MATH 3330 – Fall 2024
Assignment 2 – 50 marks – Due Nov. 8

Question 1 (16 marks): Load in the Spotify dataset into `R`. The variables are given below:

- **Popularity**: A measure of how well a track is received, often based on streaming counts and social media buzz.
- **duration_ms**: The length of the track in milliseconds, indicating how long the song plays.
- **danceability**: A score reflecting how suitable a track is for dancing, based on tempo, rhythm, and beat stability.
- **energy**: An estimate of the intensity and activity level of a track, considering factors like tempo and loudness.
- **acousticness**: A measure of how acoustic (non-electronic) a track sounds, indicating the presence of acoustic instruments.
- **instrumentalness**: A score that predicts the likelihood of a track being purely instrumental, without vocals.
- **liveness**: A measure of the presence of an audience in a track, indicating whether it feels like a live performance.
- **tempo**: The speed of the track, usually measured in beats per minute (BPM).

We are interested in the relationship between popularity of a song and the remaining variables. We continue our analysis below.

a. Carry over your regression model from assignment 1, question 1. Use a method to check the assumption that the errors are normally distributed. Do you believe the assumption is reasonable? Why? (3 marks)

b. Plot the studentized residuals (y-axis) against the fitted values (x-axis). Does this plot support any of the assumptions made in the MLR model? Why or why not? (5 marks)

c. Make a histogram of the popularity scores. Do you notice anything that could be impacting the regression model? (3 marks)

d. Refit the regression model with only observations where the popularity score was positive. Remake your plots from questions a. and b. Comment on the changes - does the model fit better now? Are there still issues with the fit? Why? (5 marks)

Question 2 (12 marks):

a. Continue with the model with only observations where the popularity score was positive. For each covariate, plot the studentized residuals (y-axis) against the covariate (x-axis). Which covariates indicate that the assumptions of the MLR are violated, and why? (5 marks)

b. Refit the model with only observations where the popularity score was positive and without the covariates liveness, instrumentalness, acousticness and duration_ms. Remake your plots from questions a. and b. Comment on the changes from Q1d - does the model fit better now? Are there still issues with the fit? Why? (7 marks)

Question 3: (8 marks)

a. Derive the variance of the ith residual in a standard MLR model, i.e., simplify $Var(\hat{\epsilon}_i)$ as much as possible. (3 marks)

b. Show that standardizing the PRESS residual, that is, dividing the PRESS residual by its standard deviation, results in $\hat{\epsilon}_i / \sqrt{\sigma^2(1 - h_{ii})}$. (4 marks)

c. How does part b. compare to the studentized residual? (1 mark)

Question 4: (14 marks) Consider following regression model: Researchers are trying to discover how much money straight boyfriends spend on Christmas gifts, based on what they bought their exes 'Ex_Gift' they play (Jewelry, Engagement Ring, Tech, Homemade, Nothing), whether they are employed (Yes, No), how much they love their girlfriend 'Love' (a continuous variate, 1–10) and their 'Cheating_probability' (a continuous variate, 0–1). They fit a normal MLR model, regressing 'spend ($)' against the described variates. The following is the resulting `R` output:

	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$	VIF
(Intercept)	50	7.6	3.4	0.0009	-
Ex_Gift: Tech	-100	5.4	3.9	0.002	1.23
Ex_Gift: Homemade	-200	2.1	1.3	0.0004	1.43
Ex_Gift: Nothing	-400	2.1	1.2	0.002	2.25
Ex_Gift: E. Ring	-5	4.1	4.1	0.2	2.25
Employed: Yes	60	0.03	2.3	0.0001	2.1
Love	10	3.4	3.9	0.005	2.85
$Cheating\_probability$	750	0.2	1.7	0.2	2.56
Ex_Gift: Homemade $\times$ *Love*	15	0.2	-0.1	0.32	2.51
Ex_Gift: Tech $\times$ *Love*	5	0.1	-0.1	0.72	2.2
Love$\times$ $Cheating\_Probability$	0	0.1	-0.1	0.52	1.56

a. Write out the dummy variables used to include the 'Ex_Gift' variable in this regression model. (3 marks)

b. Interpret the coefficient for Ex_Gift: Tech (4 marks)

c. Interpret the coefficient for Ex_Gift: Tech $\times$ *Love*. (4 marks)

d. What is wrong with including the following in the model instead of the dummy variables defined in part a:

$$Ex\_Gift = \begin{cases} Jewelry & 1 \\ Tech & 2 \\ EngagementRing & 3 \\ Homemade & 4 \\ Nothing & 5 \end{cases}$$

(3 marks)