Final Project

Evelyn Delph, Thomas Janes, & Nick Mobley

5/1/2020

Introduction

Abstract

Are undergraduate students receiving the upmost quality of education from their professors at Indiana University? Since instate students pay $10,948 for tuition and out of state students pay $36,512, it is important to determine if such expenditures are worth the cost. For the purposes of this report, we are examining if an instructors' grade distrubition for undergraduate courses is indicative of instructors' salaries. For this study, the quality of education that an instructor provides is determined by their average GPA. We are interested in studying grade distrubtion and salary across other variables including: the department, the course level, the instructors' level of teaching, and the year. The goals of this report aim to answer the following questions:

- Does GPA impact an instructor's salary?
- Do GPA and salary relationships vary across course levels?
- Do GPA and salary relationships vary across years?
- Do GPA and salary relationships vary across different departments?

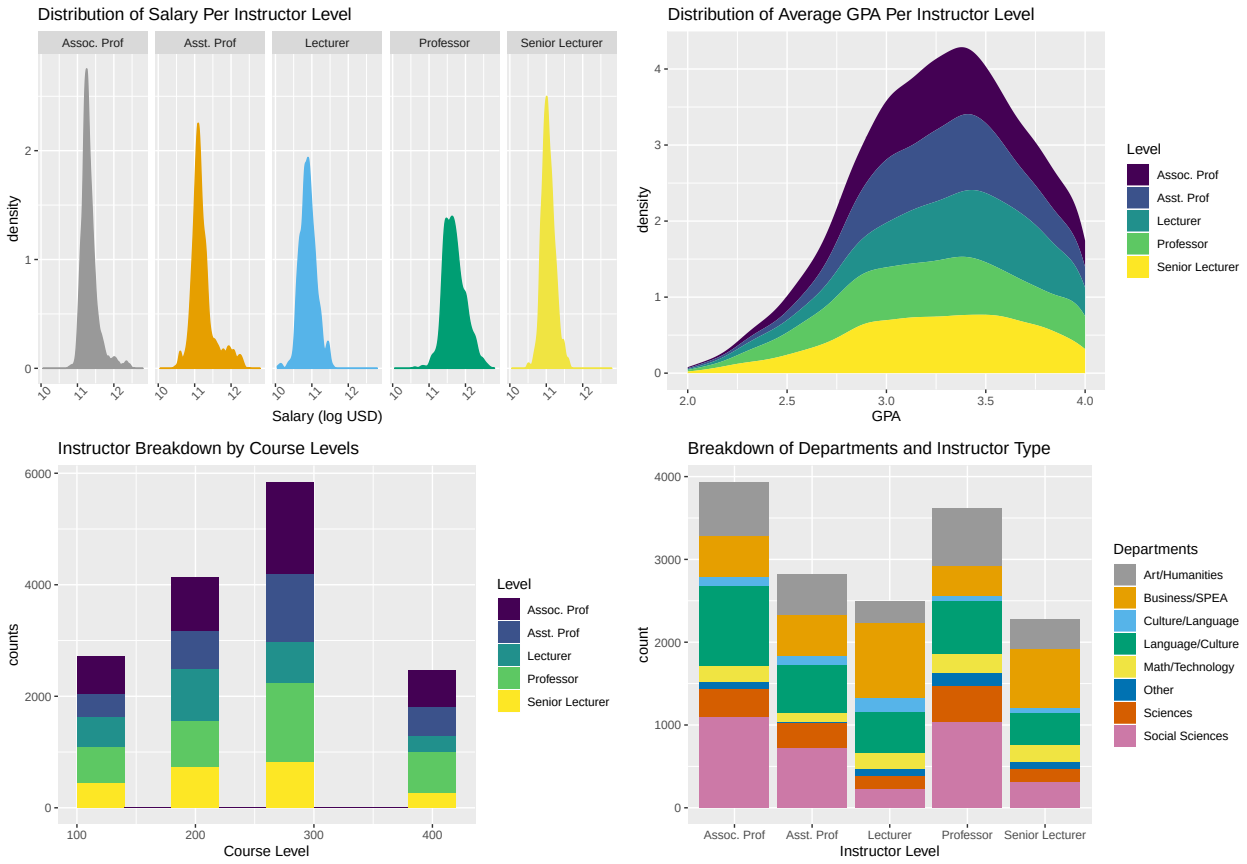
We will answer these questions with numerous EDA techniques to study the relationships among these variables. We aim to achieve a better understanding if the quality of education impacts instructors' salaries.

Description of Data

Our master data file consists of 9 grade distribution and 9 salary files from the IU Registrar between the years 2010-2018. The datasets were joined by the instructor's name. The variables of interest pulled from the datasets include: Name of Instructor, Department, Instructor Level, Salary, Year, Average GPA, Number of Students, and Course Level.

```
DESCRIPTION
Total Rows: 15,163
Total Columns: 7
1. Name of Instructor: 1,942 unique instructors (string)
2. Departments: 100 departments at IU categorized into 8 groups (string)
3. Instructor Level: The title of an instructor (string)
4. Salary: The annual salary of a professor (integer)
5. Average GPA: Average GPA grouped by course and instructor, between 2.0-4.0 (double)
6. Number of Students: Number of students grouped by course and instructor (integer)
7. Course Level: Denotes 100, 200, 300, and 400 level classes (integer)
```

Below are four graphics to better visualize and understand some of the variables of interest.

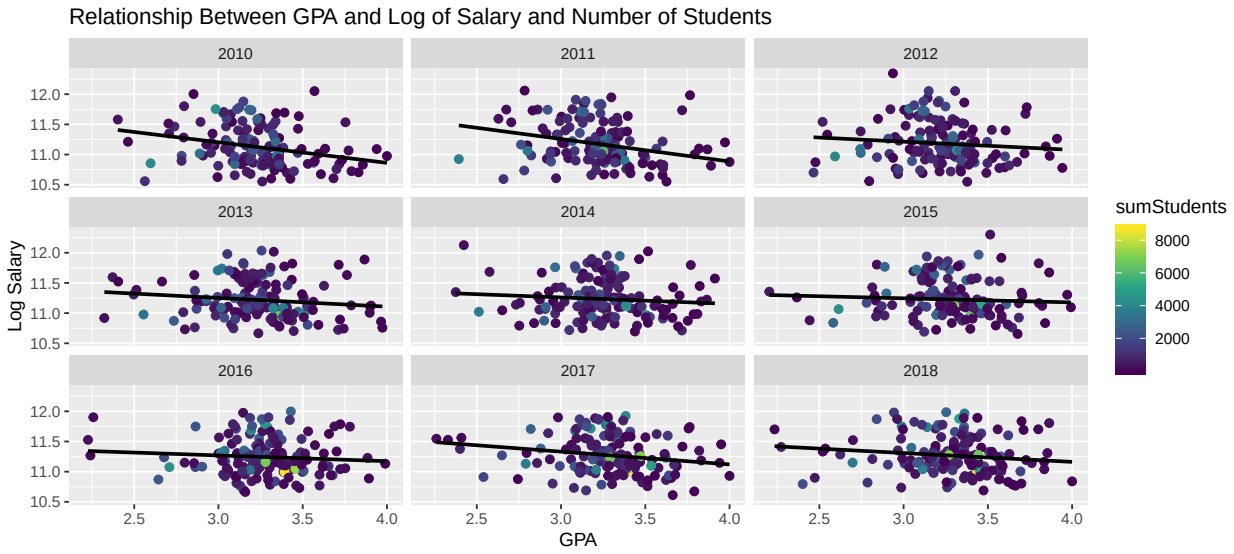


Based on these simple visualizations, from our data we learn that professors will have the highest log salaries with lecturers earning the least. Associate and assistant professor appear to have a similar mean, as well as senior lecturers and lecturers. The density plot indicates that average GPA across all instructor levels center close to 3.5. This may indicate that average GPA and instructor level do not significantly impact one another. The breakdown of instructor level by course level shows there's more instructors in 300 level courses and the least in 400 level courses. It will be interesting to see if outcomes in 300 level courses are much different from the others due to the volume in instructors. The last visualization shows the breakdown of instructor levels against various departments. There does not appear to be a disinct pattern in instructor level against departments.

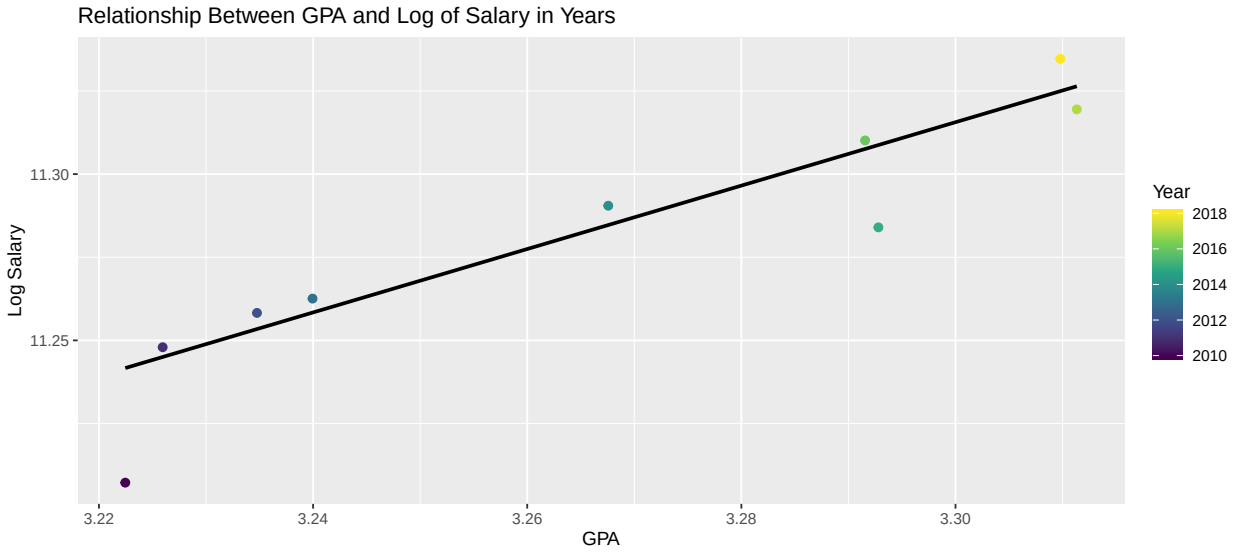
The next step is to take a deeper dive analyzing the interactions of these variables with more robust visualizations.

Exploratory Data Analysis

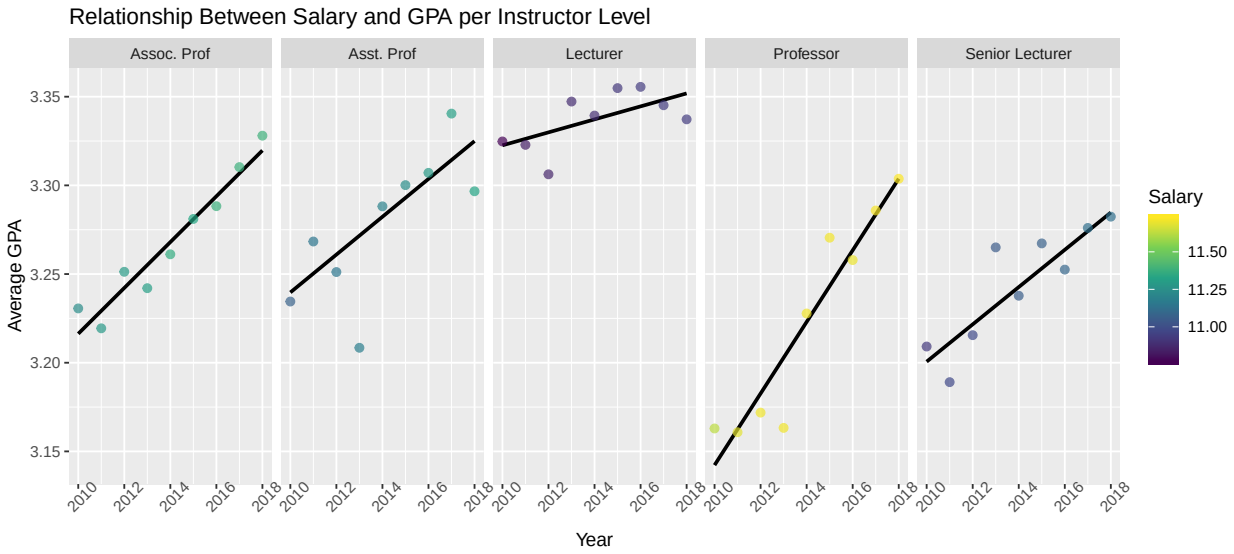
The next visualizations show the relationships of salaries and GPA over the number of students taught by instructor. The purpose of these visualizations is to determine if there is a visible relationship between salaries, GPA, and the number of students taught.



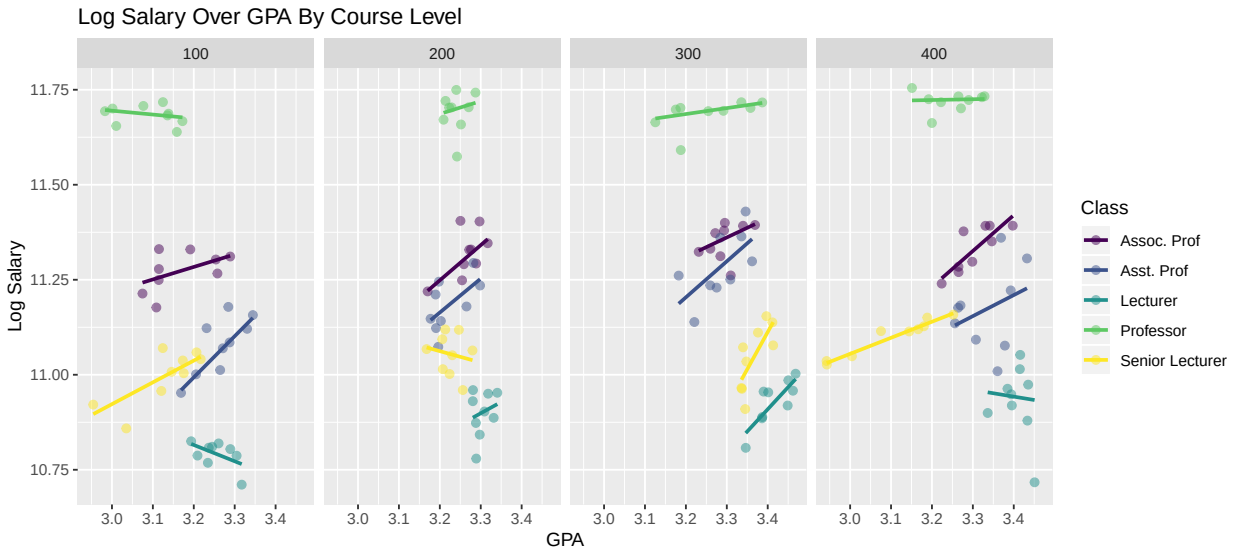
If there was a relationship present, we'd see the colors go in a specific direction. In this case, we do not see any indication that the number of students an instructor taught impacts the relationship between salary and gpa. The next visualizations show the relationships be log of salaries and GPA over instructor level, department, and year.



This first graphic looks at log of salary and average GPA over through 2010-2018. It appears that at a high level GPA and salary increased. It will be more beneficial to analyze other interest variables to determine the importance of year. Year alone cannot adequately determine if this increase in salary is related to increase in GPA. The next visualization also factors in instructor level for a deeper understanding of these relationships.



This shows the relationship between average GPA and salary over the nine year period by instructor level. We also learn that average GPA increased throughout the years per instructor level. By examining salary on a log scale (the color) and average GPA (y-axis), there appears to be a positive relationship between them. For associate and assistant professors and senior lecturers, the color becomes lighter as it progresses throughout the year, indicating an increase in salary. Professors and lecturers appear to have some relationship with GPA and salary, but not as much compared to the other instructor levels. This could indicate a relationship between GPA and salaries.



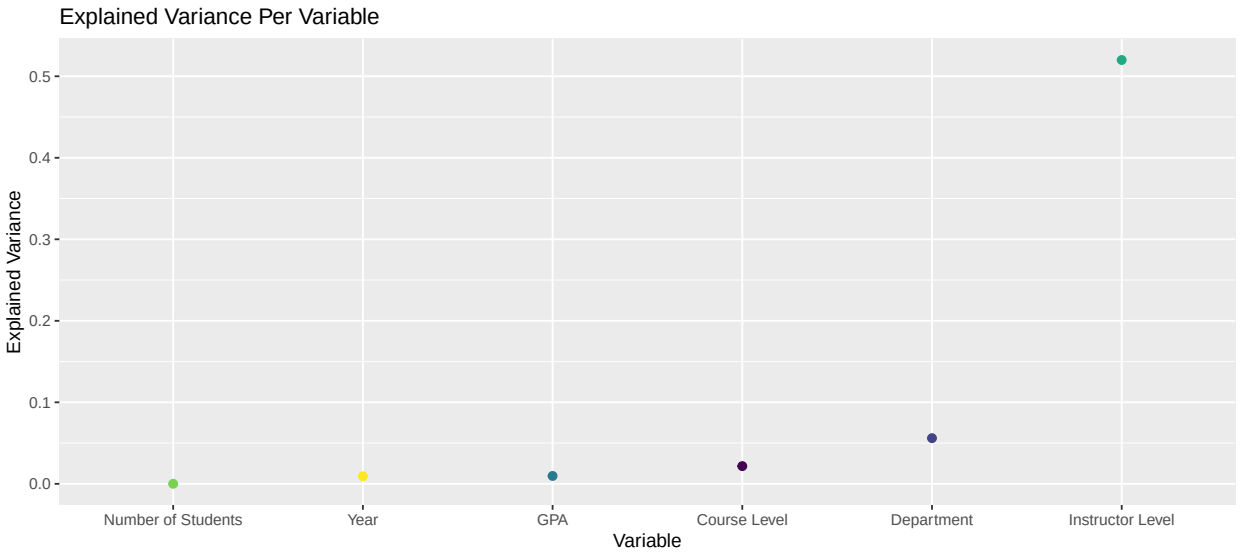
This shows the log salary over average GPA by course level and instructor type over the nine year period. There is one exception with lecturer in the 100 level course having a negative slope, but that is most likely due to the outlier. Overall this could point to salary and GPA being postively correlated to eachother. Although it helps to see the breakdown among course levels, it appears that the trend between salary and average GPA is consistent.

This shows the log salary over average GPA by department and instructor type over the nine year period. Most of the instructor types over the nine years show a slight positive relationship between the average GPA and salary. Social sciences is the anomaly where there's a handful of negative slopes. This could mean there were more newer lecturers or senior lecturers starting in later years who had higher grade distributions. But in general, when breaking up GPA and salary by department, we continue to see positive relationships between GPA and salary.

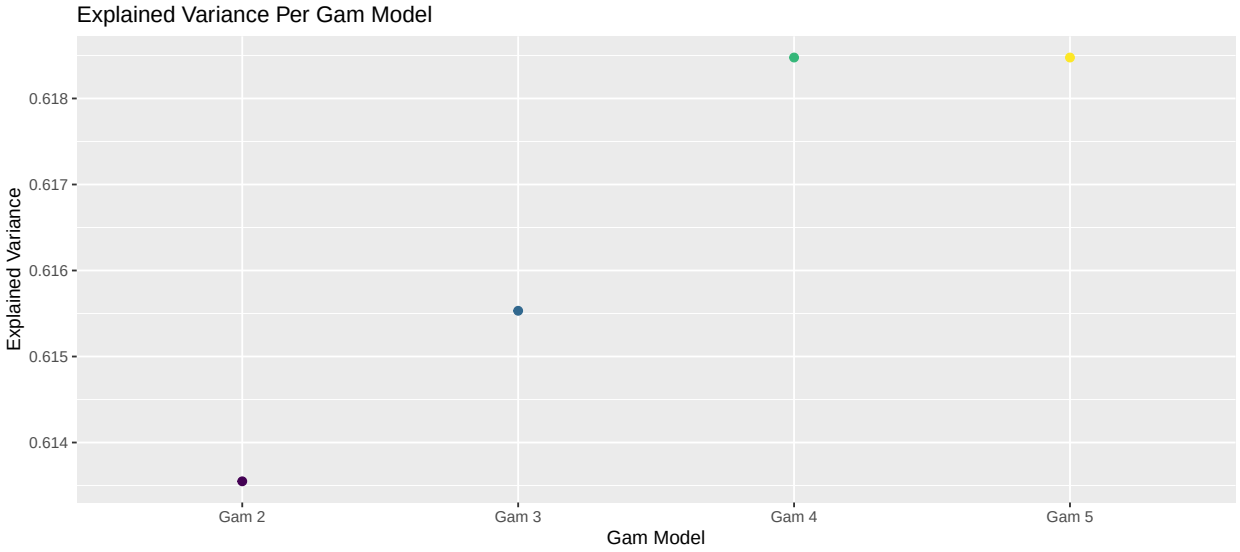
After examining these four visualizations, we can see that year, department, course level, and instructor level may impact the relationship between GPA and Salary. The next step is to construct predictive models to see which variables best predict salary.

Model

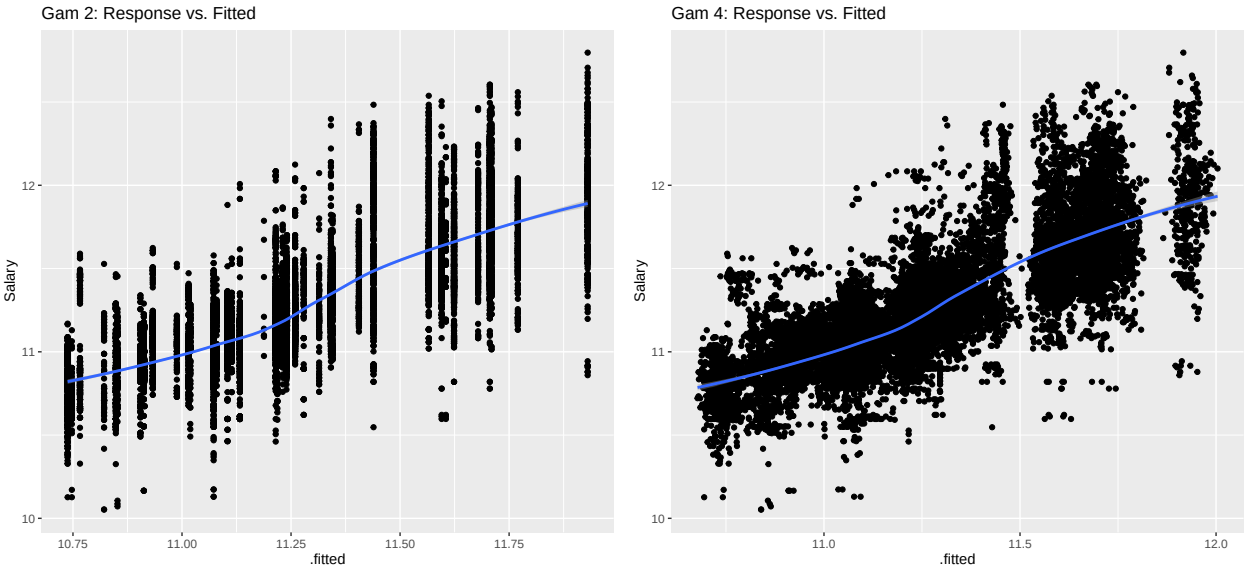
To model the relationship between salary and our feature variables of interest, a generalized additive model (GAM) was fitted on the data. First, five gam models were constructed consisting of one variable and Salary as the predictor to calculate the explained variance. This determined feature importance.



This shows the explained variance across single variable gam models. This shows the level of variance each variable captures when predicting salary. The single variable gam models indicate that the order of most important variables for predicting salary are instructor level, department, course level, GPA, and year. With regards to salary, on its own, GPA plays an insignificant role. Even though GPA on its own cannot predict or explain salary, it can be incorporated into a multi-variable gam model to optimize predictions. The next step is to incrementally add variables into the gam model and evaluate the results to determine the best model. The variables will be added in descending order of explained variance. Due the number of students being 0, we omitted this variable from the gam models.



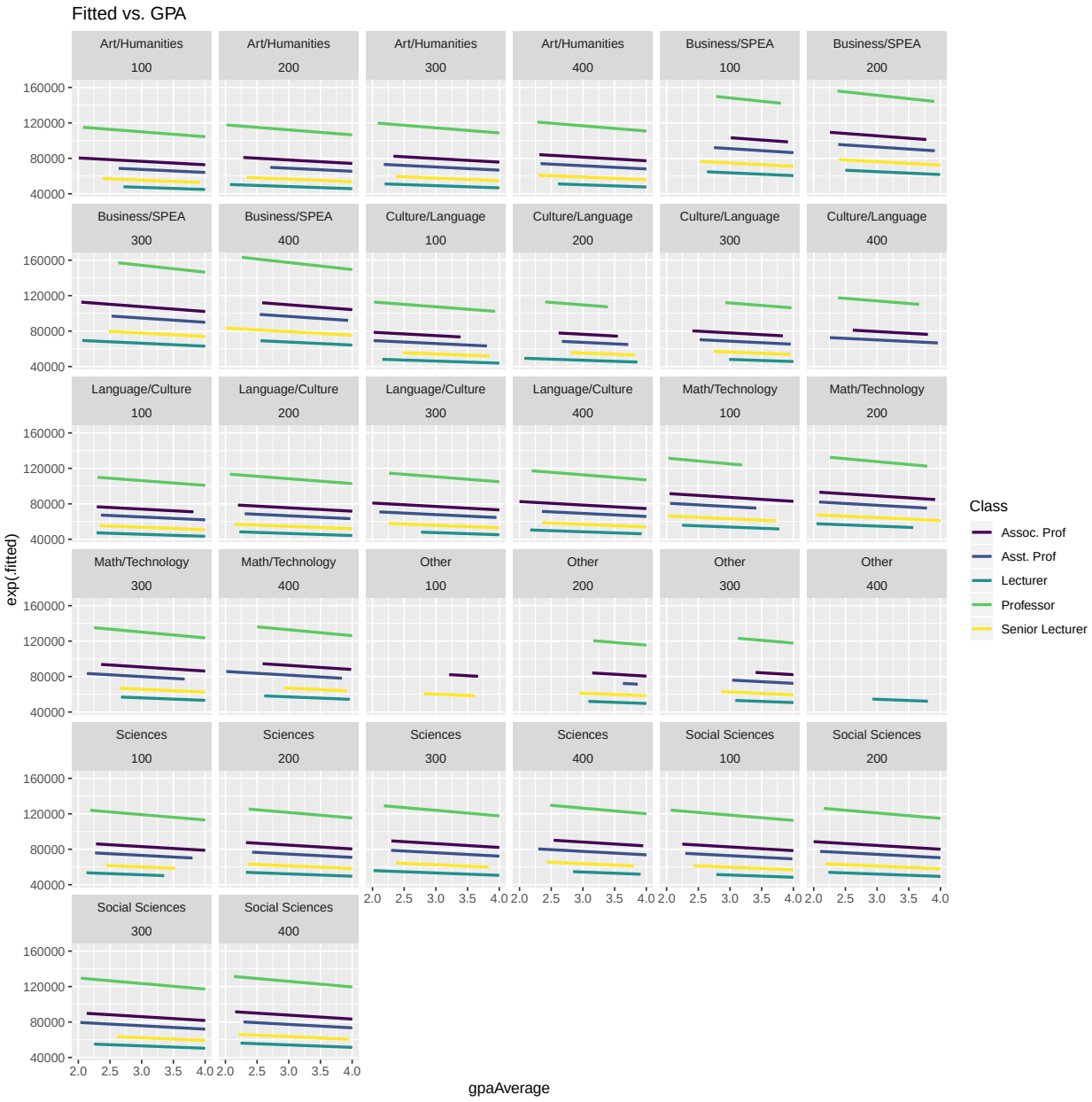
It appears that the explained variance stops increasing when the top four variables (instructor level, department level, course level, and GPA) are included. Year makes little to no impact on the model. This means the model captures the most variance with these four variables. GPA led to an increase in explained variance compared to the gam models excluding it (models 2 and 3). The next step is to examine the fitted residuals over salary to see if the fit becomes smoother while adding more variables.



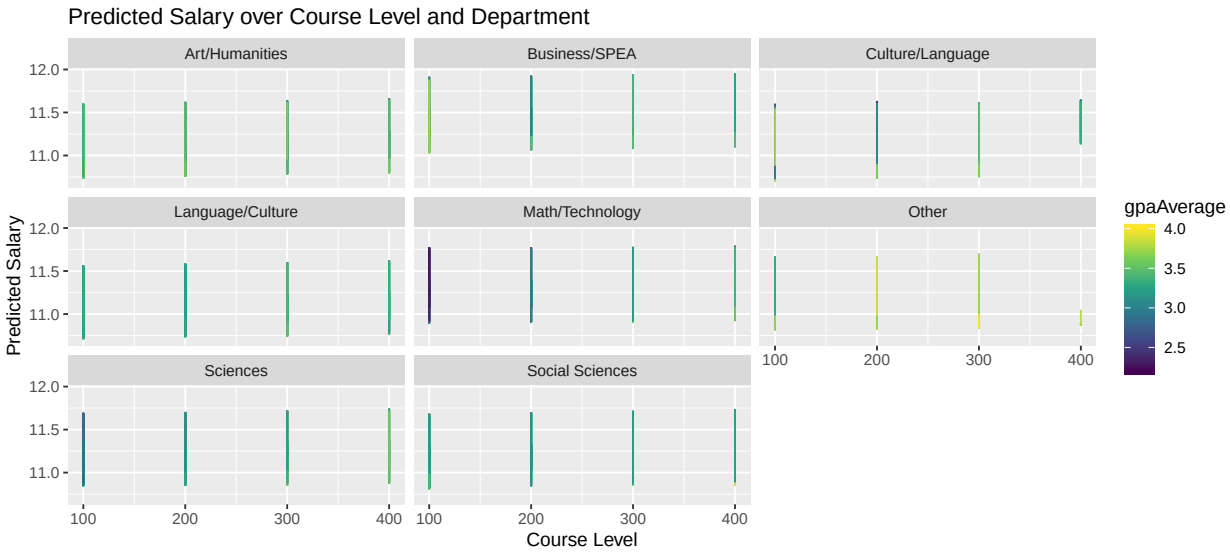
There's a noticeable difference between gam models 2 and 4. It appears the residuals become smoother when adding variables. Similar to the explained variance visualization across the four multi-variable gam models,

when we added a fifth variable there wasn't a distinct difference between the two models (Gam 5 omitted).

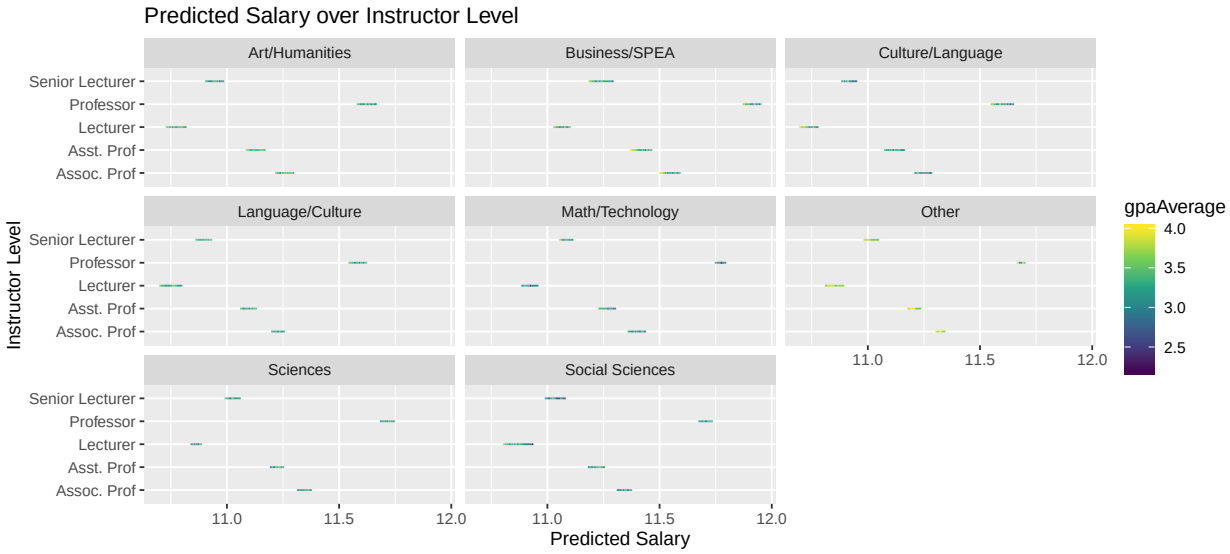
A Closer Look at Gam Model with Four Variables



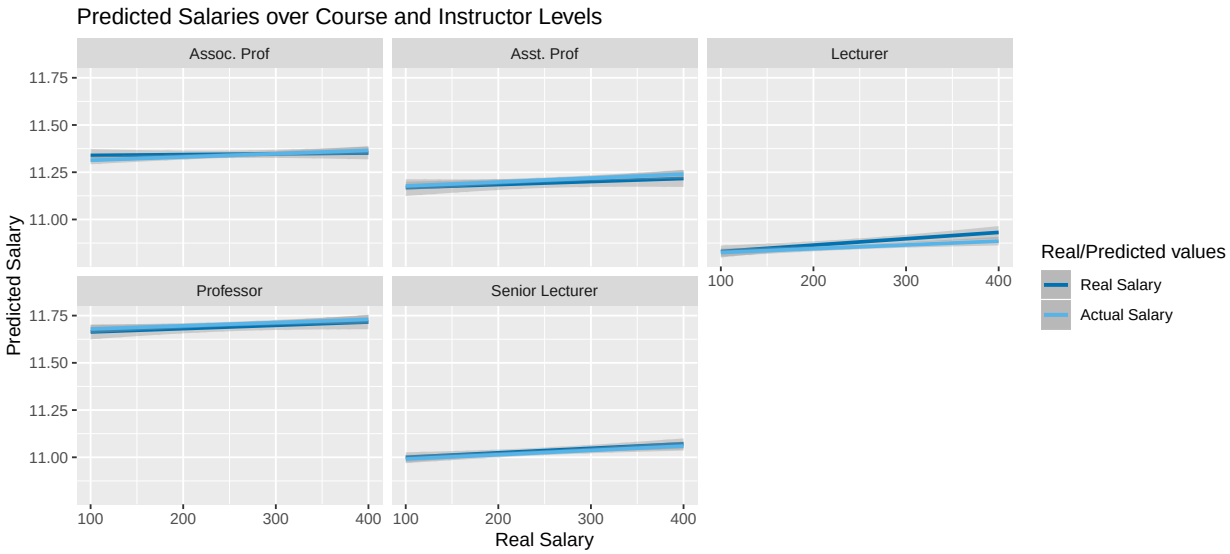
Based on this graph, it would seem our model predicts that as the average GPA of a course increases, the instructor's salaray actually decreases. However, average GPA is not the only variable determining this trend, or even the most influential. The course's department, level, and level of instructor all explain variation in the salaray of a course's instructor.



Next, we visualized our optimal gam model by predicting salary over the entire original dataset. This graphic shows predicted salary over the course level and department, with GPA as the color scale. The lines are vertical and parallel. GPA appears to vary at the course level. This could be due to the varying GPA averages among each department. This could indicate the little to no impact course level effects salary.



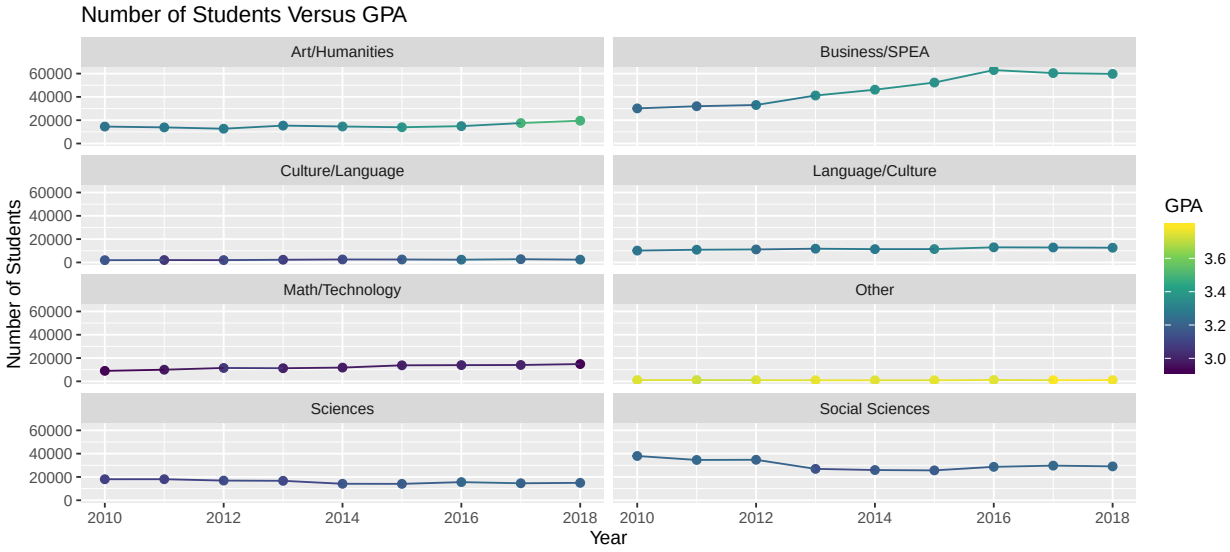
This visualizes the average GPA and predicted salary for instructor level by department. Similar to before, the lines are parallel to one another. However, there's much more variation in where the salarys range from due to the instructor level. This could indicate that the instructor level is probably the most important variable when predicting salary.



This shows the actual versus predicted salaries over course and instructor levels. You see that the gam model with four variables closely matches the actual salaries. This does indicate our model is sufficiently accurate when predicting salaries. Based on the previous visualizations, we have seen the little impact the variables of interest (other than instructor level) impacts GPA. Now that we had an indepth look at our gam model, let's explore why some variables did not greatly impact salary predictions.

A Closer Look at Number of Student and Year Variables

After determining the optimal model did not include number of students and that year explained little variation, we wanted to speculate more closely. We examined the number of students over department and year as well as department, number of students, and GPA over year. We wanted to see if more students were switching to departments with higher GPAs and if there was a decrease in enrollment in more difficult disciplines. If this was the case, it may explain why year and number of students did not impact salary, but showed significant in previous graphs.



This shows the trends in number of students and GPA throughout the departments over the nine yaer period. Here, we see that departments like SPEA/Business and Arts/Humanities increase in enrollment and GPA. Fields like Sciences and Social Sciences go back and forth increasing and decreasing in enrollment and the

GPA slightly increases. Language/Culture and Math/Technology increase in enrollement but GPA does not change.

This reveals year and number of students could have been important when factoring GPA and salaries if we were only considering business and SPEA courses. Year and number of students probably explained little variance because this trend was not present among the other six categories.

Conclusion

After this analysis, we hoped to answer whether or not GPA impacts salaries for instructors. Although our gam models improved after including average GPA, there were more important factors that contributed. Of course, we expected salary to vary across instructor level. However, we wanted to know if there was more that could impact salary. Our analysis revealed that salaries varied across departments and schools. Salaries also could have varied due to the course levels. GPA seemed to have little impact compared to the other variables we considered.

We wanted to answer whether or not students were getting quality education with respect to GPA and professors. For the purposes of this research we defined high quality education to be directly linked with GPA. This research did not include other campus resources like libraries, labs, and educational aid into consideration when factoring the quality of education. Essentially, examining GPA and salaries might not be enough to answer this question. More factors like the ones stated need to be taken into consideration.

Further and more refined research can help answer the question of quality education. For IU, analyzing quality of professor and education should occur at the department level. What's considered a high GPA varies across disciplines. Next, the services each department and college provides for undergraduates must be taken into consideration. Our research only scratches the surface when trying to answer this question.