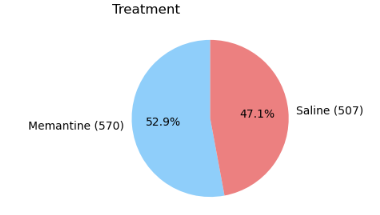
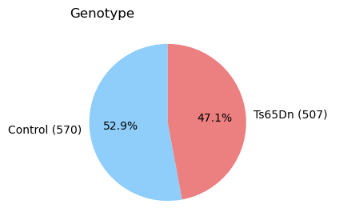
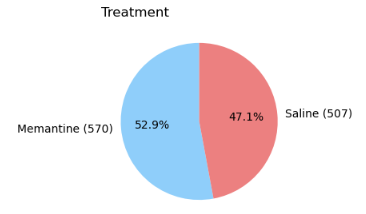
Name: Chu Jia Sing (P7480385)

Explore data

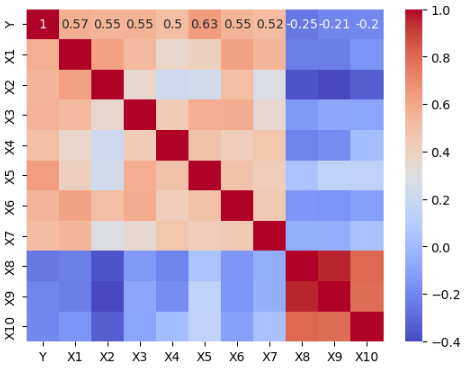
Besides the response variable (y), there are 10 quantitative predictors and 3 qualitative predictors.

There are 1077 rows of data with no null values for any of the variables.

The count of the qualitative predictors are as such:



As shown in the heatmap below, there is some correlation between pairs of quantitative predictors (correlation coefficient > 0.5) which could potentially cause multicollinearity issues if these pairs of quantitative predictors are included in the multiple linear regression (MLR) model.

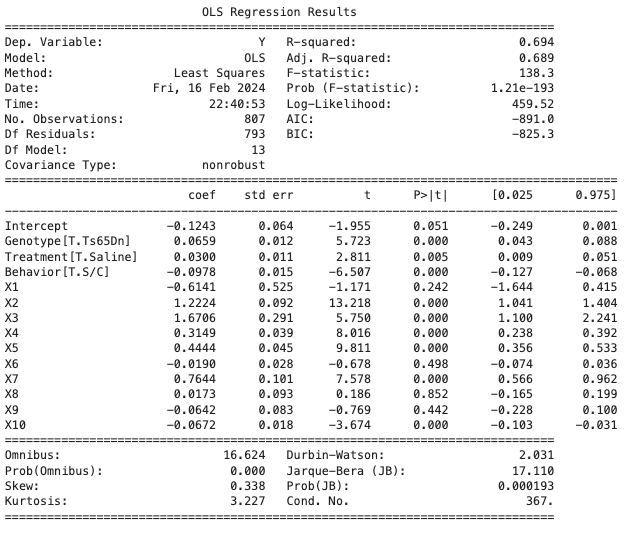


Fit model

*Note: I set the random state to 80385 as I received the following error “leading zeros in decimal integer literals are not permitted; use an 0o prefix for octal integers” when I set state to 0385.*

Fitted model on training data:

| ŷ = -0.124 + 0.066 \* Ts65Dn + 0.03 \* Saline - 0.098 \* S/C - 0.614 \* X1 + 1.222 \* X2 + 1.671 \* X3 + 0.315 \* X4 + 0.444 \* X5 - 0.019 \* X6 + 0.764 \* X7 + 0.017 \* X8 - 0.064 \* X9 - 0.067 \* X10 |
| --- |



Evaluate Model

1. *Model fit*

Adjusted R2 = 68.9% which is relatively high. The model has quite a good fit, even though it could be better.

1. *Prediction accuracy*

Mean Squared Error (MSE) = 0.01907

MSE measures the average squared difference between the predicted ŷ and the actual y. 0.01907 seems to be relatively low given that the mean of y in the full dataset is 1.548413. The model is relatively accurate in its prediction.

1. *Model/predictor significance*

t-test on the significance of predictor coefficients at 5% significance level

H0: β*j*= 0

H1: β*j* ≠ 0

where j = 1, 2, 3, …, 13

Except for the predictor coefficients for X1, X6, X8, X9, most p-values are <0.05 which we will reject H0 for these predictors.

Conclusion: All predictors in the MLR model are statistically significant predictors of Y, after adjusting for other predictors, except for X1, X6, X8, X9.

F-test on the significance of the overall model at 5% significance level

H0: β*1* = β*2* = … = β*13* = 0

H1: at least one β*j* ≠ 0

where j = 1, 2, 3, …, 13

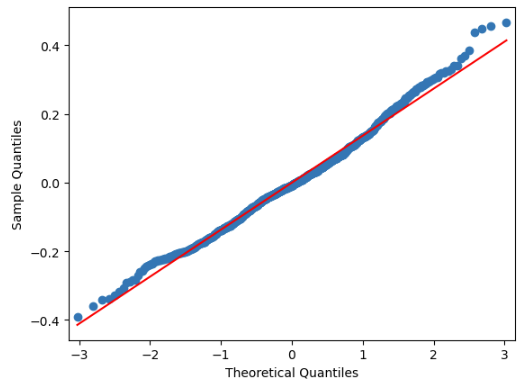
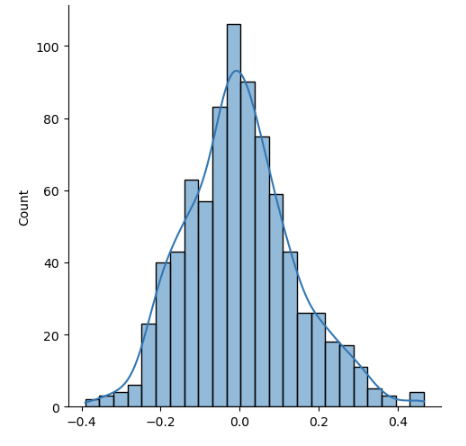
p-value = 0.000 < 0.05. Reject H0.

Conclusion: At least one predictor contributes significantly to the model.

1. *Checking of assumptions*

Are residuals normally distributed?

The residuals seem to be skewed slightly right based on the histogram and normality plot and do not follow a normal distribution. We can conduct the Jarque-Bera test to verify.



*Jarque-Bera test*

H0: Residuals are normally distributed.

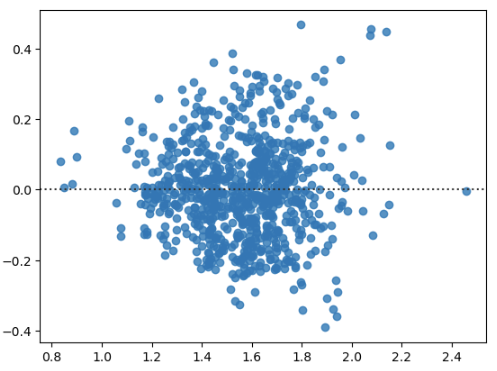
H1: Residuals are not normally distributed

p-value = 0.000193 < 0.05. Reject H0.

Conclusion: Residuals are not normally distributed. Assumption is violated.

Are residuals homoscedastic?

The residuals vs fitted values plot show a slight funnel effect and it is unlikely that the variance is constant. We can conduct the Breusch Pagan test to verify.



*Breusch Pagan test*

H0: Residuals are homoscedastic, i.e. have constant variance.

H1: Residuals are heteroscedastic.

p-value = 0.000 < 0.05. Reject H0.

Conclusion: Residuals are not homoscedastic. Assumption is violated.

Are residuals independent?

*Durbin-Watson test*

H0: Residuals are independent, i.e. autocorrelation does not exist.

H1: Residuals are not independent, i.e. autocorrelation is present.

Durbin-Watson test statistic = 2.030 ≈ 2.

Conclusion: No autocorrelation. Assumption is not violated.

Is there a multicollinearity problem?

Condition number: 367.36 > 30 which indicates a severe multicollinearity problem.

VIF of individual predictors:

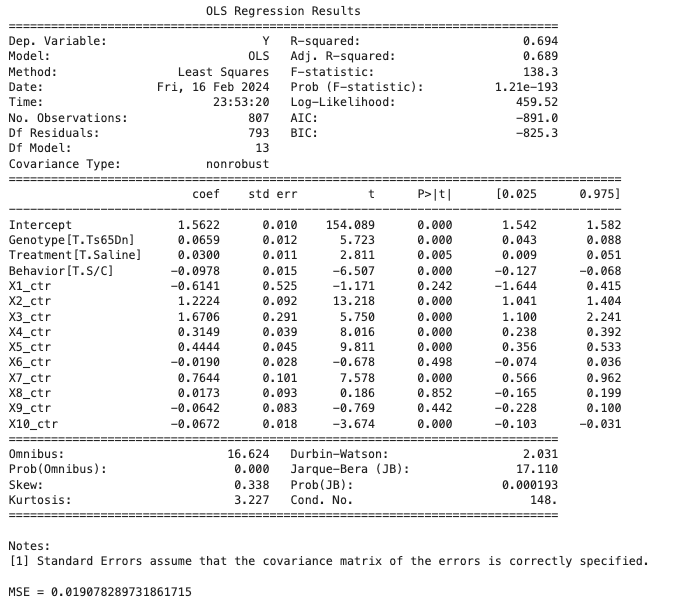
| Intercept  Ts65Dn  Saline  S/C  X1  X2  X3  X4  X5  X6  X7  X8  X9  X10 | 170.95  1.4  1.2  2.38  2.56  2.44  2.22  2.3  2.56  2.11  1.79  17.11  18.29  3.08 |
| --- | --- |

The VIFs of the intercept, X8 and X9 > 10 which indicate that these predictors contribute greatly to the multicollinearity problem.

Improve Model

1. Mean-centering of variables

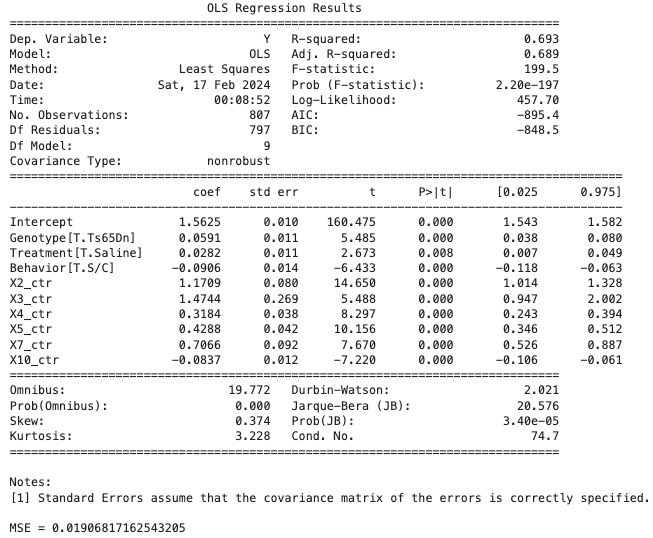
This improves the multicollinearity problem caused by the intercept.



Condition number = 147.84 < 367.36 and VIF of the intercept = 4.35 < 170.95 which indicate an improvement to the multicollinearity problem.

1. Variable selection

Apply backward stepwise selection to remove variables which are not statistically significant (with the highest p-value for t-test) until all remaining predictors are statistically significant (i.e. p-value for t-test < 0.05).

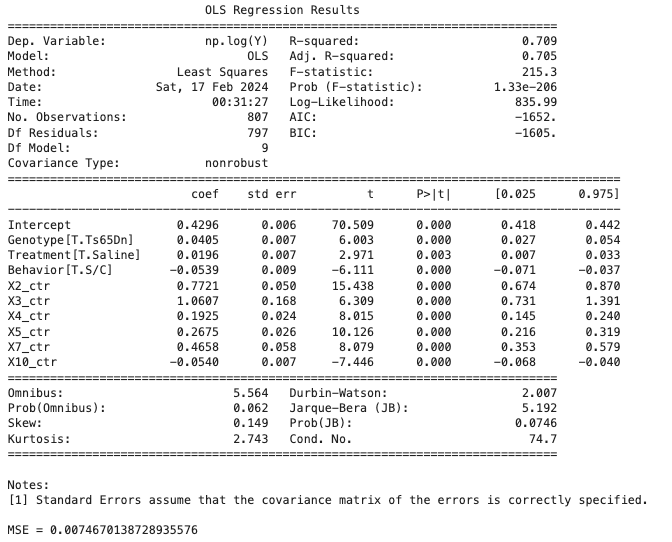


This ensures that we keep the model parsimonious. At the same time, we have also removed some of the predictors that cause multicollinearity problems, e.g. X8 and X9.

1. Transformation of Y

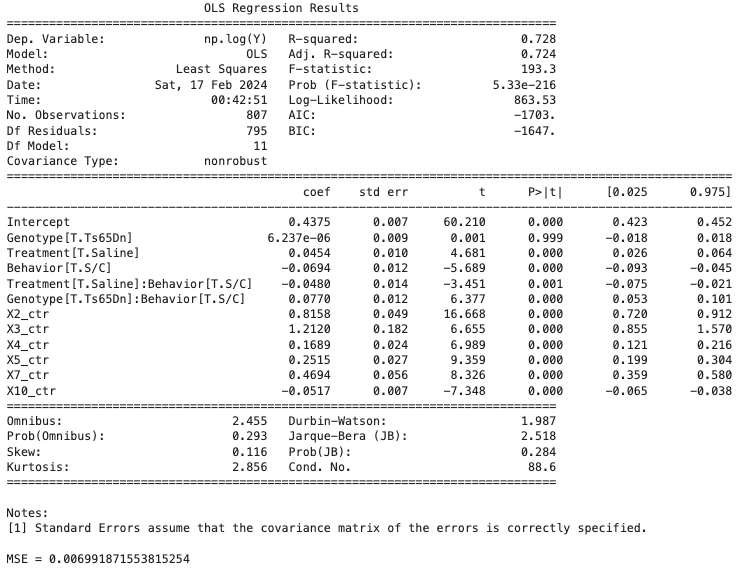
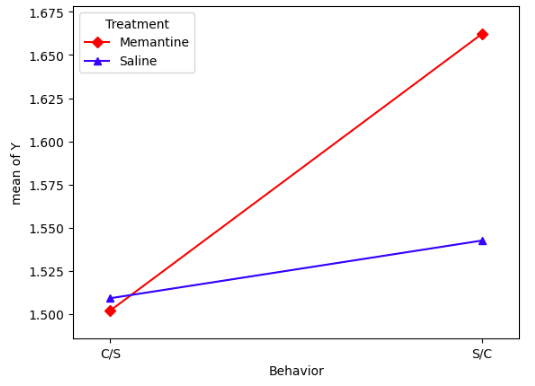
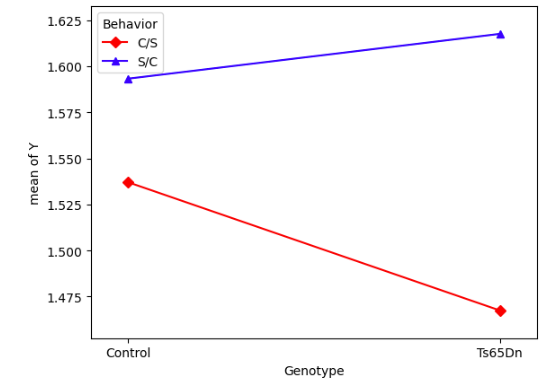
To address non-constant variance and non-normality of residuals.

After experimenting with transformation of Y to 1/Y, sqrt(Y), ln(Y), Y2, the transformation to ln(Y) was the most suitable as the p-values of the Jarque-Bera test = 0.0746 and Breusch Pagan test = 0.108 which are > 0.05. The models now fulfill the assumptions that the residuals are normally distributed and homoscedastic.



1. Interaction of variables

Non-parallel slopes in the interaction plot indicate that interactions exist between Treatment and Behavior, and Genotype and Behavior.



Final fitted model:

| ln(ŷ) = 0.4375 + 6.237 \* 10-6 \* Ts65Dn + 0.0454 \* Saline - 0.0694 \* S/C - 0.0480 \* Saline \* S/C + 0.0770 \* Ts65Dn \* S/C + 0.8158 \* X2ctr + 1.2120 \* X3ctr + 0.1689 \* X4ctr + 0.2515 \* X5ctr + 0.4694 \* X7ctr - 0.0517 \* X10ctr |
| --- |