STAT 474 Paper II - Generalized Additive Models

Author: Juan Manubens

Professor: Richard Berk

Prompt - ***It is well known that in most occupations, women earn less than men. There can be many reasons: differences in seniority, different jobs and other factors, including outright gender discrimination. Academic institutions have undertaken studies to determine if there are salary differentials by gender and if so, what might explain those disparities. For this analysis, you are to consider these issues with an academic dataset collected during the 2007-2008 academic year. The data can be obtained from the car library using the name "Salaries."***

1 Introduction

I would like to mainly investigate possible differences in the mean salary between male and female professors, conditional to gender as well as the other four factors in the dataset. I will investigate conditional mean differences from a non-legal standpoint - I will not make any claims related to discrimination, and will not extrapolate beyond the sample at hand. I will explore and examine the univariate and multivariate statistics and assess whether moving to a Level II analysis is reasonable. That is, I will begin with a description of the data at hand, followed by a *Generalized Additive Model* that is relatively assumption-free. The validity of any statistical inference made will fundamentally depend on how the data was generated, and thus I will only tread into Level II as I test different assumptions.

2 The Data

2.1 Overview

The data used in this paper comes from the ***Salaries*** dataset from the built-in ***car*** R package, originally featured in the book, ***An R Companion to Applied Regression***[1]. It contains a 2008-09 sample of nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members. There is some ambiguity regarding the dataset - for

instance, it is unclear whether the variable **Years of Service** represents the number of years a Professor has taught in his career or simply in this institution (given the information at hand, we are assuming the latter). I will address these concerns as well as others with more detail below.

2.2 Format and Variables

The data frame contains a total of $N = 397$ observations. The following variables are used:

1. **Rank** - a factor with 3 levels of academic titles (*Assistant Professor, Associate Professor, and Professor*);
2. **Discipline** - a factor with 2 levels: *A ("theoretical" departments) or B ("applied" departments)*; **Discipline - a factor with 2 levels: *A ("theoretical" departments) or B ("applied" departments)*;**
3. **Yrs.since.phd** - a discrete numeric variable representing years since the academic earned their PhD;
4. **Yrs.service:** - a discrete numeric variable representing years of service in this particular data set. As mentioned previously, there is no other information about this variable beyond what has been described so far;
5. **Sex** - a factor with 2 levels (*Female and Male*);
6. **Salary,** - a discrete numeric variable nine-month salary for each academic, in dollars.

2.3 Univariate and Multivariate Statistics

Sex	Statistic	N	Mean	St. Dev.	Min	1Q	Median	3Q	Max
Female	Years Since PhD	39	16.5	9.8	2	10	17	23.5	39
	Years of Service		11.6	8.8	0	4	10	17.5	36
	Salary*		$101	$25.95	$62.90	$77.25	$103.75	$117.00	$161.10
Male	Years Since PhD	358	23	13	1	12	22	33	56
	Years of Service		18.3	13.2	0	7	18	27	60
	Salary*		$115.10	$30.44	$57.80	$92.00	$108.00	$134.90	$234.60
Both	Years Since PhD	397	22.31	12.9	1	12	21	32	56
	Years of Service		17.6	13	0	7	16	27	60
	Salary*		$113.70	$30.29	$57.80	$91.00	$107.30	$134.20	$234.60

*Salary in '000s of Dollars (USD$)

Table 1: Descriptive Statistics for Years Since PhD., Years of Service and Salary

At a glance, the first striking statistic is the gender imbalance - less than 10% of our sample is female professors. Another observation is that the sample is small, which is not ideal for inference as we cannot safely assert asymptotic properties with far too few observations in the data. Additionally we can see in Figure 1., Years of Service and Years since PhD and Salary are all slightly right skewed.

We also see a slightly even split in discipline, with about 45% of the professors in Theoretical Departments and the rest in Applied Departments. Rank shows a much higher proportion (roughly two thirds) of Professors compared to Assistant and Associate Professors. From Table 1, we see that on average, Female professors tend to be younger (in terms of their time as professors and their years since obtaining their PhD). As I mentioned previously, the caveat here is that it is not exactly clear whether years of service refers to their time in this particular university, or their time working as professors throughout their career. We can see this in more detail in the Appendix.

Sex	Discipline*	Frequency
Female	Theoretical	18
Male		163
Female	Applied	21
Male		195

*By Department, (A) Theoretical; (B) Applied

Table 2: Gender Frequency by Discipline

Sex	Rank	Frequency
Female	Assistant Professor	11
Male		56
Female	Associate Professor	10
Male		54
Female	Professor	18
Male		248

Table 3: Gender Frequency by Rank

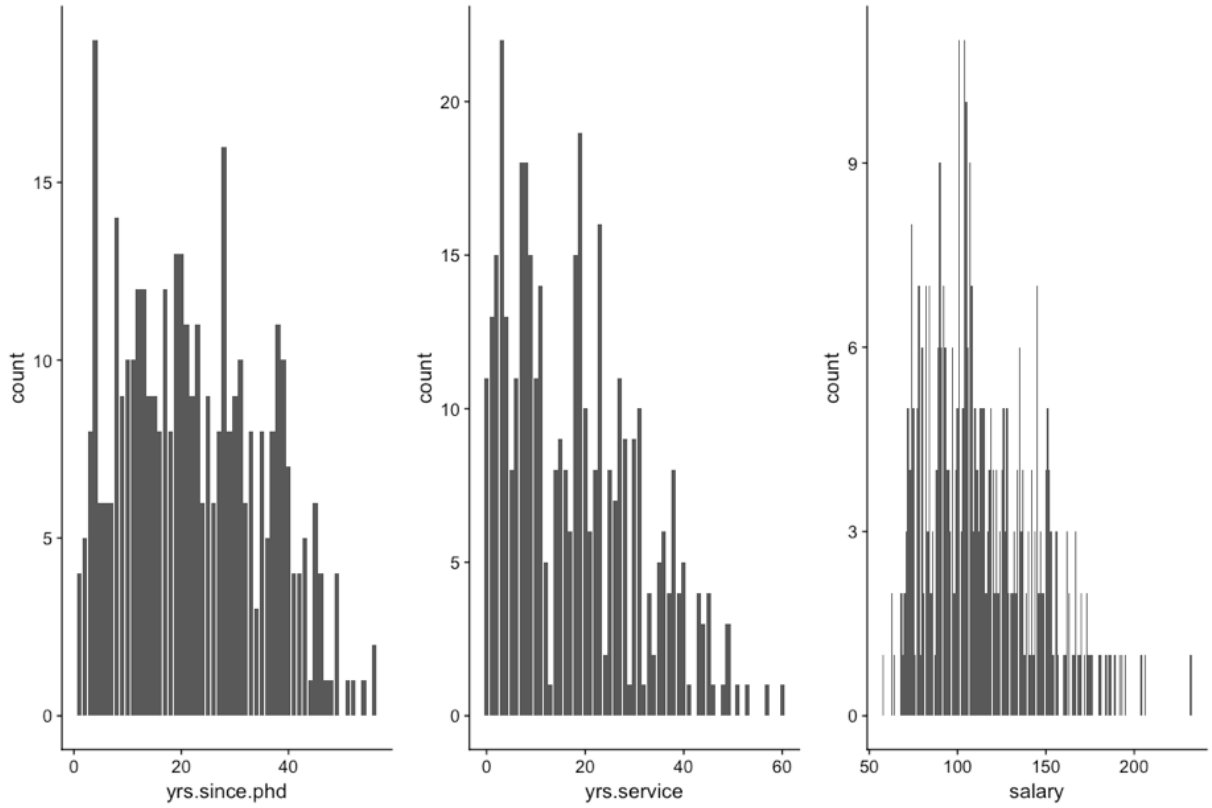


Figure 1: Figure 1: Distribution of Discrete Numerical Variables

The split within genders by discipline is even across genders, between 83.5% and 85% (Table 2). The same cannot be said about rank (Table 3). Here, the 2-way frequencies show a much higher of proportion of Professors in the case of males: over 62%, in contrast to around 46% for the female academics in the sample (a more detailed split including Discipline can be found in the Appendix).

Discipline	Frequency
Theoretical	181
Applied	216

Table 4: Discipline Frequency

Rank	Frequency
Assistant Professor	67
Associate Professor	64
Professor	266

Table 5: Rank Frequency

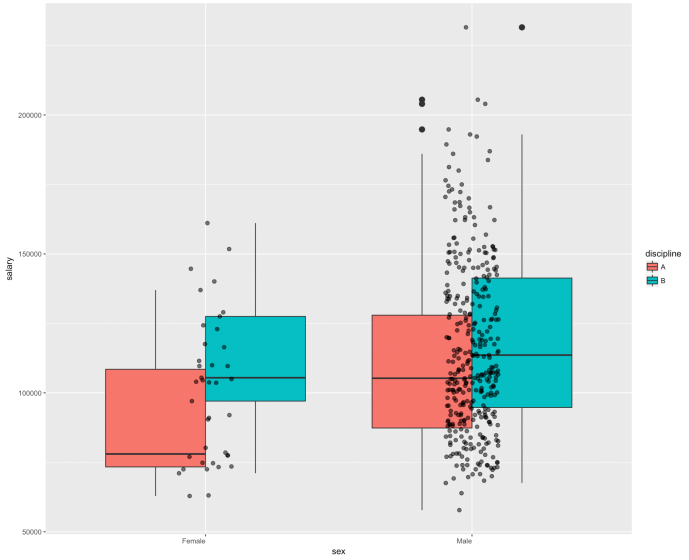


Figure 2: Boxplots with Scatter by Gender, Discipline

The above observations are also highlighted in Figures 2 and 3, which extend the Gender and Salary relationship. An important observation here is the presence of outliers, indicated as slightly larger dots above and below the boxplots. This can be a major problem when running a regression with training and test data, as data from the tails may be excluded from either set. We are relying too heavily on the luck of the draw, and an unlucky random split could result in a unrepresentative test set. GAM will likely offer more flexibility, but the sample size is too small and there are too many caveats to justify a Level II analysis.

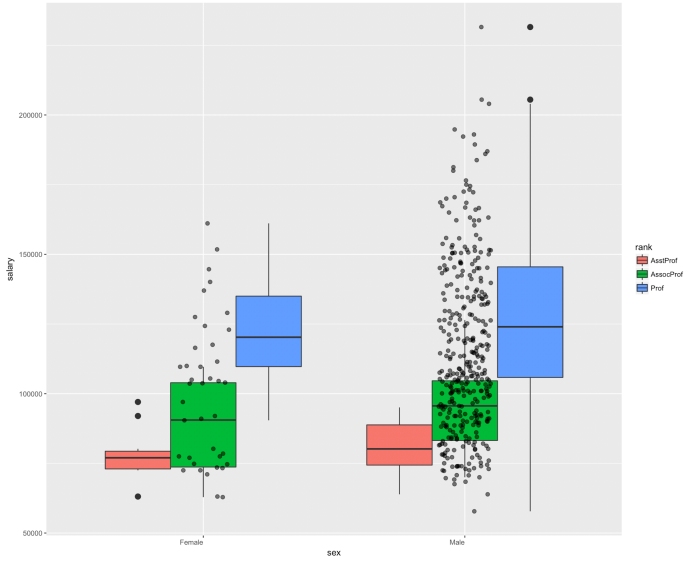


Figure 3: Boxplots with Scatter by Gender, Rank

3 Generalized Additive Model (GAM) Analysis

Using the *mgcv* R package, we fit the GAM model using all predictors [1]. Here $p > 2$, and we assume that the conditional mean of our response is a linear combination of functions $f_i(X_i)$ of predictors X_i. Note that we are running a GAM Model from a Level I standpoint - we are not attempting to estimate the true response surface and are operating on a wrong model perspective. We have already detected enough problems with the data to bar us from a Level II analysis, so we will interpret the estimated mean function as misspecified. Thus, we interpret the residuals ϵ_i as simply the variability of the fitted values. As we do not assume our model is first order correct, we do not assume that the probabilities associated to our statistical tests are correct. The GAM summary can be seen in Table 6 below.

Table 6: GAM Model Summary

	Estimate	RSE	t-value	$Pr(>\mid t \mid)$
Intercept	$68,773^{***}$	7,138	9.635	$< 2e-16^{***}$
f_1 : rankAssocProf	$16,506^{*}$	7,197	2.294	0.0224^{*}
f_2 : rankProf	$44,553^{***}$	8,386	5.313	$1.86e-07^{***}$
f_3 : disciplineB	$14,464^{***}$	2,292	6.311	$7.80e-10^{***}$
$s\{f_4$: yrs.since.phd$\}$	N/A	N/A	N/A	0.000389^{***}
$s\{f_5$: yrs.service$\}$	N/A	N/A	N/A	0.001308^{***}
f_6 : sexMale	5,047	3,782	1.335	0.1828
Observations	397			
Adjusted R^2	0.486			
Deviance Explained	51.3%			

Note: *p<0.1; **p<0.05; ***p<0.01

We find that our Level I GAM model using all available predictors accounts for 51.3% of the deviance. The model finds that sex is not a significant variable - on average, a Male Professor will earn $5,000 more than a Female colleague, holding all other remaining predictors constant. The professor rank has the biggest impact, with an average salary bump of $28,000 relative to the preceding rank. The caveat, as we saw previously, is the uneven gender proportion among professors. So while the model sees a gap in the average salary conditional to sex, it seems to be influenced more heavily by other factors. Nevertheless, these conclusions can only be stated for the data at hand in a very cautious fashion - we would require more information about the data itself and substantial subject matter expertise to reach any broader conclusions.

[1] The code can be found in the Appendix. The model is fitted applying spline smoothers without tuning on both continuous variables, estimating out fitted values using the penalized regression sum of squares (PRSS) method

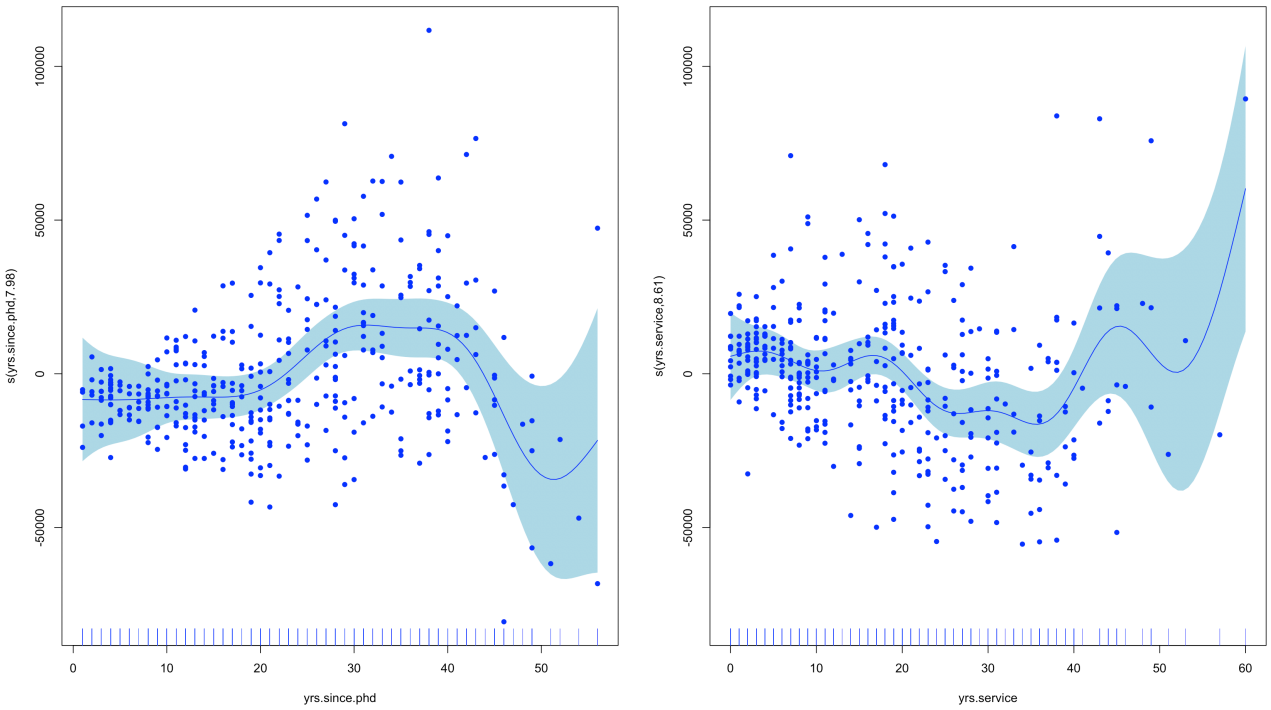


Figure 4: GAM Analysis of Professors's Salaries

In terms of Academic Discipline, professors Applied fields earn roughly $14,500 more on average, holding all other remaining predictors constant.

From the regression model from my previous paper found that rank and discipline had high significance, and the F Statistic told us that the model as a whole is significant. We also found evidence of collinearity in the case of Years Since PhD and Years of Service, with square root of GVID above 2 (2.76 and 2.5 respectively). With GAM, our plot in Figure 4 above, while we see some similarity in the overall fit of $X_i, f_i(X_i)$, the plot shows a very variable conditional relationship. Once again, we cannot make strong assertions given our doubts regarding the data and lack of subject matter expertise.

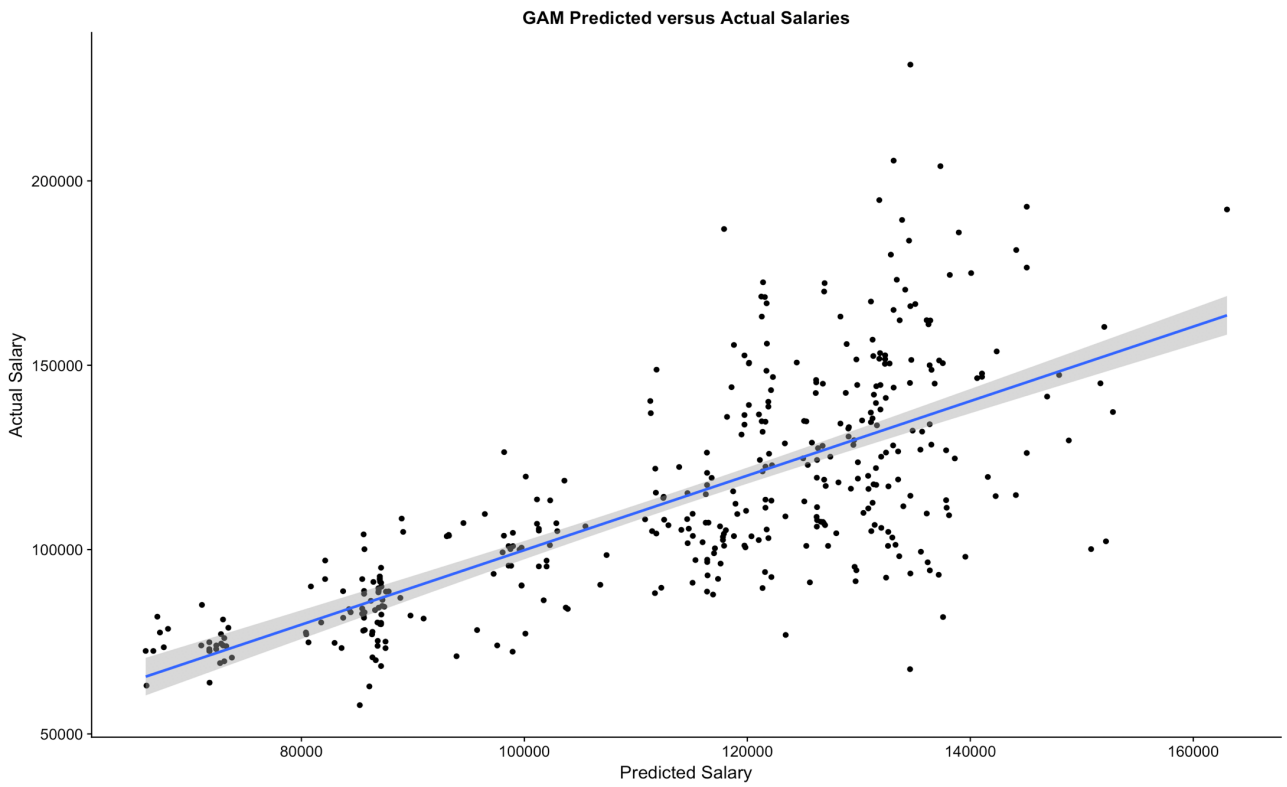


Figure 5: GAM - Predicted versus Actual Salaries

The fact that GAM shows rank and discipline to be the most significant. Can, to an extent, be explained intuitively - applied fields are known to pay more, and it would be rare to find Professors paid less than Associate Professors, and even rarer, paid less than Assistant professors. It also makes sense that the more years an academic is in the field, the higher the odds of them having a higher rank, which could explain the multicollinearity from our OLS regression with Years of Service and Years Since PhD. Nevertheless, these are only observations and possible interpretations from a Level I perspective. If we wanted to conduct a level II analysis, we would need to make the case that each observation in the dataset was independently realized from a relevant joint probability distribution. The uneven proportions of rank and gender were a red herring to this not being the case, particularly at the professor level, with 93.2% of Professors being male running a chi-square test on both predictors rejects independence at $p = 0.0141$.

4 Conclusion

There are numerous problems with the data on several dimensions, from the small sample size, skewness of numerical discrete predictors, to the source itself among others. Data transformation using *log*, for instance, was not possible for the continuous variables given the presence of observations where the value of either attribute was $x_i = 0$. The presence of professors with a higher value for The documentation is sparse, and there is no information regarding how this data was collected (the *salary* data set is also around 10 years old). There is no way to ensure that this sample contains a representative sample from this anonymous university, making it impossible to rule out participation bias. We also have a very limited number of predictors - other predictors such as a Professor's own research portfolio performance, hours taught during this 9-month period, among others, could prove to be highly significant. A larger sample encompassing more universities could provide a much larger insight, so with the data at hand, what we see is what we get.

While the average mean salary conditional to gender is different, the difference cannot be attributed to gender alone. Younger female academics in the sample could explain why there are less Professor-rank females, which in turn could have an impact in the salary.

It would be interesting to explore whether gender and other variables such as age, as well as less ambiguous variables relating to years in the field of academia, are related to Rank and Discipline, which in turn relate to Salary.

5 Appendix

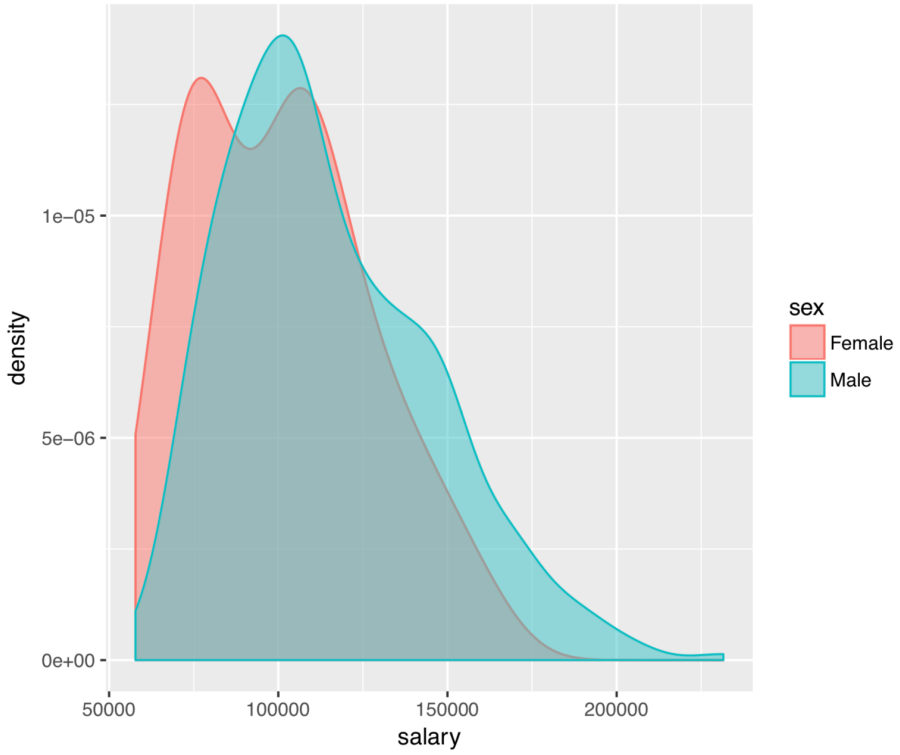


Figure 6: Figure 1.1: Density Distribution of Salary, by Gender

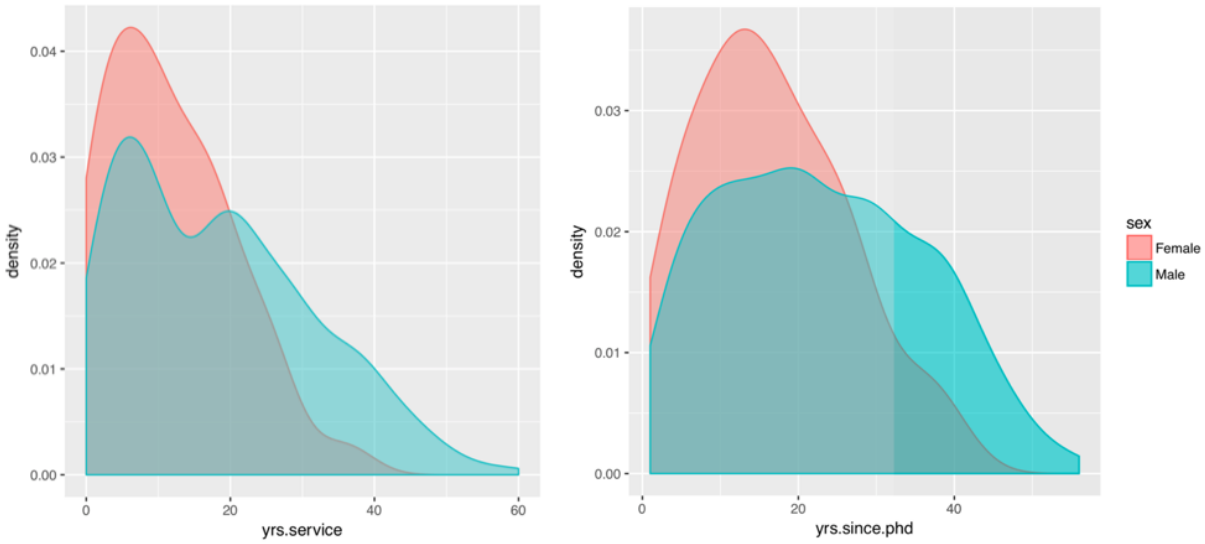


Figure 7: Figure 1.2: Density Distribution of Years of Service and Since PhD, by Gender

Sex	Discipline	Rank	Frequency
Female	A	Assistant Professor	6
	B		5
Male	A		18
	B		38
Female	A	Associate Professor	4
	B		6
Male	A		22
	B		32
Female	A	Professor	8
	B		10
Male	A		123
	B		125

Table 7: Gender Frequency by Rank and Discipline, $N = 397$

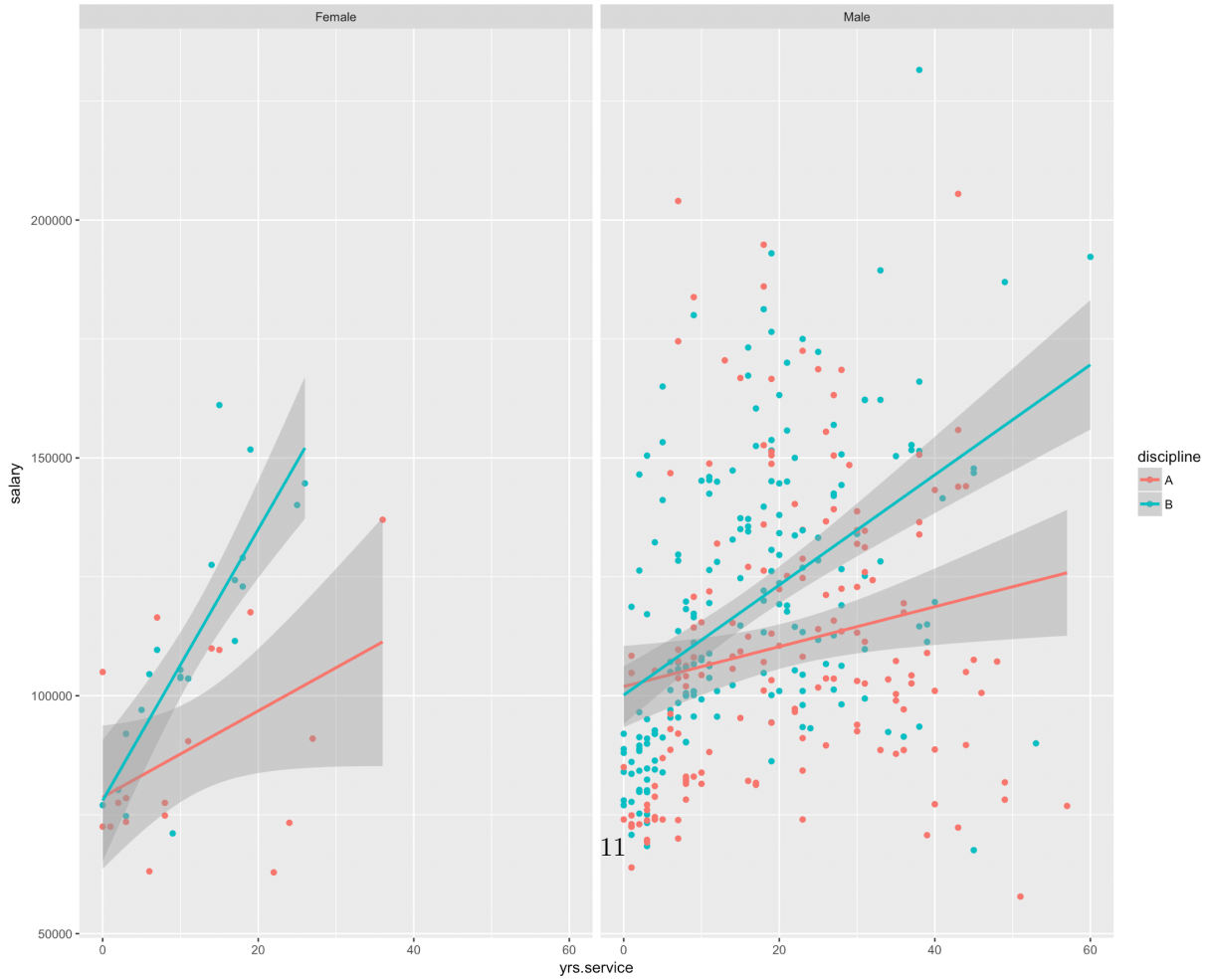


Figure 8: Figure 1.3: Years of Service versus Salary, by Gender

Figure 9: Years of Service versus Salary, by Rank

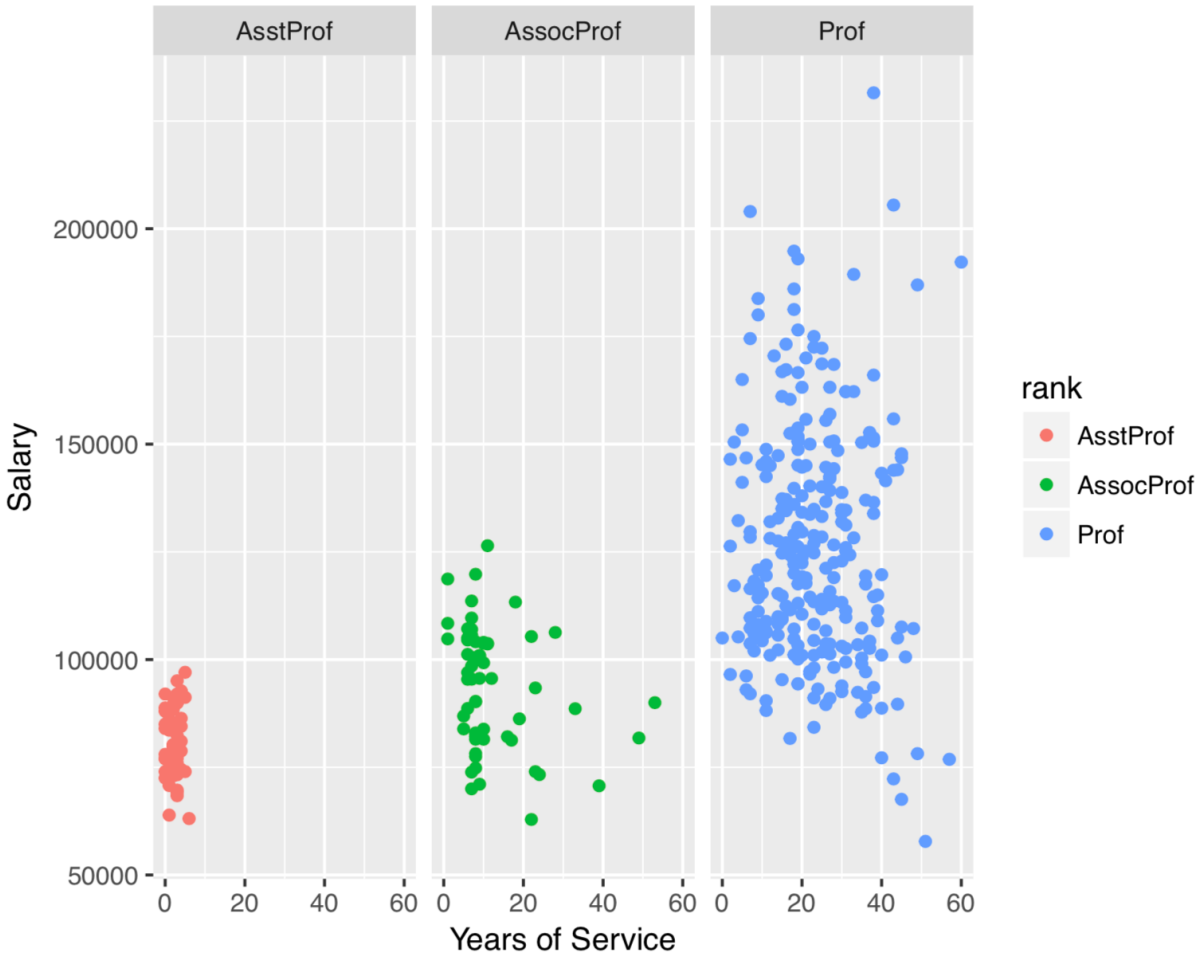
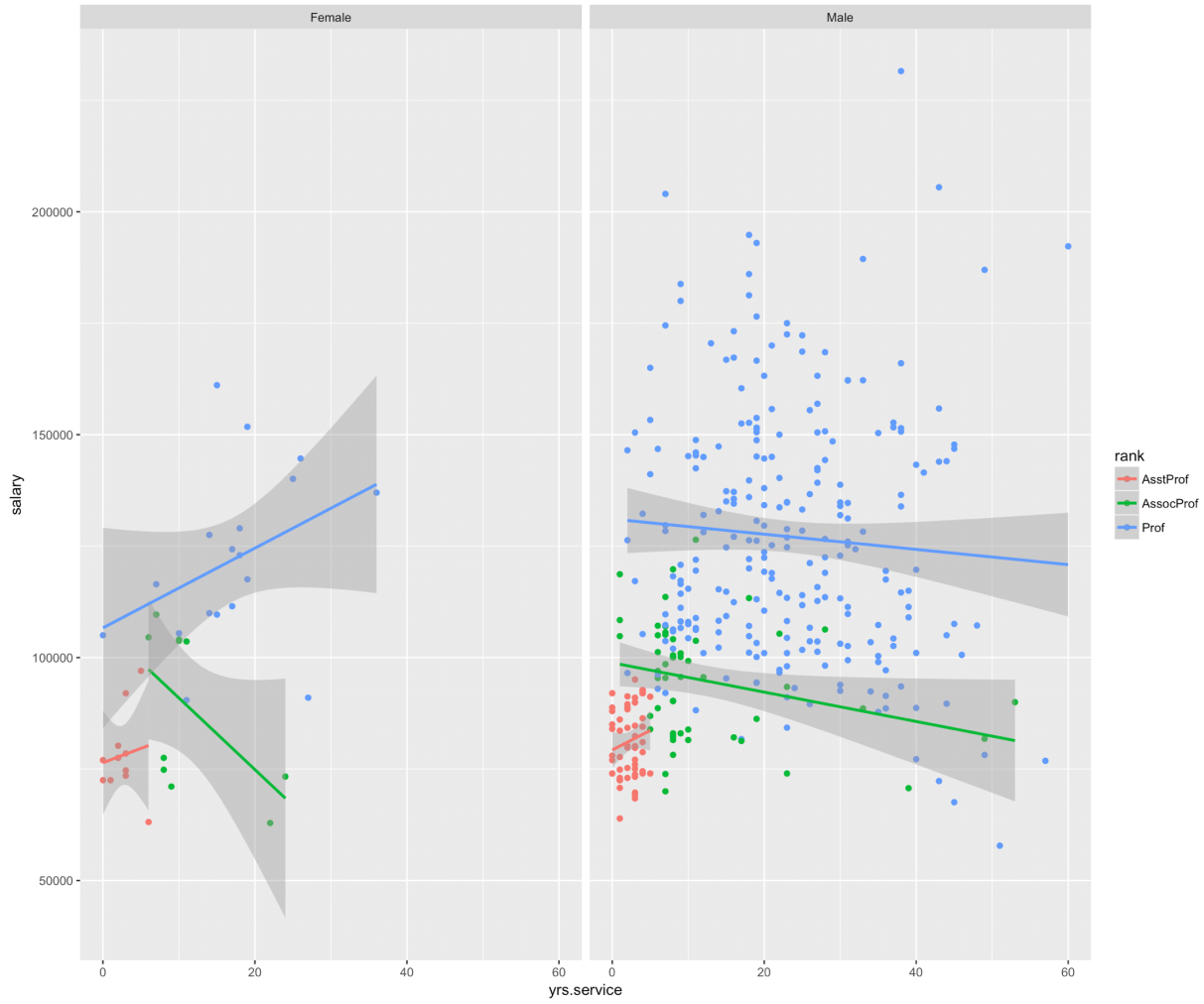


Figure 10: Years of Service versus Salary, by Gender and Rank



6 R Code - GAM

```r
## A Wrong Generalized Additive model
gam1 <-  gam(salary ~ rank + discipline + s(yrs.since.phd) + s(yrs.service) + sex, data = df)


summary(gam1) # GAM Summary

## Getting Generalization Error Estimate
preds_gam <-predict(gam1, newdata=df) # Get test data fitted values
df$preds_gam1 <- preds_gam

# Plot fit (ggplot2)
gg_df_gam <- df %>%
  ggplot(aes(x = preds_gam1, y = salary)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  xlab("Predicted Salary") +
  ylab("Actual Salary") +
  labs(title = "GAM Predicted versus Actual Salaries")

gg_df_gam

par(mfrow = c(1,2))
plot(gam1, se = T, residuals = T, pch = 19, cex = 0.8, col = "blue", shade = T,  shade.col = "light blue")

sg <- summary.gam(gam1)
sg
```

References

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