MECH481A6: Engineering Data Analysis in R

Chapter 11 Homework: Modeling

Brad Portouw

11 December, 2022

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?

model1 is a linear model with waist size (circumference in cm) as the independent variable and height (cm) as the dependent variable with coefficients of:

Intercept: 155.5, slope: 0.11

model2 is a linear model with mass (kg) as the independent variable and height as the dependent variable with coefficients of:

Intercept: 150.6 slope: 0.191

model3 combines the predictor variables of mass and waist circumference into one independent variable and their relationship to height with the coefficients being:

Intercept: 177.46 waist: -0.634 mass: 0.639

Here the coefficients seemingly cancel each other out, and are greater in magnitude than in models 1 and 2. This would suggest that waist size and mass are colinear. This becomes a problem becuase it is difficlut to determine which value is the determining factor for height as taller people both likely weigh more and have larger waist sizes, but both values can vary because higher waist sizes are also likely correlated with greater mass.

It may be hard to use model 3 as the two predictor variables. The coefficients given are not useful in this case unfortunately. However it may give a potentially accurate answer. Depending upon the need for either coefficients or just outputs, the model may be vaild or unusable.

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`.

The waist value is the waist circumference of the individuals recorded, which is diameter of an individual times pi assuming a perfect circle. In this case, the area of a circle is (pi/4)*d^2, meaning if the circumference is substituted the diameter will become (circumference/pi) or (waist/pi).

Model 4, gives coefficients depending upon the calculated volume of each individual with:

Intercept: 25.46 kg Coefficient B1: 4.135*10^-4

This would translate to:

(Mass in kg)= 25.46kg + (4.135*10^-4)(Volume in cm3)

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)

```
##
## Call:
## lm(formula = bodySize$mass ~ bodySize$height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.457 -14.051  -3.331  10.310 158.689
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -83.3645     4.6492  -17.93   <2e-16 ***
## bodySize$height   0.9969     0.0279   35.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.69 on 5432 degrees of freedom
##   (99 observations deleted due to missingness)
## Multiple R-squared:  0.1903, Adjusted R-squared:  0.1902
## F-statistic:  1277 on 1 and 5432 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = bodySize$mass ~ bodySize$waist)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -28.810  -6.721  -0.231   6.354  52.621
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -34.371523   0.816004  -42.12   <2e-16 ***
## bodySize$waist   1.164661   0.008029  145.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.948 on 5178 degrees of freedom
##   (353 observations deleted due to missingness)
## Multiple R-squared:  0.8025, Adjusted R-squared:  0.8025
## F-statistic: 2.104e+04 on 1 and 5178 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = bodySize$mass ~ bodySize$volume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.818  -5.217  -0.417   4.821  57.583
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.546e+01  3.149e-01   80.86   <2e-16 ***
## bodySize$volume 4.136e-04  2.150e-06  192.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.841 on 5173 degrees of freedom
##   (358 observations deleted due to missingness)
## Multiple R-squared:  0.8774, Adjusted R-squared:  0.8773
## F-statistic: 3.701e+04 on 1 and 5173 DF,  p-value: < 2.2e-16
```

Initially looking at the summaries of the linear models made for comparing mass to height and mass to waist circumference, the R squared values are:

mass-height R2: 0.1903

mass-waist R2: 0.8025

mass-volume R2: 0.8774

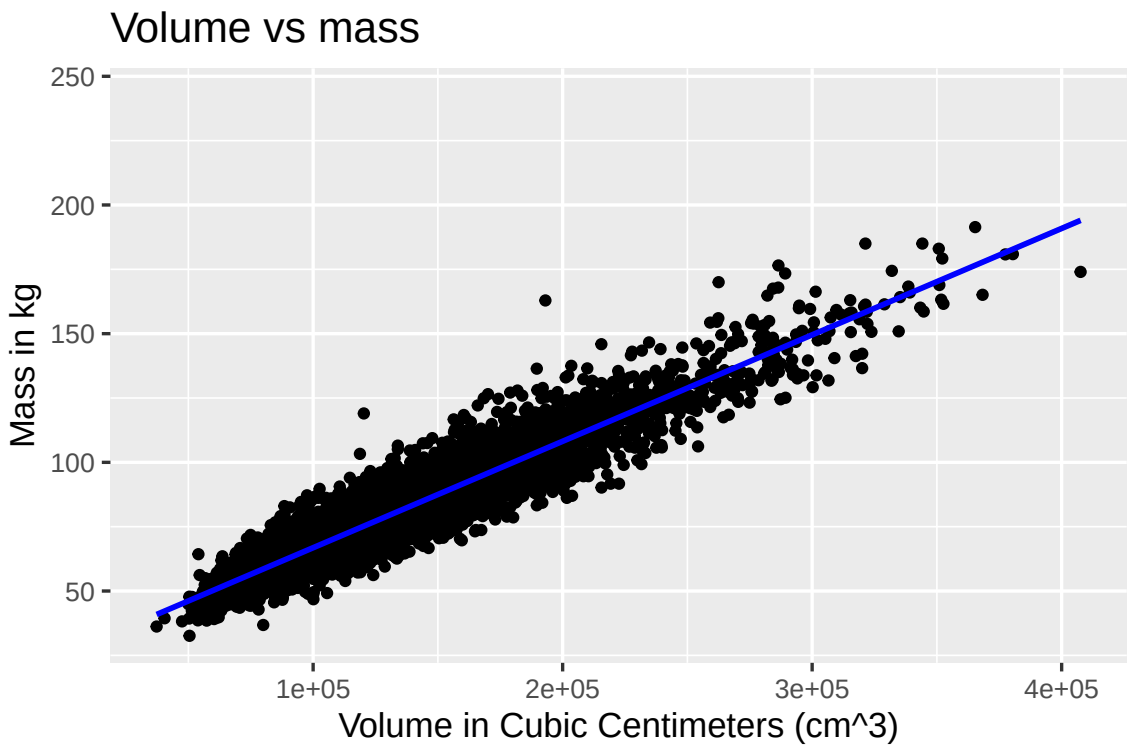
These R-square values suggest that height isn't a great predictor of mass while waist size may have a greater correlation for determining mass. The volume model has a considerably larger R-squared compared to waist size, suggesting that including the height variable helps explain more of the variance in values observed for mass.

When determining mass from a person's geometry, we would multiply the volume of a person by their density. Volume of a cylinder is determined by the cross sectional area (found from waist), times the height. With the area being a squared value, it makes sense that its value would carry more weight in determining volume overall versus height. This may be explained by a tall skinny person having lower mass than a shorter person who is more round.

Volume = (pi/4)(d^2)(height)

Overall model 4 likely is a better model than using either height or waist size for determining mass as it is probably more correctly specified

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

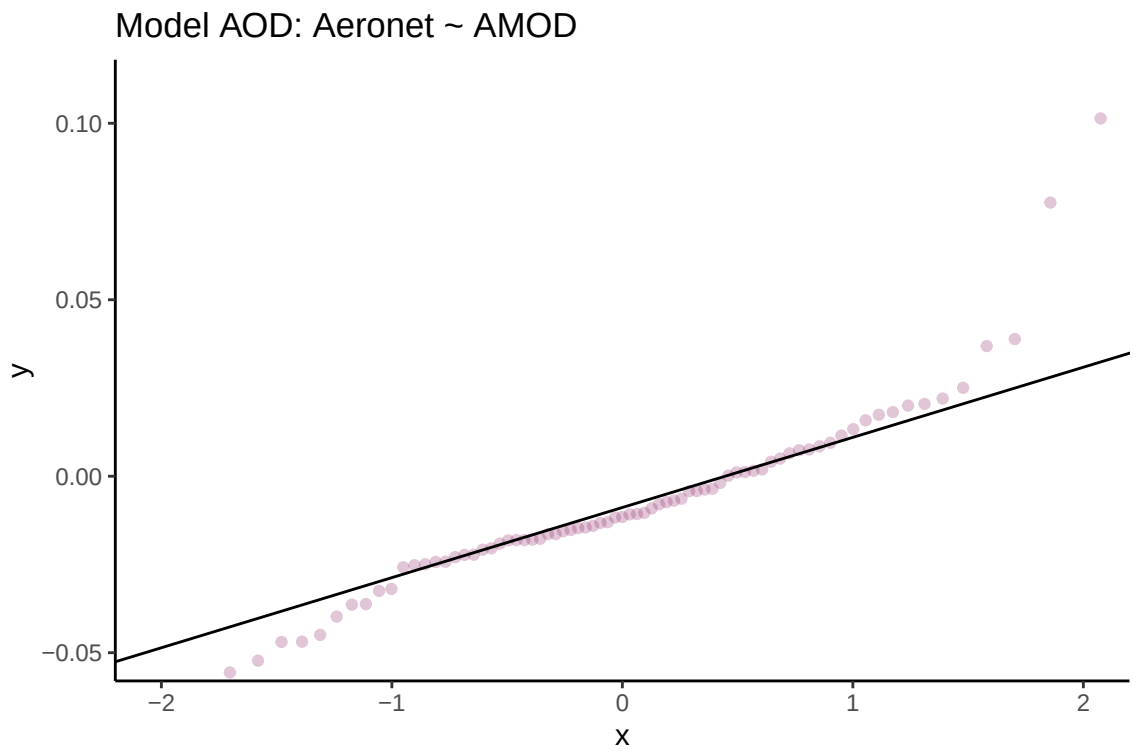


Question 3

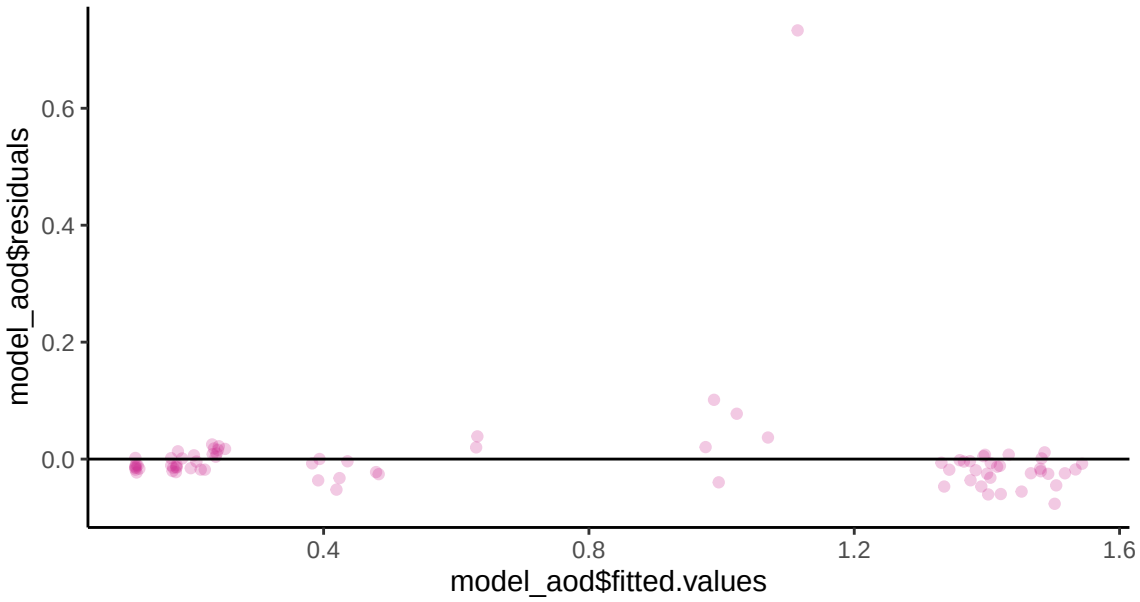
Load the `cal_aod.csv` data file and fit a linear model with `aeronet` as the independent variable and `AMOD` as the independent variable.

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

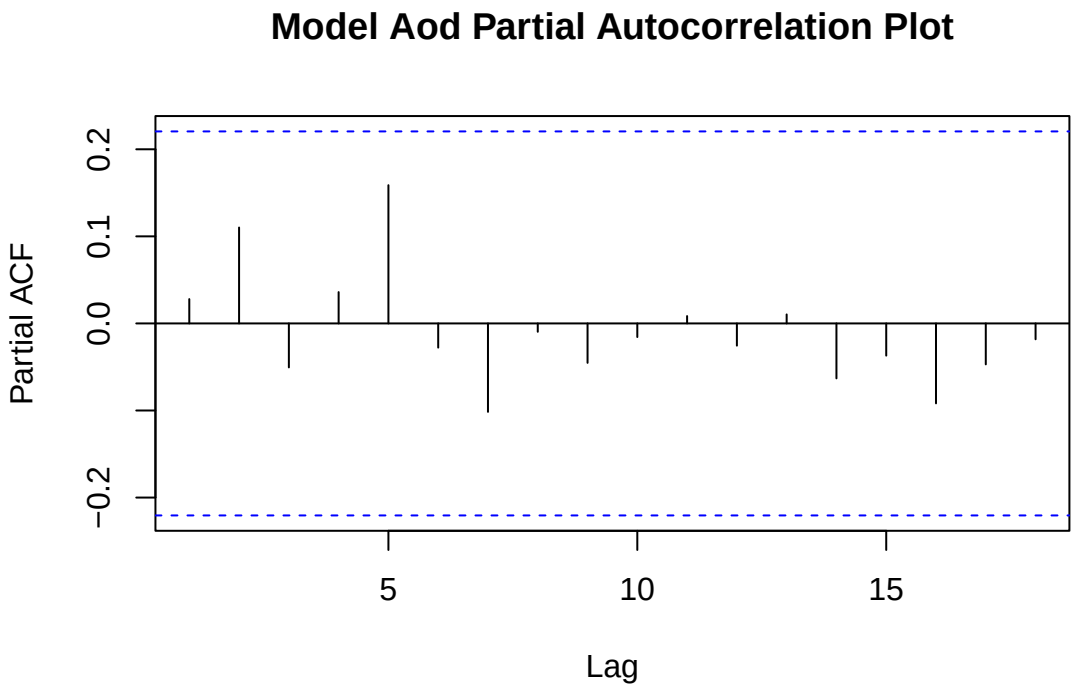
Assumption 4 wants to see if the mean of the residuals is equal to zero. In this case the variable res_mean returns a number to the -18th power, which is very small and is essentially zero. Here It can be said it passes assumption 4.



Looking at plot p5, the majority of the datapoints fall under what would be expected for a normal distribution, however some points on the extreme ends deviate, especially towards the right end. I will likely say that this is acceptable for assumption 5 and say generally the residuals are normally distributed, but it definitely could fit the distributon better, and could use a second look potentially taking into account other factors.



The residuals seemingly do not change much with different values of AMOD, however two points are interesting. First around an AMOD of 1.1, there exists a value much greater than the others which reside near zero. It is possible that this point is an outlier, but it isn't clear for now. Just to the left of this point there are a grouping of points that have higher variance in their residuals than the other values on the plot. This may combine with the outlier point to suggest that near values of 1.0 to 1.2, more variance in AMOD occurs. Overall I may say that the majority of the distribution of residuals are homoscedastic, but in the 1.0 to 1.2 range the behavior is suspect. Hopefully more data could be found to increase the sample size.



Looking at the autocorrelation plot all values seemingly fall within the bounds, meaning that there is a good chance there is no autocorrelation umong the residuals taking place, therefore I may say that assumption 7 is valid.