Midterm

PSTAT 120C

Summer 2022 Session B

Instructions: This exam is open book and open note and has no strict time limit. You can use any course materials; you are also allowed to use results in the book, lecture notes, and past homework without repeating proofs or derivations. Please do not consult with other students until after the submission deadline has passed.

You may use statistical software, including (but not limited to) R or Python, to help answer these questions, or you may solve them manually. If you do use software, you *must* also submit your code.

By submitting your work, you are acknowledging that your work is entirely your own.

Background

The Environmental Protection Agency, or EPA, regularly publishes data on automotive trends by year; it has maintained its database since 1975 and is updated annually to include the most up-to-date data available for all model years.

Real-world miles per gallon (*mpg*) refers to an EPA-calculated weighted average of city and highway miles per gallon. Engine displacement (*displacement*) is measured in cubic centimeters (cm^3); it is considered an expression of engine size, or a representation of the power an engine is capable of exerting and the amount of fuel it can be expected to consume.

For this exam, you'll investigate the relationship(s) between *mpg*, weight in pounds (*weight*), and *displacement*. You'll use a random sample of 15 vehicles from the summary automotive trends data, which includes information about vehicle attributes for model years ranging from 1975 to 2021.

Below, Figure 1 shows the distribution of *mpg* in a random sample of 15 vehicles.

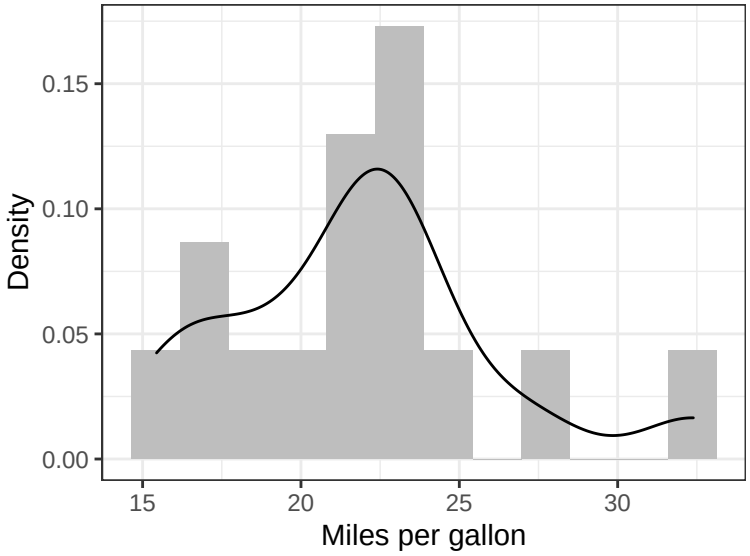

Figure 1: Histogram and density curve for miles per gallon of 15 randomly sampled vehicles.

Figure 2 is a matrix of the correlation coefficients (r^2) between *mpg*, *weight*, and *displacement*. Note that the diagonal is made up of the correlations between each variable and itself, or 1.

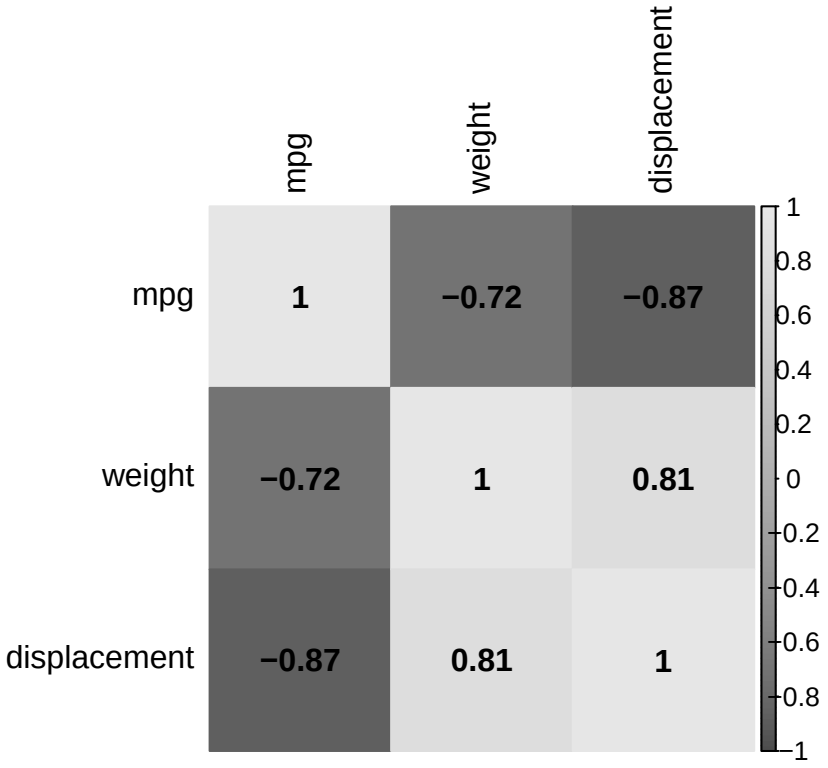

Figure 2: Correlation plot of miles per gallon, weight in pounds, and engine displacement.

Finally, the data themselves are presented in Table 1 at the end of this document. They are also available as a .csv file for download on GauchoSpace.

The overall question of interest throughout this exam is: **Should miles per gallon be predicted based on weight alone, or on the linear combination of weight and displacement?**

1. Answer the following based on a *simple* linear regression, predicting *mpg* (y) with *weight* (x_1).

   (a) Fit the specified model. Write the model equation, including your estimates.

   (b) Create a scatterplot of *mpg* and *weight*. Add a line representing the model, with 95% confidence bands. Does the model appear to fit the data?

   (c) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. **Do not** interpret the conclusion of this test.

2. Answer the following based on a *multiple* linear regression, predicting *mpg* with *weight* (x_1) and *engine displacement* (x_2).

   (a) Fit the specified model. Write the model equation, including your estimates.

   (b) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Interpret the conclusion of this test at $\alpha = 0.05$.

   (c) Consider $x_1^* = 3000$ and $x_2^* = 150$. Calculate a 95% confidence interval for $E[Y|x_1 = x_1^*, x_2 = x_2^*]$. Calculate a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$. Interpret both of these intervals in context.

   (d) Which model constitutes the "complete" model and which the "reduced" model? Can x_2 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

3. Consider your answers to the previous questions, then answer the following.

   Suppose that the true population relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

   Further suppose that there is a relationship between x_1 and x_2, given by:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

   where γ_1 and β_2 are non-zero.

   (a) Find the expected values of β_0 and β_1 if the independent variable x_2 is omitted from the regression.

   (b) Calculate the bias (if any) of β_0 and β_1 when x_2 is omitted.

   (c) What values of γ_1 and β_2 would result in β_0 and β_1 remaining unbiased?

   (d) In light of the above:

      i. What assumption of linear regression is being violated in Question 1? Is this assumption met in Question 2?

      ii. In Question 1, are the estimates of β_0 and β_1 BLUE? Why or why not?

Table 1. Raw data for a random sample of 15 vehicles from the EPA Automative Trends Database.

Model Year	MPG	Weight	Displacement	Class
2015	21.54716	4124.129	178.5575	Truck
2007	17.02911	4736.041	236.0139	Truck
2003	19.33781	3777.898	179.4107	Truck
1986	23.02399	3174.024	190.2972	Car
2017	22.54566	4650.112	164.4554	Truck
Prelim. 2021	32.38923	3194.868	114.4701	Car
2000	22.51440	3400.909	168.2990	Car
2014	22.18444	4458.683	208.4433	Truck
1999	21.50476	3879.585	197.3525	Car
2012	27.21958	3450.740	137.7964	Car
1990	23.73371	2929.358	122.0215	Car
1998	24.57349	3304.248	142.4937	Car
2007	19.09633	4461.215	218.8619	Truck
2002	15.44052	4987.675	302.1571	Truck
1988	16.42429	4357.654	239.6896	Truck