

Project Log

1. Fitted initial Poisson loglinear model with all main effects and $\log(\text{units})$ as an offset
2. Performed goodness of fit of model which showed evidence of poor fit
3. Checked model for overdispersion and found strong evidence for overdispersion
4. Ran main effect models for NB2, QL ($\text{var} = \mu$), QL ($\text{var} = \mu^2$) to address overdispersion
5. Compared goodness of fit statistics between the 4 models
6. NB2 and QL($\text{var} = \mu^2$) had lowest X^2 and were very close so I continued with NB2 model (both had evidence on poor fit)
7. Ran stepwise model selection in both directions for NB2 model using AIC but model still had evidence of poor fit
8. Examined plots for studentized residuals, hat values, cooks distances and residual vs fitted plots
9. Found extreme outlier in data with 10 customers from 19 units (obs 11). This is likely a misinput since the other 109 observations all have units > 100 . With this in mind, I felt comfortable with removing obs 11 from the analysis.
10. Performed steps 4-5 again with observation 11 removed
11. NB2 and QL($\text{var} = \mu^2$) had best X^2 test statistics
 - a. QL model seemed to be extremely low
12. QL ($\text{var} = \mu^2$) had evidence of underdispersion so I decided to move forward with NB2 model
 - a. Dispersion parameter = 0.372 which is less than 1
 - b. All residuals w/in 2 SD of 0
 - c. Null deviance = 66 on 108 df which indicates the model fits the data very well with no predictors
13. Performed stepwise model selection in both directions using AIC for NB2 model with obs 11 removed
14. Model contained interaction so examined VIFs and had some > 10 which is very high
15. Centered all predictor variables (not including the offset)
16. Performed stepwise model selection in both directions using AIC for NB2 model with centered predictors and obs 11 removed (same set of predictors was produced)
17. Examined VIFs, highest was 2.363 which is fine
18. Began hypothesis testing on model from stepwise selection to decide on final model
19. Removed interaction since it was not sig at 5% significance level
 - a. Current final model: income, storedist, compdist (all centered)
20. Performed LRT checking for significance of age which was left out. Found age did not significantly improve model
21. Performed LRT comparing final model to model with all 2-way interactions. Found no significant improvement

22. Performed LRT comparing final model to model with higher order squared terms. Found no significant improvement. Some evidence that income^2 was associated with the response.
23. Found no significant improvement so final model was still: income, storedist, compdist (all centered)
24. Performed goodness of fit test showing no sig evidence of the model fitting the data poorly
25. Examined plots for studentized residuals, hat values, and cooks distances for final model
 - a. Found observations 14 and 93 to be have high leverages but not influential
 - b. These obs corresponded to high income neighborhoods and were kept in the model since they were not influential
26. Computed final model predictions