Edie and Dari

MA-23c

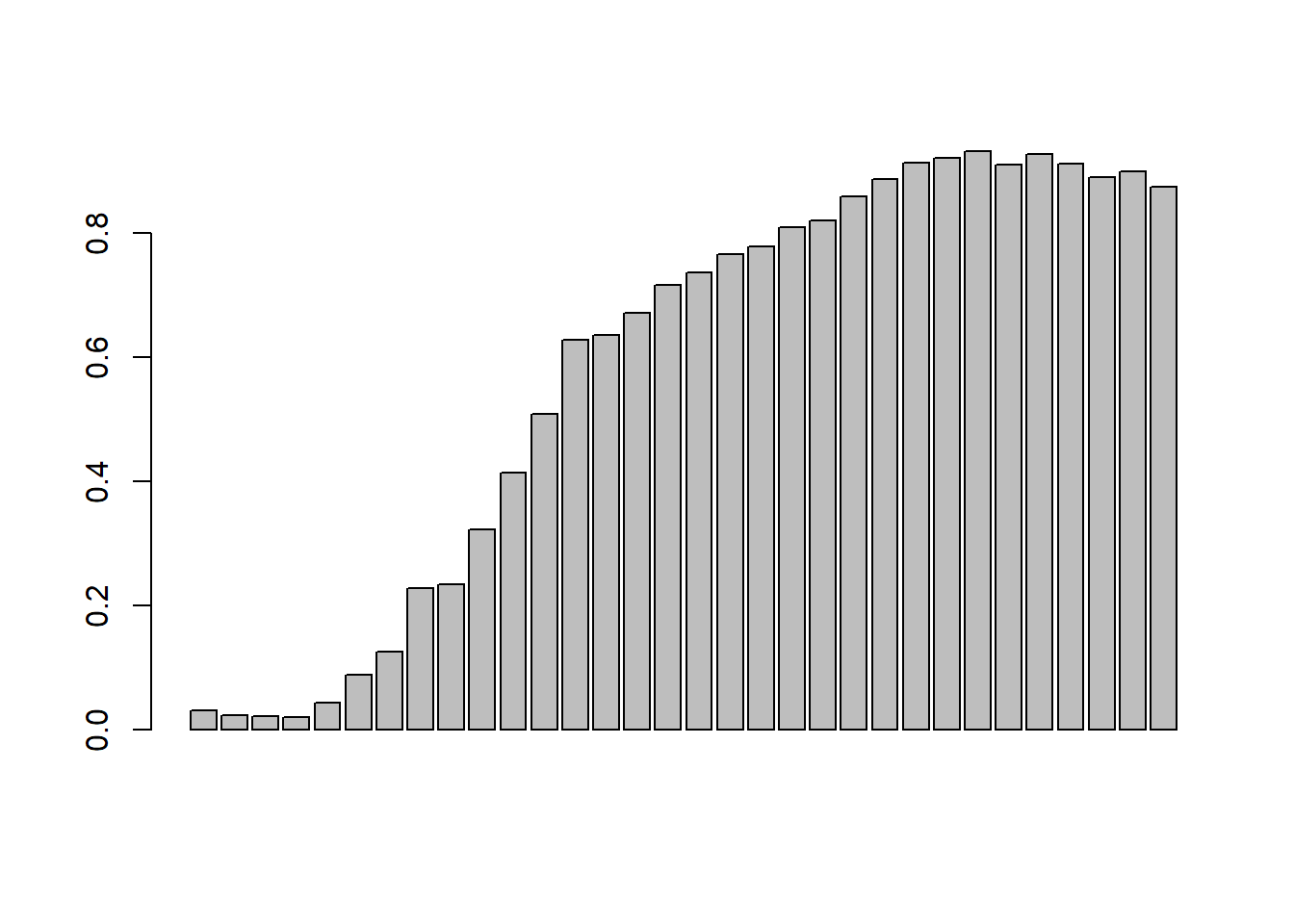
R Project Written Summary (LONG)

OVERVIEW-

Dari coded a population simulation (<https://math-23c-final.netlify.app/>) which models generations of the populations that consider adjustable variables, including dominant and recessive phenotypes that influence survivability. For our dataset, we used the specific inputs: 0% for the start percentage of dominance, 50% chance for the dominant phenotype to live, 75% chance for the recessive phenotype to live, and a 3% mutation rate.

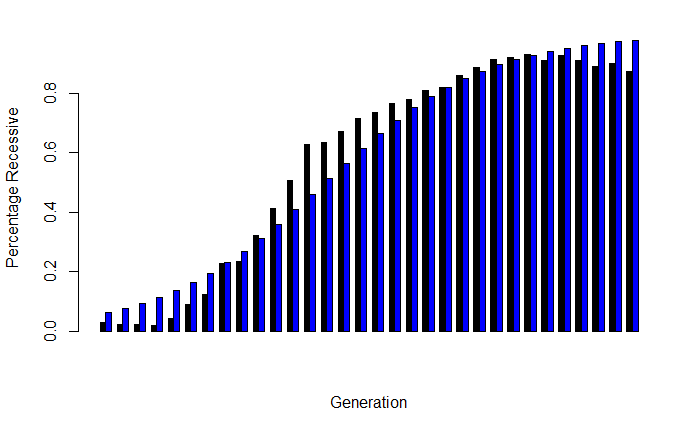
ANALYSIS-

Our data analysis focused on the dominant and recessive genes, which we labeled as gene A and gene B, respectively. We extracted the columns of the amount of homozygous dominant, heterozygous (dominant), and homozygous recessive cells, created a new column for the total number of cells and created variables for the total number of each gene (in each generation). Then, we generated bar plots of the total amount of cells with the dominant phenotype, the total amount of cells with the recessive phenotype, the percentage of cells with the dominant phenotype, and the percentage of cells with the recessive phenotype, and the percentage of geneB (Figure 1).

**

*Figure 1.*

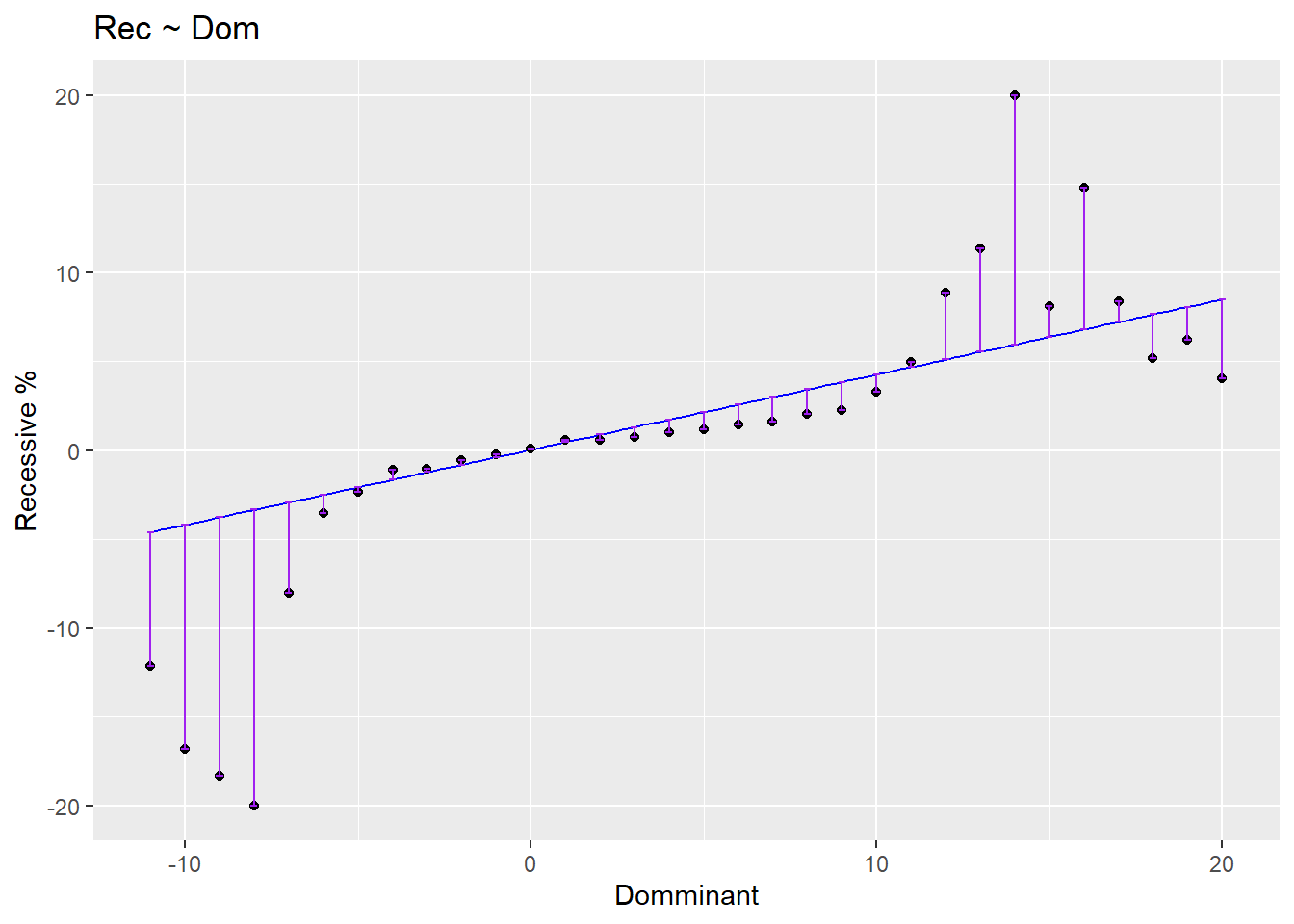
Using maximum likelihood estimation (MLE) of minus log-likelihood (MLL), we modeled the total percentage of recessive genes as a logistic function with the coefficients of -2.895979 and 0.2105899 and resulted in an MLL of 13.702. We overlayed the logistic regression on the appropriate barplot which can be seen below in Figure 2.



*Figure 2.*

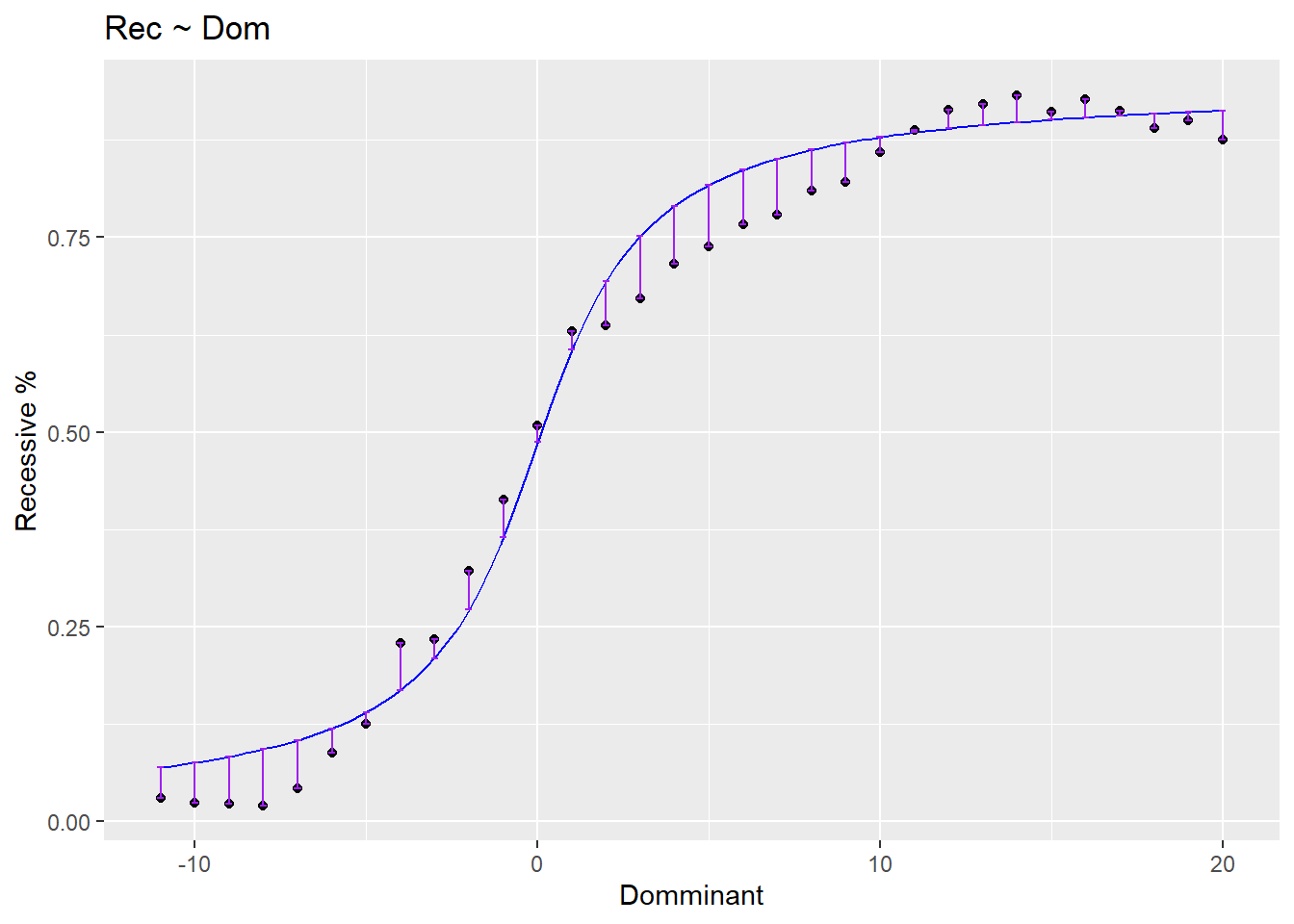
Lastly, we ran 2 χ2 test of the logistic regression, the first of which (which we scaled the values by multiplying by the number of data points) returned a p-value of just under 0.1. Using the typical test (p <= 0.05), we can say that the p-value is not low enough to reject the null hypothesis that they come from the same distribution. However, 0.1 is still a relatively small p-value and, therefore, this test implies it is still quite unlikely that they did come from the same distribution. The second test was done by scaling so that the total area under the curve was equal to the number of data points. This version should be more accurate and yielded a P-value of 1 which implies that they definitely came from the same distribution.

We then took the same data and modeled it as an arctangent function. First, we had to normalize the data and center our x-values. Using our centered x-values, we were able to then center our y-values which we performed a linear regression, as seen below in Figure 3.



*Figure 3.*

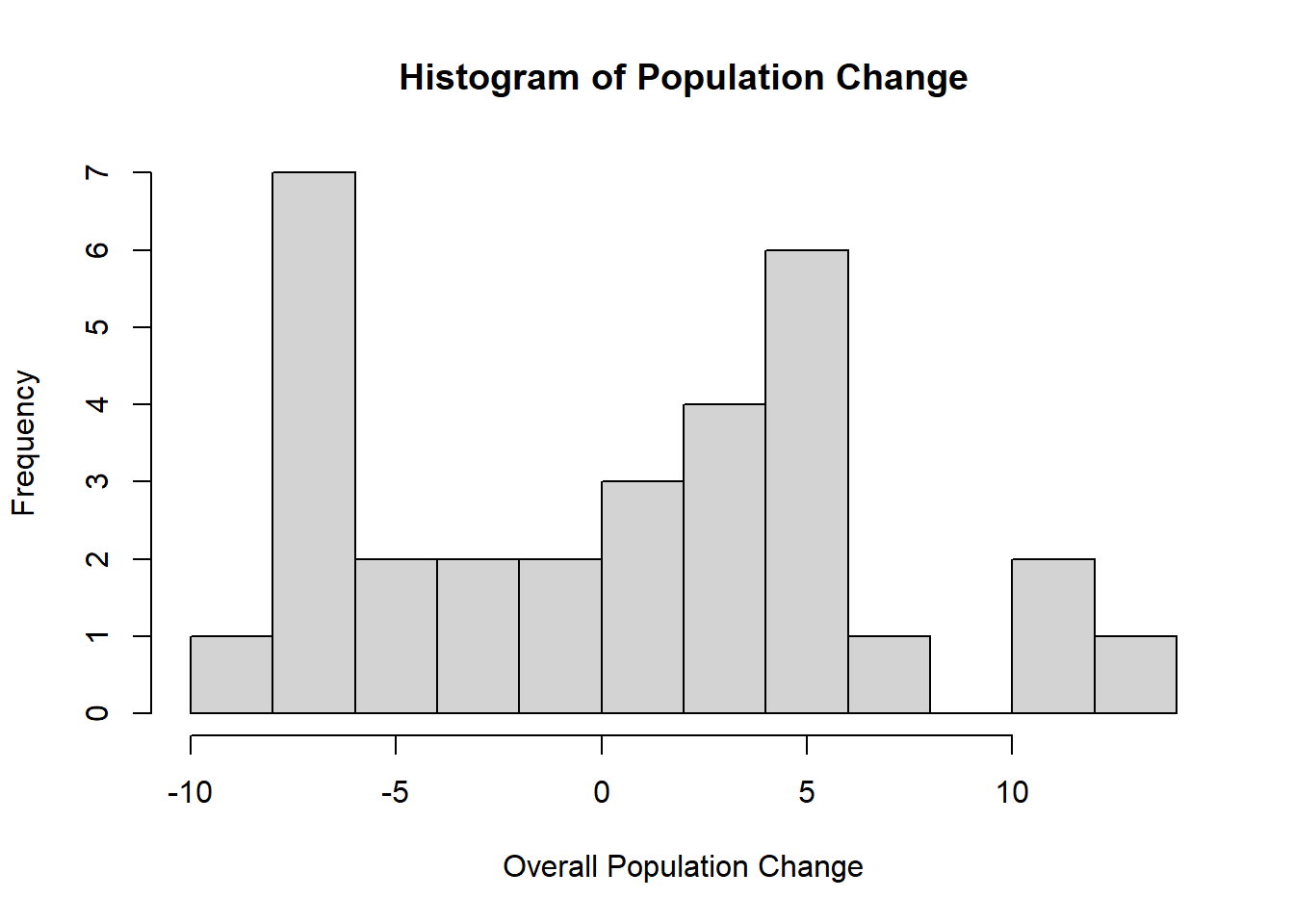
To fit our data, we took the arctangent of those values. We denormalized our y-values and overlaid the fitted arctangent function, which is shown in Figure 4.



*Figure 4.*

Looking at Figure 4, we can see that this model is a good fit. However, to confirm, we ran a χ2 test on our arctangent model using both of the previous scaling factors. The first returned a relatively high p-value of 0.7869044 and proves that our model is an even better fit (comparing it to the first test from the prior distribution). This is very interesting since logistic regression is the proper model for populations, but our arctangent model is a better fit for our data. The second χ2 represents

Next, we created a matrix where each column is representative of the change of the population from the previous generation. We also added boolean columns, shows an interesting spike around -8 and 5, but a gap near 8.

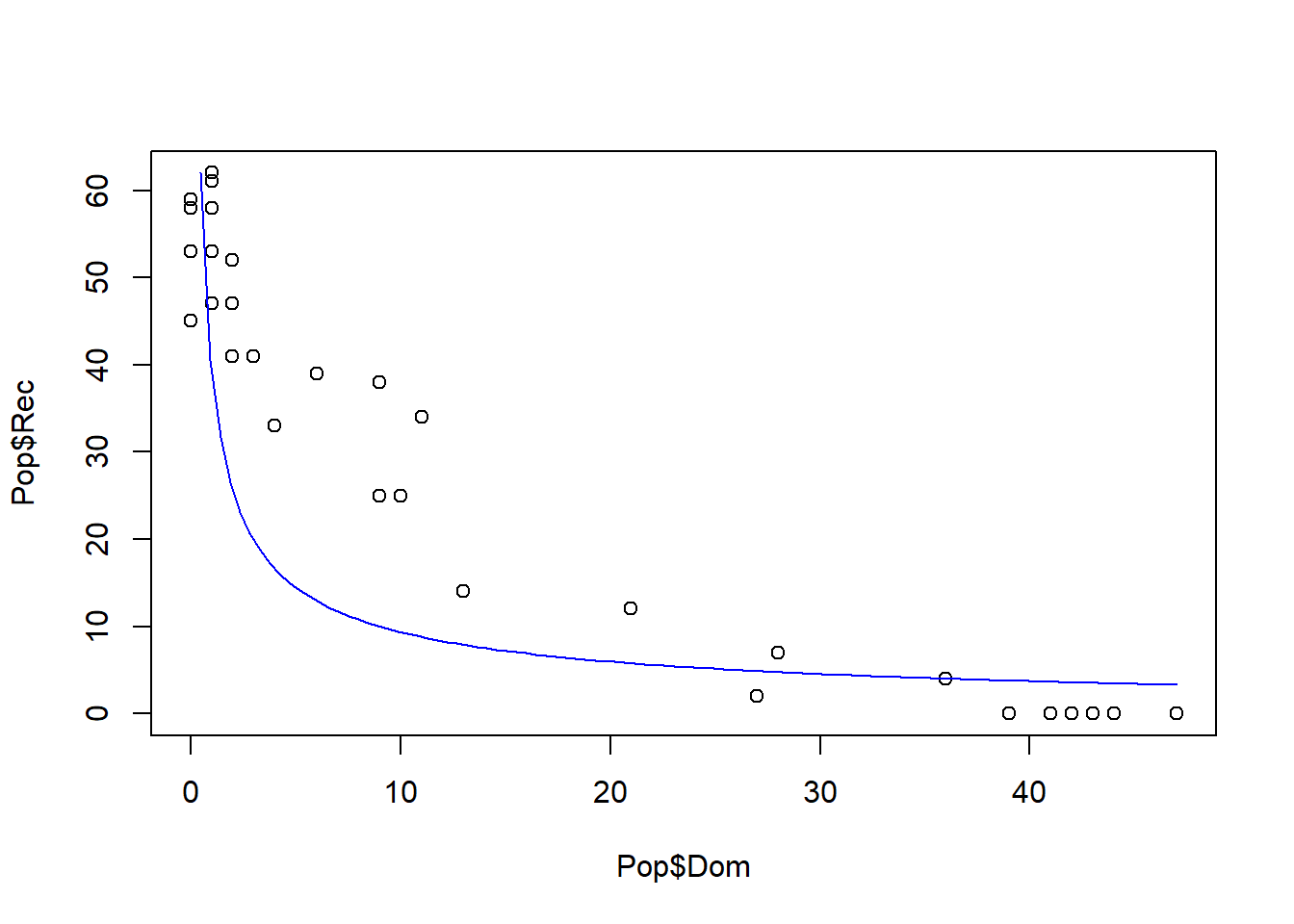


*Figure 5.*

We then made a contingency table of the increase of recessive and heterozygous phenotypes. We ran a permutation test which confirmed our hypothesis that the two are correlated. We used a permutation test because we had boolean values which we were operating over, which a permutation test is much better than a χ2 test or other classical methods. We ran the permutation test with n=1,000,000, which returned a p-value of 0.049361. While this p-value is more accurate, the code ran very slowly, so we modified the code to n=10,000 and got a p-value of ~0.05. Both of these p-values show that the result is statistically significant, according to the usual test of p <= 0.05. Thus, there is a positive correlation between recessive genes increasing and heterozygous genes increasing.

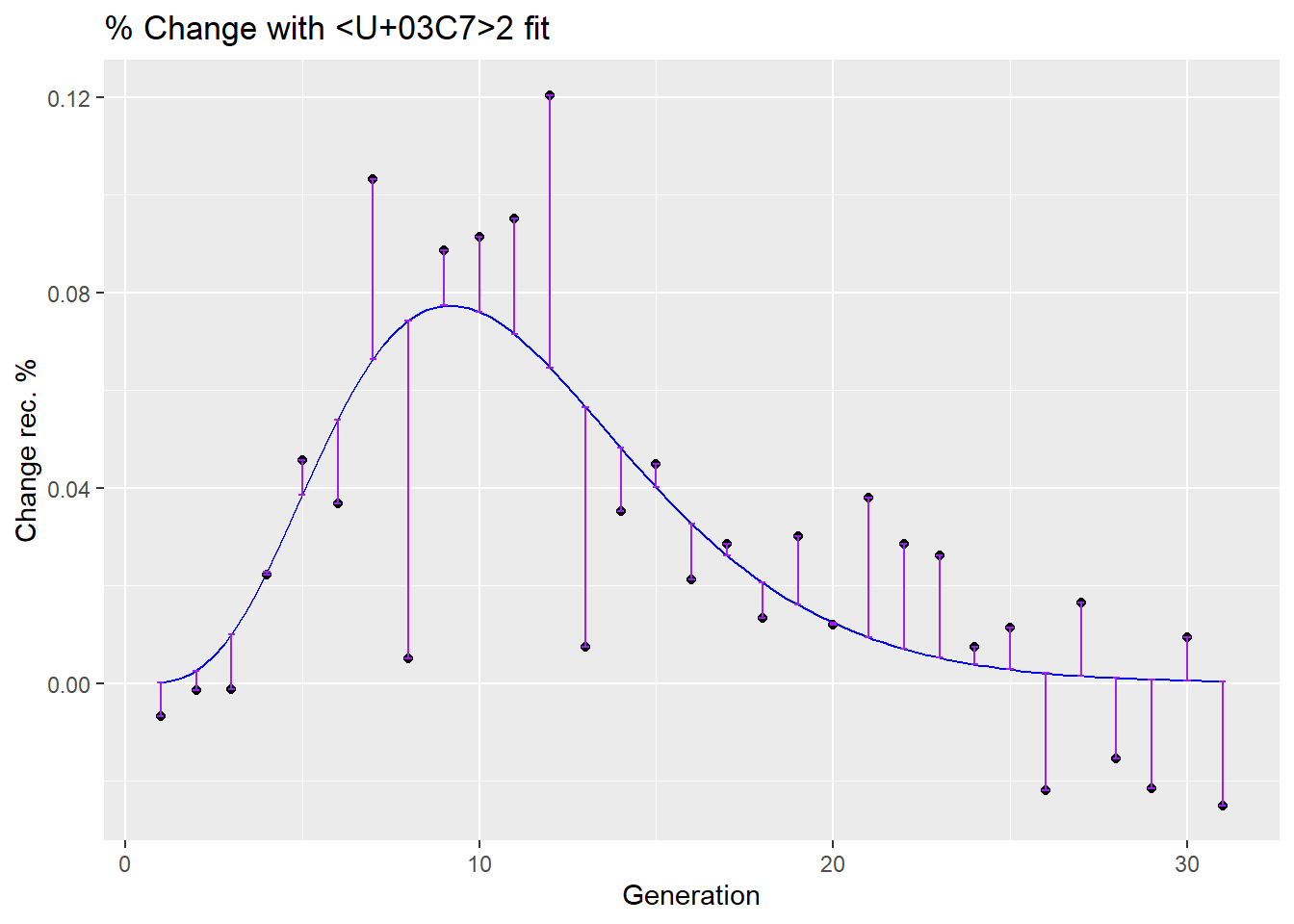
Looking at the recessive phenotypes as a function of dominant phenotypes, we should expect that the dominant and recessive phenotypes should be negatively correlated. Additionally, we should expect a high covariance because they have a large spread. When we analyzed the correlation and covariance of the variables, we found that it indeed matched our expectations.

Subsequently, we plotted the distribution of recessive phenotypes as a function of the dominant phenotype and fit it to a gamma distribution. We recreated the initial “dataset” and found its expectation and variance, which we used to find the shape and rate parameters for the Γ distribution function. In Figure 5, we overlaid our Γ distribution function over our original plot.



*Figure 6.*

Lastly, we looked at the cent increase from our population change matrix. We had to chop off the first entry since it is impossible to determine the change in population from the 0th to the first generation since the 0th generation does not exist. Following that, we normalized the data and took the mean, which should represent the degrees of freedom for the χ2 density function. To get our χ2 to be properly squared, we divided the integral of the χ2 and multiplied it by the sum of the change in population so that the total area will be the same as the area under the population change. We plotted the change in populations and overlaid our χ2 density function, which is shown in Figure 6 below. We then took a t-test and a χ2 test. The t-test returned a p-value of 0.9994, and the χ2 test gave a P-value of 1.379734\*10-17. For both of these tests, the null hypothesis is that the two data points come from the same distribution, meaning that a high P-value means that they are related, so this enormous discrepancy seems a bit odd.



*Figure 7.*

CONCLUSION-

Our simulation is based on a real model, and, thus, the simulation should be accurate. Nevertheless, we got some unintuitive results that could imply a fault inherent to the simulation. At the same time, it could be random and the result of the parameters we set at the beginning of the simulation.