MATH 3330 – Fall 2024
Assignment 1 – 50 marks – Due Oct. 2 at 11:59 pm

Question 1 (35 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.

a. Write down a multiple linear regression model relating popularity of a song (`popularity`) to the remaining variables, including all assumptions. (5 marks)

b. Using `R`, fit a multiple linear regression model relating popularity to the remaining variables (by fit, we mean estimate the coefficients). Print the `summary` of the model. (2 marks)

c. What is the interpretation of the estimated coefficient for `danceability`? (3 marks)

d. Test for significance of the regression model by obtaining the ANOVA table, use a significance level of your choice. Interpret the result in the context of the data. (4 marks)

e. Use t-tests to assess the contribution of each regressor to the model. (use the `R` model summary output) State which regressors are significant. (3 marks)

f. What extra assumption is needed for ANOVA and the t-tests to be valid? (1 marks)

g. Calculate and interpret R^2 and R^2_{Adj} (4 marks)

h. Find and interpret a 95 % CI for the regression coefficient for `liveness`. (3 marks)

i. Refit the model without `acousticness`. What happened to the `energy` coefficient? Explain why this occurred. (3 marks)

j. Estimate and interpret σ^2 (3 marks)

k. What is the predicted popularity for a song with the following attributes . (1 mark)

	duration_ms	danceability	energy	acousticness	instrumentalness	liveness	tempo
x	205594.00	0.44	0.63	0.43	0.0042	0.07	78.90

ℓ. Report and interpret a 90% prediction interval for the prediction in k. (3 marks)

Question 2: (12 marks) Assume that $Y|X \sim \mathcal{N}_n(X\beta, \sigma^2 I)$ where β has dimension $p \times 1$. For $\hat{\beta} = (X^\top X)^{-1})X^\top Y$ complete the following:

a. Explain the difference between β and $\hat{\beta}$. (2 marks)

b. Compute the expected value of $\hat{\beta}$. (2 marks)

c. Show that the covariance matrix of the vector of fitted values $\hat{Y}$ is $\sigma^2 X(X^\top X)^{-1}X^\top$. That is, show that $\text{Cov}(\hat{Y}) = \sigma^2 X(X^\top X)^{-1}X^\top$. (4 marks)

d. Note that $SSE/\sigma^2 \sim \chi^2_{n-p}$ and SSE/σ^2 is independent of $\hat{\beta}$. Let a_j be the jth entry of the diagonal of $(X^\top X)^{-1}$. Argue that $\frac{\hat{\beta}_j - \beta_j}{\sqrt{a_j MSE}} \sim t_{n-p}$. (4 marks)

Question 3: (3 marks) Suppose that $A \in \mathbb{R}^{n \times p}$ is a matrix. List a set of sufficient conditions on A, n and p which imply that $A^\top A$ is positive definite.