MECH481A6: Engineering Data Analysis in R

Chapter 11 Homework: Modeling

Ethan Rutledge

13 December, 2023

Load packages

```r
# load packages for current session
library(tidyverse)
library(gridExtra)
```

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?

```r
bodysize <- read_csv("../data/bodysize.csv")

model_1 <- lm(height ~ waist, data = bodysize)
```

```r
model_2 <- lm(height ~ mass, data = bodysize)

model_3 <- lm(height ~ waist + mass, data = bodysize)

summary(model_1)
```

```
##
## Call:
## lm(formula = height ~ waist, data = bodysize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2146  -7.3983  -0.3358   7.3422  31.1741
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.555e+02  8.123e-01  191.42   <2e-16 ***
## waist       1.100e-01  7.992e-03   13.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.9 on 5173 degrees of freedom
## Multiple R-squared:  0.03532,    Adjusted R-squared:  0.03513
## F-statistic: 189.4 on 1 and 5173 DF,  p-value: < 2.2e-16
```

```r
summary(model_2)
```

```
##
## Call:
## lm(formula = height ~ mass, data = bodysize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.6630  -6.5853  -0.0814   6.5320  28.6570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.499e+02  4.769e-01  314.23   <2e-16 ***
## mass        2.021e-01  5.593e-03   36.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.007 on 5173 degrees of freedom
## Multiple R-squared:  0.2015, Adjusted R-squared:  0.2014
## F-statistic:  1306 on 1 and 5173 DF,  p-value: < 2.2e-16
```

```r
summary(model_3)
```

```
##
## Call:
## lm(formula = height ~ waist + mass, data = bodysize)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.490  -5.153   0.171   5.345  24.064
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 177.46042    0.72108  246.10   <2e-16 ***
## waist        -0.63474    0.01378  -46.07   <2e-16 ***
## mass          0.63944    0.01060   60.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.585 on 5172 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4336
## F-statistic:  1982 on 2 and 5172 DF,  p-value: < 2.2e-16
```

Answer:

See output for coefficients of each model

When the two variables are included together the slope is negative for waist wherein the others it was always positive also the std error associated is much larger

No I dont trust model 3 based on the coefficients of model 1 and 2. They seem to have strong collinearity and so the model isnt set up well.

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`.

```r
bodysize <- bodysize%>%
  mutate(volume = waist ^ 2 * height)
```

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)

```r
model_4 <- lm(mass ~ volume, data = bodysize)
summary(model_4)
```

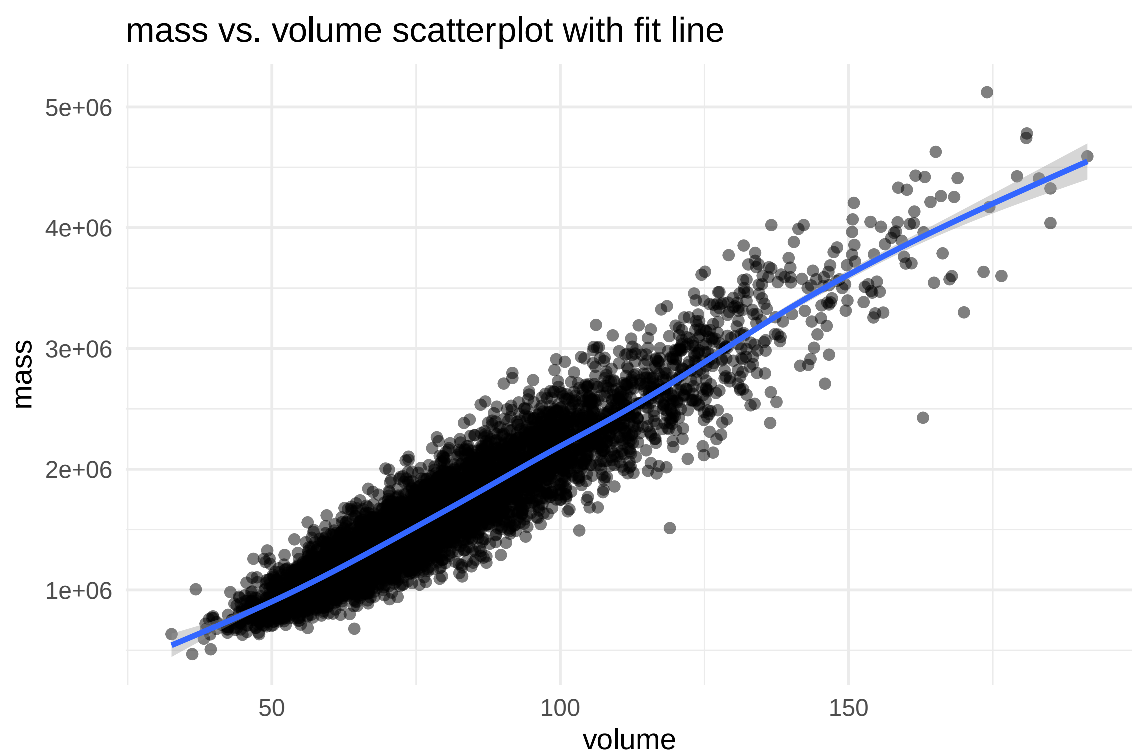
```
##
## Call:
## lm(formula = mass ~ volume, data = bodysize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.818  -5.217  -0.417   4.821  57.583
##
## Coefficients:
```

```
##               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
## 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
```

Answer: Yes, volume explains the variance in mass much better than any of the other models. This is most clear when you look at the r^2 values as this one has the highest.

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

```
ggplot(bodysize,aes(x = mass, y = volume)) + geom_point(alpha = 0.5) + geom_smooth() + labs(x="volume",
```



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.

```r
# load data
cal_aod <- read_csv("../data/cal_aod.csv")

model_cal_aod <- lm(aeronet ~ amod, data = cal_aod)
```

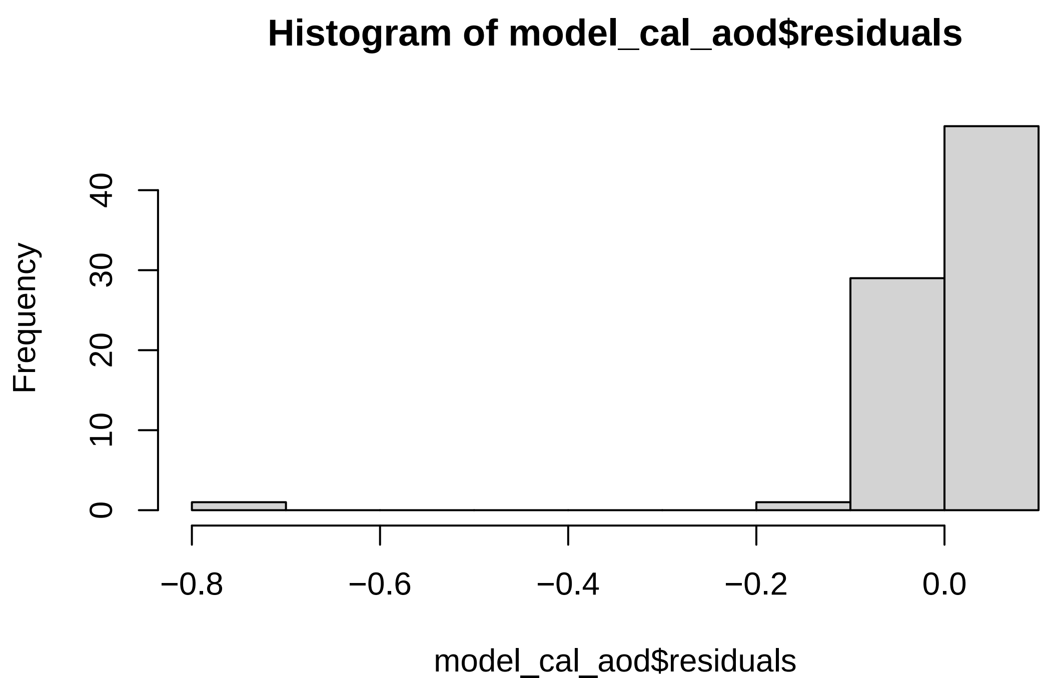
Evaluate model assumptions 4-7 from the coursebook. Are all these assumptions valid?

```r
#assumption 4: mean of residuals is zero
mean_residuals <- mean(model_cal_aod$residuals)
mean_residuals
```

```
## [1] -5.715457e-18
```

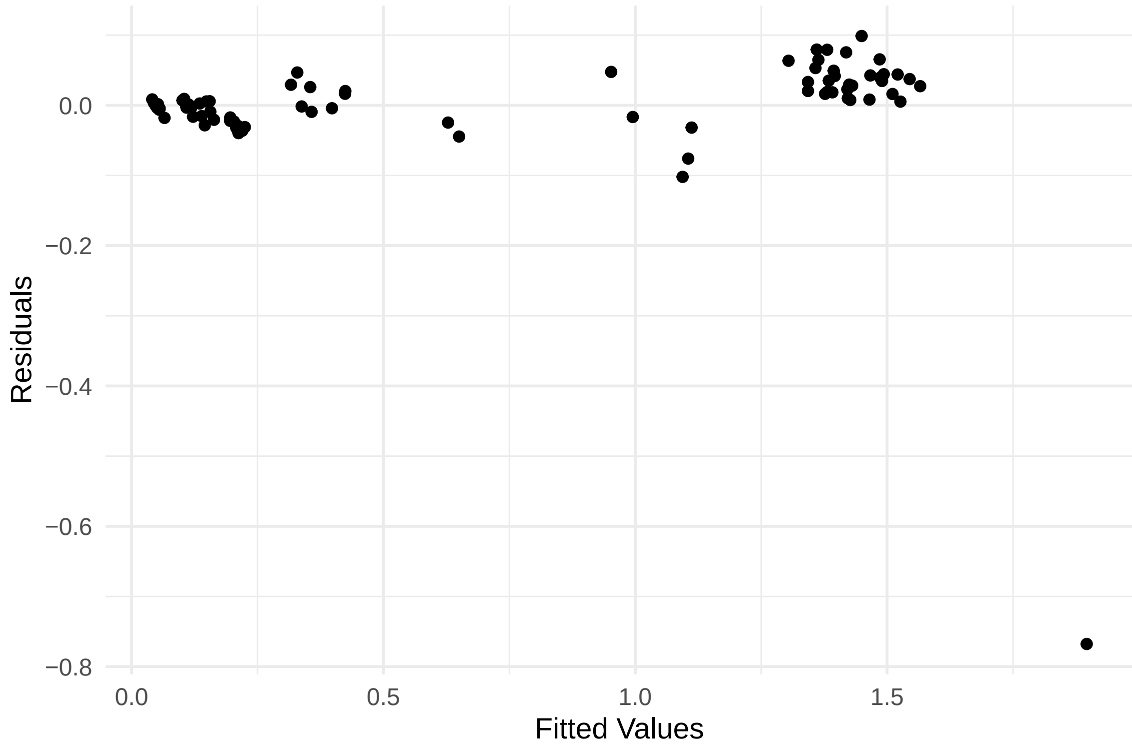
#Answer: ##Assumption4: The error term has a mean of zero ## Valid: close enought to call it zero

```r
#assumption 5: residuals are normally distributed
hist(model_cal_aod$residuals)
```



#Answer: ## Assumption 5: The error term is normally distributed ##

```r
#assumption 6: the error term is homoscedastic
ggplot() + geom_point(aes(x=model_cal_aod$fitted.values, y=model_cal_aod$residuals)) + labs(x="Fitted Va
```

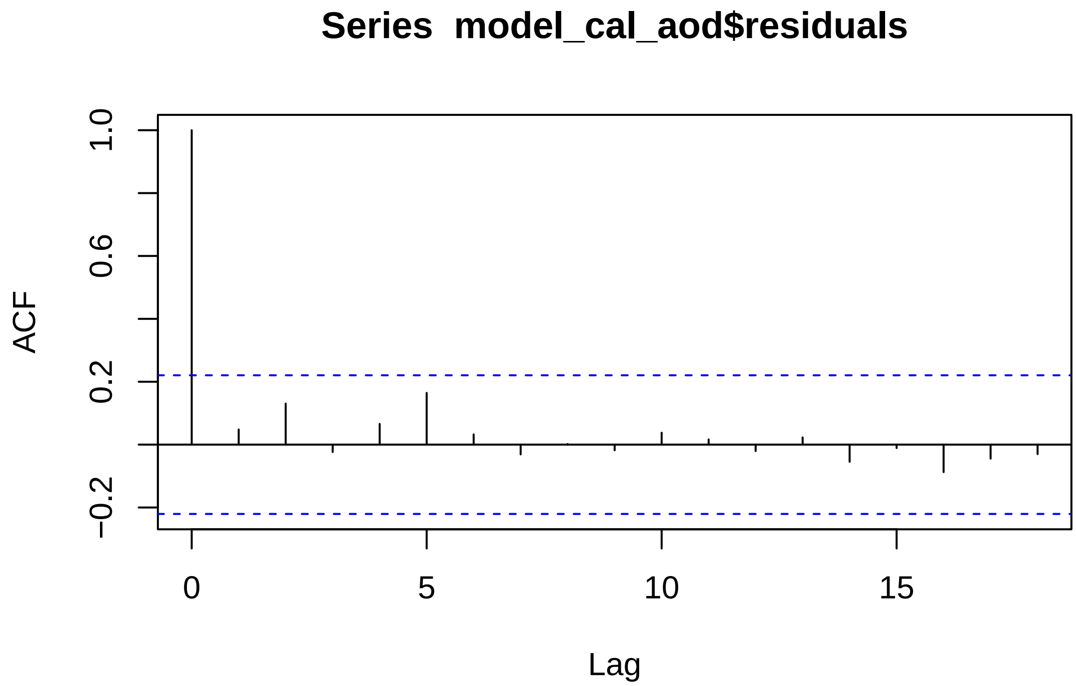


Answer:

Assumption 6: The error term is homoscedastic

##Valid: the magnitude of the residuals is constant as the fitted values increase

```
#assumption 7: no autocorrelation among residuals
acf(model_cal_aod$residuals)
```



#Answer: ## Assumption 7: No autocorrelation among residuals ## Valid: the data stays evenly around 0.0. There is one outlier that appeared also in the previous question but I believe it is an outlier and can be disregared.