

HW 7

Model Assessment and Multicollinearity

November 14, 2025

💡 Assignment Goals

1. Compare nested models using the sum of squares decomposition and nested F -tests
2. Calculate and use case influence statistics to identify unusual observations
3. Consider the implications of multicollinearity on a regression model and explore possible methods of redressing multicollinearity

Q1 (Sequential Sum of Squares Decomposition)

A number of mountains in the Adirondacks of upstate New York are known as the High Peaks, with elevations near or above 4,000 feet. The output below shows the sequential sum of squares decomposition for a model that estimates the expected time of a round-trip hike (**Time**; in hours) as a function of the vertical ascent of the hike (**Ascent**; in feet), the elevation of the highest point on the hike (**Elevation**; in feet), and the difficulty of the hike (**Difficulty**, rated from 1 (easy) to 7 (most difficult)):

$$E[\text{Time}|x] = \beta_0 + \beta_1 (\text{Difficulty}) + \beta_2 (\text{Ascent}) + \beta_3 (\text{Elevation}).$$

The total sum of squares for **Time** is $SST = 351.435$. Use this information, along with the output below, to answer questions (a) through (f).

Analysis of Variance Table

Response: Time

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Difficulty	1	230.761	230.761	99.7696	1.161e-12 ***
Ascent	1	11.262	11.262	4.8690	0.03286 *
Elevation	1	A	12.269	5.3044	0.02629 *
Residuals	42	97.143	B		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Please show all of your work.

- (a) What values belong in the cells labeled A and B?
- (b) How many mountains are part of the “High Peaks”, i.e., how many observations were there in the data set?
- (c) What is the estimated residual standard deviation, $\hat{\sigma}_\epsilon$, for the full model?
- (d) What is the model sum of squares for the full model that regresses hike time on ascent, elevation, and difficulty?
- (e) Suppose you want to test whether elevation contributes significant additional information to a model that already includes hike difficulty and vertical ascent. Use the information above to conduct the corresponding F-test. In particular, write your null and alternative hypotheses (in symbols), provide the value of the test statistic and the corresponding p -value, and state your conclusion in a sentence.
- (f) Now suppose you want to test whether elevation and vertical ascent collectively contribute significant additional information to a model that already includes hike difficulty. Use the information above to conduct the corresponding F-test. Again, please write your null and alternative hypotheses (in symbols), provide the value of the test statistic and the corresponding p -value, and state your conclusion in a sentence. *Hint: this test statistic and p -value are **not** given in the sequential ANOVA table, meaning that you will need to compute them given the information provided.*

Q2 (The Importance of Getting Some Beauty Sleep)

A team of zoologists would like to better understand the sleep habits and behaviors of mammals. To do so, they collect data on the sleep and physical characteristics of 62 different mammal species. The measured variables include:

- the total amount of sleep (in hours; **ts**) that the animal gets in 24 hours
- the brain weight (measured in grams; **brainwt**)
- the body weight (measured in kilograms; **bodywt**)
- the maximum lifespan for that species (measured in years; **life**)
- the typical gestational period (measured in days; **gp**)
- the amount of danger the animal is routinely exposed to (categorized numerically from a very low (1) to very high (5) amount of danger; **d**)

They are particularly interested in predicting mean daily sleep, and begin their analysis by fitting the following model to the data set:

```
sleep <- read_csv("https://raw.githubusercontent.com/kaitlyncook/data-sets/main/mammal-sleep.csv")

# Indicating that the amount of danger is a categorical variable
sleep <- sleep |>
  mutate(d = factor(d))

# Removing any observations with missing data in the analysis variables
sleep_cc <- sleep |>
```

```
drop_na(ts, bodywt, brainwt, life, gp)

# Fitting the zoologists' model
sleep_lm <- lm(log(ts) ~ log(bodywt) + log(brainwt) + log(life) + log(gp) + d,
  data = sleep_cc
)
```

(a) Compute the leverage for all 51 mammals in the `sleep_cc` data set. Plot the resulting statistics and identify which (if any) species have unusual or interesting leverage values. *Please identify any species by name.*

(b) Compute the studentized residuals for all 51 mammals in the `sleep_cc` data set. Plot the resulting statistics and identify which (if any) species have unusual or interesting studentized residuals. *Please identify any species by name.*

(c) Compute the Cook's distance for all 51 mammals in the `sleep_cc` data set. Plot the resulting statistics and identify which (if any) species have unusual or interesting values for Cook's distance. *Please identify any species by name.*

(d) The zoologists are particularly interested in echidnas, quill-covered mammals that (along with platypi) are among the only mammals to lay eggs. Give the value of the “raw” residual, leverage, standardized residual, studentized residual, and Cook's distance for the echidna. Then, use exploratory data analysis, data visualizations, and simple summary statistics from the `sleep_cc` data set to explain what makes *each* of these values for the echidna slightly unusual. (*For example, what about the echidna makes its residual value unusual? Its leverage? etc.*)

Q3 (An Elephant Never Forgets)

Elephants that live through droughts during the first two years of their life tend to be worse off: water scarcity leads to dehydration, which in turn leads to reduced milk production for mothers and worse health for their calves. However, does maternal experience matter? That is, are elephant calves better off if they are a firstborn calf? Or is being a firstborn harmful during a drought?

To investigate these questions, scientists collected data on 138 male elephants that lived through droughts in the first two years of their life. Each elephant's height (measured at the shoulder and recorded in centimeters) was used as a proxy for their health and overall well-being. Researchers also recorded whether the elephant was firstborn and their age (in years) at the time of the height measurement, as the researchers thought age might be a confounding factor:

```
library(Stat2Data)
data("ElephantsFB")

head(ElephantsFB)
```

	Age	Height	Firstborn
1	1.40	120	0
2	17.50	227	0

3	12.75	235	0
4	11.17	210	0
5	12.67	220	1
6	12.67	189	1

(a) Fit a multiple linear regression model of an elephant's height as a function of its firstborn status and its age, with age represented quadratically, and print out the summary table for the model. (*Don't forget to also include any lower-order terms when fitting the model!*)

(b) Calculate the Variance Inflation Factor (VIF) for each term in the fitted model. Are you surprised by what you find? Why or why not? Is the level of multicollinearity you see a concern for this model?

(c) Would you recommend that the scientists drop one or more terms from their model? If so, which one(s)? If not, why not?

(d) Use the `mutate()` function to create a mean-centered version of the `Age` variable. Then, refit the regression model from part (a), but this time using the mean-centered version of age instead. Comment on the similarities and differences between the two models: Are the model predictions (e.g., for a 15-year-old elephant) the same or different (and should they be)? Are the fitted regression coefficients the same or different (and should they be)? Is the level of multicollinearity (as assessed by the VIF) the same or different?

(e) Suppose that the scientists add a new, third variable to the regression model from (d): an indicator of the elephant's species (Asian elephant or African elephant). Suppose that this species indicator is uncorrelated with both age and firstborn status, but is related to an elephant's height. Would you expect the standard error for the coefficient of `Firstborn` to increase, decrease, or stay the same in this new model? Explain your reasoning. *Hint: how would adding the species indicator change the VIF for `Firstborn`? how would adding the species indicator change the root mean square error of the model, $\hat{\sigma}$? how are the VIF and $\hat{\sigma}$ related to the standard error of the slope for `Firstborn`?*