Hong Anh Duong Nguyen

STAT-140-01

Professor Nussbaum

2020, April 28

**Data Analysis Project**

**Data Selection**

The point of this project is to show the 2008-2009 nine-month academic salary for Assistant Professors, Associate Professors, and Professors in a college in the U.S. Using this data, the researchers hope to better recognize and monitor the salary differences between male and female faculty members.

This dataset has 397 rows and 6 columns, which includes 3 categorical variables and 3 numerical variables. The response variable in this dataset is salary. The categorical variables are rank, discipline, and sex. Rank is a factor with different levels of Assistant Professors, Associate Professors, and Professors. Discipline is a factor with levels A and B, referring to theoretical departments and applied departments, respectively. Sex is a factor with levels of Female and Male. The numerical variables include years since PhD, years of service, and salary. Salary covers a nine-month salary in dollars.

**Analysis and Summary**

1. Rank

A picture containing microwave

Description automatically generated

## AssocProf AsstProf Prof   
## 64 67 266

The first predictor is the rank of professors in a college in the U.S. In this data, this variable is divided into three levels, with AssocProf for Associate Professor, AsstProf for Assistant Professor, and Prof for Professor. According to the graph and the summary statistics, the number of Professors (266 total) is nearly double the number of Assistant Professors and Associate Professors, which are 67 and 64, respectively.

1. Discipline

A screenshot of a cell phone

Description automatically generated

## A B   
## 181 216

The second variable is the professors’ discipline in a college in the U.S. This predictor represents whether the faculty member is from level A (“theoretical” department) or B (“applied” department). According to the graph and the summary statistics, the number of Professors in “applied” departments (B) exceeds the number of Professors in “theoretical” departments (A), even though the values are roughly equivalent.

1. Years since PhD

A picture containing drawing

Description automatically generated

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 12.00 21.00 22.31 32.00 56.00

The third predictor is yrs.since.phd, which represents the professors’ number of years since PhD. This variable has recorded a minimum of 1 year, a maximum value of 56 years, and a median of 21 years. The distribution is centered around 22.31 and is a slightly right-skewed, unimodal distribution with very few unusual values.

1. Years of service

A picture containing drawing

Description automatically generated

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 7.00 16.00 17.61 27.00 60.00

The fourth predictor is yrs.service, which represents the professors’ number of years of service. The distribution is centered around 17.61 with a median of 16 and ranges from about 0 to 60. A histogram of the values is a right-skewed, bimodal distribution with peaks at around 5 years and 20 years and contains some unusual values to the right of the graph at ~60 years.

1. Sex

A screenshot of a cell phone

Description automatically generated

## Female Male   
## 39 358

The fifth predictor is the professors’ gender in a college in the U.S, which represents whether the faculty member is Female or Male. According to the graph and the summary statistics, the number of male faculty members (358 total) far outweighs that of female faculty members (39 total).

1. Salary

A picture containing clock

Description automatically generated

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 57800 91000 107300 113706 134185 231545

The final predictor is the professor’s nine-month salary, calculated in dollars. The distribution is centered around $113706 with a median of $107300 and ranges from about $57800 to $231545. It is a right-skewed, unimodal distribution with some unusual values after $200000. This variable will later be used as the explanatory variable in the simple linear regression part.

**Simple Linear Regression**

The execution of simple linear regression will frame the data analysis and help us understand the relationship between two numerical variables. Let x, the predictor variable, be yrs.service, the professors’ number of years of service in the college. Let y, the response variable, be salary, the nine-month salary, in dollars.

A screenshot of a cell phone

Description automatically generated

The relationship between years of service and salary is a moderate, positive linear relationship. There are a few potential outliers (around 8 years with $200,000, and some points to the right of the cloud of points).

The value of the correlation is 0.335, which is consistent with the graph (a weak/moderate relationship). The squared correlation is 0.112. This shows that 11.2% of the variation in salary is explained by the variation in the number of years of service.

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 99974.6529 2416.6053 41.369873 1.186078e-145  
## yrs.service 779.5691 110.4169 7.060232 7.528739e-12

The intercept shows that a professor who has 0 years of service will have a salary of $99974.6529. The slope indicates that for every extra year of service a professor has, the salary will rise by $779.5691. After constructing confidence intervals for the slope and intercept, some conclusions can be drawn about these two parameters. We are 95% confident that the slope for predicting the professors’ salary from the years of service is between 562.49 and 996.65. Since the confidence interval does not contain 0, there is some evidence that there is a true, positive relationship between years of service and salary. We are 95% confident that the intercept for predicting the professors’ salary from the years of service is between 95223.64 and 104725.70.

A screenshot of a cell phone

Description automatically generated

Before continuing, it is important to check the assumptions for the linear model. The residual plot above shows that a linear model might not fit in this data - there are lots of negative values for very large values of years of service, and lots of positive values for values of years of service near the middle.

A picture containing drawing

Description automatically generated

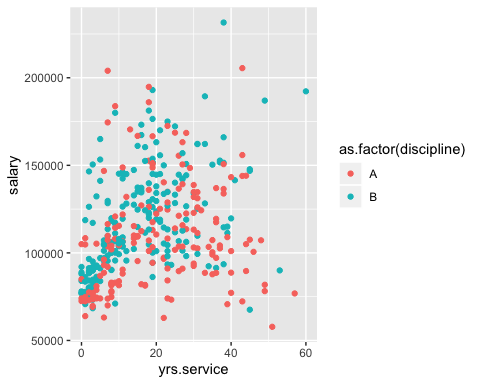
The histogram is nearly normal and centered at approximately 0. Residuals range from -75000 to 120000. There appear to be some outliers near the right tail of the graph. Thus, we proceed the analysis with caution.

In terms of constant variability, the variability of points around the least squares line remains roughly constant. Lastly, all observations are independent of each other since there are no variables that are sequentially collected or related to time-series data.

The points far from the cloud of points, but not horizontally far. Therefore, the graph does not have high leverage and thus, does not appear to strongly affect the line.

**Adding an Indicator Variable**

To further comprehend the relationship in the Simple Linear Regression part, we add a categorical variable to the model. The selected categorical variable is discipline, a factor representing whether the faculty member is from level A (“theoretical” departments) or B (“applied” departments).



The pattern of observations for A (“theoretical” departments) is a weak-moderate relationship that has some unusual values at the top of the graph. A linear model may not fit this graph. The pattern of observations for B (“applied” departments) is a fairly strong positive linear relationship that has some unusual values at the right bottom of the graph.

15.362% of the variation in the estimated academic salary can be explained by the number of years of service and discipline. The adjusted R2 is pretty low, which is consistent with the description of the scatterplots. However, this value is higher than the original R2 value for salary and years of service (R2 = 0.112), which shows that adding the professors’ discipline makes the relationship stronger.

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 91335.832 3005.361 30.391 0  
## yrs.service 862.775 109.154 7.904 0  
## disciplineB 13184.043 2846.831 4.631 0

On average, the salary for professors in a college in the U.S with discipline A (“theoretical” departments) is approximately $91335.832. For every professor in discipline B (“applied” departments), the expected salary will increase by about $13184.043. Carrying out a confidence interval, we are 95% confident that the slope for predicting the salary of professors from their discipline is between 7587.164 and 18780.92. Since this confidence interval does not contain 0, there is some evidence that there is a true, positive relationship between professors’ salary and their discipline.

**Multiple Linear Regression**

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 78862.820 4990.326 15.803 0.000  
## rankAsstProf -12907.588 4145.278 -3.114 0.002  
## rankProf 32158.411 3540.647 9.083 0.000  
## disciplineB 14417.626 2342.875 6.154 0.000  
## yrs.since.phd 535.058 240.994 2.220 0.027  
## yrs.service -489.516 211.938 -2.310 0.021  
## sexMale 4783.493 3858.668 1.240 0.216

I construct a multiple linear regression model with all of the variables in my dataset. There are five explanatory variables in this model, in which three of them are categorical variables, and the remaining two are numerical variables. The professors’ rank and discipline are the most significant predictors, while the gender is the least significant predictor.

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 82700.549 3916.719 21.115 0.000  
## rankAsstProf -12831.537 4147.669 -3.094 0.002  
## rankProf 32456.152 3534.915 9.182 0.000  
## disciplineB 14505.151 2343.418 6.190 0.000  
## yrs.since.phd 534.631 241.159 2.217 0.027  
## yrs.service -476.718 211.831 -2.250 0.025

## [1] 0.4455269

Using backward elimination with the p-value approach, I drop the predictor that has α, the significance level, over 0.05 and refit the model. Thus, I have a final, reduced model for the dataset that contains the professors’ rank, discipline, number of years since PhD and number of years of service. Furthermore, the adjusted R squared value is approximately 44.553%, pretty high for predicting salary.

A close up of a mans face

Description automatically generated

To show that the multiple linear regression is the best model for this data, I will check the assumptions for my reduced model. The residual plot above shows a moderate positive relationship with some potential outliers. For example, the observations coordinated on (105000, 3.75) at the bottom could be considered outliers. Since there are some deviations from constant variance in this model, we will proceed with caution.

A picture containing drawing

Description automatically generated

The plot above shows a nearly normal, unimodal histogram with centered at approximately around 0. The residuals range from about -25000 to 100000. There do not appear to be any outliers in this graph.

For constant variability condition, we will look at the residual plots. For categorical variables, rank, and discipline, we check that the variability does not fluctuate across groups, which they do not show any prominent differences.

A close up of a piece of paper

Description automatically generated

In the scatterplot above, even though there is less amount of variability of y when x is large, the variability of points around the least squares line remains roughly constant. The observations in this model are independent since there are no variables that are sequentially selected or related to time-series data.

The final model shows that the discipline variable has the lowest R2 (0.022); however, the total adjusted R2 of the reduced model decreases without discipline in the model. Rank has the highest R2 of 0.391, years of service has 0.110, and years since PhD has a total of 0.174. All of which have very small p-value.

**Discussion**

In conclusion, the salary of professors in a college can be predicted through numerous variables. For instance, in our dataset, we use the professor’s rank, discipline, the number of years of service, number of years since PhD, and whether the faculty member is Male or Female. Among these variables, it was found that rank, discipline, years of service, and years since PhD are the most significant predictors that have positively impacted the response variable.

The most difficult challenge I encountered was determining if the relationship can be explained through a linear model. Throughout the process of analyzing, I have run into problems such as non-constant variance, inappropriate linear model fit; therefore, I have to use model transformation and add a centered quadratic term to enhance my model.

**Citation:**

Fox J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition Sage.

https://vincentarelbundock.github.io/Rdatasets/csv/carData/Salaries.csv