MECH481A6: Engineering Data Analysis in R

Chapter 11 Homework: Modeling

Flynn Nyman

19 November, 2024

Load packages

Chapter 11 Homework

This homework will give you experience with OLS linear models and testing their assumptions.

For this first problem set, we will examine issues of ***collinearity among predictor variables*** when fitting an OLS model with two variables. As you recall, assumption 3 from OLS regression requires there be no *collinearity* among predictor variables (the X_i's) in a linear model. The reason is that the model struggles to assign the correct β_i values to each predictor when they are strongly correlated.

Question 1

Fit a series of three linear models on the `bodysize.csv` data frame using `lm()` with `height` as the dependent variable:
1. Model 1: use `waist` as the independent predictor variable:
   - `formula = height ~ waist`
2. Model 2: use `mass` as the independent predictor variable:
   - `formula = height ~ mass`
3. Model 3: use `mass + waist` as a linear combination of predictor variables:
   - `formula = waist + mass`

Report the coefficients for each of these models. What happens to the sign and magnitude of the `mass` and `waist` coefficients when the two variables are included together? Contrast that with the coefficients when they are used alone.

Evaluate assumption 3 about whether there is collinearity among these variables. Do you trust the coefficients from model 3 after having seen the individual coefficients reported in models 1 and 2?

The coefficient for model 1 is 0.11, for model 2 it is 0.202, the coefficient for waist in model 3 is -0.64, and 0.64 for the mass in model 3. In model 3 the waist coefficient has a negative sign while the mass coefficient has a positive sign. When they are used alone they both have a positive value that is much lower in magnitude from when you combine them.

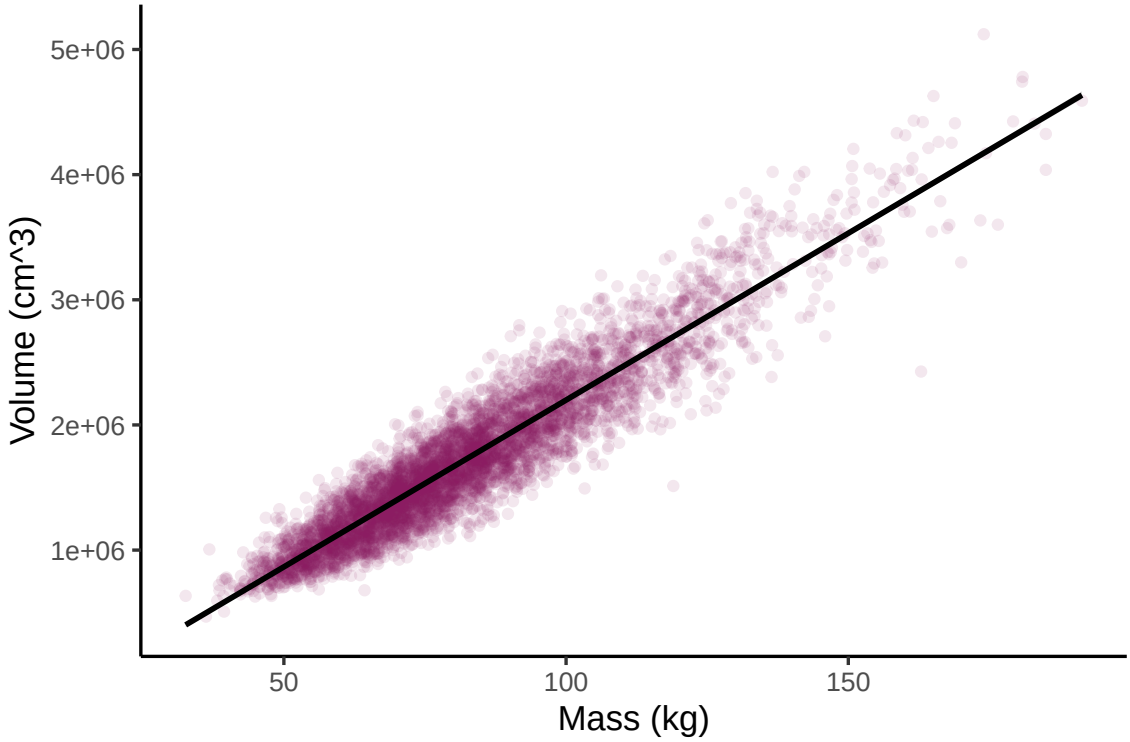
Question 2

Create a new variable in the `bodysize` data frame using `dplyr::mutate`. Call this variable `volume` and make it equal to $waist^2 * height$. Use this new variable to predict `mass`.

Does this variable explain more of the variance in `mass` from the NHANES data? How do you know? (hint: there is both *process* and *quantitative* proof here)

Yes this model likely explains more variance for a couple reasons. For one, volume and mass have a correlation that makes the most sense compared to circumference and height. As the volume of an object increases, the mass will definitely increase. Furthermore, looking at the R-squared value of model 4 shows a value of 0.877.

Create a scatterplot of `mass` vs. `volume` to examine the fit. Draw a fit line using `geom_smooth()`.



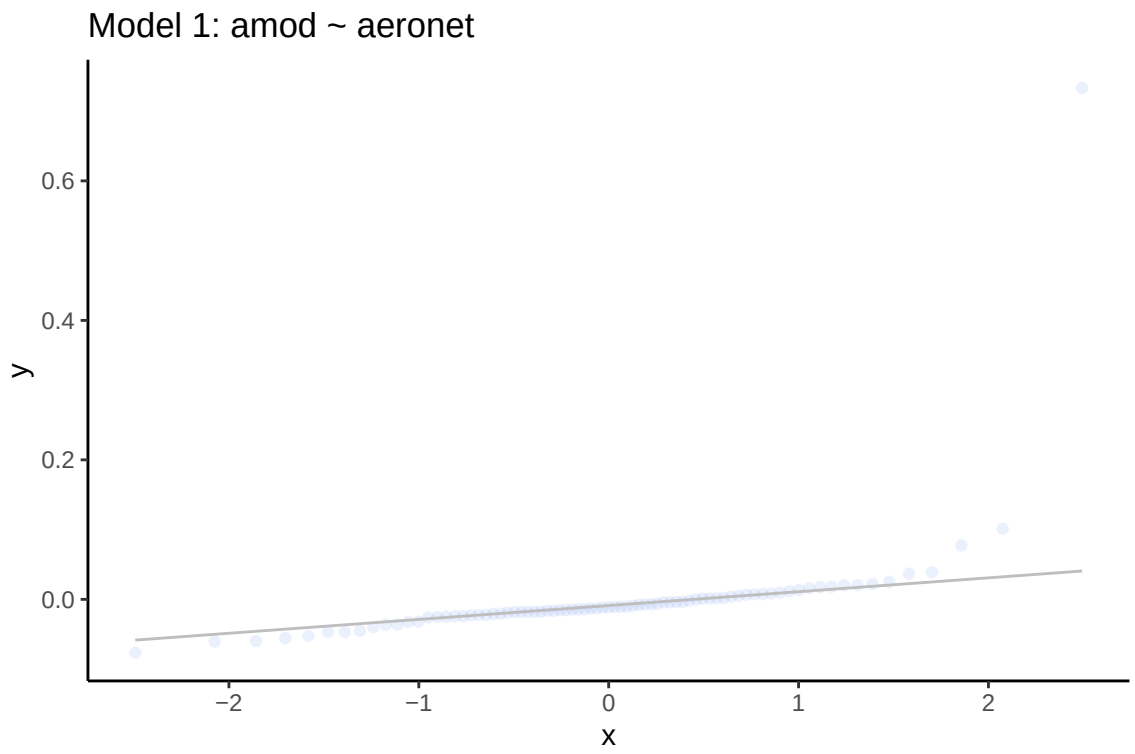
Question 3

Load the `cal_aod.csv` data file and fit a linear model with `aeronet` as the independent variable and `AMOD` as the independent variable.

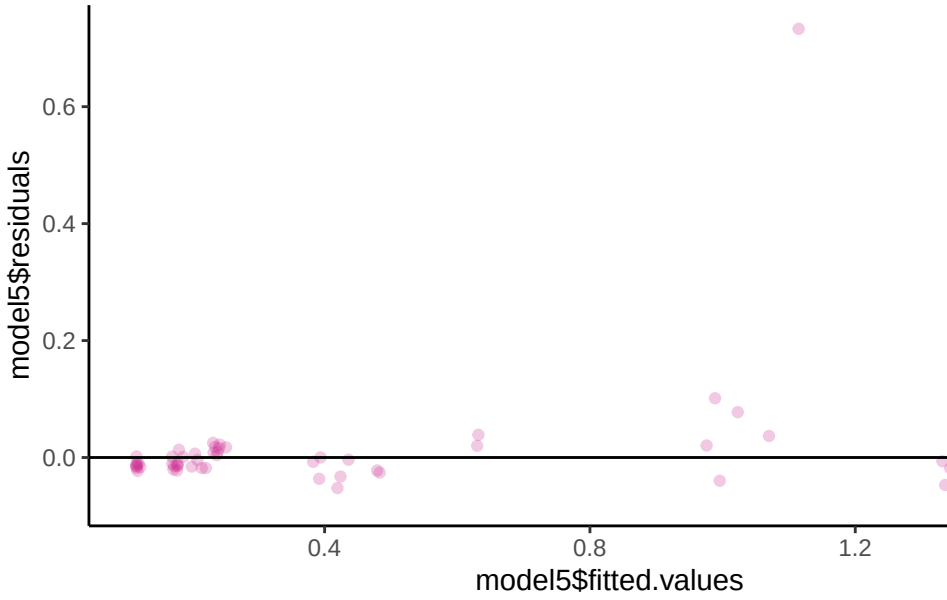
Evaluate model assumptions 4-7 from the coursebook. Are all these assumptions valid?

```
## [1] -1.129245e-18
```

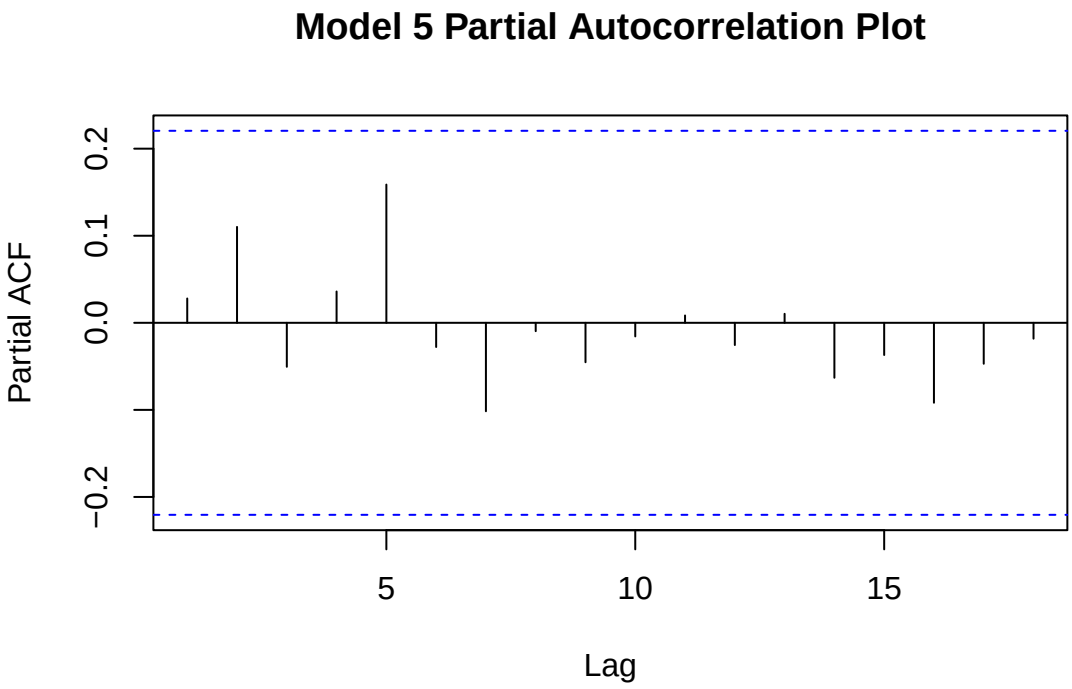
The sum of the residuals comes out to a number with an exponent of -18 which can be treated as zero.



The residuals are normally distributed enough because they follow the quantile plots within reason. There is some



stray poiunts at the upper and lower ends.
The residuals do not show major change across the range of fitted values. This means they are homoscedas-



tic.
This shows there is not autocorrelation between values in model 5.