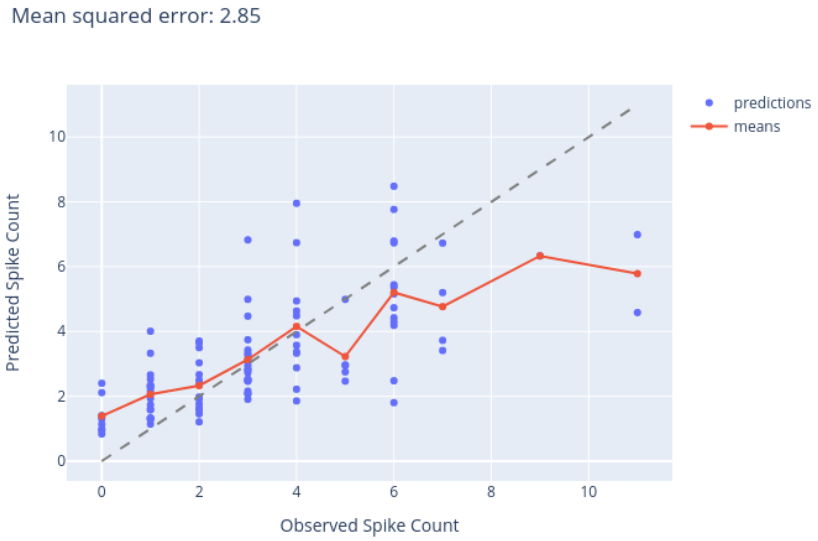
Report worksheet 7

Joaquín Rapela

January 29, 2025

Exercise 1: predict spike counts of one neuron

Figure 1a shows the predictions of the spike counts of neuron 29 from the spike counts of the other neurons. These predictions were not cross-validated, and reflect overfitting. Figure 1b is as Figure 1a but uses cross-validation and avoids overfitting.



(a) non-crossvalidated

(b) crossvalidated

Figure 1: Non-crossvalidated (a) and crossvalidated (b) predictions of response of neuron 29 from the spike counts of the other neurons. Generated with this script using the default parameters for (a) and using the parameter `--predict_on_test_data` for (b). Click on the image to see its interactive version.

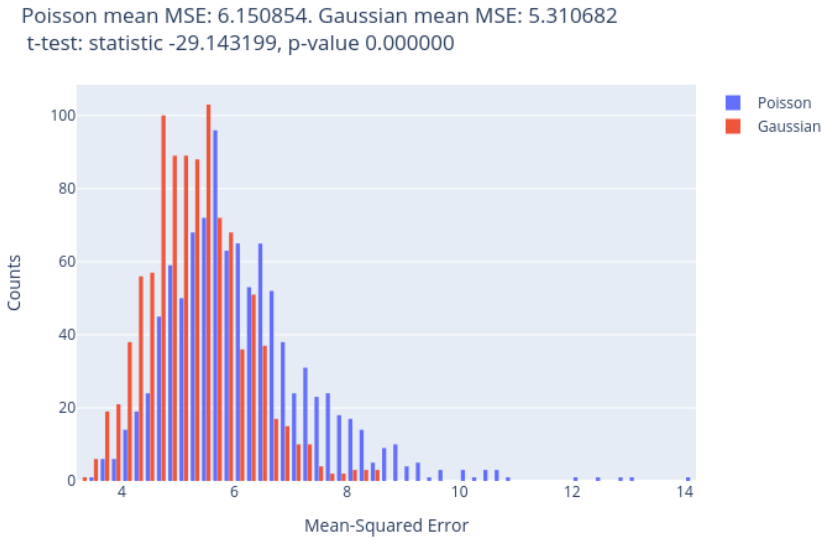
The red line is not straight because for neural responses that are too large (small) the model will tend to predict smaller (larger) responses. If by chance the response of neuron 29 is too large (small), its prediction should be smaller (larger), because too large (small) responses are uncommon and do not contribute much to the optimisation criterion (i.e.,

residuals deviance).

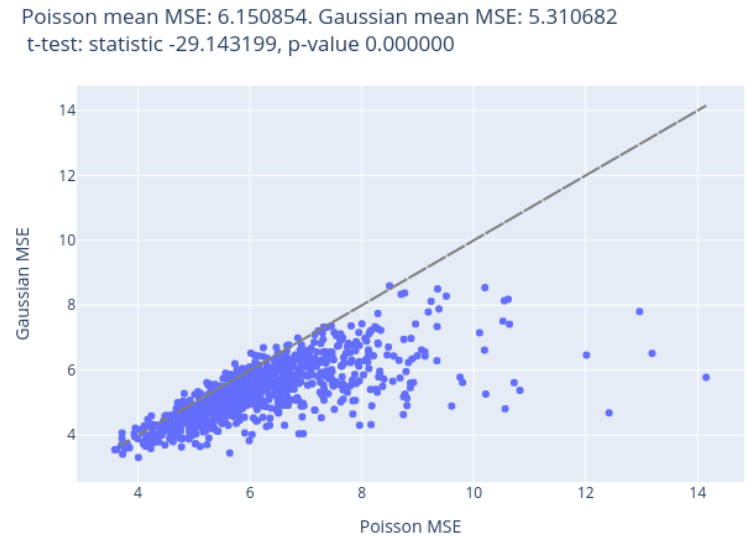
Exercise 2: compare prediction of linear regression models and Poisson GLMs

Figure 2a shows histograms of mean-squared-errors (MSEs) achieved by the Poisson and linear regression (i.e., Gaussian) models. Another visualisation for this type of data is the scatter plot of MSEs of the Gaussian versus the Poisson model shown in Figure 2b.

The title bar of these plots show the statistic and p-values of a paired t-test. In this plot there are two sources of randomness. One source is the random dataset and the other one is the random prediction of the models. Each MSE for the Poisson and Gaussian models were computed from the same randomly generated dataset. Thus, to eliminate the first source of randomness from the significance test, it is better to use a paired than a non-paired t-test. In this case both types of tests yield a significant difference.



(a) Histogram



(b) Scatter plot

Figure 2: Histogram (a) and scatter plot (b) of MSEs from a linear regression model (i.e., Gaussian) versus a Poisson GLM. Generated with this script using its default parameters. Click on the image to see its interactive version.

The linear regression model achieved a lower MSE the Poisson GLM because the former model was optimised to minimise the MSE, while the Poisson model was NOT optimise to minimise this criterion. The Poisson model was optimised to maximise the likelihood of the responses.

Another possibility is that the linear regression model was correctly estimated, but the

Poisson model not. We examine this possibility in Appendix A.

Appendix A Goodness-of-fit assessment

Were 100 trials enough to obtain a good model fit of the linear regression models or Poisson GLMs? If not the conclusion we drew in Exercise 2 could be invalid.

A useful property of linear regression models and of GLMs is that its parameters are interpretable. That is, we can read the parameters of the model and learn about the problem on which the model is applied. However, to be able to make inferences from the model parameters, it is crucial to test that the model was well fitted.

For linear regression there exist a large number of methods to assess their goodness of fit. Most of them are described in detail in Kutner et al. (2005). One should check that the errors of the model are approximately Gaussian, that the variance of the errors is the same for any group of observations (i.e., homoescedasticity), that there are not outliers or influential observations, that the estimated model fits the data better than a model without regressors, that the linearity assumption is justified, etc.

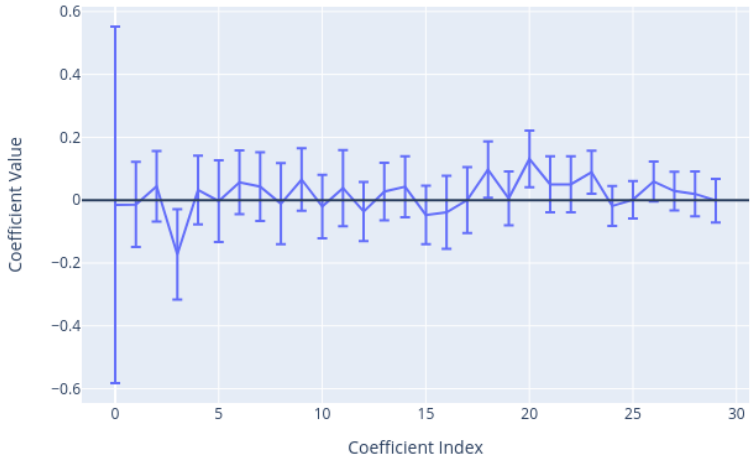
A GLM does not make as many assumptions as a linear regression model. For example, a Poisson GLM allows the variance of the observations to be different for different groups of observations (i.e., for Poisson observations the variance is equal to the mean, and observations with larger mean will have larger variance). Still, with GLMs one should also check for outliers and for nonlinearities.

Here I will only examine the estimated coefficients from the Poisson GLM. If these coefficients are not significantly different from zero, I will conclude that the estimated model is not adequate. But if some of these coefficients are significantly different from zero, I will try to interpret them.

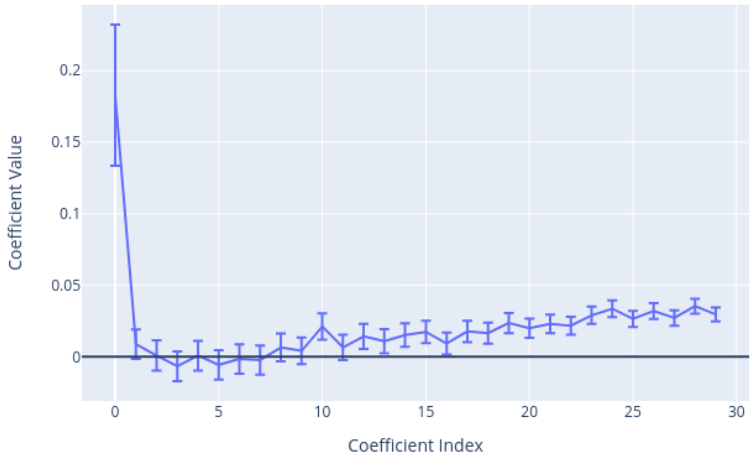
Figure 3a shows the coefficients, and their 95% confidence interval, of a Poisson GLM estimated using 100 trials. We see that the 95% confidence interval of most coefficients includes zero, which suggests that these coefficients are not significantly different from chance.

Figure 3b is the same as Figure 3a but for a model estimated using 10,000 trials. We see that with the larger number of trials the error bars of the coefficients became smaller, and most coefficients are now significantly different from zero. We learn that response of neurons close to the predicted one (neurons 15-28) contribute positively to the logarithm of mean response of the predicted neurons. Also, the closer is a neuron to the predicted one, the larger is its contribution.

It appears that the model fitted with 10,000 trials is better fitted than the one fitted with 100 trials. Will this better fitted model yield worse predictions than a linear regression model? Figure 4 shows that that this is still the case. Moreover, for better fitted models the linear regression model appears to be much better than the Poisson GLM. So, the conclusion we drew in Section 4 holds.

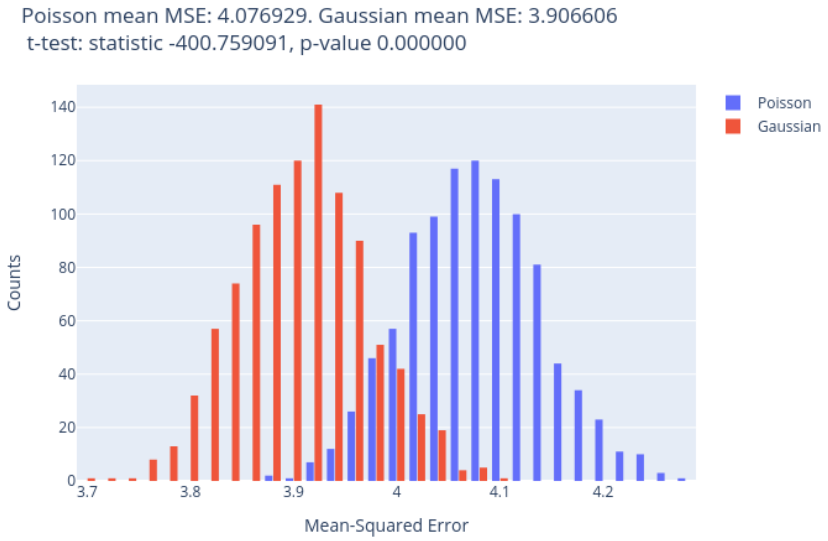


(a) 100 trials.

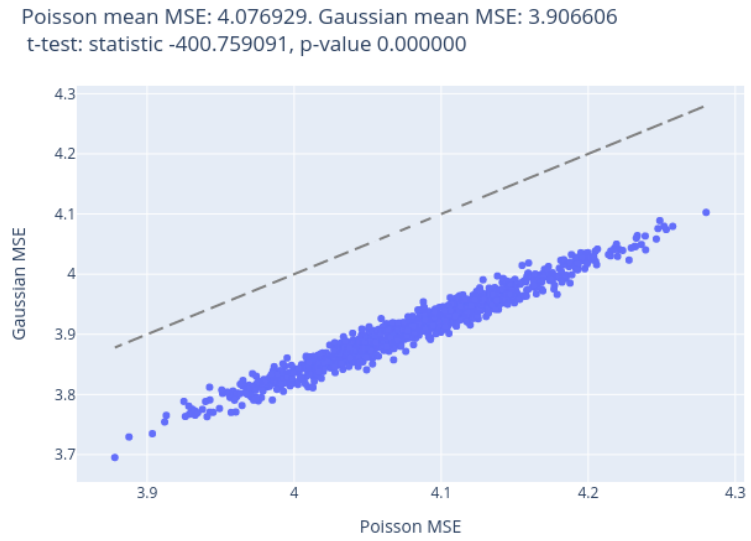


(b) 10,000 trials.

Figure 3: Estimated coefficients of the Poisson GLM fitted to 100 (a) and 10000 (b) trials. Generated with this script using parameter `--n_trials=100` (a) and `--n_trials=10000` (b). Click on the image to see its interactive version.



(a) Histogram



(b) Scatter plot

Figure 4: Histogram (a) and scatter plot (b) of MSEs from a linear regression (i.e., Gaussian) versus a Poisson GLM. Generated with this script using parameter `--n_trials=10000`. Click on the image to see its interactive version.

References

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin, New York, NY, 5th edition.