Linear models 3

Models behaving badly

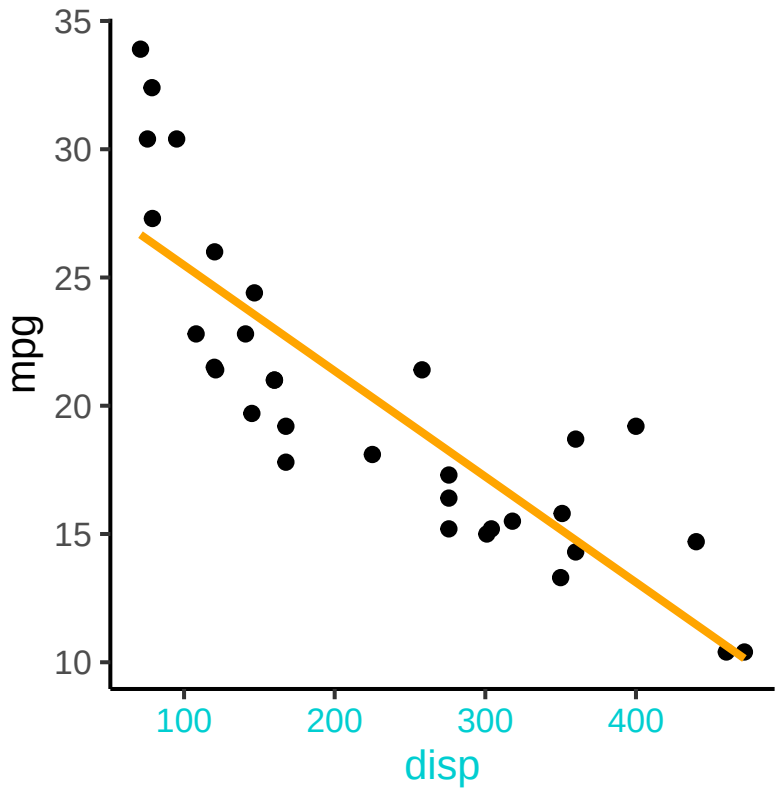
Samuel Robinson, Ph.D.

October 15, 2020

Motivation

- Are my model results reliable?
  - Residual checks
  - Transformations
  - Scaling
  - Collinearity
- How do I tell if terms are important or not?
  - Drop-1 ANOVA
  - Wald t-tests
- How much stuff should I put into my model?
  - Causal modeling vs Machine learning
  - Avoiding a fishing expedition (model weights, stepwise selection)

Are my model results reliable?



There are 3 main assumptions to this model:

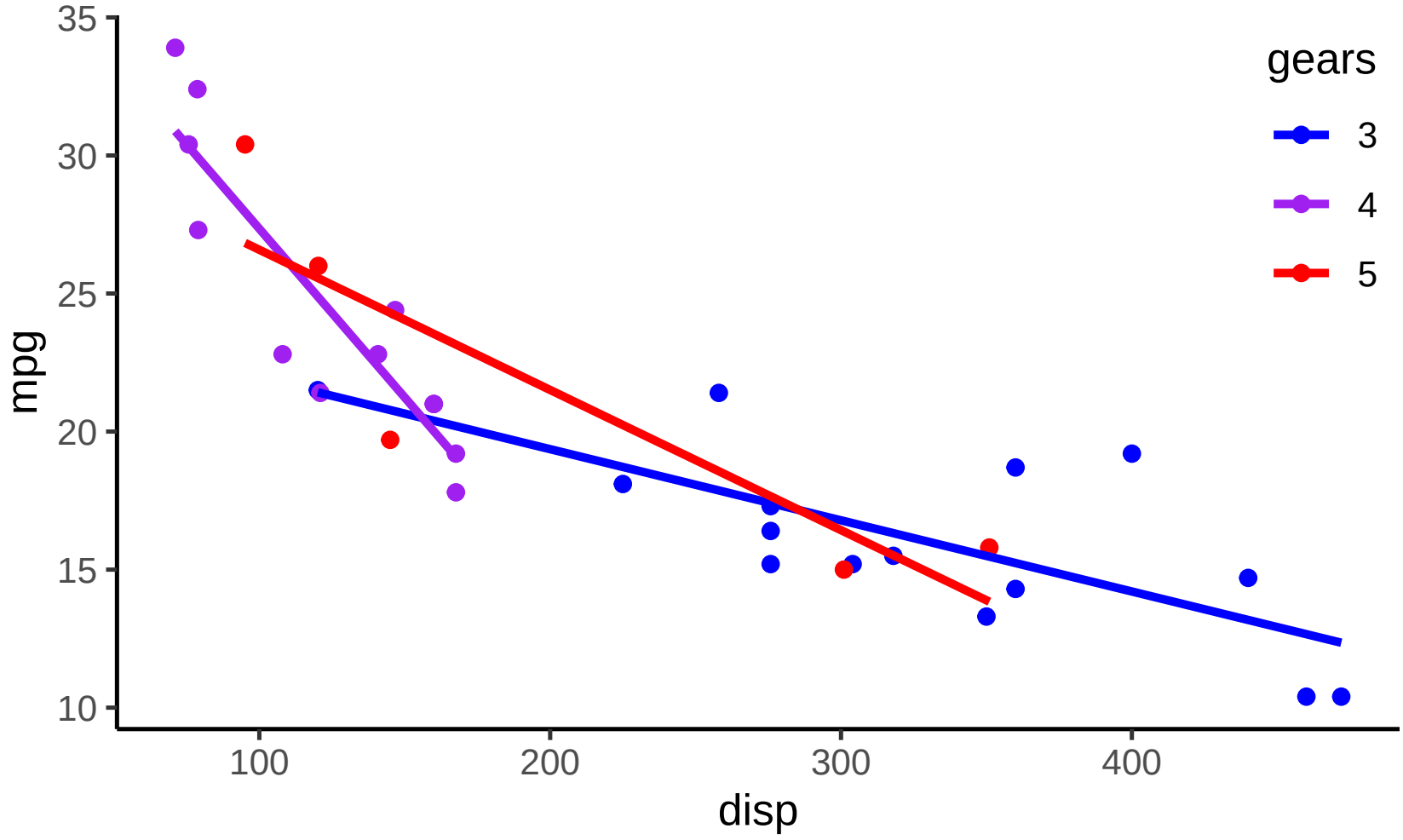
1. The relationship between *disp* and *mpg* is linear
2. *mpg* (the data) is Normally distributed around $\hat{mpg}$ (the line)
3. σ is the same everywhere

This is pretty easy to see if you only have 1 variable, but...

$$\hat{mpg} = b_0 + b_1 disp$$

$$mpg \sim Normal(\hat{mpg}, \sigma)$$

What if I have many variables?

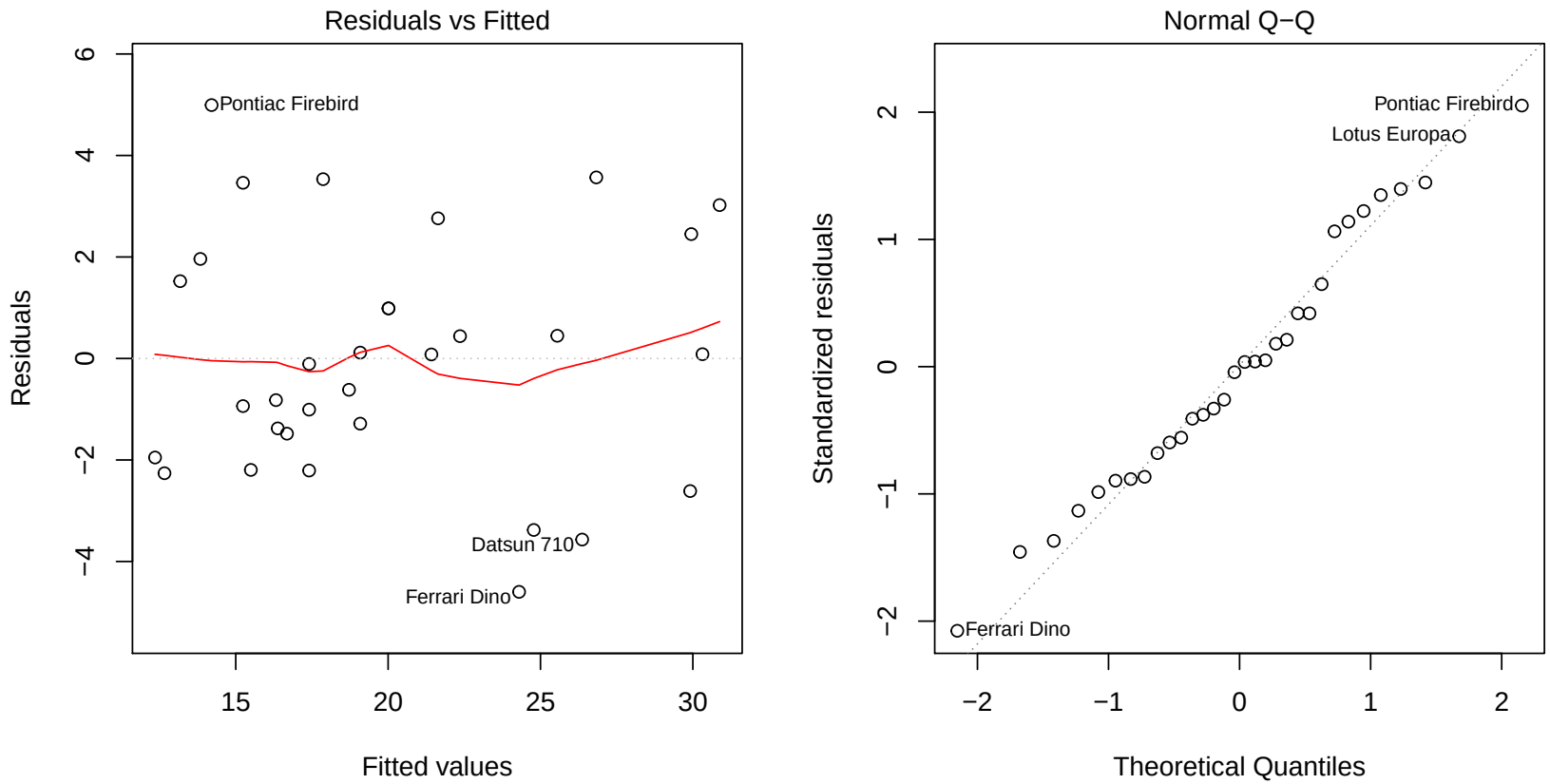


Difficult to see if the assumptions are met

Solution: residual checks

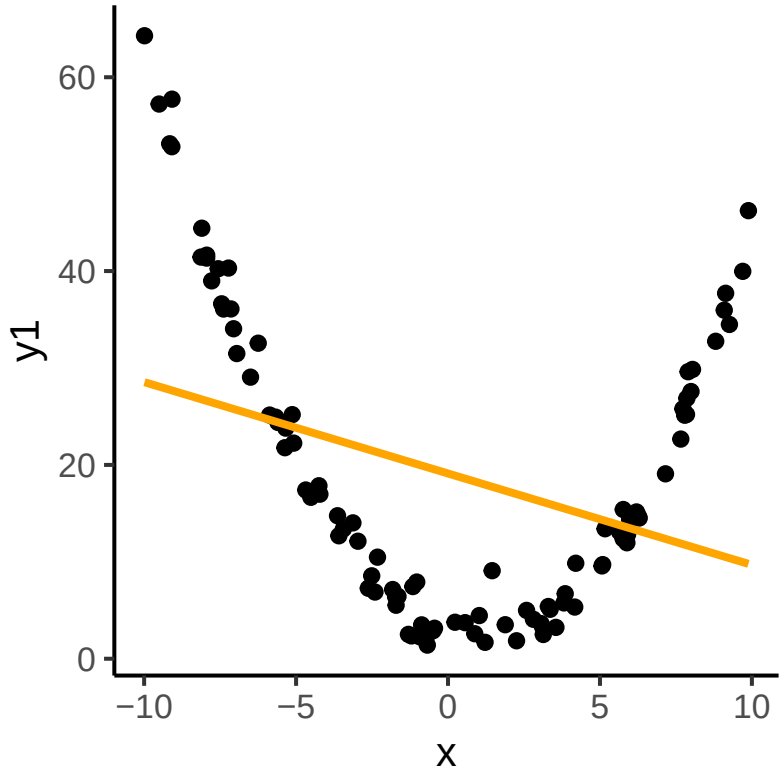
Some common ways of checking the assumptions: **residual plots**

```
mod1 <- lm(mpg~disp*factor(gear),data=mtcars)
par(mfrow=c(1,2),mar=c(3,3,1,1)+1)
plot(mod1, which=c(1,2))
```



1. Points in Plot 1 should show *no pattern* (shotgun blast)
2. Points in Plot 2 should be *roughly* on top of the 1:1 line

Problem 1: Non-linear relationship



$y1$ clearly follows a hump-shaped relationship, not a linear one