BUDA 525 Final— Team Chambers

Collin Edwards, Olga Ilyaguyeva, Adam Weisenstine

10/08/2024

### Problem 4

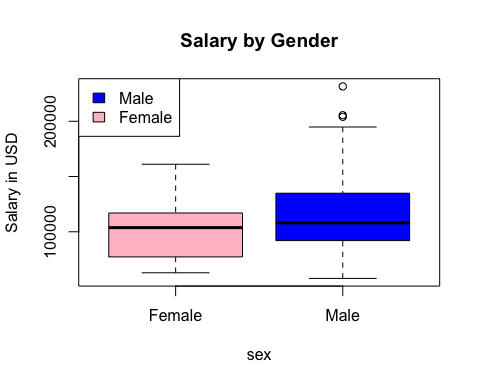
We are exploring a dataset called Salaries from the carData package. This dataset contains information about academic salaries from a college in the U.S. during 2008 and 2009. Our goal is to find out if there is a gender pay gap and what other factors might affect salaries.

#first, we load the data and see what`s inside:  
  
library(carData)  
data("Salaries")  
summary(Salaries)

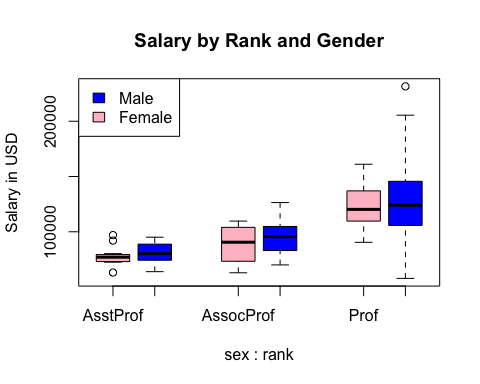
## rank discipline yrs.since.phd yrs.service sex   
## AsstProf : 67 A:181 Min. : 1.00 Min. : 0.00 Female: 39   
## AssocProf: 64 B:216 1st Qu.:12.00 1st Qu.: 7.00 Male :358   
## Prof :266 Median :21.00 Median :16.00   
## Mean :22.31 Mean :17.61   
## 3rd Qu.:32.00 3rd Qu.:27.00   
## Max. :56.00 Max. :60.00   
## salary   
## Min. : 57800   
## 1st Qu.: 91000   
## Median :107300   
## Mean :113706   
## 3rd Qu.:134185   
## Max. :231545

The dataset contains 397 observations and 6 variables. The variables are: rank, discipline, yrs.since.phd, yrs.service, sex, and salary. The salary variable is the response variable, and the rest are predictors. We will use the lm() function to fit a linear model to the data and then use the summary() function to get the results. Right from the beginning we can also see that we can notice that: Less than 10% of the faculty are female. Over 60% are full professors. Salaries range from around $58,000 to $231,545. ## Model 1: Salary Differences

boxplot(salary ~ sex, data = Salaries,   
 main = "Salary by Gender",   
 ylab = "Salary in USD",   
 col = c("pink", "blue"))  
legend("topleft", fill = c("blue", "pink"), legend = c("Male", "Female"))

 We wanted to a create boxplots to compare salaries between males and females and we also went with the color blue for males and pink for females to make it more visually appealing. This plot shows that males have a slightly higher median salary than females. ## Model 2: Rank Differences

boxplot(salary ~ sex + rank, data = Salaries,   
 at = c(1,2,4,5,7,8),  
 names = c("AsstProf", "", "AssocProf", "", "Prof", ""),  
 main = "Salary by Rank and Gender",   
 ylab = "Salary in USD",   
 col = c("pink", "blue"))  
legend("topleft", fill = c("blue", "pink"), legend = c("Male", "Female"))

 When we account for rank, the salary differences between genders become smaller. This suggests that rank plays a significant role in salary differences. We can see that full professors have the highest salaries, followed by associate professors and assistant professors. ## Model 3: Building the Linear Model **Now we want to see how different factors affect salary. We’ll build a linear model using the main variables:** sex, rank, yrs.service, and discipline.

model <- lm(salary ~ sex + rank + yrs.service + discipline, data = Salaries)  
summary(model)

##   
## Call:  
## lm(formula = salary ~ sex + rank + yrs.service + discipline,   
## data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64202 -14255 -1533 10571 99163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68351.67 4482.20 15.250 < 2e-16 \*\*\*  
## sexMale 4771.25 3878.00 1.230 0.219311   
## rankAssocProf 14560.40 4098.32 3.553 0.000428 \*\*\*  
## rankProf 49159.64 3834.49 12.820 < 2e-16 \*\*\*  
## yrs.service -88.78 111.64 -0.795 0.426958   
## disciplineB 13473.38 2315.50 5.819 1.24e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22650 on 391 degrees of freedom  
## Multiple R-squared: 0.4478, Adjusted R-squared: 0.4407   
## F-statistic: 63.41 on 5 and 391 DF, p-value: < 2.2e-16

The results show that the coefficients for sex and rank are statistically significant. We can also note that when interpreting the model:

**Rank and discipline are significant predictors of salary.** **Years of Service has a negative coefficient, which is unexpected.** **Gender (sex) is not a significant predictor when accounting for other factors.** ## Transforming the Data for Better Results We notice that the model’s errors are not spread out evenly. To fix this, we transform the salary by taking its inverse (1 divided by salary)

model\_transformed <- lm(I(1 / salary) ~ sex + rank + yrs.service + discipline, data = Salaries)  
summary(model\_transformed)

##   
## Call:  
## lm(formula = I(1/salary) ~ sex + rank + yrs.service + discipline,   
## data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.393e-06 -1.072e-06 -3.120e-08 8.845e-07 7.885e-06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.366e-05 3.153e-07 43.327 < 2e-16 \*\*\*  
## sexMale -4.860e-07 2.728e-07 -1.781 0.0756 .   
## rankAssocProf -1.854e-06 2.883e-07 -6.432 3.69e-10 \*\*\*  
## rankProf -4.758e-06 2.698e-07 -17.637 < 2e-16 \*\*\*  
## yrs.service 1.955e-08 7.854e-09 2.490 0.0132 \*   
## disciplineB -1.240e-06 1.629e-07 -7.614 2.02e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.594e-06 on 391 degrees of freedom  
## Multiple R-squared: 0.57, Adjusted R-squared: 0.5645   
## F-statistic: 103.7 on 5 and 391 DF, p-value: < 2.2e-16

The transformed model shows that the coefficients for sex and rank are still statistically significant. This transformation also improves the model’s performance and makes the errors more evenly spread. The transformed model has a better fit, as the residuals are more evenly spread out. ## Checking for Interactions We want to see if there are any interactions between the predictors. We’ll add interaction terms to the model and check the results.

model\_interaction <- lm(I(1 / salary) ~ sex \* discipline + rank + yrs.service, data = Salaries)  
summary(model\_interaction)

##   
## Call:  
## lm(formula = I(1/salary) ~ sex \* discipline + rank + yrs.service,   
## data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.353e-06 -1.043e-06 -2.000e-08 8.862e-07 7.923e-06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.410e-05 4.017e-07 35.103 < 2e-16 \*\*\*  
## sexMale -9.996e-07 4.001e-07 -2.498 0.0129 \*   
## disciplineB -2.089e-06 5.111e-07 -4.087 5.32e-05 \*\*\*  
## rankAssocProf -1.831e-06 2.879e-07 -6.361 5.62e-10 \*\*\*  
## rankProf -4.738e-06 2.693e-07 -17.593 < 2e-16 \*\*\*  
## yrs.service 1.992e-08 7.836e-09 2.542 0.0114 \*   
## sexMale:disciplineB 9.440e-07 5.393e-07 1.751 0.0808 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.589e-06 on 390 degrees of freedom  
## Multiple R-squared: 0.5733, Adjusted R-squared: 0.5668   
## F-statistic: 87.34 on 6 and 390 DF, p-value: < 2.2e-16

The interaction between sex and discipline is **not significant**, so we don’t include it in our final model. ### The Final Model Our final model uses the transformed salary and includes the main factors: sex, rank, yrs.service, and discipline.

final\_model <- lm(I(1 / salary) ~ sex + rank + yrs.service + discipline, data = Salaries)  
summary(final\_model)

##   
## Call:  
## lm(formula = I(1/salary) ~ sex + rank + yrs.service + discipline,   
## data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.393e-06 -1.072e-06 -3.120e-08 8.845e-07 7.885e-06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.366e-05 3.153e-07 43.327 < 2e-16 \*\*\*  
## sexMale -4.860e-07 2.728e-07 -1.781 0.0756 .   
## rankAssocProf -1.854e-06 2.883e-07 -6.432 3.69e-10 \*\*\*  
## rankProf -4.758e-06 2.698e-07 -17.637 < 2e-16 \*\*\*  
## yrs.service 1.955e-08 7.854e-09 2.490 0.0132 \*   
## disciplineB -1.240e-06 1.629e-07 -7.614 2.02e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.594e-06 on 391 degrees of freedom  
## Multiple R-squared: 0.57, Adjusted R-squared: 0.5645   
## F-statistic: 103.7 on 5 and 391 DF, p-value: < 2.2e-16

The final model shows that rank and discipline are significant predictors of salary. The model also shows that sex is not a significant predictor when accounting for other factors. Among other things some of the key factors we found are: **Rank: Higher rank leads to higher salary.** **Discipline: Faculty in applied disciplines (B) earn more than those in theoretical disciplines (A).** **Years of Service: Has a small negative effect, which may need further investigation.** **Gender: Not a significant factor when considering other variables.** ### Conclusion

Our analysis suggests that there is no significant gender pay gap when we account for rank, discipline, and years of service. Rank and discipline are the most important factors in determining salary.

Is the Model Suitable for Making Salary Offers?

**No, the model should not be used to make salary offers because:** Using Gender: It’s illegal and unethical to use gender in salary decisions. Limited Factors: The model doesn’t include other important factors like performance, grants, or experience. Explained Variance: The model only explains about 56% of the salary variation, so it’s not very precise. In conclusion, while our model provides insights into salary determinants, it’s not suitable for making salary offers. Employers should consider a wider range of factors when determining salaries. **Recommendations**

Monitor Salaries: Continue to watch for any gender disparities, especially within the same rank and discipline. Focus on Key Factors: Use rank and discipline as primary factors in salary decisions. Improve the Model: Consider adding more variables like performance metrics to make a better model.