STAT550 Project Analysis Log

Hailey Ernest

2026-03-14

0. Load dataset (n = 100). Structure, dimensions, and variable types.
1. Summary statistics for continuous variables (falls, balance, strength): means, SDs, ranges.
2. Frequency tables and proportions for binary variables (exercise, gender).
3. Bar chart and histogram of fall counts. Right-skewed with many zeros.
4. Fit Poisson to falls by max likelihood. Estimated rate parameter $\hat{\lambda} = 2.3$
5. Plot empirical vs. theoretical PMF and CDF. Visually mismatched, so Poisson may not fit well.
6. Mean, variance, and variance/mean ratio of falls by exercise group. Ratios > 1, shows overdispersion.
7. Zero falls by exercise group. High number of zeros, possible zero inflation.
8. Create boxplots of falls by exercise group. Exercisers had less falls than non-exercisers.
9. Create boxplots of falls by gender. Difference by gender small, & smaller than by exercise.
10. Plot falls vs. balance by exercise group. Weak positive trend for non-exercisers and little change for exercisers.
11. Plot falls vs. strength by exercise group. Same as 10.
12. Correlation matrix for falls, balance, and strength. Balance & strength uncorrelated, r=0.049.
13. Check VIF from Poisson GLM. Max VIF = 1.064, no multicollinearity concern.
14. Check correlation and VIF for balance and strength. r = 0.049 & VIF < 4, so no centering needed.
15. Model 1 (Poisson): GLM with all predictors, falls ~ exercise + gender + balance + strength.

16. Model 2 (Negative Binomial): Same predictors, allows overdispersion w/ dispersion parameter θ.
17. Model 3 (Quasi-Poisson): GLM with variance $\varphi\mu$ to account for overdispersion.
18. Model 4 (Quasi-μ^2): GLM with variance $\varphi\mu^2$.
19. Model 5 (ZIP): Zero-inflated Poisson with same predictors in count and zero components.
20. Model 6 (ZINB): Zero-inflated Negative Binomial model.
21. Model 7 (Hurdle-Poisson): Binomial hurdle with truncated Poisson count component.
22. Model 8 (Hurdle-NB): Binomial hurdle with truncated NB count component.
23. Compare observed fall count frequencies to model-expected frequencies for Models 1-8. Poisson underpredicted zeros counts, NB/ZIP/ZINB/hurdle models aligned more with observed.
24. Compute deviance/df ratios for Models 1 & 2 as descriptive GoF checks. Model 1 showed a higher deviance, so more lack of fit than Model 2.
25. Simulated fall-count distributions: Models 1 & 2 using fitted means. Poisson simulation created too few zeros and too many middle counts compared to observed data (NB closer).
26. Overdispersion test: Compare Model 1 & Model 2 using a likelihood-ratio test. Significant overdispersion detected.
27. Extract dispersion estimates for Models 3 & 4. Both $\varphi > 1$, agree with overdispersion in fall counts.
28. Compare observed zeros to model-expected zeros across Models 1-8. Observed zeros=39, predictions: Poisson = 30.1, NB = 32.7, ZIP/ZINB/hurdle models predict observed count better.
29. Examine residual diagnostics for Models 1 & 2. NB residuals showed improved dispersion and pattern from Poisson, & Q-Q plots show less deviations in the tails. Studentized residuals showed no extreme outliers.
30. List influential obs. Cook's distance (threshold 4/n). Poisson 12 observations flagged for review, NB 6.
31. Visual Cook's distance for Models 1 & 2 to find potentially extremely influential observations.
32. Review values of influence for flagged observations. Influence moderate but not extreme.
33. Check for extreme influence (Cook's D $\geq$ 1). None found, all observations retained for model comparison.

34. Compare Models 1-8 using AIC/BIC: hurdle-Poisson had the lowest AIC=340.24, and NB had the lowest BIC= 361.41.

35. Model 9: NB as candidate for variable selection. Overdispersion exist and NB had lowest BIC overall. Hurdle-Poisson doesn't model overdispersion & NB is simpler (6 params vs 10).

36. Test exercise:gender, exercise:balance, and exercise:strength interactions in the NB model. None significant (p = 0.756, 0.816, 0.379).

37. Use Type II LR tests on the full NB model. Gender was not a significant predictor, p=0.499.

38. Remove gender and fit Model 9 (NB): falls ~ exercise + balance + strength. LR test showed no significant loss of fit, chi-sq = 0.456, df = 1, p = 0.499, AIC lower than full model.

39. Apply Bonferroni and Benjamini-Hochberg corrections to the final model p-values. All retained predictors remained significant.

40. Check fit of Model 9 residual deviance and deviance/df. Fit was improved from full Poisson.

41. Check overdispersion for Model 9 using odTest. Overdispersion still there, so NB family is fit.

42. Check VIFs for Model 9. No multicollinearity.

43. Examine residual diagnostics for Model 9. No extreme observations are visible, no clear patterns.

44. Check influence for Model 9. Few observations had moderate influence, but no extreme.

45. Check leverage for Model 9 (3p/n). 3 obs. above threshold. Observation 92 high leverage & influence. Obs. not extreme.

46. Check outliers for Model 9 (abs(residuals) > 3). No extreme outliers in model.

47. Obs. 92 showed high leverage and influence. Sensitivity check without obs. gave very similar results. Observation retained.

48. Compute rate ratios (RR = $\exp(\beta)$) and 95% CI's for predictors in the final NB model.

49. Compute predicted fall counts for exercising & non-exercising participants at mean balance and strength values.