Predicting Income and Attrition

An analysis of the IBM HR Analytics Employee dataset

Nathan England (nle4bz), Jon Gomez (jag2j), Michael Langmayr (ml8vp), and Yihnew Eshetu (yte9pc)

Layman's Summary

Objectives

We worked with the IBM HR Analytics Employee Attrition & Performance dataset. The data consider both professional (e.g., job performance) and personal (e.g., marital status) features. We were interested in answering two specific questions:

How can we predict monthly income?

How can we predict whether an employee is likely to leave the company?

The Data

The data consists of employee information crafted by data scientists at IBM to be used in developing analytical models[1]. In particular, it deals with variables that could be realistically measured and features a variety of employees with heterogeneous qualities, and the kinds of data include both continuous quantities and discrete factor variables. Most factors primarily take the form of scaled response that could be hypothetically obtained using surveys. For example, Job Satisfaction ranges from 1 (Low) to 4 (Very

[1] IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

High). A few fields, such as EducationField, could be free text, but given the small number of unique value in the data, we treat as categorical factors.

Outcomes

For the *attrition* predictions, we recommend the following:

- For the general population, we provide two models that can classify whether or not an employee is going to leave the company with an accuracy of ~80-82%.
- Our approach is premised on the idea that we prefer to have a classifier with higher sensitivity in predicting whether or not an employee is going to leave.

For *monthly income* predictions, we recommend the following

- For the general population, we provide two models, one that maps the job level to constant scores and one that provides a more sophisticated result. These are not suitable, however, to answer hypothetical questions.
- For specific subsets (job levels 3 and 4), we provide two models that provide more granular results and which can be used to pose and answer hypothetical questions about the data.

Significance of Results

These results provide powerful tools. The attrition model provides a way for employers to assess whether a high-performing employee might be contemplating a change in job, providing an opportunity to find ways to retain the employee. The monthly income models provide ways to examine and predict what factors most contribute to differences in management choices about rewards to employees.

Data Cleanup

We started by identifying continuous quantities and factors. We also looked at histograms and summary statistics using factor plots and R’s *summary* function. From this analysis, we found constant-value or irrelevant variables, which we excluded from further consideration, labelling them as columns to drop. Based on this analysis, used the tidyverse[2] library to apply the types when loading the data. We also set reference levels to minimum values where factors appeared to be on an ordered scale.

Models

1. The Attrition Model

We already knew from our data investigation that the data included the binary field Attrition, coded as “Yes” indicating the employee left the company, or “No” indicating that the employee did not leave the company. Having this knowledge, we decided to build a logistic regression model in an attempt to see if we could accurately predict whether or not an employee was likely to leave the company. At the start, we knew Attrition would be our binary response variable but needed to decide which of the 31 predictors to include in the model. We first checked to see if any of the binary or quantitative variables were correlated with Attrition but ended up finding that none had a strong correlation with Attrition.

[2] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

From there, we decided that using an automated search procedure utilizing the stepAIC()[3] function would be an easy way of reducing the number of predictors to include in our models. Before beginning this process, we started with a null model only including the intercept and a full model including all of the potential predictors, seeing as we knew these would be necessary inorder to run forward selection, backward elimination, and stepwise regression automated selection procedures.

All three procedures use AIC as a form of model selection criteria in the variable selection process which we thought would be appropriate considering AIC is suitable for more complicated modeling methods such as generalized linear models and adding additional regressors does not automatically improve the model as does in the case of R^2. We first tried to build a simple model using forward selection. The forward selection model begins with just the intercept (i.e the null model) and then adds the independent variables one by one to improve the model based on the AIC. However in our case, the procedure produced a model that only included the intercept, thus we determined this would not be a good model to fit to the data.

We then tried using backwards elimination. The backward elimination model begins by including all the independent variables (i.e. the full model) and each one is deleted one at a time if they do not improve the model's AIC. In our case the procedure dropped 9 predictors from the full model leaving us with a model with 22 predictors. Finally, we used stepwise regression which is a modification of forward selection where after each step in which a variable was added, the procedure calculates the AIC and

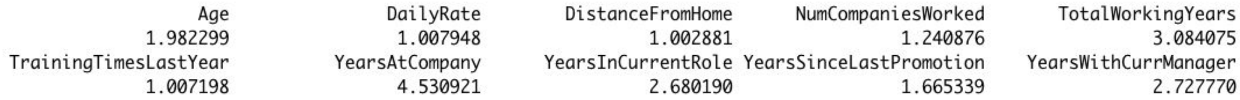
[3] From the MASS library: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

adds or removes the predictor if it ends up lowering the AIC until you are left with the model with the lowest possible AIC given the predictors. This procedure left us with the exact same model as the backwards elimination process. We therefore determined that this stepwise model would be a good candidate to fit to the data.

Stepwise AIC based model:

- Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome + EducationField + EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager

Following this process, we thought that 22 predictors was still a lot to work with and considered if any of the quantitative variables, and we calculated their VIF scores to check for any instances of multicollinearity. All quantitative variables had VIF scores of less than 5, so we decided to keep all 22 of the predictors in the model.



```
                    Age               DailyRate        DistanceFromHome      NumCompaniesWorked       TotalWorkingYears
               1.982299                1.007948                1.002881                1.240876                3.084075
  TrainingTimesLastYear          YearsAtCompany      YearsInCurrentRole YearsSinceLastPromotion    YearsWithCurrManager
               1.007198                4.530921                2.680190                1.665339                2.727770
```

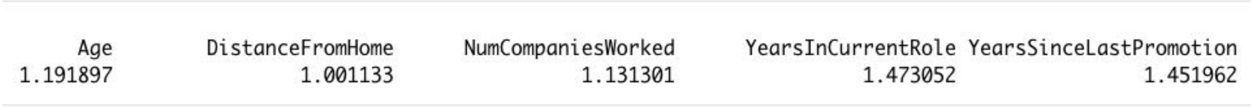
From there we decided we would like to assess at least one other model based on a different model selection criterion. We reused our procedure, substituting BIC as the model selection criteria. We thought this was appropriate since BIC, like AIC, is more suitable for more complicated modeling methods such as generalized linear models and penalizes additional regressors more harshly than AIC. In particular, we hoped for fewer predictors using this criterion.

The forward selection again produced a model with only the intercept, and the backwards selection and stepwise methods again produced equivalent models as each other. However, the backwards and stepwise procedures dropped 18 predictors from the full model leaving us with a model with 13 predictors. We therefore determined that this stepwise model would also be a good candidate to fit to the data.

Stepwise BIC based model:

- Attrition ~ Age + BusinessTravel + Department + DistanceFromHome + EnvironmentSatisfaction + JobInvolvement + JobLevel + JobSatisfaction + NumCompaniesWorked + OverTime + StockOptionLevel + YearsInCurrentRole + YearsSinceLastPromotion

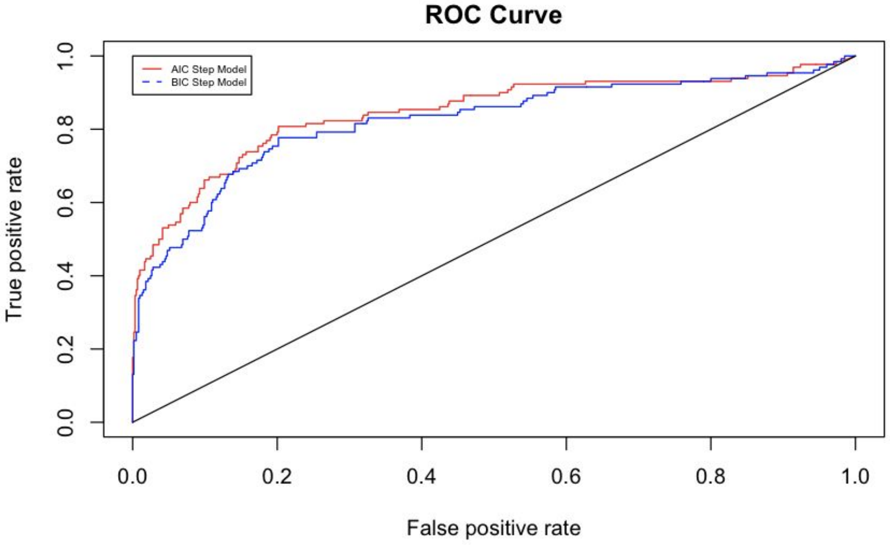
Again we checked to see if any of the quantitative variables were collinearly related so we calculated their VIF scores to check for any instances of multicollinearity. All quantitative variables had VIF scores of less than 5, so we decided to keep all 13 of the predictors in the model.



```
     Age        DistanceFromHome       NumCompaniesWorked       YearsInCurrentRole YearsSinceLastPromotion
1.191897                1.001133                 1.131301                 1.473052                1.451962
```

We then performed cross validation to assess the predictive capability of our models. We randomly divided the data into two samples of equal size to create our training and test data. We then fit both models under consideration with the training data and used these fitted models to predict the responses over the test set data. Finally, we calculated the values for the confusion matrices, plotted the ROC curves, and calculated the area under the curve (AUC) for each of the two models.

The ROC curve plots true positive rate against the false positive rate for varying cutoff values. If the logistic regression models do no better than random guessing, then the true positive rate is equal to the false positive rate and both models will be plotted on average along the diagonal line.

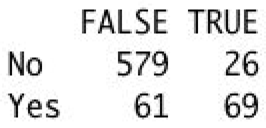


Since the ROC curves for both models lie on the top-left side of the diagonal line, we conclude that that both of our logistic regression models do better than random guessing. The red curve represents the AIC-based stepwise model while the blue line represents the BIC-based stepwise model. The red curve is slightly closer to the point (0, 1) than the blue line suggesting that our AIC-based model is a stronger classifier than our BIC-based model.

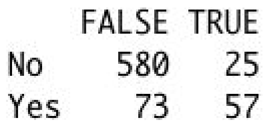
This conclusion is further confirmed when looking at the AUC of both models. The AUC of the AIC-based model was `0.8456834` while the AUC of the BIC-based model was `0.8217673`. Again both models got AUC scores above 0.5 so we can conclude that they classify better than random guessing but have scores less than 1 and are thus not perfect classifiers. Since the AUC of the AIC-based model is higher than that of BIC-based model, this suggests that the AIC-based model is a stronger classifier than the BIC model.

In order to fully evaluate the performance of our logistic regression models we had to look at them in terms of how well they did in predicting outcome, so we began by creating a confusion matrix for each of them using the default cutoff value of 0.5.

AIC-Based Model Confusion Matrix   **BIC-Based Model Confusion Matrix**



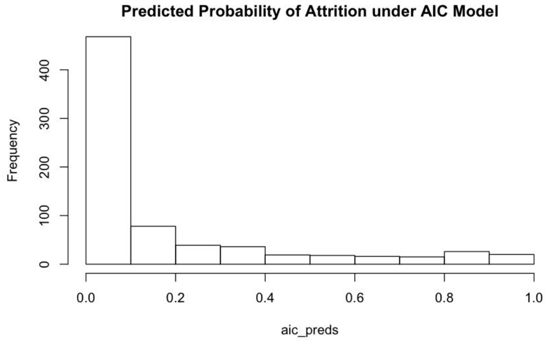
```
     FALSE TRUE
No     579   26
Yes     61   69
```

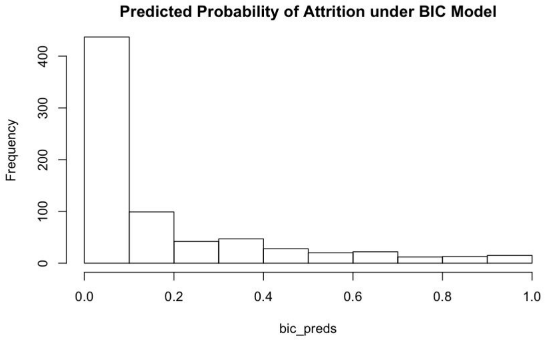


```
     FALSE TRUE
No     580   25
Yes     73   57
```

From these matrices we see that the accuracy of our AIC-based model is 0.8816327 with a misclassification rate of 0.1183673. This model also has a sensitivity/recall value of (true positive rate) 0.5307692 meaning that when the true attrition value is “Yes” it predicts that value as “Yes” 53% of the time. This model has a precision value of 0.7263158 meaning that when the predicted value of the model is “Yes”, it is correct 73% of the time.

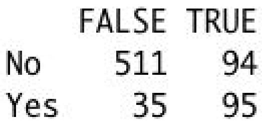
The confusion matrix of the BIC-based model shows an accuracy of 0.8666667 with a misclassification rate of 0.133333. The recall of this model is 0.4384615 and the precision of this model is 0.6951220. This goes agrees with the previously found conclusions surrounding the models and further reaffirms that the AIC based model is a more accurate model than the BIC based model when fit to the data. Lastly we investigated if the cutoff value was appropriate being set at 0.5 so we decided to plot a histogram of predicted probability of being classified as leaving the company to determine where to draw the cutoffs.



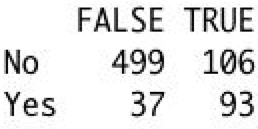


Both models suggest a cutoff of 0.2 might be more appropriate considering that attrition rates are typically low. Under the 0.2 the cutoff value the models had the following confusion matrices.

AIC-Based Model Confusion matrix **BIC-Based Model Confusion Matrix**



```
     FALSE TRUE
No     511   94
Yes     35   95
```



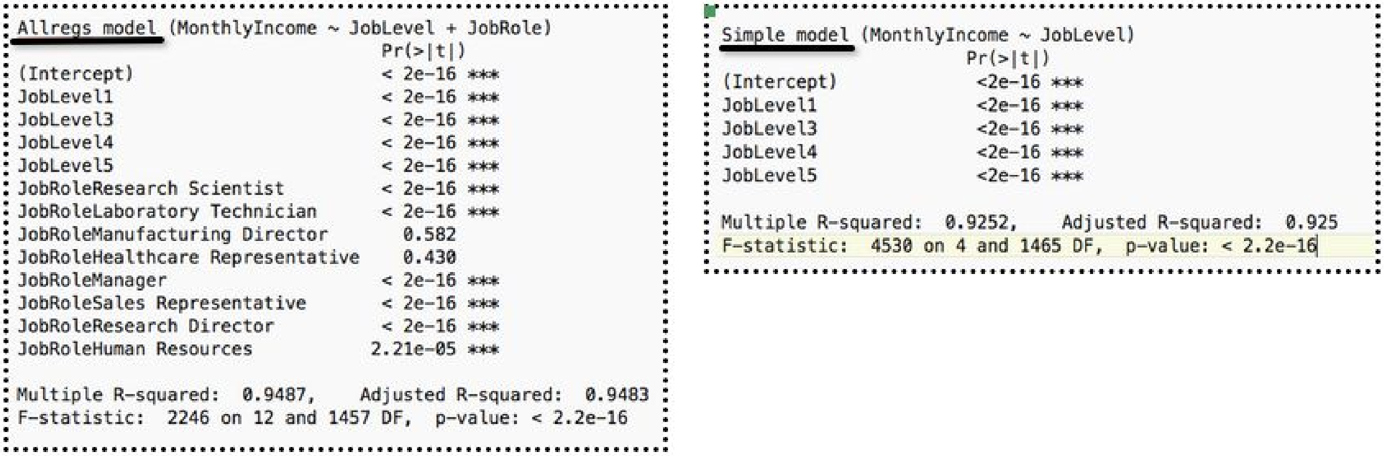
```
     FALSE TRUE
No     499  106
Yes     37   93
```

Under the new cutoff value of 0.2, the accuracy of the AIC model is `0.8244898,` a misclassification rate of `0.1755102,` a recall value of `0.7307692,` and a precision value of `0.5026455.` Under the new cutoff value of 0.2, the accuracy of the BIC model is `0.8054422,` a misclassification rate of `0.1945578,` a recall value of `0.7153845,` and a precision value of `0.4673367.` While this new cutoff value leads to a lower accuracy in both models, it increases the recall of the AIC model and decreases the recall of the BIC model, but decreases the AIC model's precision and increases the BIC model's precision. While we sacrifice some accuracy under this new cutoff, this cutoff is preferable under the AIC model because we want a model with high recall so that it over-classifies instances of attrition, because in a workplace it is preferable to classify someone is going to leave and they do not, rather than classifying someone as staying, when they in fact do leave. For this same reason, it is preferable to use a cutoff value of 0.2 under the BIC model.

2. The Income Model

Developing the recommended models

We started by looking at high-correlation pairs of data, and we found that monthly income and job level have a Pearson correlation of 0.76. Using this baseline, we set the *simple model* to be Monthly ~ JobLevel. We then performed exhaustive selection and investigated the best model for each number of predictors, up to thirty predictors. We found every criterion was optimal[4] for the same model, MonthlyIncome ~ JobLevel + JobRole. We called this the *allreg model*. The summary output for models was very



```
Allregs model (MonthlyIncome ~ JobLevel + JobRole)
                                  Pr(>|t|)
(Intercept)                       < 2e-16 ***
JobLevel1                         < 2e-16 ***
JobLevel3                         < 2e-16 ***
JobLevel4                         < 2e-16 ***
JobLevel5                         < 2e-16 ***
JobRoleResearch Scientist         < 2e-16 ***
JobRoleLaboratory Technician      < 2e-16 ***
JobRoleManufacturing Director       0.582
JobRoleHealthcare Representative    0.430
JobRoleManager                    < 2e-16 ***
JobRoleSales Representative       < 2e-16 ***
JobRoleResearch Director          < 2e-16 ***
JobRoleHuman Resources           2.21e-05 ***

Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9483
F-statistic:  2246 on 12 and 1457 DF,  p-value: < 2.2e-16
```

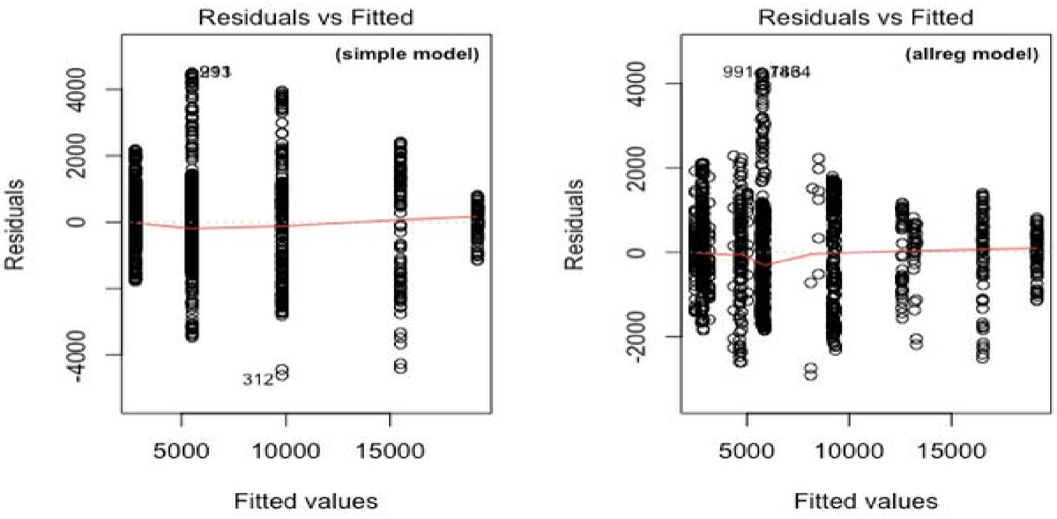
```
Simple model (MonthlyIncome ~ JobLevel)
                     Pr(>|t|)
(Intercept)           <2e-16 ***
JobLevel1             <2e-16 ***
JobLevel3             <2e-16 ***
JobLevel4             <2e-16 ***
JobLevel5             <2e-16 ***

Multiple R-squared:  0.9252,    Adjusted R-squared:  0.925
F-statistic:  4530 on 4 and 1465 DF,  p-value: < 2.2e-16
```

(Figure 2a)

favorable, with high R^2 and adjusted R^2 values, significant coefficients (excepting some factor levels), and high F scores (Figure 2a).

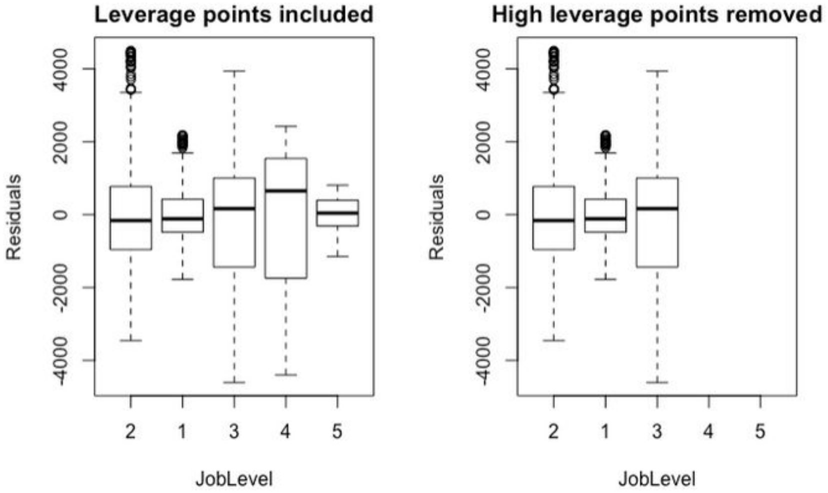
The residual plots showed a large problem with non-constant variance (Figure 2b) that appeared to represent a rough bellows shape. We investigated Box-Cox plots and based on the reported intervals elected to use a square root transform and 0.7 power transform on the response variables for the simple and allreg models. We called the refitted models over the transformed data *simple.transformed* and *allregs.transformed*, respectively. The transformed models had comparable adjusted R^2 values (*transformed:* simple 0.90, allreg 0.93) , but the residual plots showed little improvement (not shown).



(Figure 2b)

We then considered a leverage analysis to see if any outliers were causing significant problems. Such outliers could be hypothetically assessed for potential removal. We found the points in the simple model for which the leverage was large[5]. We then removed these and replotted the residual plots.

We also changed our plot to show the residuals by job level in a box-and-whiskers plot instead of a scatter plot against fitted values. This was based on the observation that the factorized data caused the fitted values to clump into subgroups around the factor levels. Comparing the plot with and without the leverage points, we observed that the removal of leverage points destructively erased entire job levels (Figure 2c). Entire data from some job levels were characterized as high leverage points when fitting the generalized model! We concluded from this that different job levels potentially represented distinctive sub-structures in the data.



(Figure 2c)

We then divided the data up into subsets by job level and fitted these individually exhaustive selection. The best fit models were chosen by maximizing adjusted R^2 and

[5] We used the *2*p/n* cutoff. See *Linear Regression Analysis*, Fifth Edition by Montgomery et al. for details.

were quite different. The fitted models for job levels 3 and 4 had the highest adjusted R^2 scores, 0.68 and 0.80 respectively

- *Job level 3 model*:

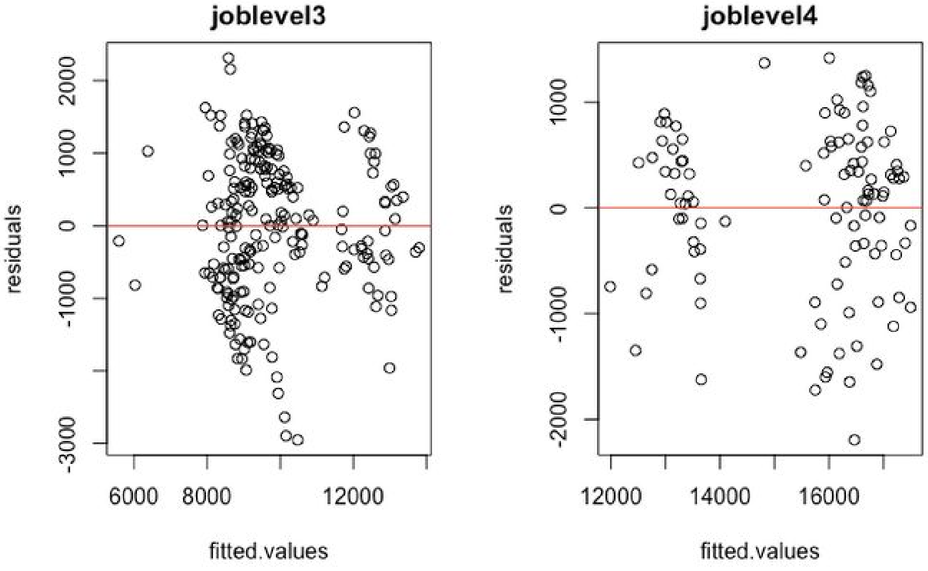
  MonthlyIncome **~** BusinessTravel + Education + JobRole + NumCompaniesWorked + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears
- *Job level 4 model:*

  MonthlyIncome **~** Education + EnvironmentSatisfaction + JobRole + MaritalStatus + MonthlyRate + PerformanceRating + RelationshipSatisfaction

The model assumptions appeared much better met by these specific models. In particular, the residual plots showed less sign of skewing (Figure 2d). The ACF plots and normality plots also showed no signs of a major problem. These models would therefore likely offer effective models for testing hypotheses in the relevant subsets of the population.

We concluded that two sets of models should be recommended for predicting monthly income, depending on the specific (hypothetical) requirements for the models.

On the one hand, the *simple* and *transformed allreg* models showed high R^2 performance and provided an overall view of the data. The *simple* model provided a very simplistic model bird's eye view. In fact, it replaces each job level with an average value of the monthly income for the population subset in question. The *transformed allreg* loses some simplicity but has some improvement to the R^2 value. Additional analysis would be reasonable, over more data, to assess its utility.



(Figure 2d)

On the other hand, the two models for job levels 3 and 4 have better behaviour and meet the major model assumptions. If specific models are viable, these models would serve well for the subsets in question.

Additional attempts to improve the models

We also considered other ways of improving the general model without resorting to subsets: (1) stepwise selection and (2) predictor transformations.

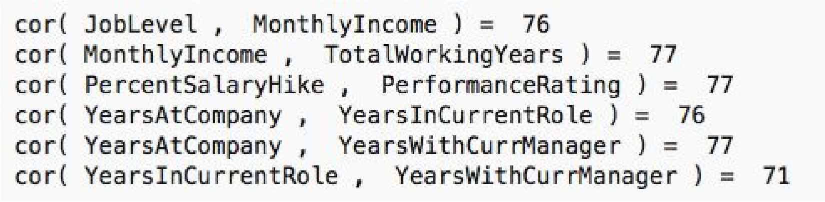
The first attempt was motivated by similar ideas to the ideas behind shrinkage methods. We reasoned that stepwise selection might not pick the most performant method, but it might have the flexibility of picking a reasonable but more complex method that could provide more granular predictive performance than the simpler models. However, we found that while the stepwise model found using the ‘both’ method fitted eight predictors, it suffered from the same innate problems.

We also looked at two transforms to the predictors, given the bellows shape. However, these did not resolve the problem. Additional work may be worthwhile here.

Future work

Several techniques may provide useful insights into the data, including principal component analysis and additional fitting methods.

Principal component analysis may suggest new approaches to the data. In addition to examining the model assumptions during our exploration of the data, we also checked the variable inflation factors and considered the individual coefficient t tests in summary reports. None of these indicated that a problem with multicollinearity was likely. Since our initial correlation analysis nevertheless showed several predictors with high correlation, it would be worth a deeper examination.



```
cor( JobLevel ,  MonthlyIncome ) =  76
cor( MonthlyIncome ,  TotalWorkingYears ) =  77
cor( PercentSalaryHike ,  PerformanceRating ) =  77
cor( YearsAtCompany ,  YearsInCurrentRole ) =  76
cor( YearsAtCompany ,  YearsWithCurrManager ) =  77
cor( YearsInCurrentRole ,  YearsWithCurrManager ) =  71
```

(Figure 2e)

We also considered other ways of fitting the data using generalized methods, rather than simple linear model fits, that could better handle heteroscedasticity. We leave to future work the discussion of how best to approach this.