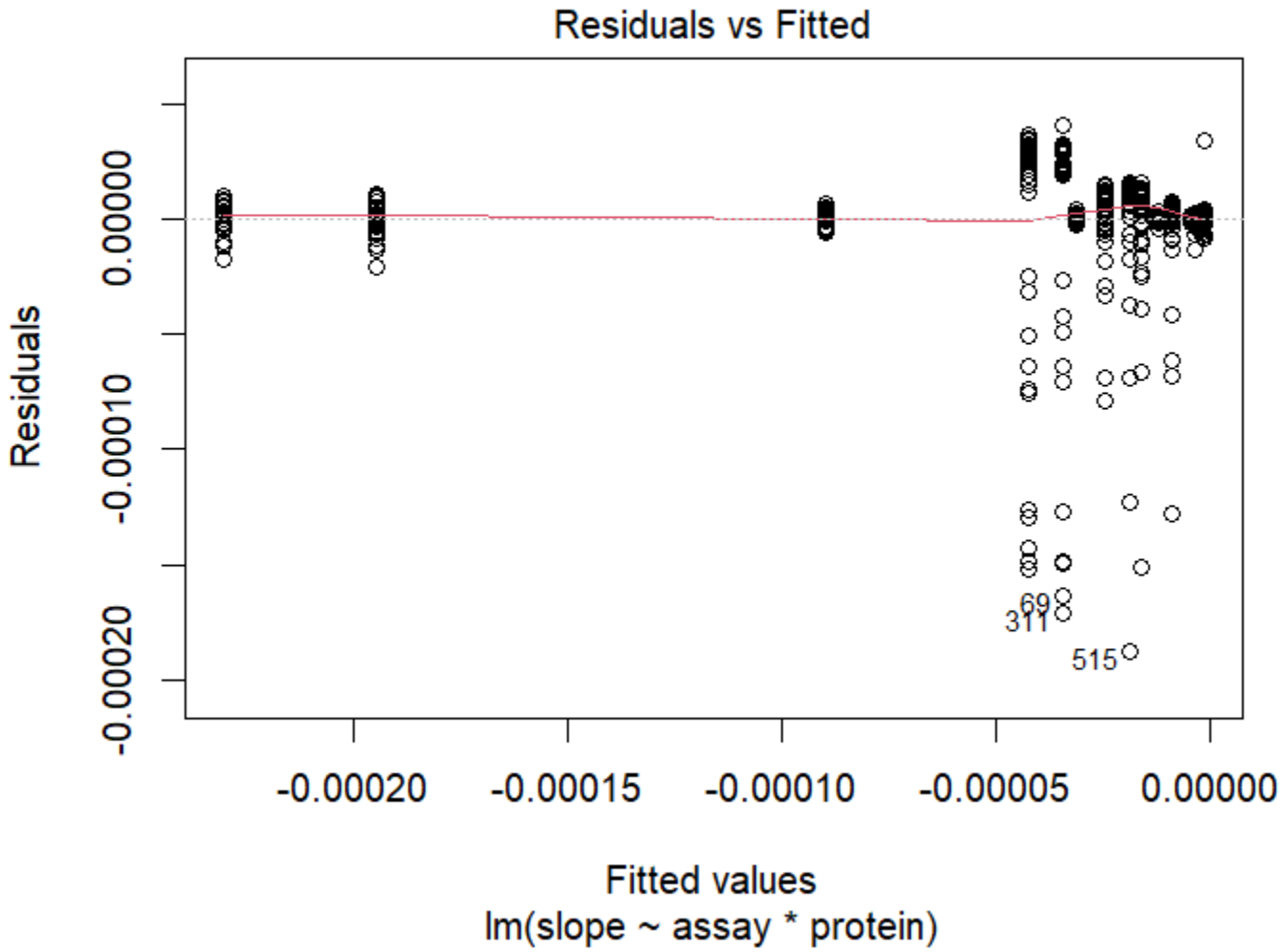
Hypothesis: The hypothesis for the data is that the quercetin chemical will result in elevated activity for all the CYP proteins (CYP 1A, 1B1, 1C1, 1D1) and that CYP1A will have the greatest activity then 1B1, 1C1 and 1D1 will have the least activity.

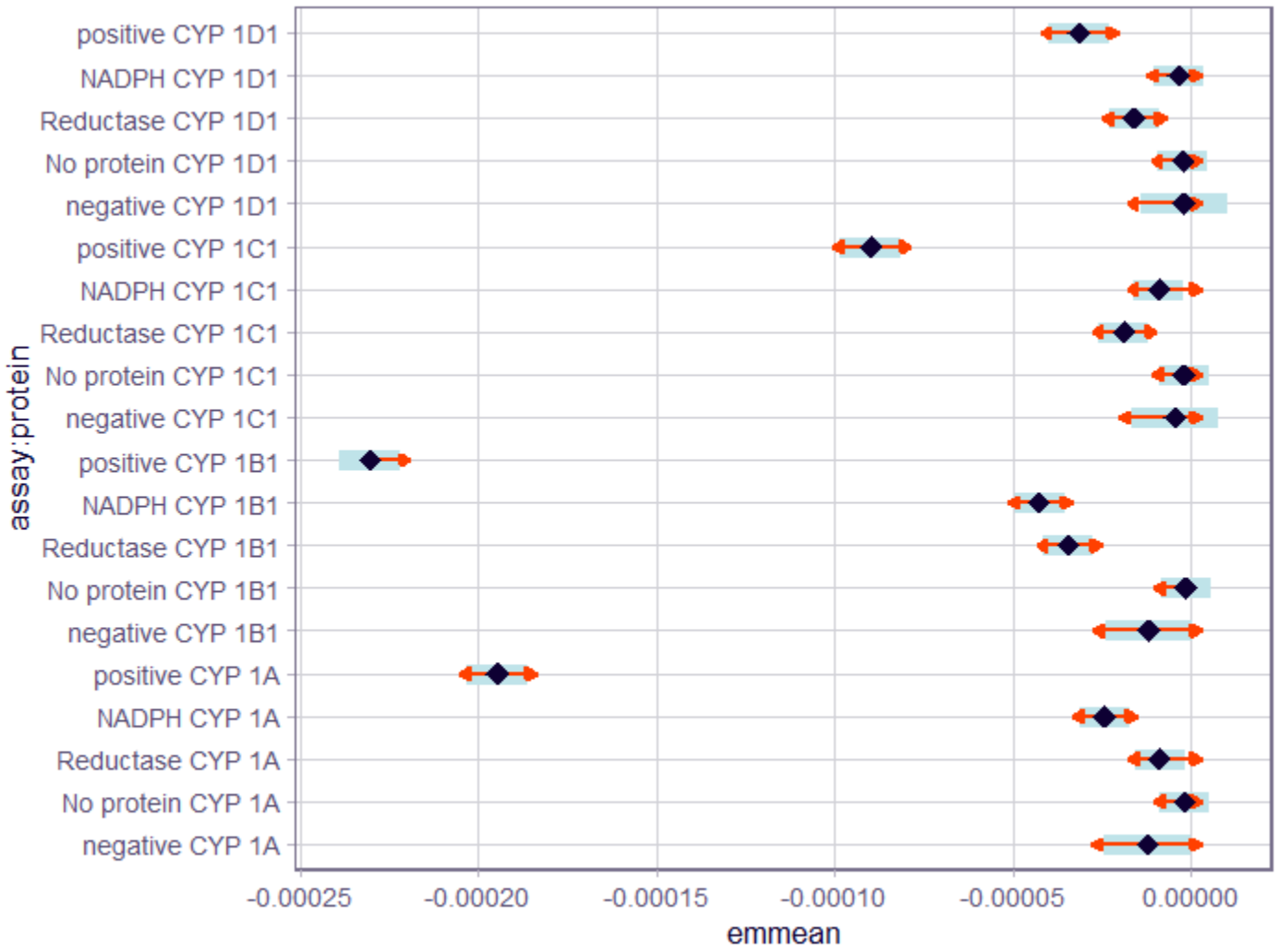
The linear model that I made for my data is based on the interaction between the assay type and the protein.

lm_all<- lm(slope~assay*protein,data=slopes)



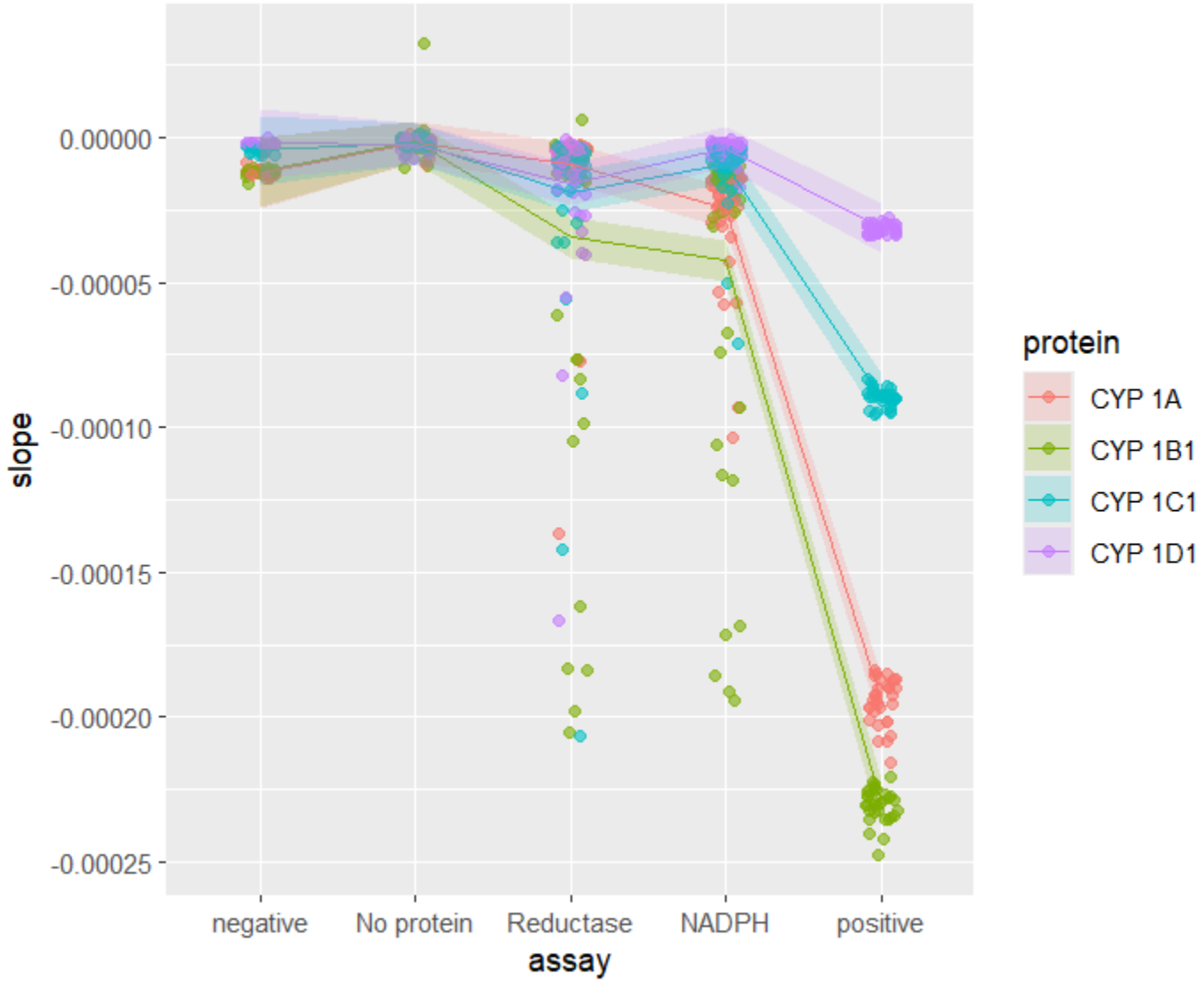
From the residuals vs fitted diagnostic plot, we see that the points are distributed pretty evenly on either side zero. The red trend line is relatively flat, suggesting that linearity is true for the model. For the points with a less negative slope there are greater negative residuals

and a greater spread between them indicating some heteroscedasticity. The assumptions of the data being linear and not being heteroscedastic appear to be reasonable based on this diagnostic plot.



This emmeans comparison chart shows the estimated marginal means and 95% confidence interval for each combination of assay and protein. We see the assay types, negative, no protein, redustase and NADPH, all having very similar and overlapping confidence intervals which was intended as each of these assays didn't have all the required components for proper enzyme activity. In the positive assays we see large differences in the slopes for CYP1A, 1B1 and 1C1. The fact that the arrows on arrows don't overlap with other combinations show that the slopes in the model are statistically significantly different from other combinations involving the same protein. This difference in slope is being used as a proxy for enzyme activity so from this we can confidently say

that quercetin is a ligand for the CYP1A, 1B1, and 1C1 proteins. For CYP1D1, it has some overlap with its own reductase assay and a very small difference in the marginal means, from this we can't say that it is activated by quercetin as there is also a generally used 25-30% increase in activity to say that an enzyme is activated by a ligand. When comparing the CYP1A, 1B1, and 1C1 proteins we can also see that CYP1B1 is activated the most followed by 1A and then 1C1.



This prediction plot overlays the original raw data points on to a chart with a line graph of the means of the predicted dataset from the model and its confidence intervals. From this chart we see that the confidence intervals encapsulate most of the raw data points in the negative, no protein and positive assays for each of the proteins. But we see that in the reductase and NADPH assays there is a larger proportion of the raw data points that are not likely to appear in the predicted dataset. These points are the source of the

heteroscedasticity in the diagnostic plot and for these assays the model doesn't fit the greatest for what we saw in the experiment.

There is proof of activation of 3 of the 4 CYP genes (not CYP1D1) by quercetin as activity is measured as a proxy of change in absorbance over time and the protein with the greatest activation was CYP1B1 then CYP1A followed by CYP1C1.