Exam 2 Practice

MATH 2025

Answer the questions in the spaces provided. Show your work wherever practicable. You are permitted to use a calculator but not any other resources.

The data set for this exam is a subset from the Spotify Songs Tidy Tuesday data set. The data were originally obtained from Spotify using the **spotifyr** R package. **It is the same data set we used for our activities.**

It contains numerous characteristics for each song. This analysis will focus specifically on the following variables:

variable	class	description
track_popularity	double	Song Popularity (0-100) where higher is better
energy	double	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
valence	double	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
danceability	double	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
playlist_genre	character	The genre of the song (`edm`, `latin`, `pop`, `r&b`, `rap`, `rock`)

```r
spotify |>
  select(track_popularity, energy, valence, playlist_genre) |>
  head() |>
  kable()
```

track_popularity	energy	valence	playlist_genre
80	0.859	0.520	pop
80	0.694	0.216	pop
83	0.777	0.706	latin
83	0.499	0.511	rap
89	0.892	0.478	pop
83	0.880	0.534	pop

For questions 1-7, consider `model1` below:

```
model1 <- lm(danceability ~ valence + energy, data=spotify)
model1 |> tidy() |> kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.643	0.022	29.694	0.000
valence	0.260	0.027	9.693	0.000
energy	-0.125	0.037	-3.410	0.001

1. (5 points) Use the values above to write down the equation that represents the regression model. Use the variable names to represent the variables rather than Y's and X's.

2. (10 points) Give a one sentence interpretation (in context) of the first two numbers in the estimate column (i.e. 0.643 and 0.260).

3. (5 points) Which of the following represents the predicted `danceability` for a single observation that has a `valence` of 0.5 and an `energy` of 0.25?

   A. $0.643 + 0.260 - 0.125 = 0.778$

   B. $0.643 + 0.260 \times 0.5 - 0.125 \times 0.25 = 0.7418$

   C. $29.694 + 9.693 \times 0.25 - 3.410 \times 0.5 = 30.4123$

   D. $0.260 \times 0.25 - 0.125 \times 0.5 = 0.0025$

   E. Not enough information

4. (5 points) What is the residual of this estimated `danceability`?

   A. -0.240

   B. 0.143

   C. 0.375

   D. Not enough information to say

5. (5 points) Suppose you generate a 95% prediction interval for this prediction. If we change this prediction interval to a confidence interval (that is make a prediction for the average of your response rather than a single prediction) the width of the interval would:

   A. increase

   B. decrease

   C. remain unchanged (approximately)

6. (15 points) Suppose you use bootstrapping to generate a 95% confidence interval for the slope of `valence` in `model1`. For each of the following changes, indicate whether the width of your interval would increase, decrease, or approximately remain the same (keeping all else constant):

   (a) Increase sample size:

   A. increase

   B. decrease

   C. remain unchanged (approximately)

   (b) Increase confidence level:

   A. increase

   B. decrease

   C. remain unchanged (approximately)

   (c) Increase the number of bootstrapped replicates:

   A. increase

   B. decrease

   C. remain unchanged (approximately)

7. (15 points) Consider a hypothesis test to determine whether there is a relationship between `valence` and `danceability` in `model1`:

   (a) State the null and alternative hypothesis in both words and symbols.

   (b) Give a rough definition of a p-value in the context of this problem.

   (c) What would be the result of the test in the context of the problem if $\alpha = 0.05$?

For question 8, consider `model2` below:

```
model2 <- lm(danceability ~ valence + energy + playlist_genre, data=spotify)
```

8. (5 points) Recall that `playlist_genre` has 6 distinct genre's. How many *additional* parameters (i.e. β's) would `model2` include that `model1` would not?

   A. 5

   B. 6

   C. 10

   D. 12

   E. 15

   F. 18

9. (10 points) For the constant variance assumption of a linear model, sketch a plot that would show that this assumption would be violated. Make sure to clearly label your axes so I know what the plot represents. Feel free to also describe what I should be looking for in words if you feel that your plot is not sufficiently clear.

The code below will be helpful for the rest of the exam.

```
spotify <- spotify |>
  mutate(pop_indicator = as.factor(ifelse(playlist_genre == "pop", "pop", "not pop")))
```

10. (8 points) Consider the following model:

```
model1 <- lm(danceability ~ energy*pop_indicator, data=spotify)

model1 |> tidy() |> kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.6478241	0.0312549	20.727138	0.0000000
energy	0.1053857	0.0489308	2.153772	0.0317316
pop_indicatorpop	0.0477146	0.0464676	1.026837	0.3049900
energy:pop_indicatorpop	-0.1266868	0.0710332	-1.783487	0.0751086

This model contains two different lines. One for `pop` songs and one for `not pop`. Write down the equations of both lines. For partial credit you may just write down the equation of the model.

(a) `pop` line:

(b) `not pop` line:

For questions 11-18, consider the following models:

```
model1 <- lm(danceability ~ energy*pop_indicator,
             data=spotify)
model2 <- lm(danceability ~ valence, data=spotify)
model3 <- lm(danceability ~ valence + energy + pop_indicator, data=spotify)
model4 <- lm(danceability ~ valence + track_popularity + energy*pop_indicator, data=spotify)
```

11. (5 points) Which of the following metrics is the LEAST useful for comparing these four models:

    A. R^2

    B. R^2_{adj}

    C. AIC

    D. BIC

    E. C_p

12. (5 points) For the answer you selected above, which of the following is true (select all that apply):

    A. It is fine to use for comparing models with the same number of parameters.

    B. Using it can lead to overfitting because increasing the model complexity will always lead to a "better" number.

    C. We only use it when we are looking at a model with one predictor.

    D. Even though it isn't useful for comparing models of different sizes, it is useful for describing a model's quality once a model is chosen.

    E. It penalizes models for being more complex.

13. (6 points) Which of the following models are nested in one another (select all that apply):

    A. model1 in model2

    B. model1 in model3

    C. model1 in model4

    D. model2 in model3

    E. model2 in model4

    F. model3 in model4

For questions 14 - 18 consider the following output in addition to the code at the beginning of the previous page:

Model	r.squared	adj.r.squared	AIC	BIC
model1	0.023	0.017	-623.009	-601.857
model2	0.139	0.137	-691.201	-678.509
model3	0.169	0.164	-705.188	-684.035
model4	0.174	0.166	-704.189	-674.576

14. (2 points) Which model is "best" according to R^2?
    A. model1
    B. model2
    C. model3
    D. model4
15. (2 points) Which model is "best" according to R^2_{adj}?
    A. model1
    B. model2
    C. model3
    D. model4
16. (2 points) Which model is "best" according to AIC?
    A. model1
    B. model2
    C. model3
    D. model4
17. (2 points) Which model is "best" according to BIC?
    A. model1
    B. model2
    C. model3
    D. model4
18. (3 points) Which model do you think is "best"? If there are two candidate models select the one which is more "parsimonious"? Very briefly state your rationale.
    A. model1
    B. model2
    C. model3
    D. model4