Discussion Summary (02/01)

1. Check whether data include any part-time employees
2. Conduct PPC
   1. Method 1 – Blank all for columns with missing data (more sensitive check)
   2. Method 2 – Blank all for columns with missing data of interest (yr\_1 & yr\_10 salary)

🡺 Plot Distribution of yr\_10 vs. yr\_1

🡺 Compare imputed vs. replicated (split into SES and non-SES)

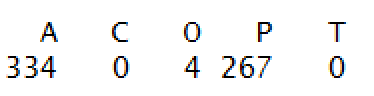
1. Simulation study to evaluate feasibility of logistics regression

Discussion Summary (02/07)

1. Use method 2 for PPC (check each dataset 20 times and include the rest in appendix)
2. Simulation study
   1. Generate a logistic regression model based on true data (P(SES = 1) = pi)
   2. Simulate outcome variable Y based on pi
   3. Refit logistic regression model based on the new simulated outcome variable Y
   4. Compare estimated parameter B with the actual one

Discussion Summary (02/15& 02/16)

1. Refine the population
   1. Combine factor for Grade (ask Alex for dealing with special pay plan)
      1. Combine grade into four bins (1 - 3,4 - 6, 7 - 9, 10 - 12, 13 - 18, Others)
      2. For people whose pay plan are non-ES in year 10 and never get promoted, find the baseline levels for year 10 grade, and toss out those sub-population.
      3. Rethink pay\_plan & grade
   2. Toss out Clerical and Technical occupation category



* 1. Toss out Hawaiian & Pacific Islander (none of them are promoted)
  2. Special occupation (AD) has very high salary distribution, which produces bias for the model
  3. Toss out people whose pay plan = ES in year 1

1. Compare the simulated data Beta with actual data

🡺 Pretty similar

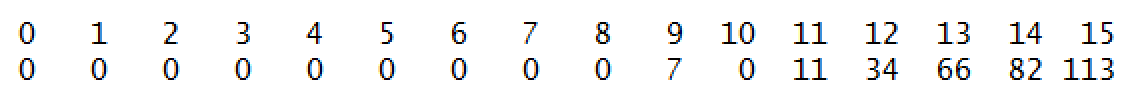
1. Redefine Estimated Start Year

Discussion Summary (02/22)

1. Add variable svsrstat (indicator of supervisor/manager)
   1. Remove those who got promoted in the first 10 years
   2. Remove estimated start year as a predictor
2. EDA
   1. Binned plot to check the relationship between percentage of each group who get promoted given independent variable
   2. Possibility to apply spline to continuous variable
3. Model fit
   1. Binned residuals to check model fit
   2. Hosmer-Lemeshow goodness of fit test

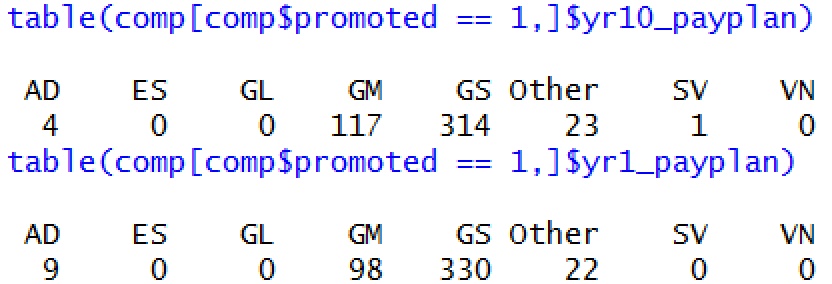
Discussion Summary

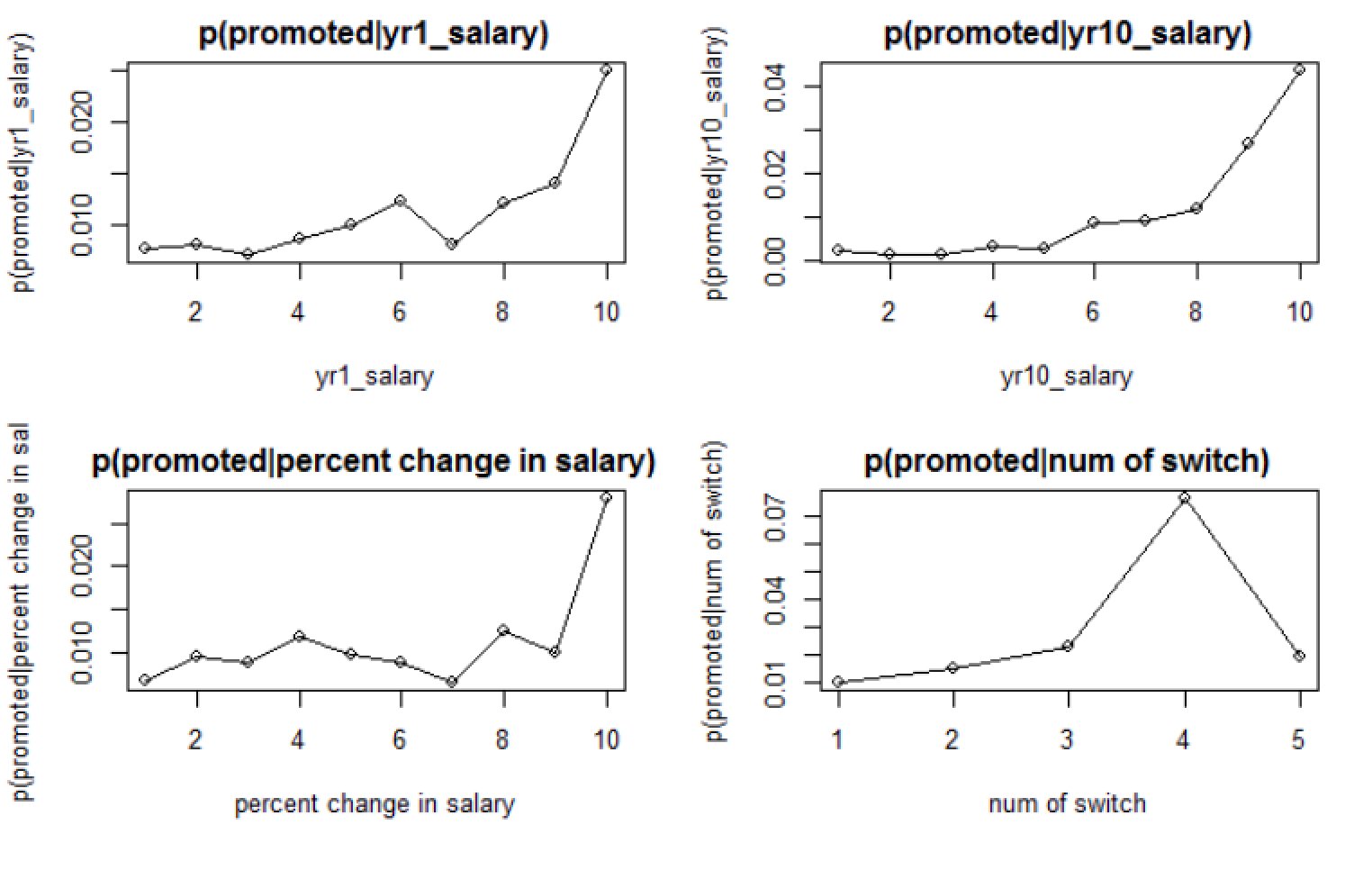
1. Remove people who were promoted into SES btw yr1 – yr10 and did re-imputation



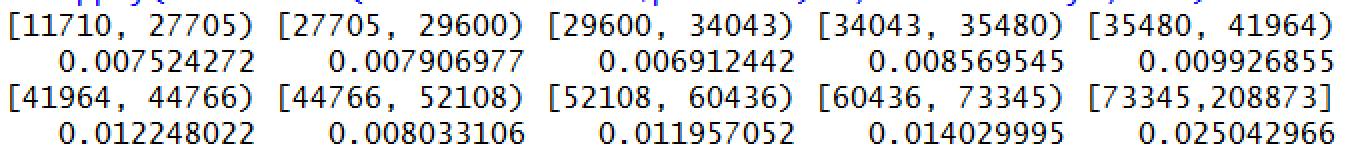
Distribution of grade with GS pay plan in year 10 for those who got promoted

🡺 shows that we can toss out people whose grade is below 9 with GS pay plan in year 10

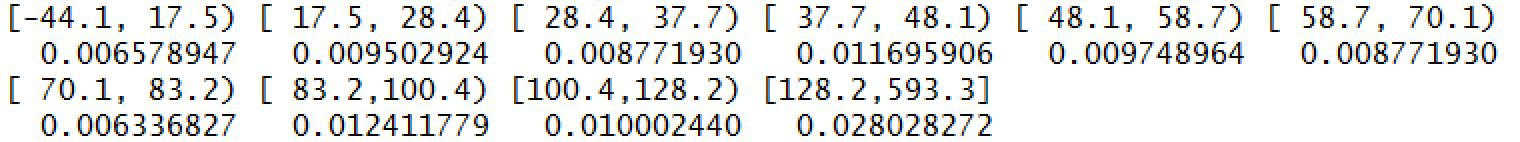
1. Combine AD, GL, VN, and SN to other 🡺 GS, GM, Other
2. Combine service status from 8 categories to supervisor vs. non-supervisor



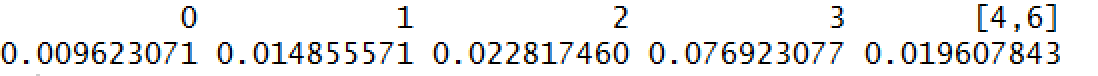
yr1 salary: cubic spline through variable selection



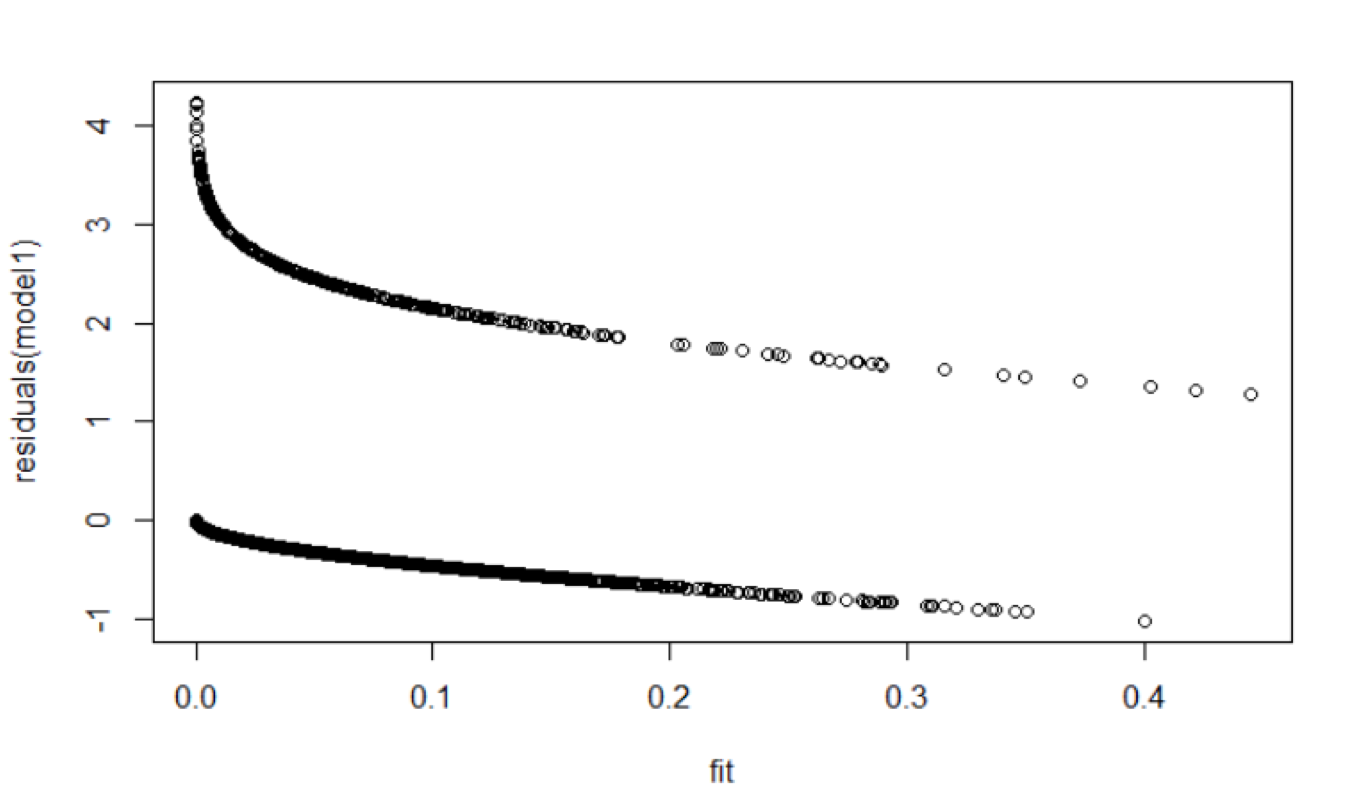
yr10 salary: y10\_salary + yr10\_salary^2



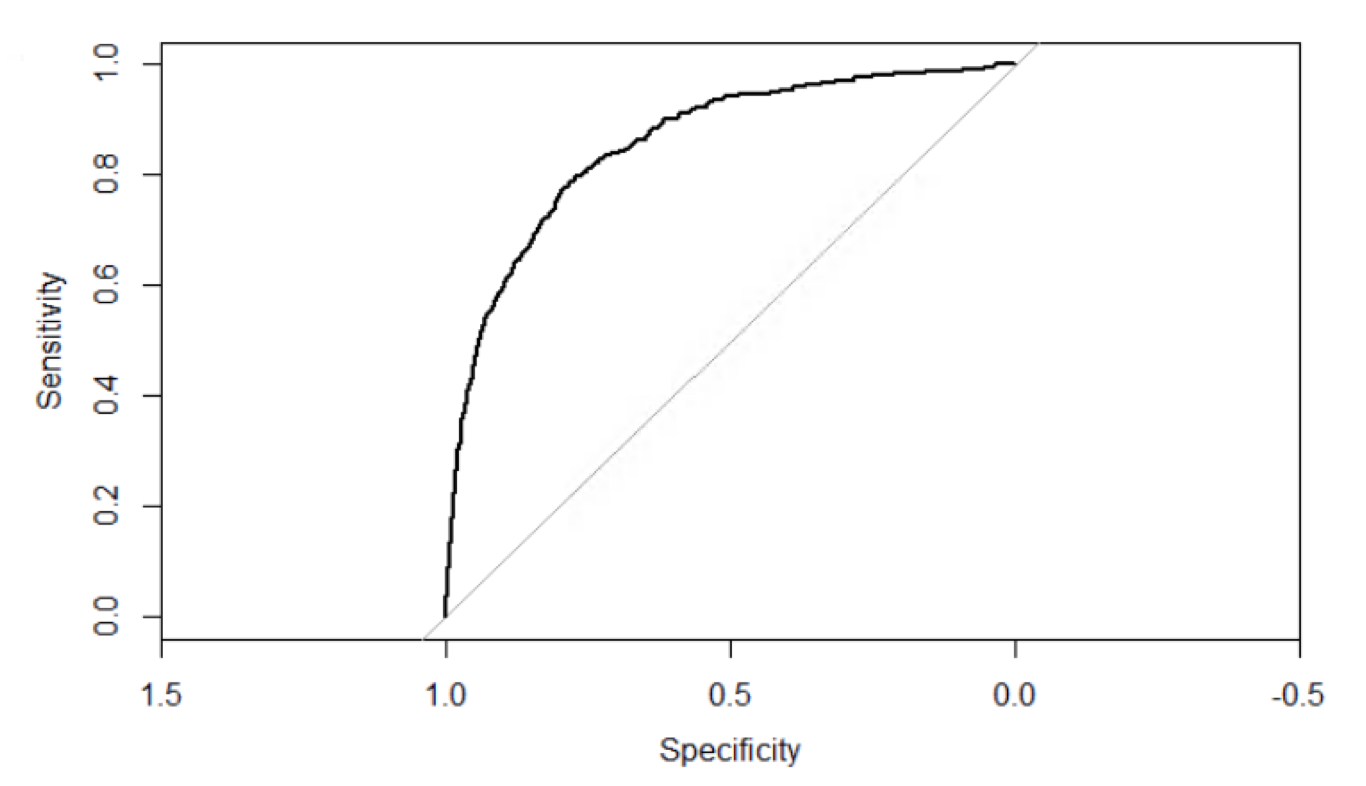
Num\_of\_switch: linear splines with two knots (0-2, 2-3, 3-6)



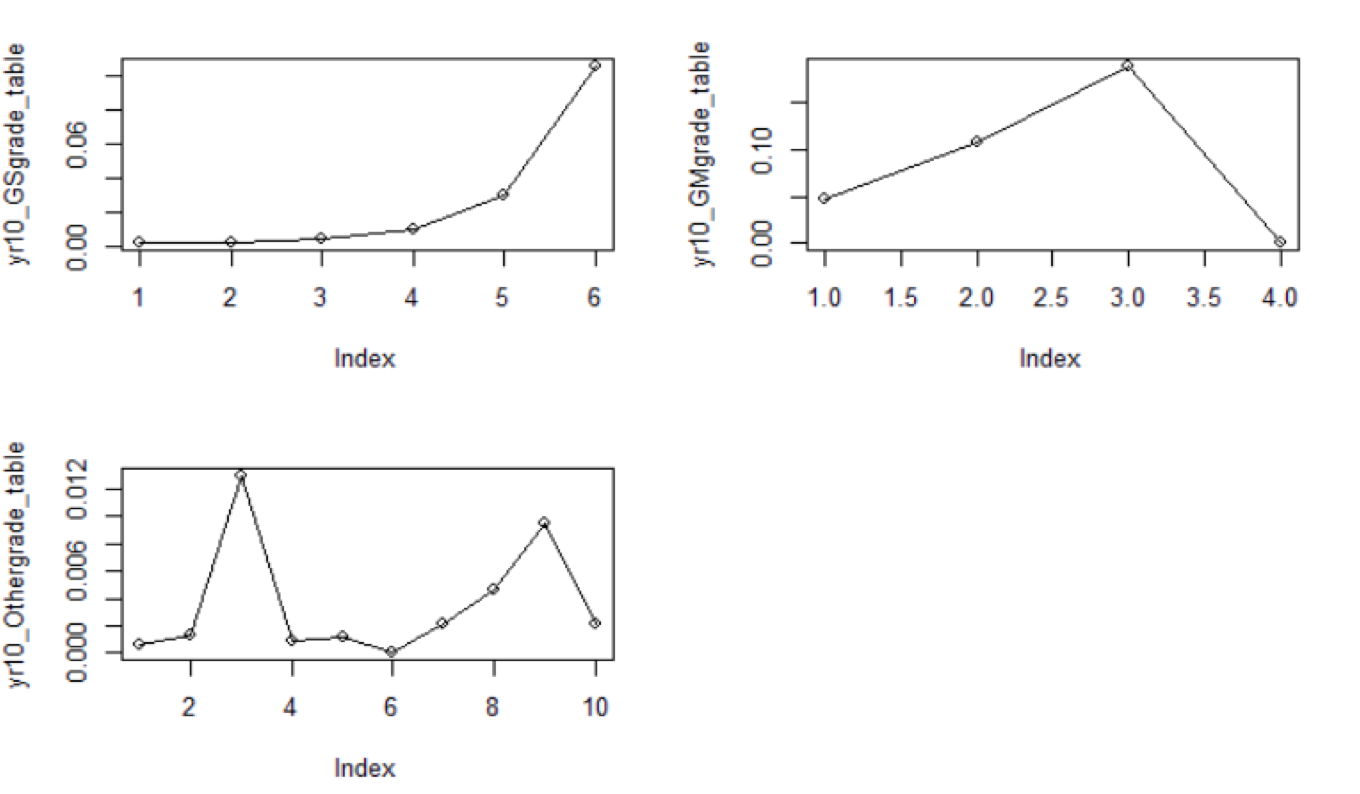
1. Variable selection techniques for knots
2. Model Diagnostic
   1. Hosmer-Lemeshow test
      1. Chi-square statistic is small 🡺 model is prob not a good fit for the data
   2. Deviance Residual plot



* 1. ROC
     1. Very High AUC: 0.875
     2. ROC curve suggest that the model has good predictive probability. However, it is ROC on the entire dataset instead of testing data.



1. Make Grade into numeric and apply smoothing-splines mixed-effect model based on three pay plans



* + - SME packages
    - Nelder –Mead simplex search to find the optimal smoothing parameters
    - Use BICn (Bayesian information criterion across the number of subjects in each group) as penalty for over-smoothing

Discussion (03/22)

1. Correct change & salary baseline (comment that % change in sensitive to imputation model) (DONE)
2. Scale salary 🡺 (yr10\_salary – mean(yr10\_salary where yr10\_grade = 13))/10000

🡺 Pick a constant number for knots (DONE)

1. Model Diagnostics (residual binnedplot). Other techniques for categorical variable (Residuals for all categorical variables are small)? Not really
2. Analysis
   1. Predictive probability
   2. Odds