Simon Lee

Problem set 3

Question 1:

Effect coding and dummy coding are two commonly used techniques in statistical modeling, particularly in the context of categorical variables. Effect coding (-1, 1) involves representing categorical variables through contrast coding, where each level of the variable is compared to a reference level. This method provides insights into the differences between each level and the reference level. On the other hand, dummy coding (0,1) creates binary variables for each level of the categorical variable, except for one reference level. This technique facilitates the comparison of each level with the reference level separately, making it particularly useful for examining specific group differences. Both effect coding and dummy coding play crucial roles in statistical analyses, offering distinct perspectives on categorical data and enabling researchers to draw meaningful conclusions from their models.

We are therefore asking to compute the means, as well as X'X and the inverse of X'X for both coding schemes. We present the following below

Effect Coding:

1	-1	-1	1
1	1	-1	-1
1	-1	1	-1
1	1	1	1

Means for effect coding:

1	0	0	0

X'X

4	0	0	0
0	4	0	0
0	0	4	0
0	0	0	4

Inverse of X’X:

0.25	0	0	0
0	0.25	0	0
0	0	0.25	0
0	0	0	0.25

Dummy coding:

1	0	0	0
1	1	0	0
1	0	1	0
1	1	1	1

Means for dummy coding:

1	0.5	0.5	0.25

X’X

4	2	2	1
2	2	1	1
2	1	2	1
1	1	1	1

Inverse of X’X

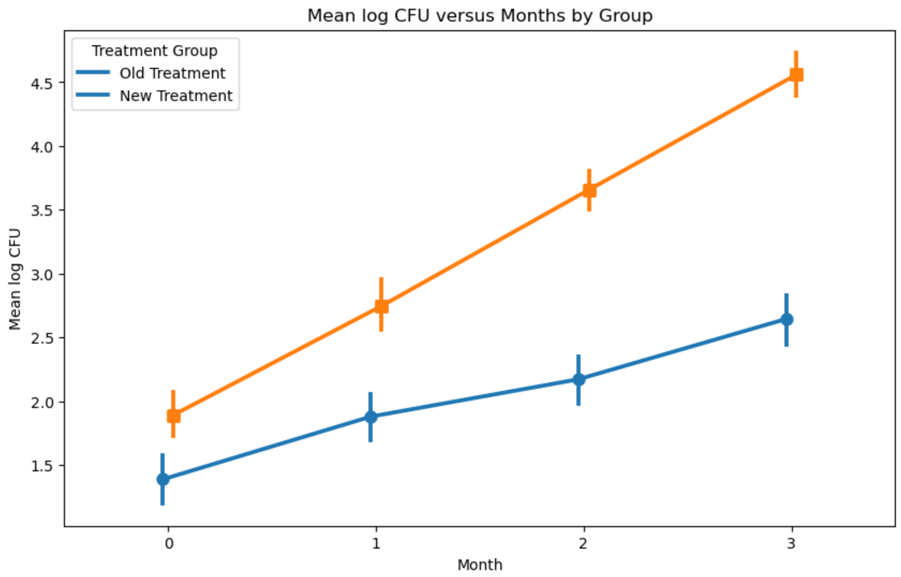
1	-1	-1	1
-1	2	1	-2
-1	1	2	-2
1	-2	-2	4

To determine whether X1, X2, and X3 are orthogonal, which means having a correlation of zero, let's consider both effect coding and dummy coding as seen in the examples above. In the case of effect coding, the correlation matrix derived from the X'X matrix is seen above. Since the off-diagonal elements are all zeros, it indicates that X1, X2, and X3 are orthogonal, suggesting that they do not correlate with one another. However, when considering dummy coding, the X’X matrix is far different. In this scenario, the presence of non-zero off-diagonal elements signifies that X1, X2, and X3 are not orthogonal, indicating that these variables have correlations among them. Thus, in summary, X1, X2, and X3 are orthogonal when using effect coding but not when using dummy coding.

In terms of calculating the betas from the table we get when using effect coding, if the means in the four groups are 0.0 for the group with no alcohol and no hypertension, 1.0 for no alcohol and hypertension, 2.0 for alcohol and no hypertension, and 3.0 for alcohol and hypertension, we can compute the regression coefficients (betas) for the relation between the means and the coded variables X. The betas computed using effect coding are [1.5, 1.0, 0.5, 0.0]. To compute the betas using dummy coding, we first need to create the dummy coded X matrix, which represents the three variables alcohol, hypertension, and their interaction. The betas computed using dummy coding are [0.0, 2.0, 1.0, 0.0]. We can see that the two sets of betas are not the same. This is because effect coding and dummy coding result in different codings of the categorical variables, leading to different interpretations of the regression coefficients.

Question 2:

I conducted a two-way ANOVA to investigate the potential significant differences in the log-transformed colony-forming units (logCFU) across different months and treatment groups. The ordinary least squares (OLS) regression model from statsmodels.formula.api was employed, with the formula 'logCFU ~ C(month) * C(tx)' specified. In this formula, the dependent variable logCFU was modeled as a function of the interaction between the categorical variables month and treatment group (tx). The C() function was used to treat month and tx as categorical variables. The model was fitted to the data (df), and the anova_lm function from statsmodels.stats.anova was applied to the fitted model to obtain the ANOVA results. We display the resulting plot and table below:



	df	Sum_sq	Mean_sq	F	PR(>F)
C(months)	3	42.150	14.051	149.431	7.633e-31
C(tx)	1	28.436	28.436	302.409	1.739 e-27
C(Months):C(tx)	3	5.950	1.983	21.091	6.562e-10
Residual	72	6.770	0.094	NaN	NaN

The two-way ANOVA results indicate significant main effects of month and treatment group (tx), as well as a significant interaction effect between month and tx on the log-transformed colony-forming units (logCFU).

For the main effect of month, the ANOVA results show a degrees of freedom (df) of 3.0, a sum of squares (sum_sq) of 42.154077, a mean square (mean_sq) of 14.051359, an F-statistic of 149.431125, and a p-value of 7.633374e-31. The extremely small p-value (< 0.05) suggests that there are significant differences in logCFU across the different months.

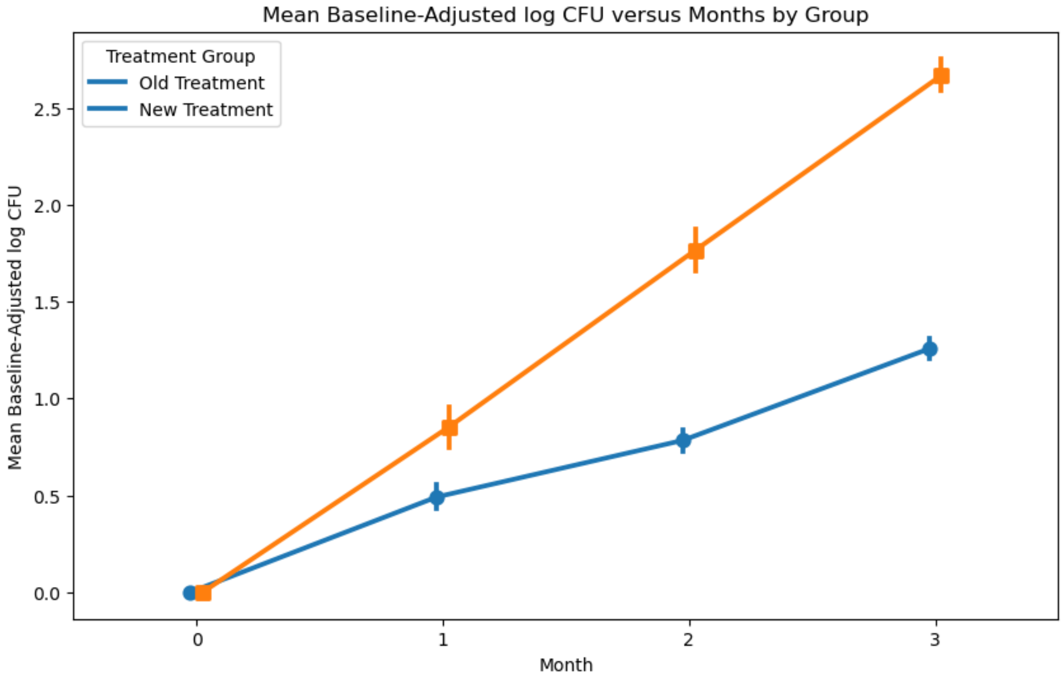
Regarding the main effect of treatment group (tx), the results show a df of 1.0, a sum_sq of 28.436225, a mean_sq of 28.436225, an F-statistic of 302.408971, and a p-value of 1.738842e-27. Again, the very small p-value (< 0.05) indicates a significant difference in logCFU between the two treatment groups.

Furthermore, the interaction effect between month and tx is also significant, with a df of 3.0, a sum_sq of 5.949672, a mean_sq of 1.983224, an F-statistic of 21.090869, and a p-value of 6.562050e-10 (< 0.05). This interaction effect suggests that the differences in logCFU between the treatment groups vary across the different months, or alternatively, that the differences in logCFU across months depend on the treatment group.

The residual row in the ANOVA table provides the degrees of freedom (72.0) and sum of squares (6.770329) for the unexplained variation in the data.

Corrections: Subtracting by Baseline

Next, baseline-adjusted logCFU values were computed by subtracting the baseline value from the corresponding logCFU measurement for each participant. This adjustment aimed to remove the influence of individual differences in baseline CFU levels, allowing for a more accurate comparison across participants and treatment groups. We then reran our two way Anove and obtained the following results below:



	df	Sum_sq	Mean_sq	F	PR(>F)
C(months)	1	9.489	9.489	659.542	5.559e-38
C(tx)	3	42.154	14.051	976.674	3.214e-58
C(Months):C(tx)	3	5.950	1.983	137.849	9.139e-30
Residual	72	1.086	0.0143	NaN	NaN

The ANOVA table presented analyzes the effects of treatments (C(tx)), months (C(month)), and their interaction (C(tx):C(month)) on a dependent variable. The treatments had one degree of freedom and explained a significant portion of the variance (sum of squares = 9.488793) with a mean square of 9.488793, resulting in an F-value of 659.542408 and a highly significant p-value (approximately 5.56×10−38). Months, with three degrees of freedom, explained a larger sum of squares (42.154077) with a mean square of 14.051359. This factor had an even higher F-value of 976.674992 and a p-value close to zero

(3.21×10−58), indicating a very strong effect. The interaction between treatments and months also showed significant effects, with a sum of squares of 5.949672, a mean square of 1.983224, an F-value of 137.848957, and a p-value approximately 9.14×10−30. Lastly, the residual variance with 72 degrees of freedom was 1.035859, with a small mean square error of 0.014387, reflecting the unexplained variation in the model. The model as a whole captures significant effects from the main factors and their interaction, suggesting complex dynamics between treatments and temporal variations across months.

Determining fit

The methodology involves several steps to analyze the data using regression models and compare them to determine the best fit. We use the data that is baseline-adjusted log of the CFU, where each measurement is adjusted by subtracting the initial baseline value. This adjusted value serves as the dependent variable in the subsequent regression analyses.

The analysis then proceeds by fitting two types of regression models: a linear model and a quadratic model. The linear model includes an interaction term between the treatment groups and months, which allows for an assessment of how the relationship between the treatment effect and time changes linearly. Following this, a quadratic regression model is also fitted, which extends the linear model by including a squared term of the month variable. This quadratic term is introduced to capture any potential non-linear effects in the relationship over time.

Finally, a likelihood ratio test is conducted to compare the fit of the linear and quadratic models. This statistical test evaluates whether the additional complexity of the quadratic model significantly improves the model fit compared to the simpler linear model. The test results in three values: the test statistic, the p-value, and the degrees of freedom associated with the test.

Analyzing the results of the likelihood ratio test, the test statistic is approximately 0.239, and the p-value is around 0.788 with 2 degrees of freedom. This high p-value suggests that there is no significant difference in the fit between the linear and quadratic models, indicating that the simpler linear model may be sufficient for modeling the relationship between the treatment groups, time, and the adjusted log CFU. Essentially, the inclusion of the quadratic term does not provide a statistically significant improvement in explaining the variability in the data over the simpler linear model.

Rescaling

Lastly we manipulate the CFU by recalculated from its logarithmic form to its original scale to facilitate an analysis in terms of actual CFU counts. This allows the statistical analysis to operate directly on the outcome measure of interest, potentially providing insights that are more interpretable in a biological context.

Two regression models are fitted to the data: a linear model and a quadratic model. The linear model explores whether the relationship between CFU and time can be adequately described using a linear function, considering interactions with treatment groups. The quadratic model extends this analysis by including a squared term of time, hypothesizing that the relationship might be non-linear and could be better described by a quadratic function. This approach tests for potential curvature in the data, which could indicate acceleration or deceleration in CFU counts over time.

A likelihood ratio test is then employed to compare these two models. This test helps determine whether the additional complexity of the quadratic model significantly improves the explanation of variability in the CFU counts compared to the linear model. The results of this test yield a statistic value of approximately 13.64, a very low p-value of about 9.07×10−69.07×10−6, and two degrees of freedom. These results indicate a statistically significant improvement in model fit with the inclusion of the quadratic term. This suggests that the relationship between CFU counts and time is not merely linear but also includes a significant quadratic component, reflecting a more complex dynamic in how CFU changes over time across different treatment groups.