LinReg\_Project

Will Mohr

2024-11-12

In this report I examine the data from the National Practitioner Data Bank, found at this url: <https://www.npdb.hrsa.gov/resources/publicData.jsp>

The purpose of the report is to determine the extent to which several general characteristics of a case that an insurance company is exposed to can be used to predict the size of a payment in the case that a payment does occur. This can combined with an additional estimate of the probability of a payment to help determine appropriate reserves.

The final model predicts the log of the payment based on the practitioner’s year of graduation from school, the year of the incident, the number of years between the occurrence of the incident and payment of the claim, the state of the practitioner, the licence field of the practitioner, the general and specific nature of the negligence that led to the claim, and the gender of the patient.

Initial transformations of the data involved trending the payments according to urban average cpi. The cpi’s were arithmetically averaged for each year according to the months for which data existed in the practitioner database and a factor was determined to on-level the payments to the 2024 October CPI.

Investigation of the data revealed that some variables did not have entries for data before 1/30/2004, or around half the entries that involved payments. For a variable such as PTGENDGER (Patient Gender), these NA values were replaced with a factor such as the character “NA”. For numeric variables, such as PTAGE (Patient Age), I chose to exclude the variable from further analysis instead of removing data from before that cutoff date from the analysis. This resulted in a feature category that in essence indicates the absense of knowledge about that variable but indicates that the date of the payment was prior to 1/30/2004.

The modelling process went as follows. First, since the outcome variable is monetary, a right-tailed distribution could be expected, which I verified first by plotting un-transformed payments compared to transformed payments.

This plots revealed that the optimal transformation may involve a Box-Cox transformation somewhere between log (lambda=0) and the sixth root (lambda = 1/6), with hope remaining that a log transformation would suffice.

I made some adjustments to the categorical predictor variables, combining categories with under 1000 occurrences into a miscellaneous category for all the categorical variables. I also created a DEV\_YEARS variable from differencing ORIGYEAR (year of payment) and MALYEAR1 (year of incident occurrence). Since this is a post-hoc variable only known at time of payment, it may be of minimal use for reserving at the time of an incident if predictions of DEV\_YEARS are unreliable. However, it may be useful for updating reserves for cases from prior years.

I then looked at the numeric variables for correlations and non-linearity. Most notably, there was some co-linearity between GRAD (Graduation Year) and MALYEAR1 (Year of Malpractice) as well as between MALYEAR1 and DEV\_YEARS

While I did not have much concerns about over fitting the data with all the categories for each variable having over 1000 observations, to increase the speed of model fitting and allow for a pure estimates of predictive power I split the dataset in half with equal allocations to train and test sets. I utilized the fastDummies package to quickly generate dummy variables for all of my categorical variables.

The naive approach to modelling resulted in a very poor r-squared of 0.1200508.

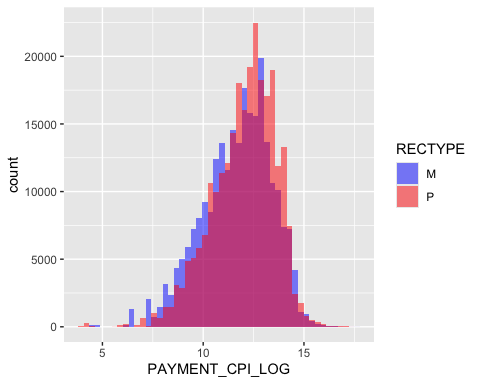
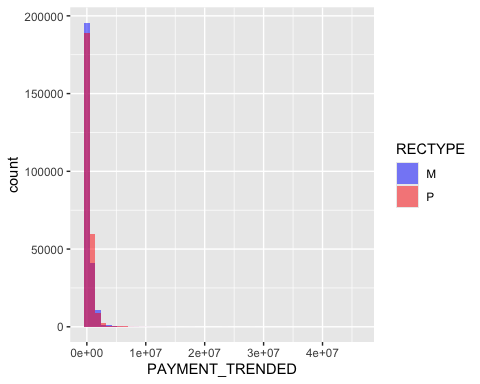
I then attempted stepwise feature selection, but the processing requirements proved too much for my local machine and so I proceeded with “lasso” regularization. This adds a penalty to the error function that is proportional to the number of parameters. This is preferrable to “ridge” regression for feature selection, as ridge operates by shrinking coefficients to near zero but not quite zero. I opted for the “1 se” aproach in selecting how severely to penalize the coefficients, as this reseulted in the removal of nearly 50 variables from

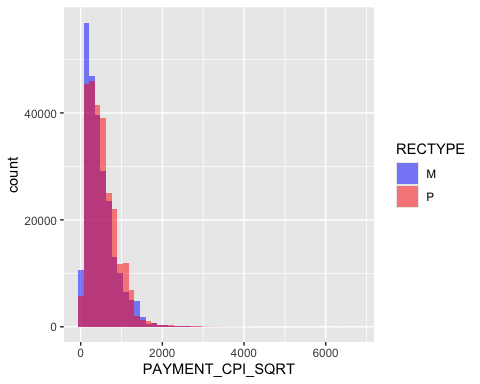
With this reduced feature set, I was able to then run backwards stepsise selection and determined that no additional parameters should be removed.

The final adjusted r-Squared as calculated by 1- SSE/(n-k-1)/(SSTO/(n-1)) was around 32%, a significant improvement over the baseline model and perhaps indicating that there is a place for computational reserving as a complement to the estimations made by claims personnel.

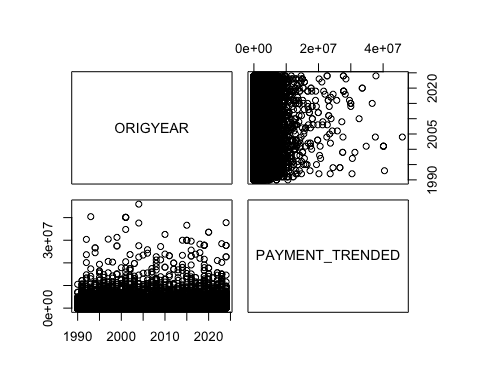
The ease of interpretability of the final model could potentially be improved by determining if further pruning of the feature space doesn’t significantly impair the predictive value of the model. Also, I could examine the STATE variable to determine if broader geographical categorization is sufficient (e.g. Pacific Northwest, New England, Southwest, etc.) Also, I would need to determine what are the baseline features included in the intercept of the model as well as document all the variables that were grouped under “miscelaneous” categories.

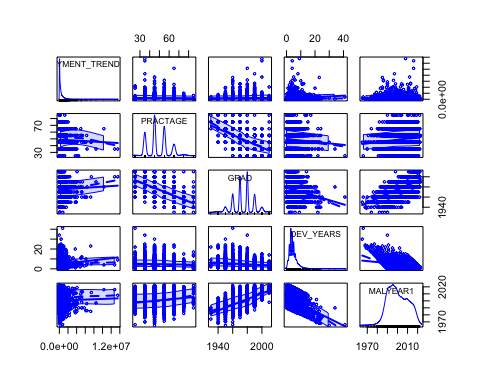
Of note when interpreting the model is that since the outcome variable is log-transformed, the final predictor variables are multiplicative. So taking the exponent, base e, of a coefficient determines the factor by which the predicted untransformed payment will change with all other variables constant.

A graph of a graph

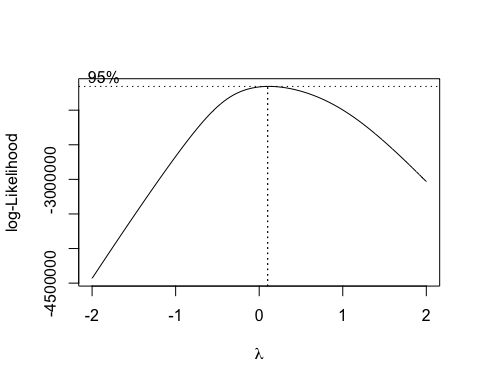
Description automatically generated

A Log box-cox transform may not be quite optimal based on the distributions above. Sixth root looks more normal, if less interpretable. We also see that record type “M” and type “P” payment distributions are similar.



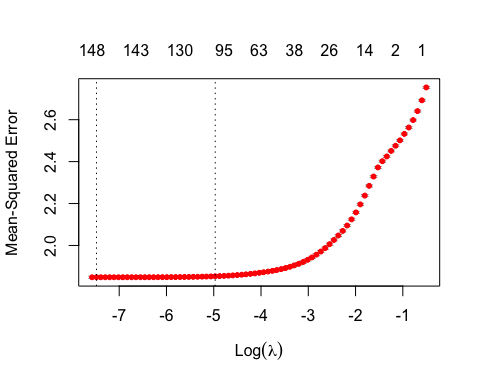


Above it is aparrent that some of the time-based features show a degree of multi-collinearity.

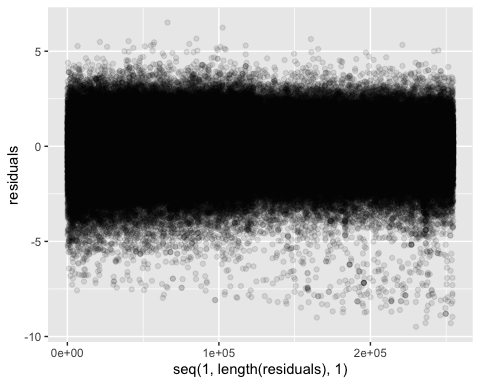


The adjusted r-squared of the un-transformed, full model is very poor at only 0.1200508.

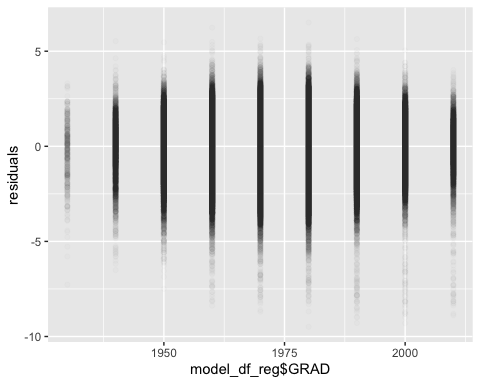
Above, we see that the optimal model has an approximately log-transformed response. This is expected for right-skewed data often seen with monetary outcomes. For ease of interpretability, this is the chosen model moving foreward.



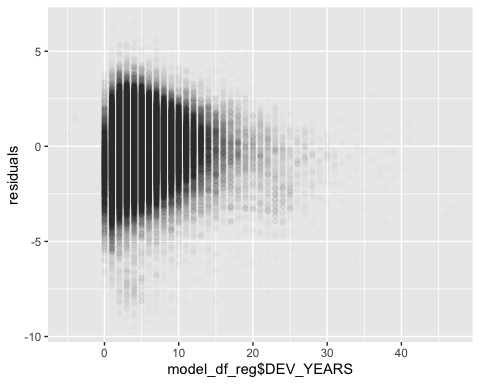
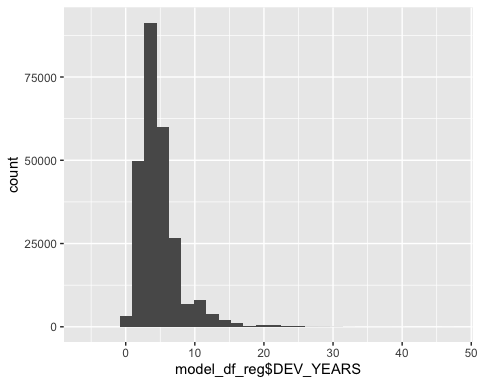
Backwards stepwise regression does not result further reduced AIC compared to Lasso Regression. A lambda resulting in a MSE 1 standard error above the minimum was selected for the feature reduction that results.

Pairwise residual analysis 

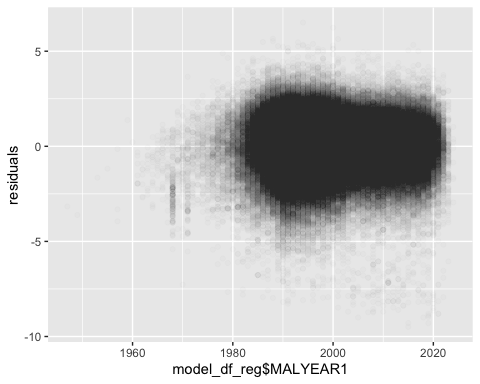
There is no clear heteroskedasticity from analysis of the residuals.



The apparent slight heteroskedasticity with respect to the graduation year is not overly concerning, and could in part be related to the small sample sizes of graduates from earlier decades leading to fewer large residuals. We may need to be careful about extrapolating the results to extremely recent graduates.



Comparing the histogram of dev\_years to the residuals plot, it appears that we can attribute the apparent heteroskedasticity to the underlying distribution of development years. However, futher modelling efforts might involve a Box-Cox transformation of this variable.

 From the above, we can see that there is no obvious non-linearity in the residuals with respect to their order in the database. From this view there may be a slight increase in the left-skew of the residuals over time.

There also is no clear non-linearity as a function of development years. The apparent heteroskedasticity reflects the histogram of the distribution by development years and so is not of immediate concern.

The significantly trimmed model is able to still explain ~32.8 percent of the variance in logged payments on adjusted basis for the r-squared.

The complete set of the coefficients is as follows:

Log(PAYMENT\_TRENDED) ~

(Intercept) GRAD DEV\_YEARS MALYEAR1 WORKSTAT\_AZ WORKSTAT\_CA WORKSTAT\_GA

-25.255165289 0.000800124 0.042762108 0.017660450 0.086123767 -0.405934032 0.183022039

WORKSTAT\_IL WORKSTAT\_IN WORKSTAT\_LA WORKSTAT\_MA WORKSTAT\_MD WORKSTAT\_MI WORKSTAT\_NJ

0.449989062 -0.424555508 -0.326687964 0.413861754 0.117139289 -0.412756121 0.220201652

WORKSTAT\_NY WORKSTAT\_Other WORKSTAT\_PA WORKSTAT\_PR WORKSTAT\_TX WORKSTAT\_VA WORKSTAT\_WA

0.295709599 -0.056789379 0.204626437 -1.476394568 -0.144817971 0.127764176 -0.069319007

LICNFELD\_10 LICNFELD\_15 LICNFELD\_20 LICNFELD\_30 LICNFELD\_50 LICNFELD\_110 LICNFELD\_120

0.462457702 0.343006583 0.376516498 -0.848723395 -1.292317958 0.232883088 0.347214279

LICNFELD\_140 LICNFELD\_430 LICNFELD\_603 LICNFELD\_636 LICNFELD\_99999 ALGNNATR\_1 ALGNNATR\_30

-0.250437447 -0.459671409 -0.442784384 -0.259592483 -0.089318193 0.089302369 -0.330745427

ALGNNATR\_40 ALGNNATR\_50 ALGNNATR\_60 ALGNNATR\_70 ALGNNATR\_80 ALGNNATR\_90 ALGNNATR\_100

-0.284960307 0.388480108 -0.224399250 -0.162290687 -0.734006391 -0.432268070 -0.175509268

ALEGATN2\_0 ALEGATN2\_104 ALEGATN2\_200 ALEGATN2\_201 ALEGATN2\_203 ALEGATN2\_305 ALEGATN2\_322

-0.107589894 0.266427886 0.158715504 0.253357829 0.205221113 0.146853734 -0.597271962

ALEGATN2\_330 ALEGATN2\_706 ALEGATN2\_707 ALEGATN1\_104 ALEGATN1\_106 ALEGATN1\_112 ALEGATN1\_201

-0.440714993 -0.807234687 -0.247157962 0.305566630 -0.077899074 0.149310038 0.253225009

ALEGATN1\_202 ALEGATN1\_203 ALEGATN1\_303 ALEGATN1\_304 ALEGATN1\_305 ALEGATN1\_306 ALEGATN1\_311

0.087406444 0.211983034 -0.108600830 0.179322531 0.130477942 0.120384528 -0.260356673

ALEGATN1\_312 ALEGATN1\_318 ALEGATN1\_319 ALEGATN1\_320 ALEGATN1\_322 ALEGATN1\_327 ALEGATN1\_329

-0.340383900 0.290972795 -0.477679196 0.214437151 -0.859611667 -0.221135045 -0.476465972

ALEGATN1\_330 ALEGATN1\_334 ALEGATN1\_400 ALEGATN1\_404 ALEGATN1\_500 ALEGATN1\_504 ALEGATN1\_601

-0.657964212 -0.176855344 0.279007771 -0.160328802 -0.194261844 -0.332334513 0.180786551

ALEGATN1\_706 ALEGATN1\_707 ALEGATN1\_999 ALEGATN1\_9999 PTGENDER\_U OUTCOME\_1 OUTCOME\_2

-0.318574572 -0.318591478 -0.271017429 -0.070042677 0.324667919 -0.924682718 -1.506310114

OUTCOME\_3 OUTCOME\_4 OUTCOME\_5 OUTCOME\_6 OUTCOME\_7 OUTCOME\_8 OUTCOME\_9

-0.960247945 -0.208420316 -0.249868341 0.372517126 0.641328155 0.959596774 0.187478681

OUTCOME\_10

-0.510484203