-squared-error of a model. The smaller, the better, so compare the two values from our models...

```
mean(mod1$residuals^2)
```

```
## [1] 9.911209
```

```
mean(mod2$residuals^2)
```

```
## [1] 29.02048
```

We'll abandon our second model for now because it sucks (relative to the first one) and look at how to make predictions using it.

The add_predictions() function from the modelr package lets us take our data frame and our model and look at what values our model assigns to our response variable (mpg). This is looking at ACTUAL vs PREDICTED values. If they are close enough for comfort we can move on and make predictions for unknown values in our model.

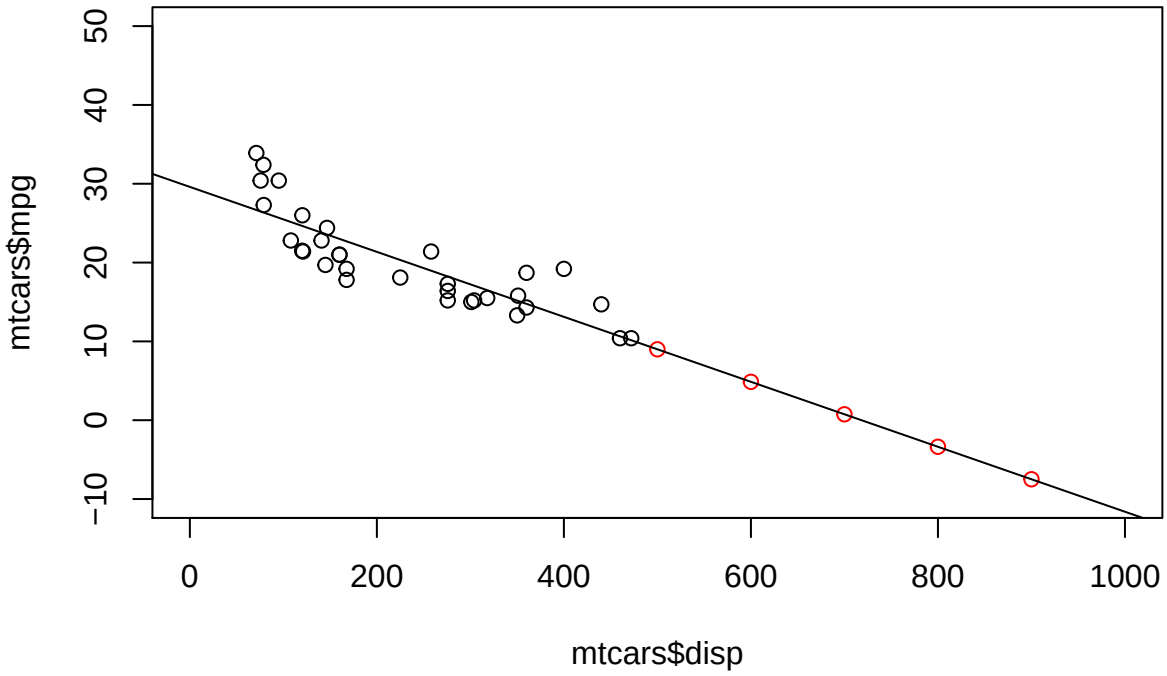
```
preds = add_predictions(mtcars, mod1)

preds[1:5,c(1,12)]
```

```
##                     mpg     pred
## Mazda RX4          21.0 23.00544
## Mazda RX4 Wag      21.0 23.00544
## Datsun 710         22.8 25.14862
## Hornet 4 Drive     21.4 18.96635
## Hornet Sportabout 18.7 14.76241
```

```
# Make a new dataframe with the predictor values we want to assess
# mod1 only has "disp" as a predictor so that's what we want to add here
newdf = data.frame(disp = c(500,600,700,800,900))
predictions = predict(mod1, newdata = newdf)

# plot those predictions on our original graph
{plot(mtcars$mpg ~ mtcars$disp,xlim=c(0,1000),ylim=c(-10,50))
points(x=newdf$disp,y=predictions, col="Red")
abline(mod1)}
```



Note a few things about our model's predictions (in red). They fall right on the prediction line, as expected... but some of them are negative! Can a car have negative mpg?

At last, your assignment!

- Make a new Rproj and Rscript in your personal Assignment_7 directory and work from there.
- Write a script that:
  1. loads the "/Data/mushroom_growth.csv" data set
  2. creates several plots exploring relationships between the response and predictors
  3. defines at least 2 models that explain the **dependent variable "GrowthRate"**

    - One must be a lm() and
    - one must be an aov()
  4. calculates the mean sq. error of each model
  5. selects the best model you tried
  6. adds predictions based on new values for the independent variables used in your model
  7. plots these predictions alongside the real data
- Upload responses to the following as a numbered plaintext document to Canvas:
  1. Are any of your predicted response values from your best model scientifically meaningless? Explain.
  2. In your plots, did you find any non-linear relationships? If so, do a bit of research online and give a link to at least one resource explaining how to deal with this in R