Final Assignment – Introduction to R – Yuval Even Zahav

A – Defining the research question

1. I selected a dataset from an ongoing research project in my lab– **"War and Time Estimation: How Conflict Alters Memory for Event Timing"**. I chose this dataset because I experience a sense of distorted time perception during the war and am curious about this phenomenon in relation to time perception theories.

   The data were collected in 2 studies: one conducted before the war (75 participants) and one during the war (190 participants). Participants provided demographic information (age, gender, ethnicity, etc.), and completed three questionnaires:

- An **events questionnaire,** where they estimated the timing of news events. Higher error size (the difference between the actual time when the event occurred and its estimated time) indicates lower accuracy or altered estimation of time.
- The **Perceived Stress Scale (PSS-10)**, with scores ranging from 0 to 56, where higher scores indicate greater stress levels).
- The **Generalized Anxiety Disorder Scale (GAD-7)**, with scores ranging from 0 to 28, where higher scores indicate greater anxiety levels).

   Initially, the data consisted of four files (separate Arabic and Hebrew datasets for each study), which were later merged into two CSV files containing all relevant variables and time estimations for each event.
   * GLMM analysis is more suitable for the data, but for the current assignment I will use Multiple Linear Regression and Logistic Regression.

2. See script "TimeEstimation1.R".
   Exploratory data display:

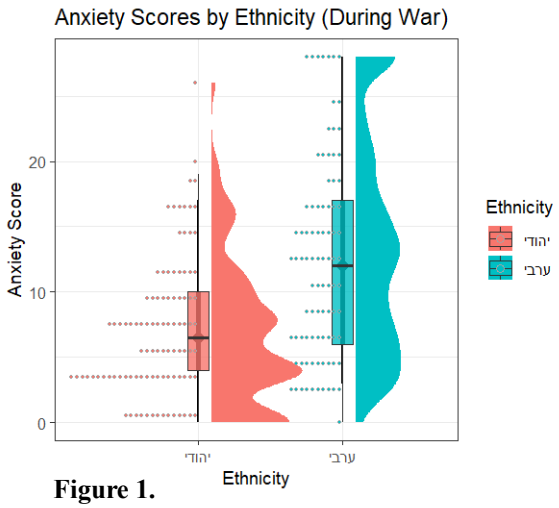

Figure 1.
Raincloud Plot of Anxiety Scores by Ethnicity in Study 2 (during war).

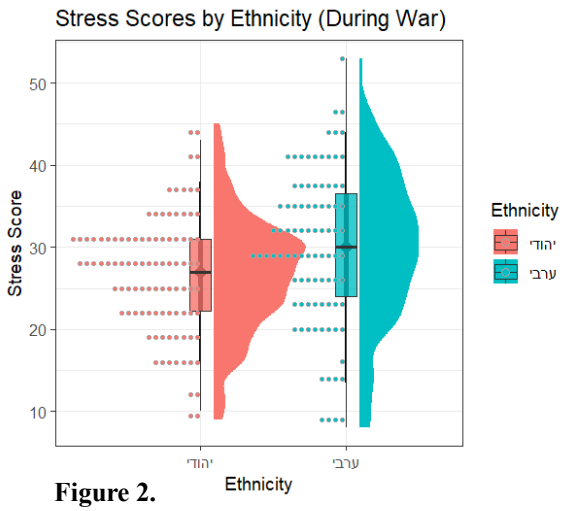

Figure 2.
Raincloud Plot of Stress Scores by Ethnicity in Study 2 (during war).

Figure 3.
Boxplot of Stress Scores by Gender in Study 2 (during war).

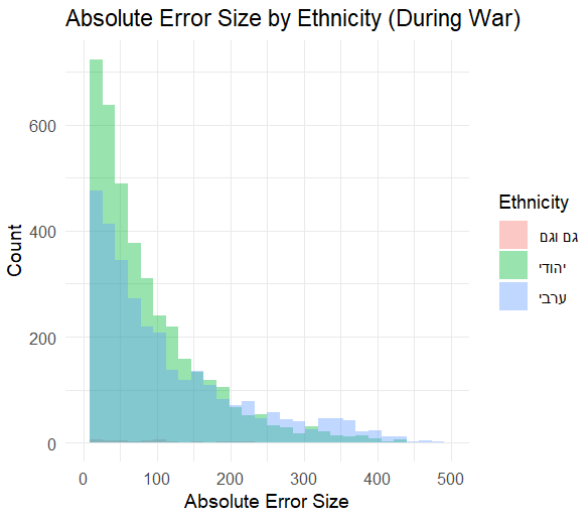

Figure 4.
Histograms of the Absolute Error Size of different ethnic identities in Study 2 (during the war).

3. The current study examines whether the Israel-Hamas war influenced individuals' time estimation and stress levels. Additionally, we explore the effects of Relevancy and Ethnicity on time estimation.
   First, we hypothesize that time estimation was more accurate before the war (i.e., lower Absolute Error Size) compared to during the war.
   Second, we hypothesize that stress levels increased during the war compared to before the war.

B – Pre-Processing

1+2. See script "TimeEstimation2.R".

After the pre-processing, the following variables remained in the filtered dataset:

- **Participant** – Unique participant IDs were assigned: "S1_x" for Study 1 and "S2_x" for Study 2. Defined as a factor for use as a random effect in a GLMM model (in a further analysis for the lab's ongoing research).
- **Event** – Events were labeled by study ("x_S1" for Study 1, "x_S2" for Study 2) and defined as a factor, also crucial for further GLMM analysis.
- **Study Period** – Converted from *Study No.* for clearer interpretation.
- **Ethnicity** – Participants were Jewish, Arab, or Other. Used as a predictor in multiple linear regression analysis.
- **Gender, Age and Anxiety Score** – Demographic and psychological factors that may be relevant for additional analyses.
- **Absolute Error Size** – The dependent variable in the Multiple Linear Regression. Represents estimation error, irrespective of direction. Higher values = lower accuracy. In time estimation.

- **Positive & Negative Error Size –** Distinguish between errors where participants estimated events as closer (*positive*) or farther (*negative*) than their actual occurrence. Relevant for further analysis.
- **Relevancy** – **Event-specific relevancy** as rated by participants (0–5 scale). Used as a predictor in the linear regression model.
- **Stress Score** – The participant's perceived stress level (0-56 scale).
- **Stress Binary variable (Stress_Bin)** – A binary variable for logistic regression, classifying Stress Score as "high" (> 27) or "low" (< 27) based on a median split.

- Since Stress Score and Stress Bin were duplicated across multiple events per participant, a separate dataset was created for the logistic regression, aggregating participants to prevent inflation of stress scores.

3. See script "TranslationFunction.R" and "TimeEstimation2.R".
4. See script "TimeEstimation2.R" ("forcats" package – as_factor, line 27)

C – Data Analysis

1. Multiple Linear Regression and Logistic Regression analyses were performed (See Script "TimeEstimation3.R".
- The Multiple Linear Regression examined the effect of Study Period, Relevancy, and Ethnicity on Absolute Error Size.
- The Logistic Regression assessed whether Study Period predicted Stress Levels (High vs. Low).

2-4. The **Multiple Linear Regression** analysis revealed significant effects of Study Period, Relevancy, and Ethnicity on Absolute Error Size.

The intercept (140.72) represents the expected absolute error size for Jewish participants before the war, when Relevancy is at its minimum value (0).

The main finding suggests a shift in time estimation before and during the war. However, contrary to our hypothesis, participants demonstrated greater accuracy in estimating event timing during the war, as indicated by a 40.6-day decrease in absolute error size ($p < .001$) compared to the pre-war period (See Figure 5A & 6).

Additionally, higher Relevancy was associated with greater accuracy, as each 1-point increase in Relevancy reduced Absolute Error Size by 6.24 days ($p < .001$; See Figure 5B).

Ethnicity also influenced time estimation, with Arab participants showing larger errors (22.64-day increase, $p < .001$) compared to Jewish participants (See Figure 5C).

However, the model explained only 4.2% of the variance ($R^2 = .042$), suggesting that additional factors influence time estimation accuracy.

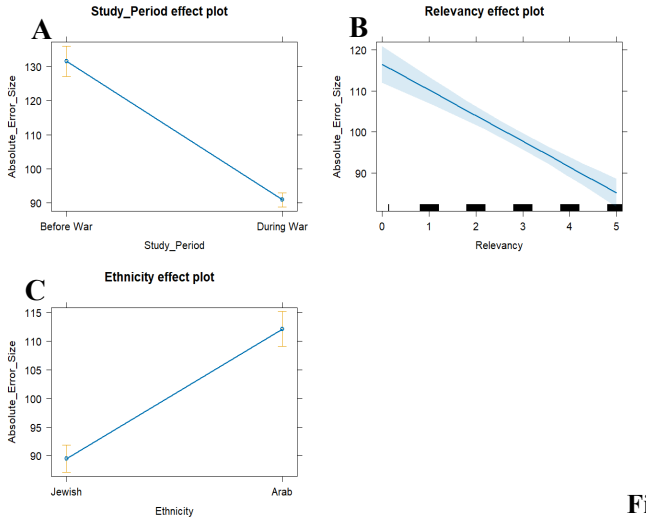

Figure 5.
(A) Effect of Study Period on Absolute Error Size.
(B) Effect of Relevancy on Absolute Error Size.
(C) Effect of Ethnicity on Absolute Error Size.

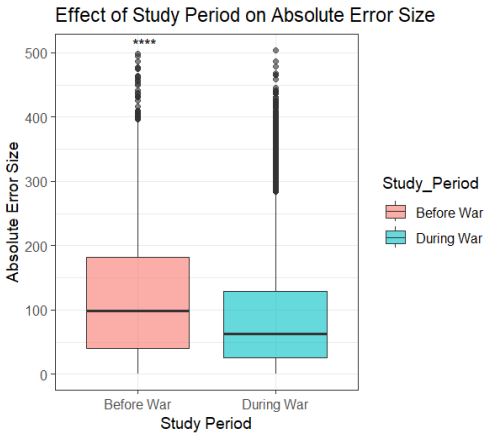

Figure 6. Boxplot of Absolute Error Size (days) before and during the war. Errors were significantly lower during the war ($p < .001$), indicating greater accuracy. The plot highlights the median, variability, and outliers.

The **Logistic Regression** analysis revealed a significant effect of Study Period on the likelihood of high stress levels.
The intercept (log-odds = -1.48, OR = 0.23) represents the odds of high stress before the war. Participants during the war were significantly more likely to experience high stress (log-odds = 1.64, OR = 5.18, $p < .001$), , indicating that the odds of high stress were over five times greater than before the war.
Additionally, predicted probabilities indicate that before the war, the probability of high stress was ~0.18, whereas during the war, it increased to ~0.54 (See Figure 7).
However, the model's predictive ability was weak (AUC = 0.61, see Figure 8), suggesting that Study Period alone does not strongly predict stress levels.

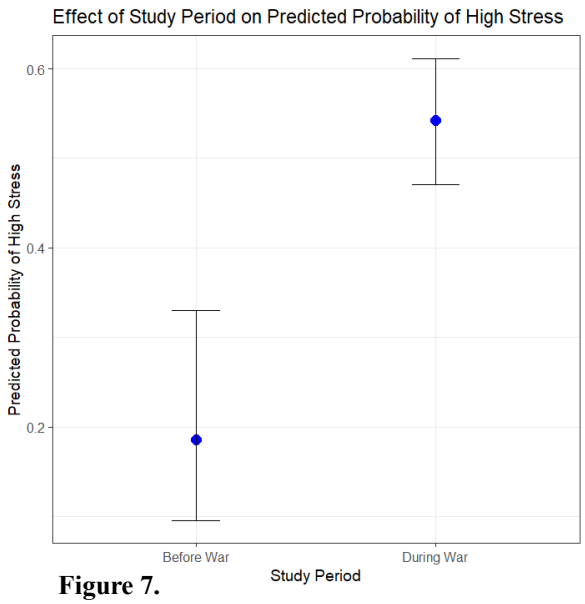

Figure 7.
Predicted probability of high stress before and during the war, showing a significant increase in stress levels during the war.

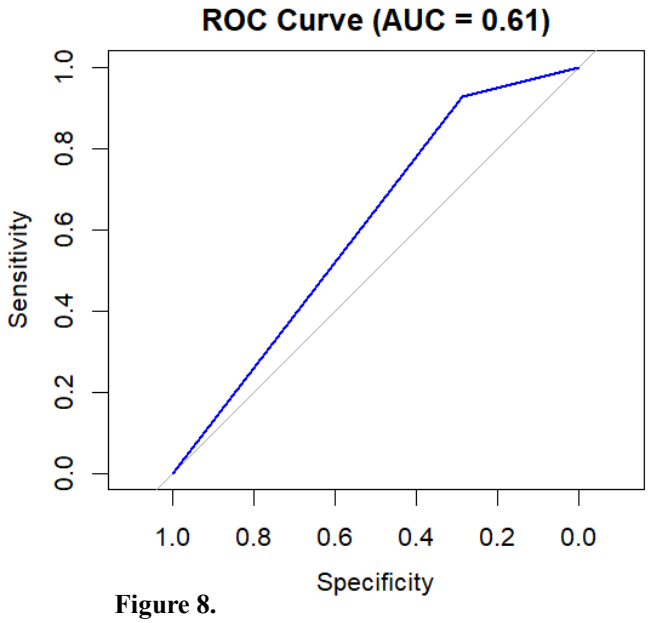

Figure 8.
ROC Curve for the Logistic Regression model assessing high stress classification.