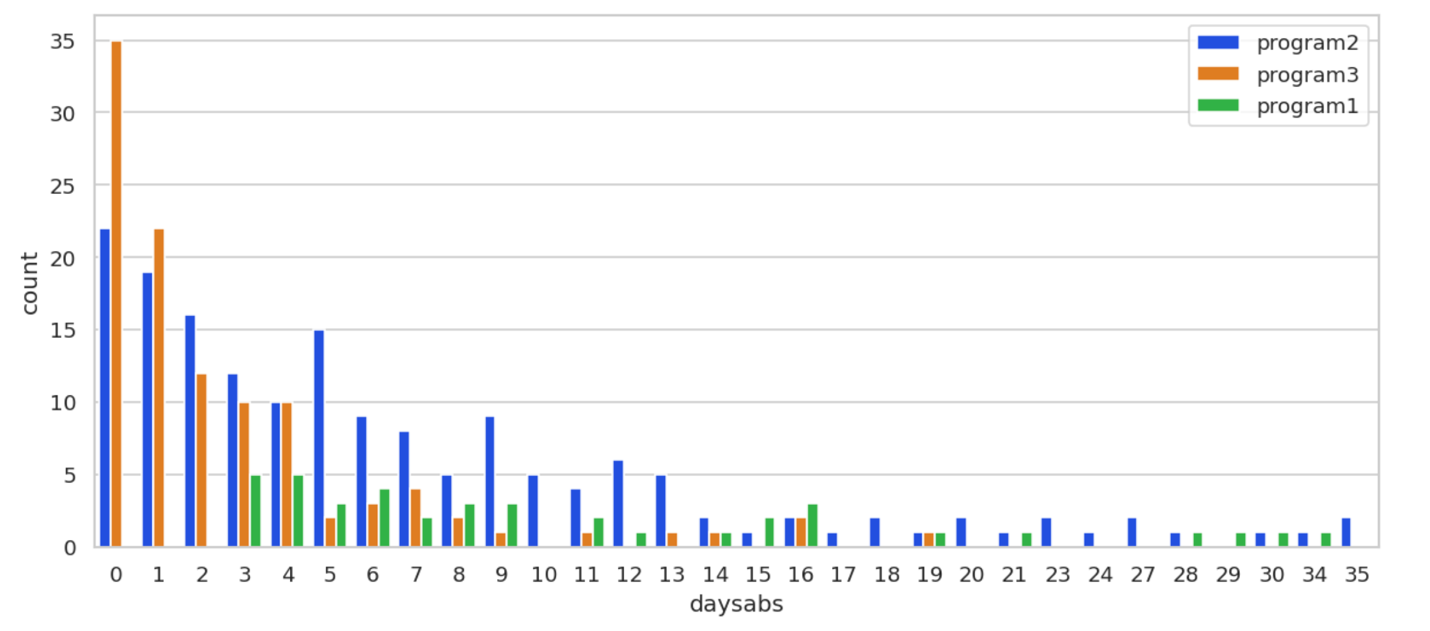
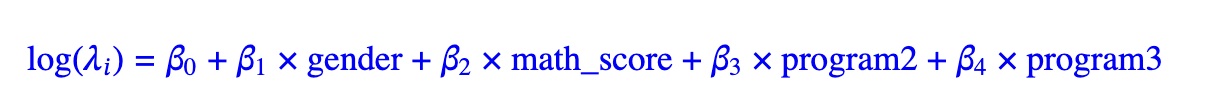
[i] First aggregate the data at program level and plot a bar chart for each program for the number of absent days. Do you see a difference in absent days across programs? [5]



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.



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

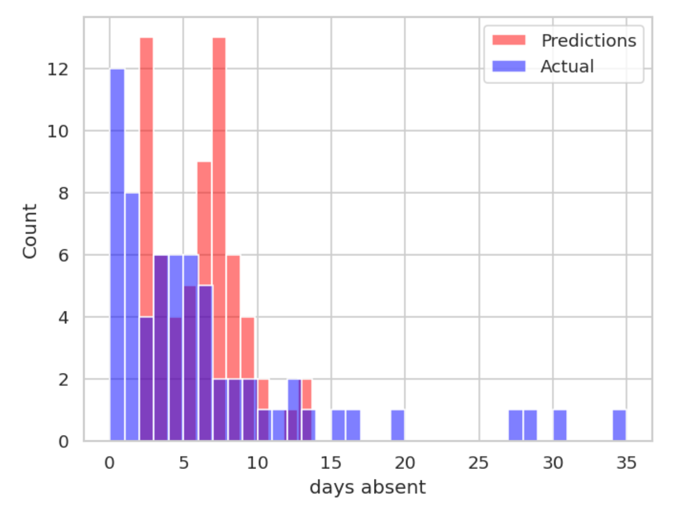
Predict the results for a hold-out sample and plot the actual data and prediction. [3]

A table of numbers and letters

Description automatically generated

**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

A graph with red and blue dots

Description automatically generated

[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

A black and white document with numbers and letters

Description automatically generated

A graph of different colored bars

Description automatically generatedA graph with red and blue dots

Description automatically generated