Homework 6 – Stat 230 – Fall 2022

Due date: Friday, October 21

Complete the following exercises and submit your assignment via gradescope (linked on the course webpage).

Problems to start after class Oct 14

Q1

Chapter 2 exercise E.16

```
wings <- read.csv("https://aloy.rbind.io/kuiper_data/WingLength2.csv")
```

> Polynomial regression in R
>
> You can include polynomial terms in your regression model by calculating the higher-order terms within your regression equation. For example, if your data set has columns `y` and `x`, then you can fit the model $\mu(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$ using the following code:
>
> `lm(y ~ x + I(x^2), data)`
>
> The key is to put your polynomial terms inside an `I()`.

Q2

The `milk` data set contains comparative primate milk composition data taken from Table 2 of Hinde and Milligan (2011) *Evolutionary Anthropology* 20:9-23.

```
milk <- read.csv("https://aloy.rbind.io/data/milk.csv")
```

(a) Create a scatterplot matrix displaying the kilocalories per gram of milk (`kcal.per.g`), the percent fat (`perc.fat`), and percent lactose (`perc.lactose`). Describe the associations you see.

(b) Fit a multiple linear regression model to predict the kilocalories per gram of milk (`kcal.per.g`) using the percent fat (`perc.fat`) and percent lactose (`perc.lactose`) measurements and report table of coefficients produced by `tidy()` in the {broom} package.

(c) You should have found that only one of the slopes was statistically discernibly different from 0 based on the individual t-tests. Your friend, who hasn't taken Stat 230 is surprised by this, since they were convinced that both the percent fat and the percent lactose would be important predictors based on the scatterplot matrix. Explain to your friend why this isn't surprising.

(d) Calculate the variance inflation factor for each predictor. Are there any indications of multicollinearity?

Q3

Explain why multicollinearity it not a problem for researchers who are trying to develop a model that accurately predict their response.

Problems to start after class Oct 19

Q4

Suppose that you are a wine enthusiast and wish to understand what factors impact the price of fine wine, which would allow you to find good deals. You collect data on 72 wines, recording the following variables:

Variable	Description
wine	Name of the winery (character vector)
price	price (in pounds sterling) of 12 bottles of wine
parker	Robert Parker's rating (out of 100)
coates	Clive Coate's rating (out of 20)
p95	Is Parker's score above 95? 1 = yes, 0 = no
first.growth	Is the wine a first growth? 1 = yes, 0 = no
cult	Is the wine considered to be a cult favorite? 1 = yes, 0 = no
pomerol	Is the wine from Pomerol? 1 = yes, 0 = no
superstar	Is the wine a vintage superstar as awarded by Parker? 1 = yes, 0 = no

To load the data set, run

```r
wine <- read.csv("https://aloy.rbind.io/data/bordeaux.csv")
```

(a) Develop a single multiple regression model that enables you to do the following:

  - estimate the percentage effect on price of a 1% increase in Parker Points and a 1% increase in Coates Points
  - control for the impacts of first growth, cult favorite wines, and wines from the Pomerol region
  - comment on the following claim from Eric Samazeuilh (a courtier): > “Parker is the wine writer who matters. Clive Coates is very serious and well respected, but in terms of commercial impact his influence is zero.”

  Are the model assumptions reasonable for your model? To answer this question include only a single “final” model that you feel adequately fits the data and show that it is adequate (or close enough to adequate).

(b) Report the fitted regression equation for your multiple regression model.

(c) Using your multiple linear regression model, estimate the percentage effect on price of a 1% increase in Parker Points and a 1% increase in Coates Points.

(d) Using your multiple linear regression model to justify your answer, comment on the above claim from Eric Samazeuilh.

(e) Identify the wines in the data set which, given the values of the predictor variables, are (i) unusually high priced, and (ii) unusually low priced.