Apply Statistical Analysis

Michael Dang

University of Missouri - Kansas City May 8, 2025

Written Project

Solutions are due on 8 May.

The full code can be found here: https://github.com/micho0802/STAT5551/blob/main/src/Written_project.sas

1 Data

- The dataset will be used for this project is included under this competition on Kaggle: https://www.kaggle.com/competitions/stanford-rna-3d-folding/overview
- This dataset is about a 3-dimensional ribonucleic acid (3D RNA) structure prediction for the Kaggle competition affiliated with Stanford University. Kaggle is a community for machine learning and data science.
- In this project, we will focus only on the train_label.csv dataset, which is a part of the training data together with train_sequences.csv.
- After filtering, the data contains target_id, sequence, x_1, y_1, z_1 features and can be seen in Figure 1 below. Note, we will remove the sequence column to make the dataset more feasible for a two-way ANOVA, as seen in Figure 2 below. [!h]

2 Two-way ANOVA

2.1 Treatment

- This is a two-factor experiment.
- Treatment structure: $3 \pmod{\times 2}$ (RNA squences) = 6 (treatment combinations).



Figure 1: The dataset has been filtered to contain only 2 target_id and the corresponding sequence along with their coordinates. The 1RNK_A is the label ID of this RNA sequence GGCGCAGUGGGCUAGCGCCACUCAAAAGGCCCAU, with a total count of 34, and the 1SCL_A is the label ID of this RNA sequence GGGUGCUCAGUACGAGAGGAACCGCACCC, with a total count of 29.

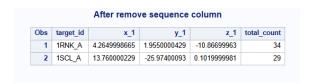


Figure 2: After removing the sequence column

- Treatment combinations are:
 - $(x_1, 1RNK_A)$
 - $(x_1, 1SCL_A)$
 - $(x_2, 1RNK_A)$
 - $(x_2, 1SCL_A)$
 - $-(x_3, 1RNK_A)$
 - $(x_3, 1SCL_A)$
- We assume that all the RNA is under a similar environment, and the (coord, RNA sequences) are assumed to be completely random.

2.2 Model

The statistical model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \cdot \beta)_{ij} + \epsilon_{ijk}$$

- Where:

 y_{ijk} = the structure of an RNA under k^{th} position of the i^{th} coordinates and j^{th} RNA sequence.

 $\mu =$ The overall mean of the RNA structure.

 α_i = The effect of the i^{th} coordinates on the response variable.

 β_j = The effect of the j^{th} RNA sequence on the response variable.

 $(\alpha \cdot \beta)_{ij}$ = The interaction effect of the $i^{\rm th}$ coordinates and $j^{\rm th}$ RNA sequence on the response variable.

 $\epsilon_{ijk} = \text{Random experimental error of the } k^{th} \text{ position under}$ the i^{th} coordinate and the i^{th} RNA sequence.

- Where:

$$i = 1, 2, 3$$

 $j = 1, 2$
 $k = 1, 2, ..., n_{ij}$

2.3 Assumptions

• Model assumption:

$$\epsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

- Constant variance assumption.
- Normality assumption.

2.4 Checking for assumption

• Constant variance assumption:

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_6^2 = \sigma^2$$
 vs $H_A: \text{Not all } \sigma_i^2$ are equal

- Levene's test:

$$< 0.0001 \text{ (P-value)} < 0.05 (\alpha)$$

- Hence, fail to reject H_0 . Therefore, the constant variance assumption is violated.
- Thus, instead of the homogenous error variance model, we could fit a heterogenous error variance model, using Satterwaite's approximation for the degrees of freedom.
- Check the model fit statistic; the smaller the AIC, the better the model fit.

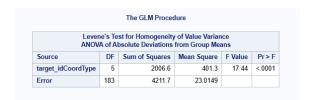


Figure 3: Levene's test

- Homogeneous error variance model (Figure 4): 1366.5
- Heterogeneous error variance model (Figure 5): 1315.3

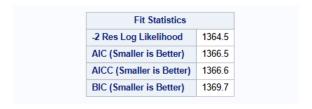


Figure 4: Homogeneous error variance model

- Hence, we go for the heterogeneous error variance model.
- Normality assumption:

 H_0 : Residual are from a Normal population H_A : Residual are not from a Normal population

- K-S test in Figure 6:

$$> 0.15 \text{ (p-value)} > 0.05 \text{ } (\alpha)$$

- Hence, the normality assumption is satisfied.

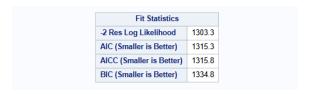


Figure 5: Heterogeneous error variance model

Goodness-of-Fit Tests for Normal Distribution							
Test	Statistic		p Value				
Kolmogorov-Smirnov	D	0.03048685	Pr > D	>0.150			
Cramer-von Mises	W-Sq	0.02494105	Pr > W-Sq	>0.250			
Anderson-Darling	A-Sq	0.21332971	Pr > A-Sq	>0.250			

Figure 6: K-S test

2.5 ANOVA

• Testing whether target_id affects values:

$$H_0: \alpha_i = 0$$
 vs $H_A: \alpha_i \neq 0$

- From Figure 7:

$$< 0.0001 \text{ (p-value)} < 0.05(\alpha)$$

- Hence reject H_0 .
- Therefore, with 95% confidence, we can conclude that the target_id affects values.
- Testing whether CoordType affects values:

$$H_0: \beta_j = 0$$
 vs $H_A: \beta_j \neq 0$

- From Figure 7:

$$< 0.0001 \text{ (p-value)} < 0.05 \text{ } (\alpha)$$

- Hence reject H_0 .
- Therefore, with 95% confidence, we can conclude that the CoordType affects values
- Testing whether there is an interaction effect between target_id and CoordType:

$$H_0: (\alpha \cdot \beta)_{ij} = 0$$
 vs $H_A: (\alpha \cdot \beta)_{ij} \neq 0$

- From Figure 7:

0.0001 (p-value)
$$< 0.05$$
 (α)

- Hence, reject H_0 .
- Therefore, with 95% confidence, we can conclude that the interaction between target_id and CoordType affects values.

Type 3 Tests of Fixed Effects						
Effect	Num DF	Den DF	F Value	Pr > F		
target_id	1	183	28.07	<.0001		
CoordType	2	183	10.42	<.0001		
target_id*CoordType	2	183	9.32	0.0001		

Figure 7: Fixed Effects