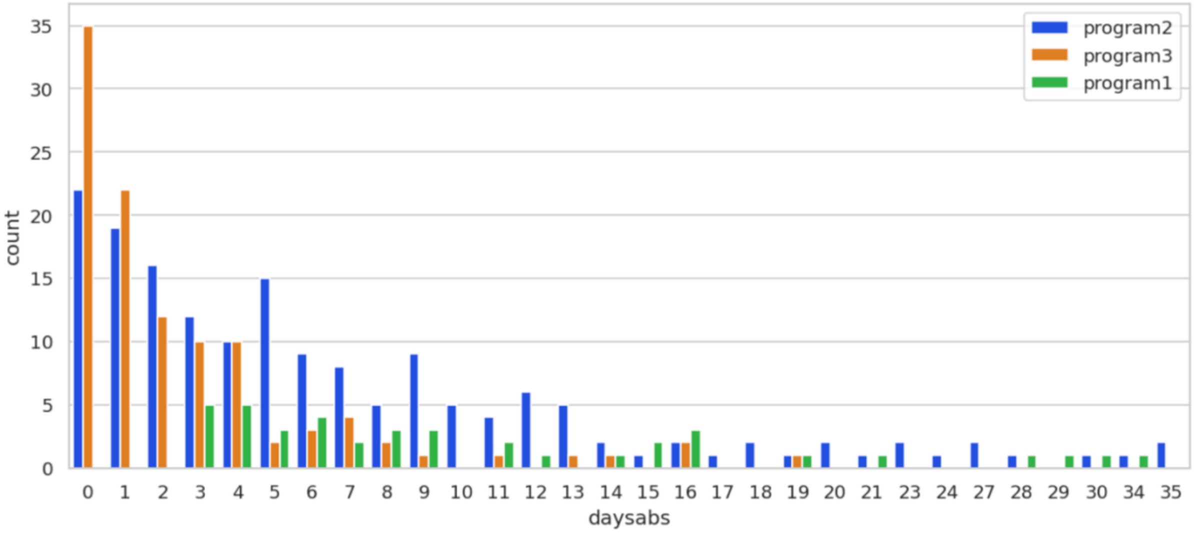
[i] First aggregate the data at program level and plot a bar chart for each program for the number of absent days. Do you see a difference in absent days across programs? [5]



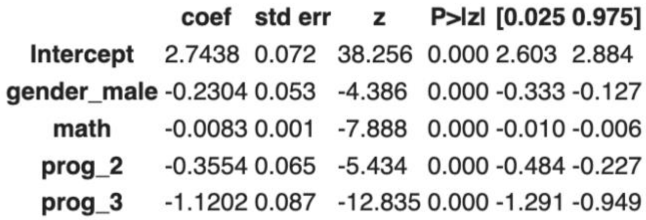
Program3 most students have very small number of absent days, Program2 has a higher distribution of absent days , while program1 distribute mostly on high number of absent days>=3

Write down the Poisson GLM equation which estimates the impact of covariates (gender, math score and program dummies) on absent days.

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{math\_score} + \beta_3 \times \text{program2} + \beta_4 \times \text{program3}$$

[ii] Now estimate a Poisson model and report the results. Split your data into train and test. Interpret the estimates on program dummies. [7]

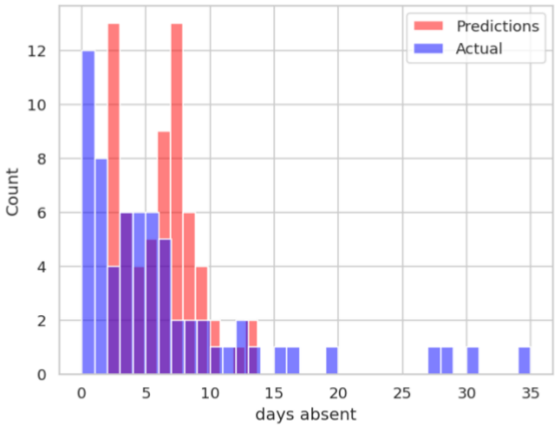
Predict the results for a hold-out sample and plot the actual data and prediction. [3]

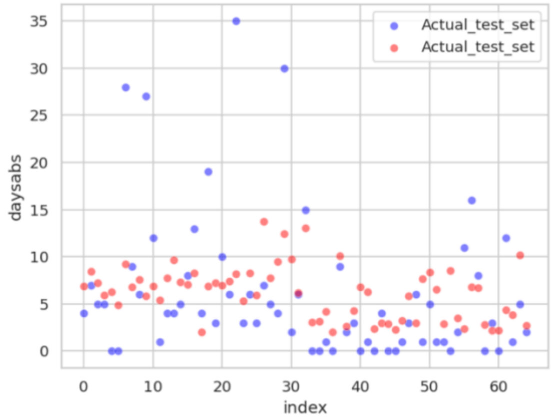


	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7438	0.072	38.256	0.000	2.603	2.884
gender_male	-0.2304	0.053	-4.386	0.000	-0.333	-0.127
math	-0.0083	0.001	-7.888	0.000	-0.010	-0.006
prog_2	-0.3554	0.065	-5.434	0.000	-0.484	-0.227
prog_3	-1.1202	0.087	-12.835	0.000	-1.291	-0.949

Prog_2(-0.355). Being in Program_2 compared to Program_1, reduces the expected log of count of absent days by 0.3554. The p value 0.0 shows it’s significant

Prog_3(-1.12). Being in Program_3 compared to Program_1, reduces the expected log of count of absent days by 1.1202. The p value 0.0 shows it’s significant



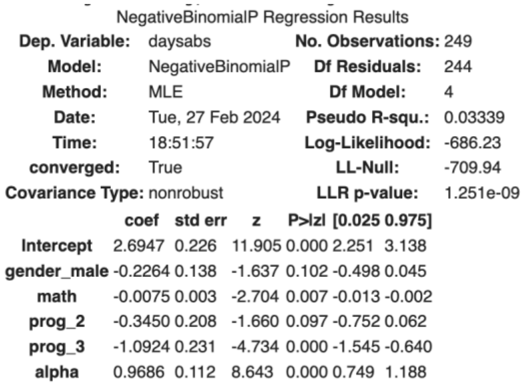


[iii] Based on your understanding of count data, you suspect that your data could be heterogenous. How will you test this? Explain the reasoning for your test. [5]Show the result of your model. Do you conclude that data is heterogenous?[5] Interpret the estimates on program dummies. [2]Predict the results for the hold-out sample and plot the actual data and prediction. [3]

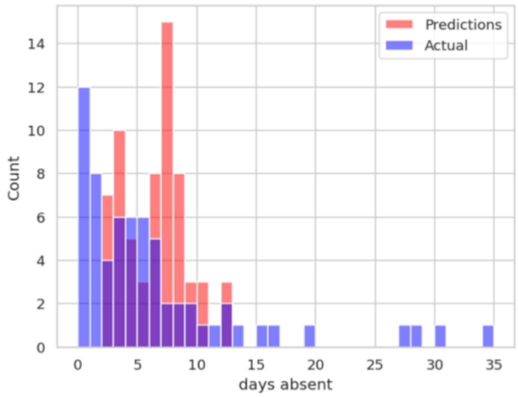
We can use the Negative Binomial model to estimate alpha. Given the results, alpha=0.9589, which suggests the presence of overdispersion, we can conclude that the data exhibits heterogeneity. It indicates that the variance of the count data significantly exceeds the mean, a scenario that cannot be adequately modeled by the Poisson distribution, therefore we can use the Negative Binomial model.

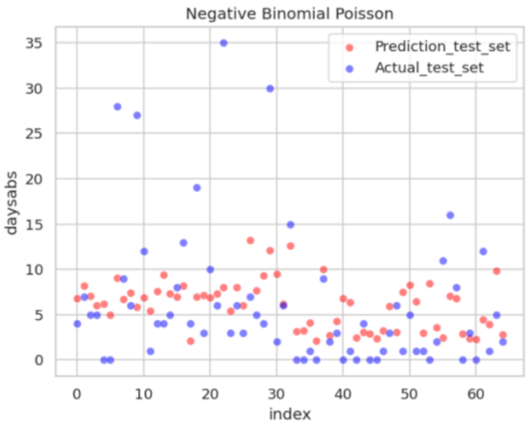
Program_2(-0.345): Being in Program_2 compared to Program_1, reduces the expected log of count of absent days by 0.345. The p value 0.097 shows it's not significant

Program_3(-1.09): Being in Program_3 compared to Program_1, reduces the expected log of count of absent days by 1.09. The p value 0.0 shows it's significant



```
                 NegativeBinomialP Regression Results
 Dep. Variable:   daysabs              No. Observations: 249
        Model:    NegativeBinomialP    Df Residuals:     244
       Method:    MLE                  Df Model:         4
         Date:    Tue, 27 Feb 2024     Pseudo R-squ.:    0.03339
         Time:    18:51:57             Log-Likelihood:   -686.23
    converged:    True                 LL-Null:          -709.94
Covariance Type: nonrobust             LLR p-value:      1.251e-09
                coef    std err   z      P>|z| [0.025 0.975]
   Intercept    2.6947  0.226    11.905  0.000 2.251   3.138
gender_male    -0.2264  0.138    -1.637  0.102 -0.498  0.045
        math   -0.0075  0.003    -2.704  0.007 -0.013  -0.002
      prog_2   -0.3450  0.208    -1.660  0.097 -0.752  0.062
      prog_3   -1.0924  0.231    -4.734  0.000 -1.545  -0.640
       alpha    0.9686  0.112    8.643   0.000 0.749   1.188
```


Predictions
Actual
Count
days absent


Negative Binomial Poisson
Prediction_test_set
Actual_test_set
daysabs
index