Homework 1

Denis Ostroushko

2023-02-05

1 Problem 1. (5 points)

Provide preliminary visualizations of the data. Provide initial assessments about what can be seen from these visualizations in relation to an Analysis of Variance analysis of the data.

1.1 Problem 1 Solution.

We begin this analysis by looking at the distribution of our response variable Y, which is the Fasting glucose levels in millimoles per liter. Measurements are recorded on the logarithmic scale.

First, we will look at the distribution of Y. Figure 1 suggests that the sample distribution of Y is approximately normal. Of course we make no assumption about the distribution of observed values, but it is good to know if there are any obvious outliers that we might want to handle right away.

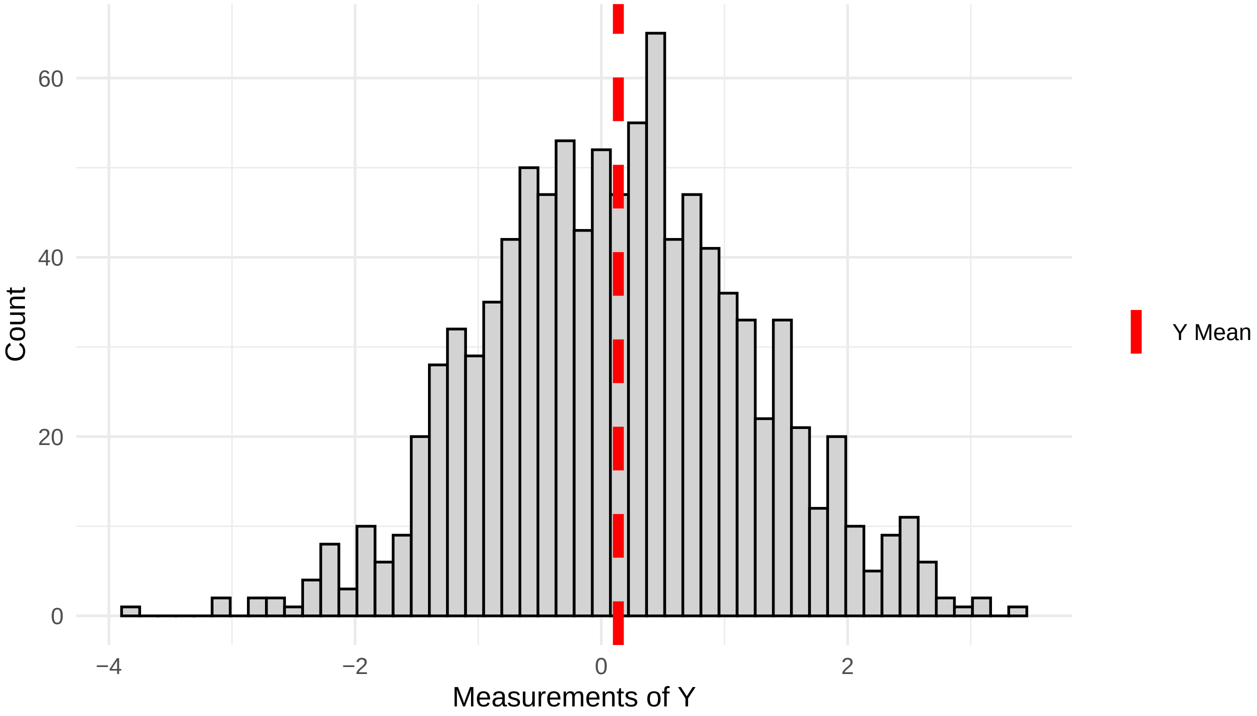

Figure 1: Distribution of Y for 1,000 observations

As we can see on Figure 2, sample average for $G_1 = 2$ is slightly higher than the other two groups, which have identical means.

Group 2 also has the lowest number of observations.

We can see that the mean and median for each group is approximately equal.

Finally, we can see that the variance in each group is approximately similar, especially for groups 0 and 1. Groups 2 apperas slightly different, however, without proper accounting for the sample size difference, it is hard to make a decisive statement. We will learn more when performing format statistical tests.

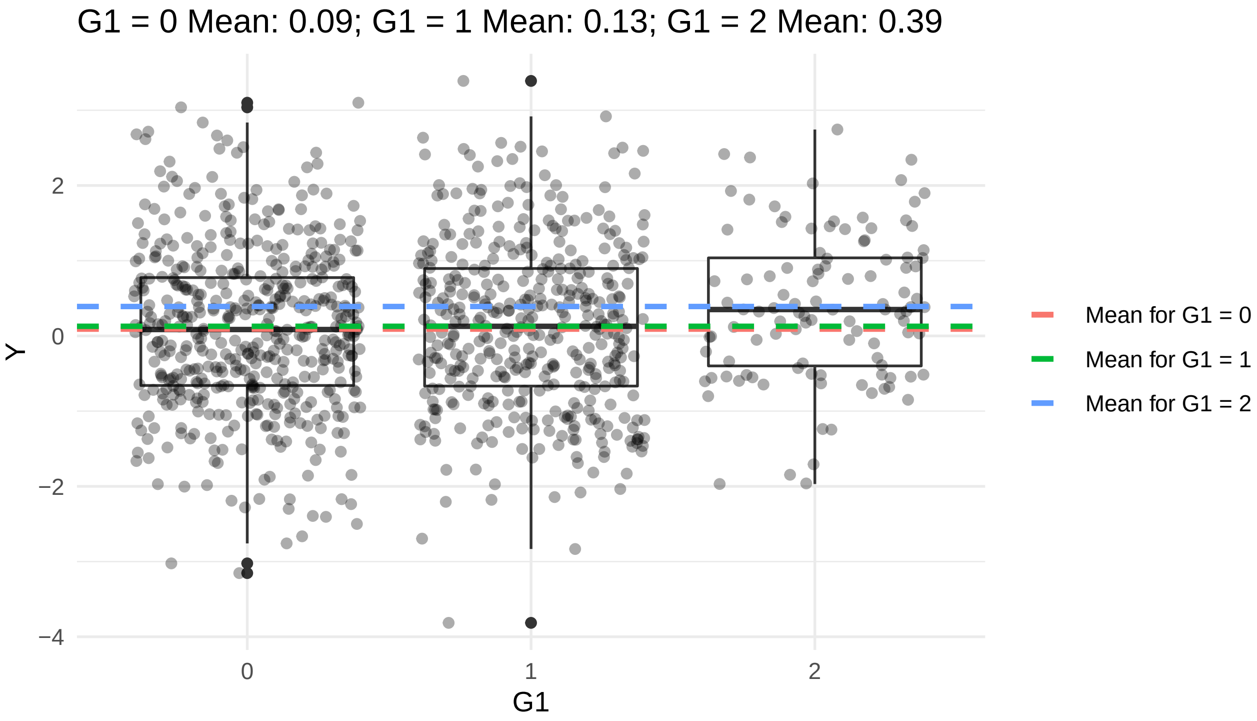

Figure 2: Distribution of Glucose Measurements over G1 Levels

Same conclusions we made about G_1 levels carry on to the G_2 levels. We can see the distributions in Figure 3. It is even harder to make preliminary statements about the distribution of Group 2 due to its sample size.

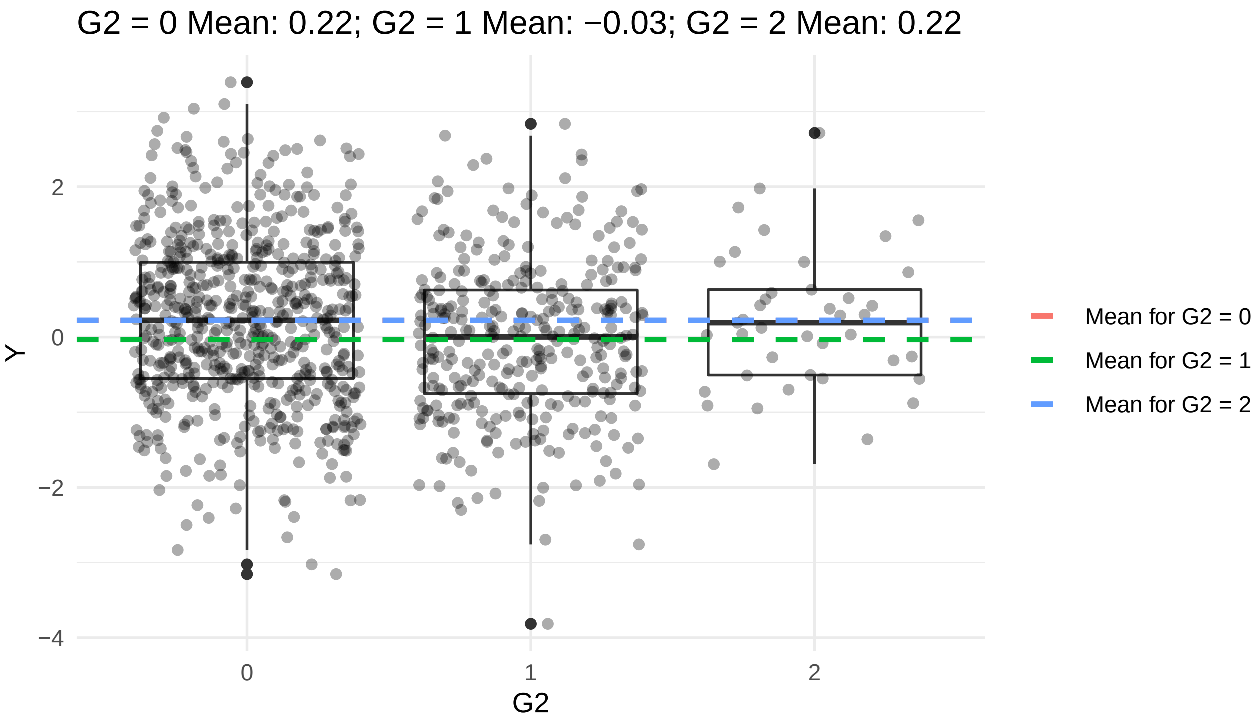

Figure 3: Distribution of Glucose Measurements over G2 Levels

It is favorable to us that while the variances for groups with smaller sample size appear unequal when compared with the other groups, we can see that the observations in group 2 for both G_1 and G_2 are more centered around its mean and median. Bigger spread of data with smaller sample size would be highly unfavorable to our analysis.

Finally we can combine G_1 and G_2 levels and create nine groups. We can see their distributions in Figure 4. I present box plots in order of their respective sample means. Group with the lowest mean is located on the left, while a group with the highest mean is located on the left part of the plot. This plot reveals that we indeed have a very unbalanced design here. Variance of each groups with comparable sample sizes appear similar visually. None of the box plots show evidence of outliers that can skew sample means.

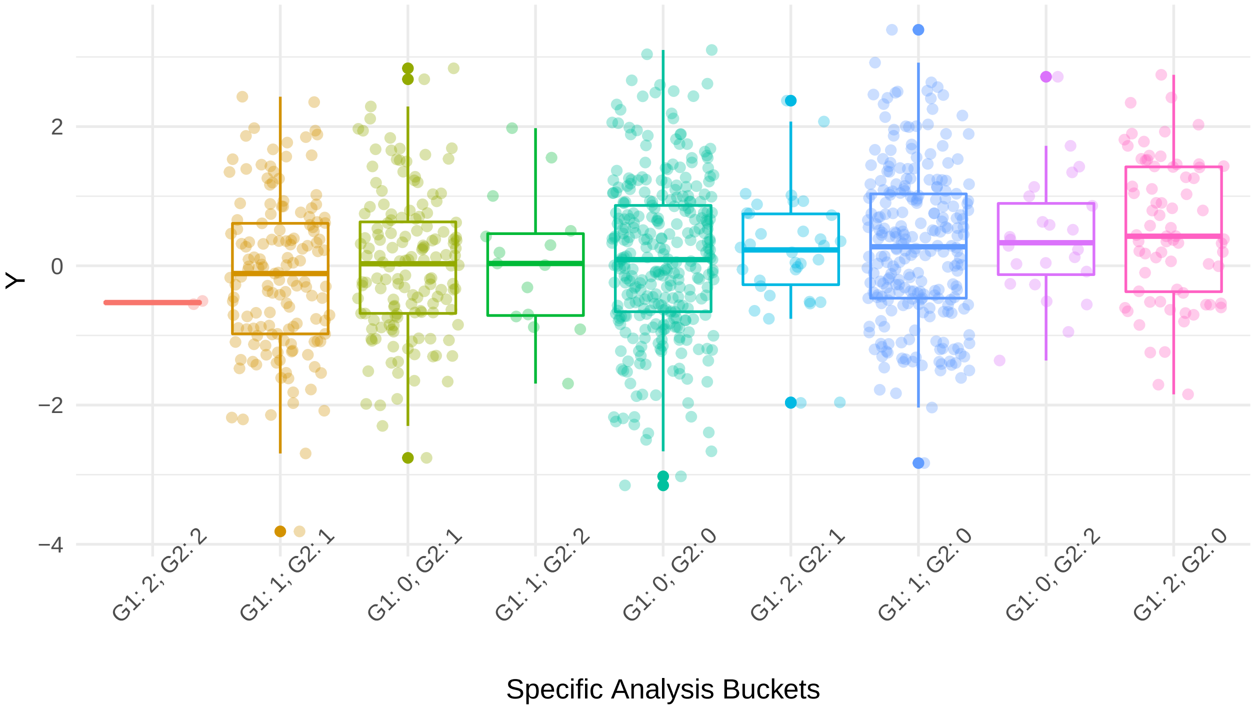

Figure 4: Distribution of Y in each specific bucket

We can summarize these data in Table 1. We present sample average and 95% confidence interval using standard errors obtained from the data. Confidence intervals are not adjusted for multiple comparisons. Estimation of confidence intervals without accounting for multiple comparison adjustments produces narrower intervals. Even with the narrow intervals, we can see that the sample variation makes there bands wide, therefore, we should not highly significant results or high values of estimated effects of G_1 and G_2 levels on measurements of Y. However, with a large sample size that we have here, it might be possible to detect significant effects.

Table 1: Sample Means with 95 Percent Confidence Intervals

G1	G2	N	Mean	Lower CL	Upper CL
0	0	310	0.10	-0.02	0.22
1	0	148	0.29	0.15	0.42
2	0	24	0.51	0.26	0.76
0	1	250	0.02	-0.16	0.19
1	1	145	-0.13	-0.31	0.04
2	1	15	0.19	-0.17	0.55
0	2	72	0.39	-0.04	0.83
1	2	34	0.05	-0.50	0.60
2	2	2	-0.53	-2.03	0.97

Before we finish our preliminary assessment of the data, we need to look at the treatment plots. Each dot represents a sample mean for a given group. We have a total of nine sample means for nine groups that are subject to analysis. Figure 5 shows that we potentially have an interaction of levels of G_1 and G_2. We can see that as G_1 levels go from 0 to 1 to 2, the average response increases for G_2 levels 0 and 1. However, when we consider the effects of G_1 when we restrict the sample for only those observations that are in G_2 level 2, then as we go from G_1 group 0 to 1 to 2, then the average response decreases. The number of observations that fall into this group is quite small, but the direction of interaction is quite strong and different. If we detect significance of interaction, it will be due to this phenomenon.

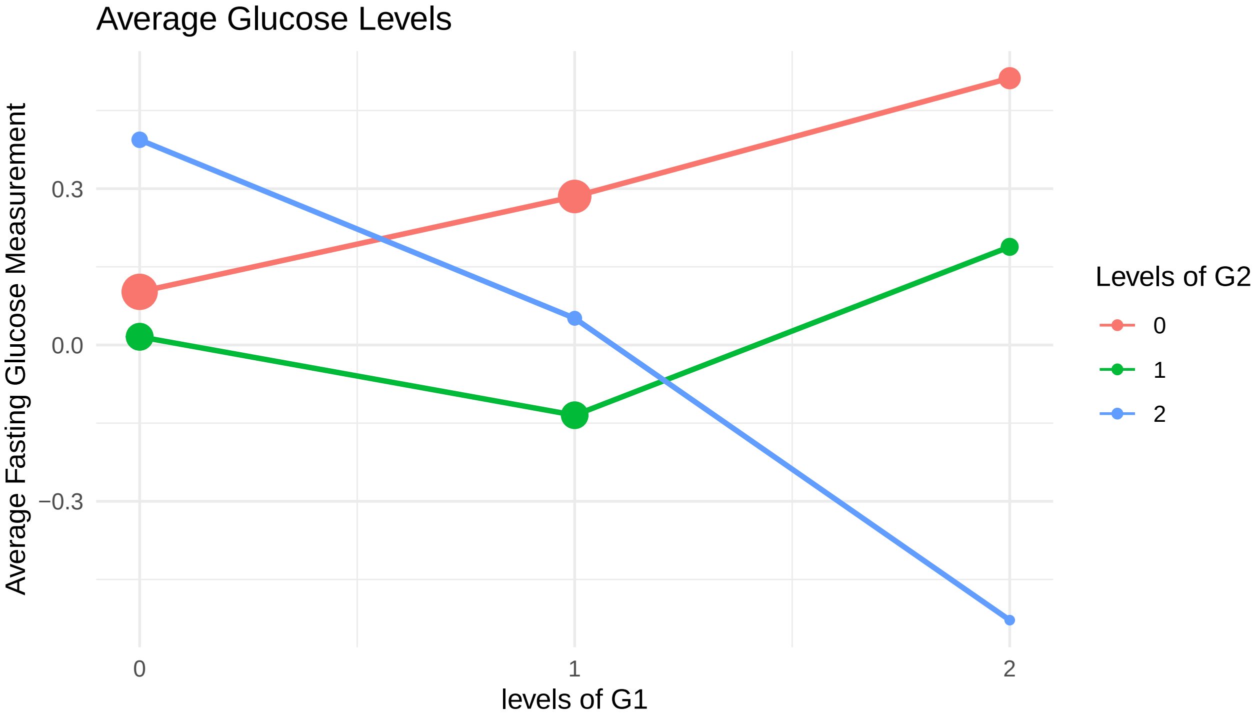

Figure 5: Treatment plot for levels of G1 and G2

2 Problem 2. (8 points)

Write down a two-factor ANOVA model using G_1 and G_2 and their interaction to explain the variation in the fasting glucose level. Define any notation you use and list all assumptions that are made by this model.

2.1 Problem 2 Solution.

These data contains 3 unique levels for G_1, and 3 unique levels for G_2. Thus, we have a total of 9 cell means.

Each cell mean will be represented in terms of the factor levels:

$$\mu_{ij} := G_{1i} + G_{2j} + (G_1G_2)_{ij}$$

Model terms are defined below:

- G_{1i} is the i^{th} level of variable G_1. Index i takes on values 0, 1, 2. This terms represents the main effect of G_1 level on the fasting glucose level.
- G_{2j} is the j^{th} level of variable G_2. Index j takes on values 1, 0, 2. This terms represents the main effect of G_2 level on the fasting glucose level.
- $(G_1G_2)_{ij}$ is the interaction term between the i^{th} level of G_1 and j^{th} level of G_2, and helps us tell if the impact of a given G_1 level on the average fasting glucose level is different across different levels of G_2 levels. Same implication can be stated for G_2 levels across varying G_1 levels.
- μ_{ij} is then the average value of the fasting glucose levels for a given combination of levels.

Once we define what factors μ_{ij} depends on, we can use that definition to write an expression for each individual observation and state model assumptions. We consider observations Y_{ijk}, that is a k^{th} observation in the i^{th} level of G_1 and j^{th} level of G_2. We assume that Y_{ijk} comes from a distribution with mean μ_{ij}, and thus, we can write Y_{ijk} as:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

where ϵ_{ijk} is a random error that we can not control for with the defined model.

Assumptions for the two factor ANOVA model are stated below:

1. All observations Y_{ijk} are independent.
2. Observations Y_{ijk} are normally distributed and independent. Y_{ijk} are independent and normally distributed within levels i and j. Thus, $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$.
3. We assume that all distributions have equal and finite variance. So, the distributions only differ in their means, a center of the distribution, and not in variances.

3 Problem 3. (4 points)

Provide the ANOVA table associated with the model defined in the previous question and explain what the different sums of squares are.

3.1 Problem 3 Solution.

We fit a linear model using `R` and present the output below. We specify the equation as `Y ~ G1 + G2 + G1:G2`, so we will interpret sequential ANOVA table by looking at the sum of squares attributed to G_1, then we will look at the sum of squares attributed to G_2 after accounting for the main effect of G_1, and finally we will look at the interaction term sum of squares.

Table 2: ANOVA Table for a two-factor cell means model

	DF	Sum of Squares	Mean Sum of Squared	F-value	P-value
G1	2	8.06	4.03	3.46	0.03
G2	2	14.30	7.15	6.14	0.00
G1:G2	4	9.19	2.30	1.97	0.10
Residuals	991	1154.04	1.16	NA	NA

- ANOVA table does not show the total sum of squares, i.e. the sum of squared differences between each individual observation and the average measurement for Y, which we refer to as $\bar{Y}$.

The total variation is the arithmetic sum of Sum of Squares from each row of the ANOVA table. In this problem the total sum of squares is 1185.6

- Sum of squares associated with G_1 is 8.06. This sum of squares shows how much total of variation can be attributed to the effect of G_1.

  While this seems like a relatively small proportion of variation, the sample size and the effect size make it such that the effects of varying levels of G_1 on the average response level are statistically significant, as indicated by the p-value.

- The sum of squares associated with G_2 is 14.3. This sum of squares shows how much additional variation can be attributed to the effect of G_2 levels, after accounting for the main effects of G_1.

Since this is a sequential sum of squares anova table, if we ask `R` to run a model where G_2 is specified before G_1, this sum fo squares quantity would be different.

In fact, if we reorder term specifications such that the model statement is `Y ~ G2 + G1 + G1:G2`, we get 14.23 as the sum of squares associated with G_2.

This number is not greatly different from the original value we stated. So, the proportion of variation in Y that is due to the effects of G_2 must not be greatly overlapping with the proportion of variation that G_1 explains.

- The interaction sum of squares measures the variability of the estimated interactions for all combination of treatments, after adjusting for the main effects of G_1 adn G_2. Since the mean of all estimated interactions is zero, the deviations of the estimated interactions around their mean is not explicitly shown. This sum of squares show how strong, or big the interactions between G_1 and G_2 levels are. As we saw previously on Figure 5, the main interaction of G_1 levels occurs with the G_2 level 2.

However, those groups have an extremely small sample size, which reduces the power of statistical tests involved in the detection of these effects.

We will show more tests and formulate these ideas more rigorously in the later sections of this assignments.

4 Problem 4. (8 points)

Provide visualizations that investigate the assumption of equal variances and the assumption of normality of the errors. Under the model assumed in question 2, conduct the Levene test to assess the equal variances assumption. Make sure to explicitly write the null and alternative hypotheses in clear statistical notation.

4.1 Problem 4 Solution.

We begin this assignment with the investigation of normality of errors. We will look at the studentized residuals to focus on the shape and quantiles of residuals' distribution.

We will look at the

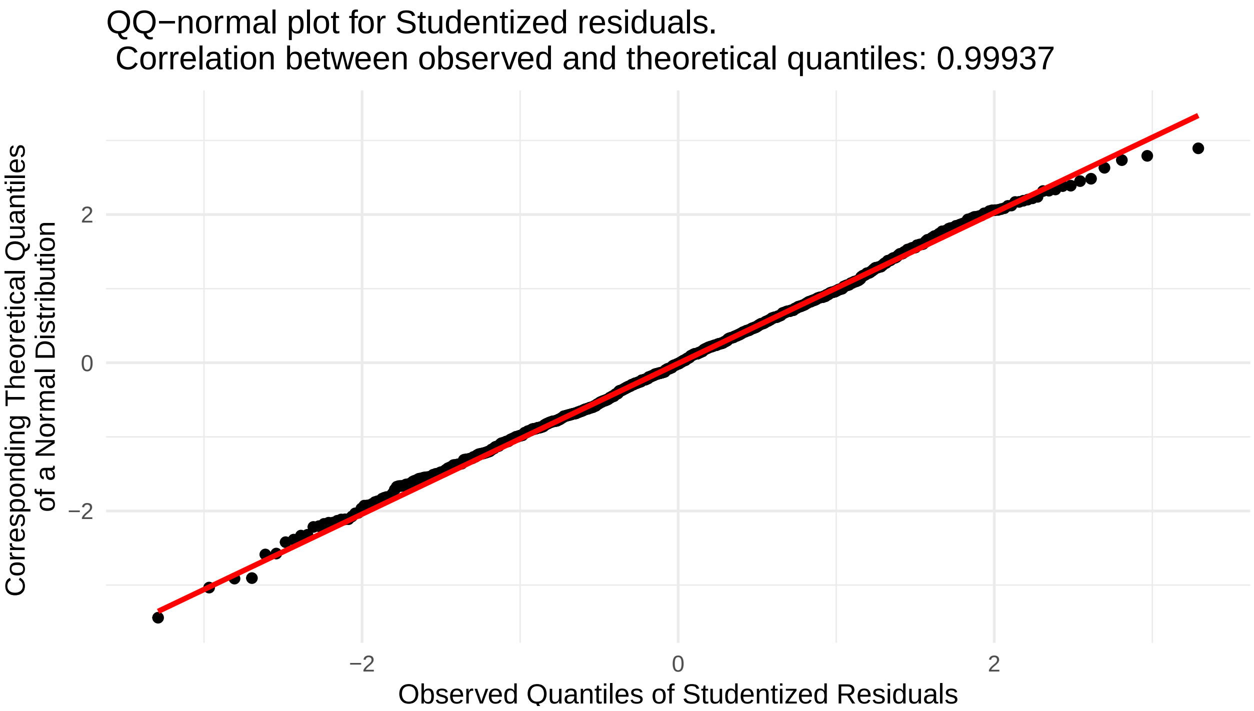

Figure 6: No evidence of deviation of residuals from normality. Approximately normal distribution of the sample measurements and absence of outliers contribute to approximately normal distribution of residuals.

Figure 6 suggests that the distribution of residuals for the entire sample is strongly approximately normal. There are no extremely strong outliers and heavy tails, so we will not investigate each subgroup and each combinations of levels of G_1 and G_2. We can continue validation of the model and assess variance of residuals.

We continue assessment of residuals by looking at the variance of studentized residuals in each treatment combination group. This figure is conceptually similar to Figure 4, however, Figure 7 orders groups by their sample size instead of means.

We now have a group with the smallest sample size on the left, and the greatest sample size on the right. This arrangement allows us to compare variance of residuals in groups with similar sample sizes, and shows how variance of residuals is affected by a changing sample size.

Based on visual evidence, we can anticipate that Levine test will suggest that the variances are not equal. However, I do not anticipate that there will be a statistically significant result of the test. In addition, there are no visible outliers that potentially can affect Levine test that relies on the absolute deviations of each residual with the mean-residual in each group.

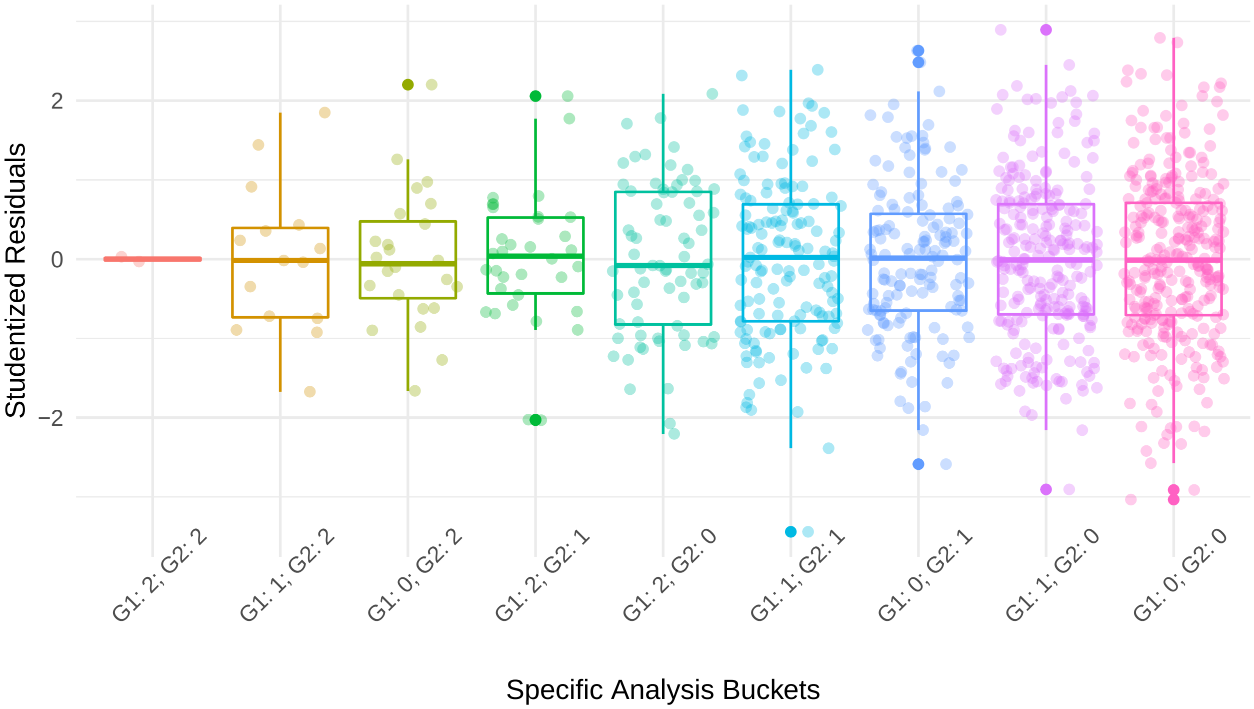

Figure 7: Groups are ordered by sample size. As we go from left to right, the sample increases

Levene Test

We consider absolute deviations $Z_{ij} = |Y_{ij} - \bar{\mu}_{ij}|$, and we perform a two-factor ANOVA with no interaction term using Z_{ij} as a response variable. Test hypotheses and results are given below:

- $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_9^2$
- H_a : at least one σ_i^2 is not equal to the rest of σ_j^2's
- Tests Results are given below in the table:

Table 3: Resuts of Levine test for variance equality between groups

	Df	F value	Pr(>F)
group	8	1.64	0.11
	991	NA	NA

- F-statistic: 1.6445 under 8 and 991 degrees of freedom
- Cutoff F^*-statistic: 1.9477 under 8 and 991 degrees of freedom
- P($F > F^*$) = 0.1081
- Conclusion: p-value is grater than the accepted significance level $\alpha = 0.05$, so we fail to reject the null hypothesis. There is not enough evidence to conclude that there is significant difference between variance of residuals between the nine groups.

  Therefore, results of further F-tests should be reliable.

5 Problem 5. (5 points)

Conduct a statistical test to determine whether the interaction of G_1 and G_2 has any impact on the mean response. What is the conclusion and how does that impact how you would model resting glucose level?

5.1 Problem 5 Solution.

Test hypotheses and results are given below:

- $H_0 : (\alpha\beta)_{ij} = 0$ for all i, j
- H_a : at least one term $(\alpha\beta)_{ij}$ is not zero
- We refer to Table 2 for test statistics. Only two rows that are relevant to the test are given below:

Table 4: Test statistics for interaction term testing. ANOVA Table for a two-factor cell means model

	DF	Sum of Squares	Mean Sum of Squared	F-value	P-value
G1:G2	4	9.19	2.30	1.97	0.1
Residuals	991	1154.04	1.16	NA	NA

- F-statistic: 1.9726 under 4 and 991 degrees of freedom
- Cutoff F^*-statistic: 2.3809 under 4 and 991 degrees of freedom
- $\text{P}(F > F^*) = 0.0966$
- Conclusion: p-value is grater than the accepted significance level $\alpha = 0.05$, so we fail to reject the null hypothesis. There is not enough evidence to conclude that all interactions terms are statistically different from zero.

  This p-value is quite close to 0.05, so perhaps is it suggestive that there are some statistically significant interactions, but most should be in fact non-significant.

  As we saw on Figure 5, there are visually interacting levels. Perhaps, as pairwise comparisons can reveal which levels are the most different.

6 Problem 6. (8 points)

Use the Tukey and Bonferroni tests to compare all pairwise comparisons of the 9 cell means. Describe and explain any differences in findings between the two approaches. What do the conclusions about the pairwise comparisons tell you about the relationship between G_1, G_2, and Y ?

6.1 Problem 6 Solution.

Tukey

We begin by looking at pairwise comparisons using Tukey Procedure for multiple comparisons adjustments. Generally, a total number of comparisons is given by $N_{groups} \times (N_{groups} - 1)\ /\ 2$. Therefore, we have a total of 0. Due to an extremely large number of comparisons, we will print only those comparisons that are significant at the Family wise significance level 0.05.

Table 5 shows all statistically significant pairwise comparisons:

Figure 8 show statistically significant comparisons, as well as all other comparisons. We can see that there are a lot of comparisons where p-value is extremely close to 1.

Table 5: Significant Tukey-adjusted pairwise comparisons

Contrast	Est. difference	Estimate standard error	P-Value
G1 Level 1 G2 Level 0 - G1 Level 1 G2 Level 1	0.420	0.113	0.006
G1 Level 2 G2 Level 0 - G1 Level 0 G2 Level 1	0.496	0.155	0.038
G1 Level 2 G2 Level 0 - G1 Level 1 G2 Level 1	0.647	0.156	0.001

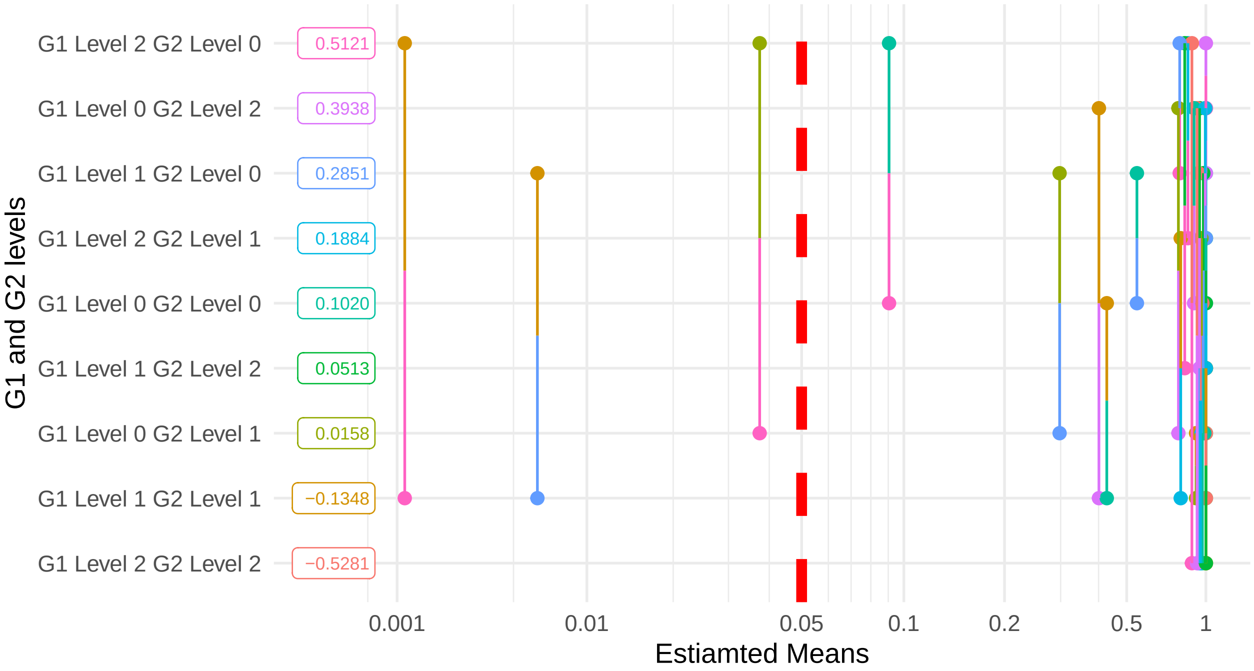

Figure 8: Red line is significance cutoff at 0.05 level

Table 5 and Figure 8 tell us following information:

- Patients in with G_1 level 2 and G_2 level 0 has the largest average response when compared with other groups. In particular, Tukey-adjusted pairwise comparisons detected three statistically significant differences.
- Members with G_1 level 2 and G_2 level 0 on average had 0.6469 more log- millimoles per liter fasting glucose levels, bounded by a (0.1632, 1.1305) 95% confidence interval, when compared with with G_1 level 1 and G_2 level 1 group. Due to the use of Tukey adjustment procedure this interval is quite wise. Figure 9 highlights two means that are subject to this comparison.

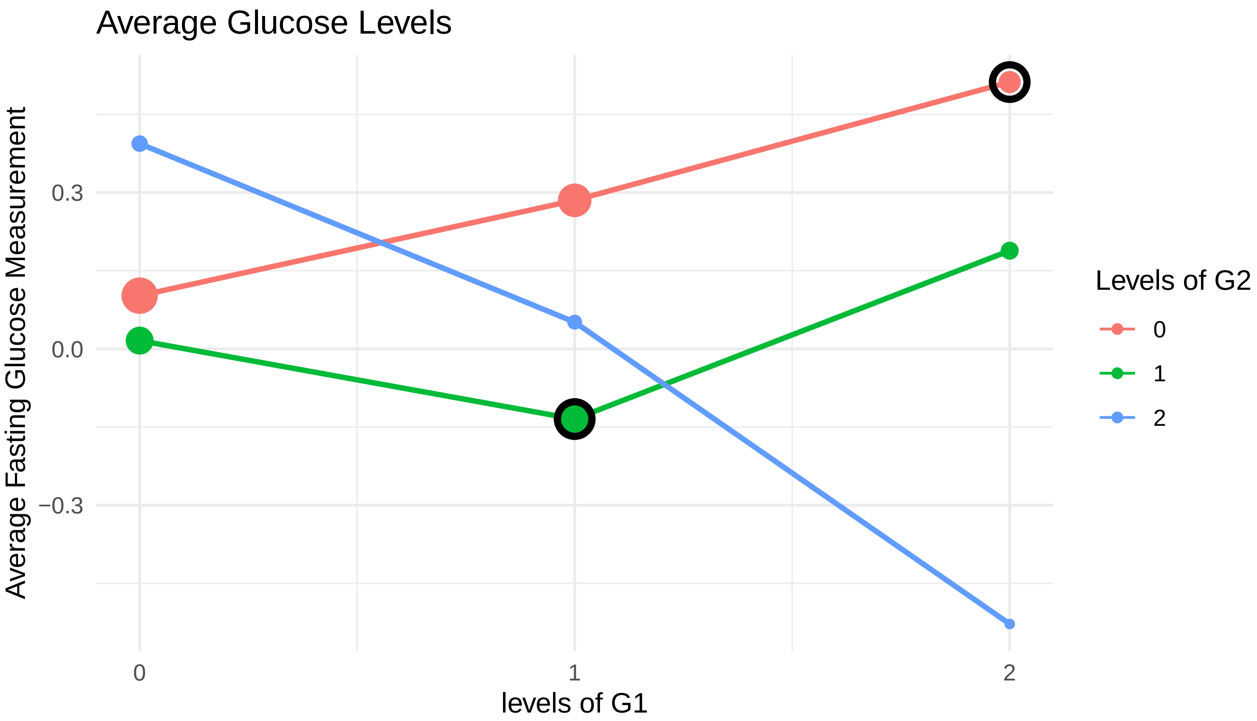

Figure 9: Caption

- Bla bla

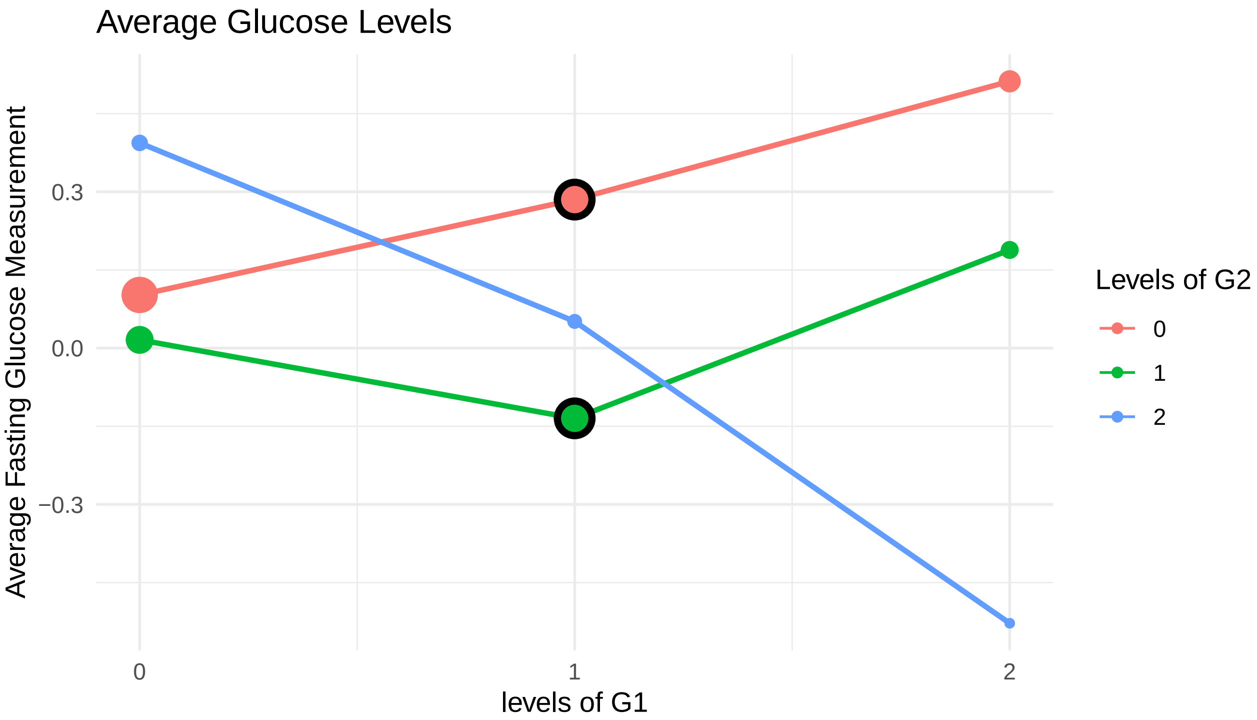

Figure 10: Caption

- Bla bla

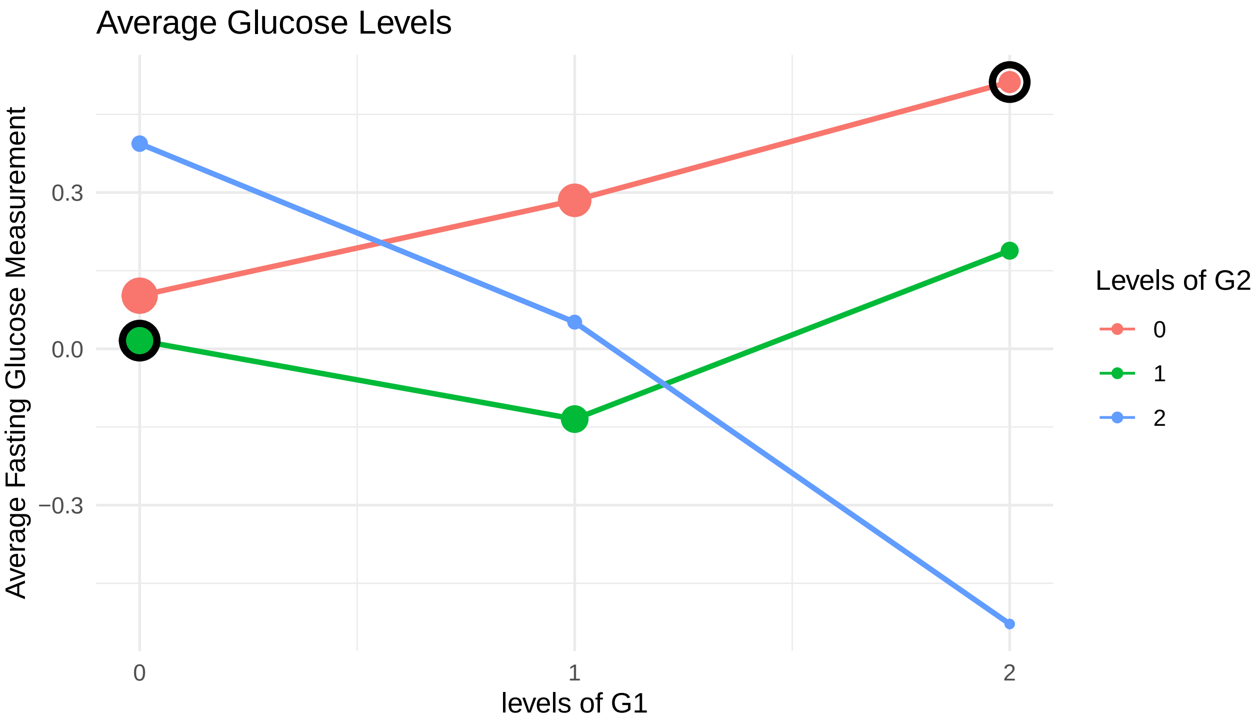

Figure 11: Caption

make a comment about difference of G1 level 2 - level 0 given that we are in G2 kevek 0

Bonferroni

Table 6 shows all statistically significant pairwise comparisons:

Table 6: Significant Bonferroni-adjusted pairwise comparisons

Contrast	Est. difference	Estimate standard error	P-Value
G1 Level 1 G2 Level 0 - G1 Level 1 G2 Level 1	0.420	0.113	0.007
G1 Level 2 G2 Level 0 - G1 Level 0 G2 Level 1	0.496	0.155	0.051
G1 Level 2 G2 Level 0 - G1 Level 1 G2 Level 1	0.647	0.156	0.001

Figure 12 show statistically significant comparisons, as well as all other comparisons. We can see that there are a lot of comparisons where p-value is extremely close to 1.

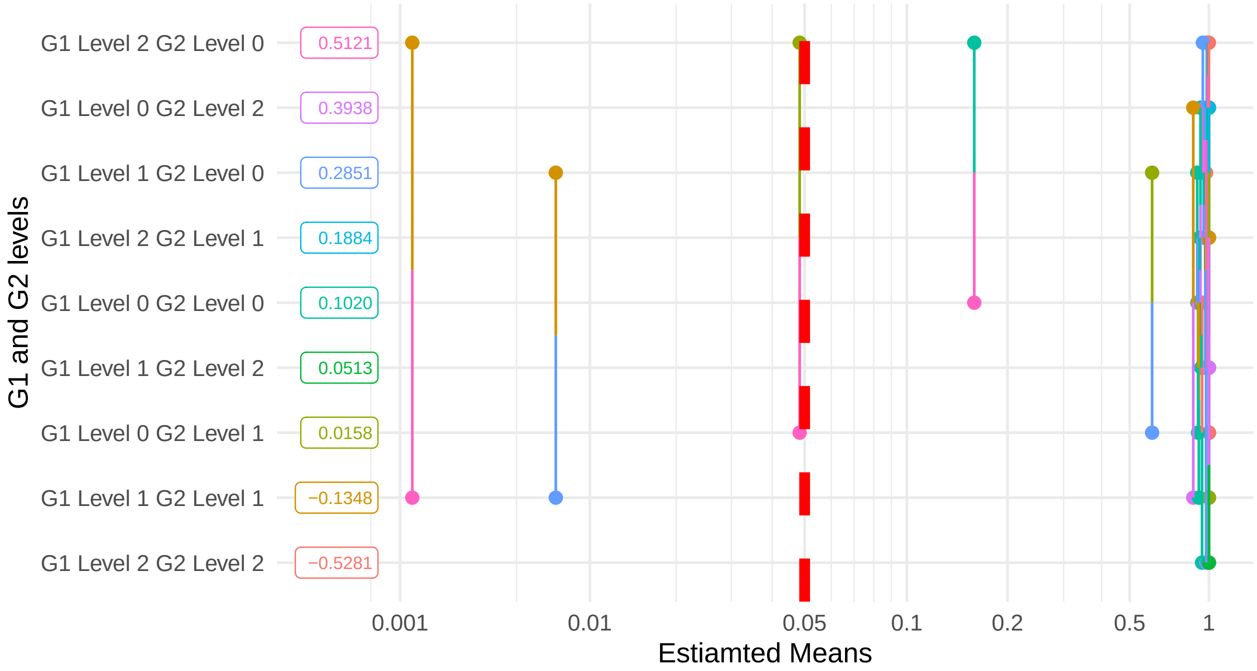

Figure 12: Red line is significance cutoff at 0.05 level