MECH481A6: Engineering Data Analysis in R

Chapter 11 Homework: Modeling

Michael Thill

12 December, 2025

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?

*** Model_1 intercept 155.4967199 coefficient 0.1099283, Model_2 intercept 150.6109083 coefficient 0.1909431, Model_3 intercept 177.4604248 coefficient -.6347416, .6394358***

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`.

```
##
## Call:
## lm(formula = mass ~ volume, data = bodysize)
##
## 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 ***
## volume      3.291e-05  1.711e-07  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
```

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)

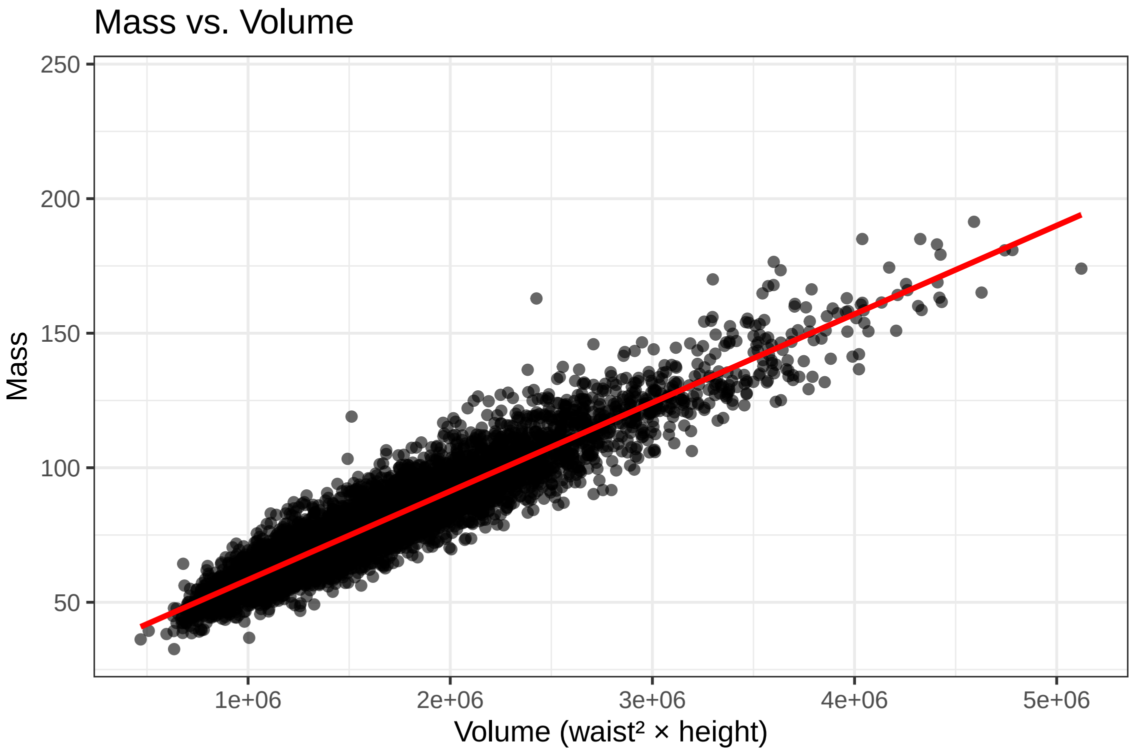
```
##
## Call:
## lm(formula = mass ~ waist, data = bodysize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.810  -6.721  -0.231   6.354  52.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.371523   0.816004  -42.12   <2e-16 ***
## 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 = mass ~ height, data = bodysize)
##
## 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 ***
## 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 = mass ~ volume, data = bodysize)
##
## 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 ***
## volume      3.291e-05  1.711e-07  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
```

*** This variable does explain the variance better then the previous models. This can be seen by the R squared value being higher in this version when compared to the other models generated and the respective R squared values they have ***

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



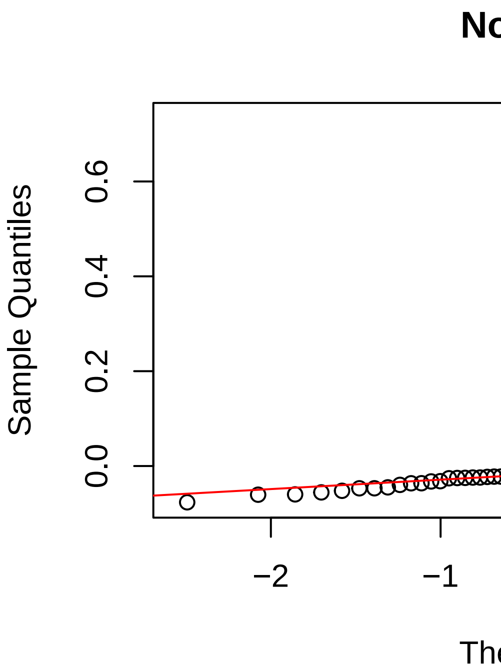
Question 3

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

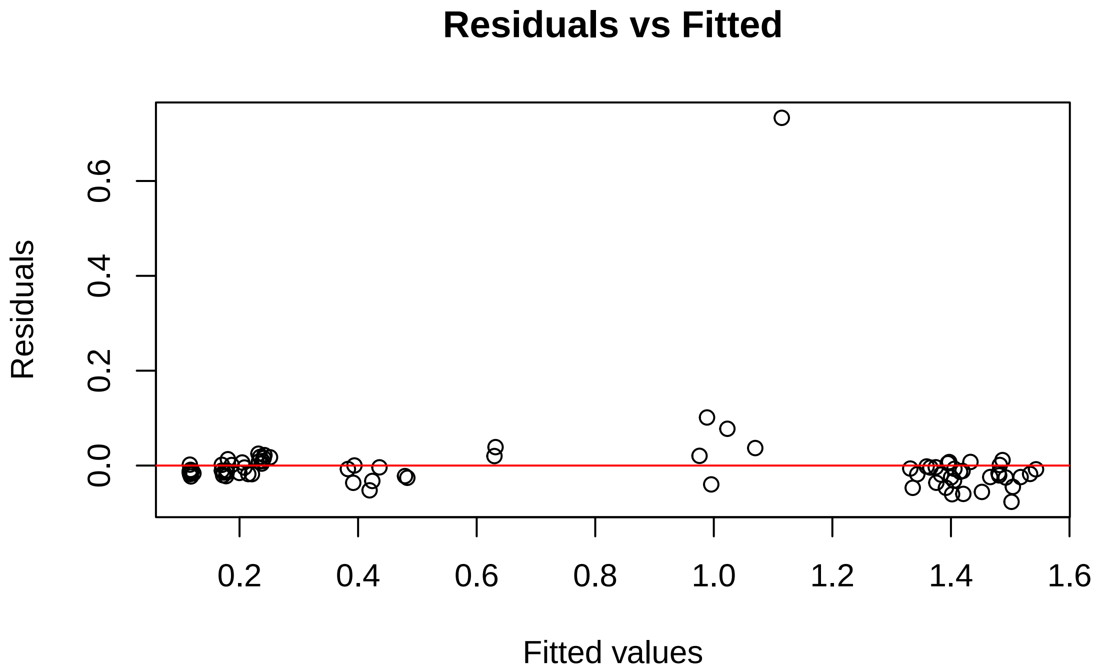
```
##
## Call:
## lm(formula = amod ~ aeronet, data = cal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07646 -0.02227 -0.01154  0.00454  0.73319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07251    0.01565   4.633 1.44e-05 ***
## aeronet      0.92344    0.01597  57.832  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08844 on 77 degrees of freedom
## Multiple R-squared:  0.9775, Adjusted R-squared:  0.9772
## F-statistic:  3345 on 1 and 77 DF,  p-value: < 2.2e-16
```

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

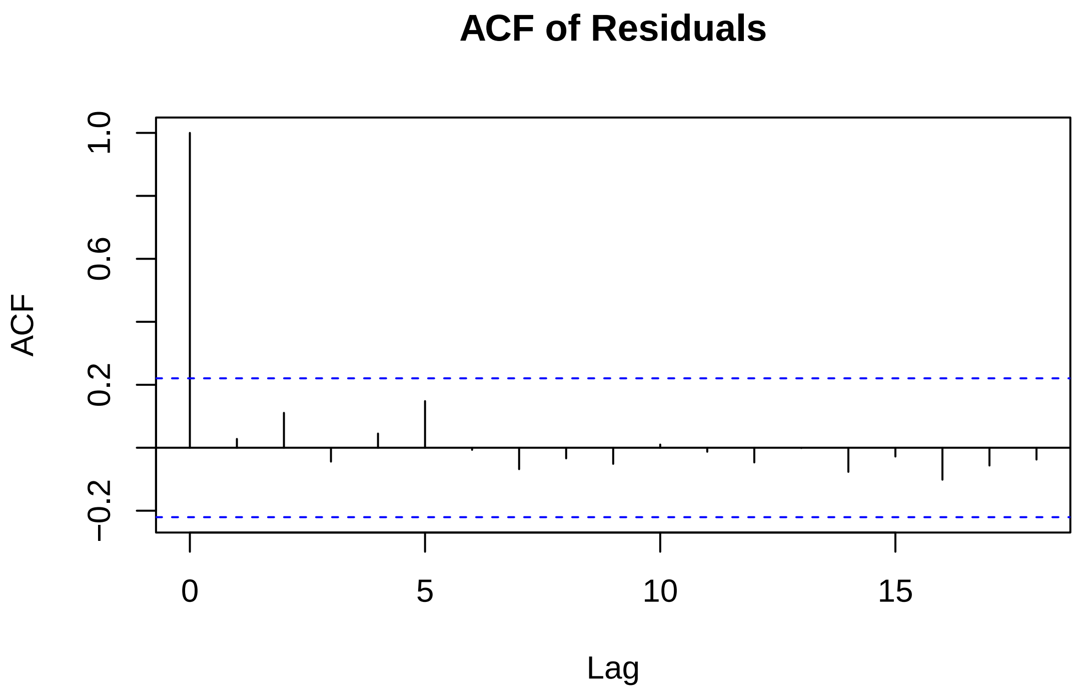
```
## [1] 2.38993e-18
```

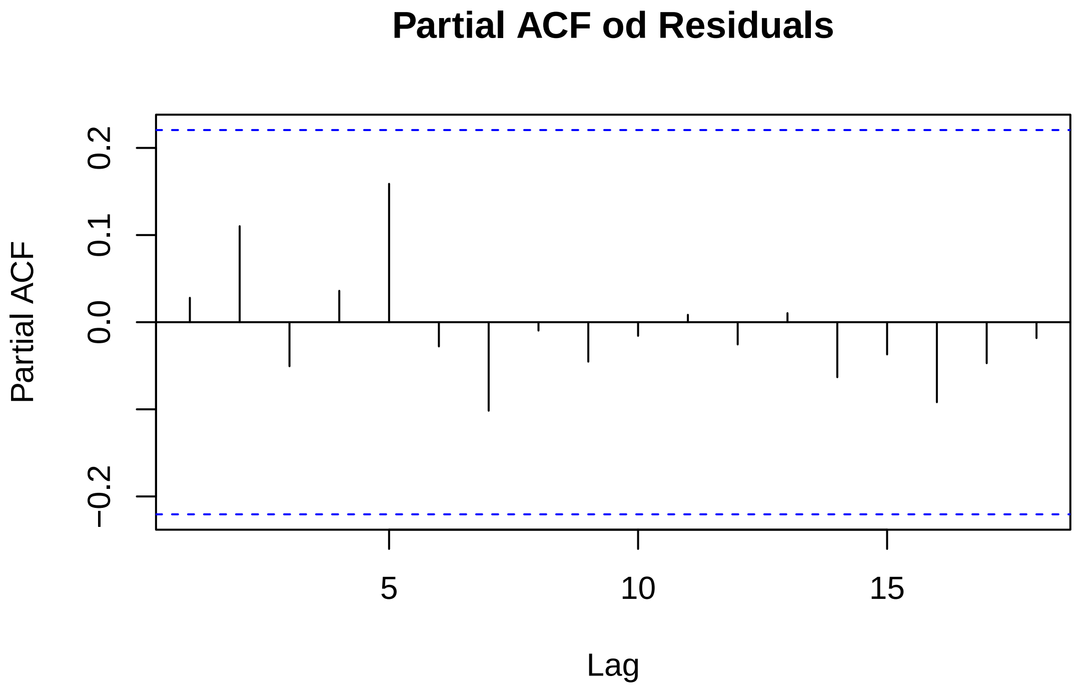


*** our assumptions are valid because the residual is basically zero for this model.
The fitted line fits the distribution plot very closely with the exception when you are looking at either end of the data where the data drops bellow on the left and and is above on the right end. ***



*** This assumption is valid for most locations as the data shows the residuals are all close to zero.





```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02797119      1.942911    0.44
##  Alternative hypothesis: rho != 0
```

*** The Durbin-Watson test is very close to zero and all values fall within the bounds except for the first one for the ACF. ***