Math 342 Workshop 4

Jonathan Lee

2/15/2022

Case Study: Salaries for Professors

The dataset consists of the 2008-2009 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 ongoing effort of the college's administration to monitor salary differences between male and female faculty members.

The variables collected include the following: rank (AssocProf, AsstProf, Prof), discipline (A = Theoretical Department, B = Applied Department), yrs.since.phd (years since PhD), yrs.service (years of service), sex (Female, Male), salary (nine month salary in dollars).

The dataset can be accessed from the car package in R, stored as Salaries. Your task is to determine the best model to predict salaries.

a. Attach the Salaries dataset and save it under the variable data. Print the first few observations in the dataset.

```
attach(Salaries)
data = Salaries
head(data)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

b. Create dummy variables for the rank variable using the following coding scheme: rank1 = 1 if Professor, 0 otherwise, rank2 = 1 if Associate Professor, 0 otherwise. Store these variables in the data.

```
data$AssocProf = ifelse(data$rank=="AssocProf",1,0)
data$Prof = ifelse(data$rank=="Prof",1,0)
```

c. Create an indicator variable for sex using the following coding scheme: sex1 = 1 if Male, 0 if Female. Add this variable in the data.

```
data$sex1 = ifelse(data$sex=="Male",1,0)
```

d. Create an indicator variable for discipline variable using the following coding scheme: disc = 1 if A, 0 if B. Add this variable in the data.

```
data$disc =  ifelse(data$discipline=="A", 1, 0)
```

e. Fit the full model to the data using only the newly created dummy variables and indicator variables, years since phd, and years of service to predict salary. Print the coefficients column and use the output to determine if the overall model is significant. If it is, test the significance of each coefficient.

```
head(data)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary AssocProf Prof
## 1      Prof          B            19          18 Male 139750         0    1
## 2      Prof          B            20          16 Male 173200         0    1
## 3  AsstProf          B             4           3 Male  79750         0    0
## 4      Prof          B            45          39 Male 115000         0    1
## 5      Prof          B            40          41 Male 141500         0    1
## 6 AssocProf          B             6           6 Male  97000         1    0
##   sex1 disc
## 1    1    0
## 2    1    0
## 3    1    0
## 4    1    0
## 5    1    0
## 6    1    0
```

```
model = lm(salary ~ yrs.since.phd + yrs.service + sex1 + disc + AssocProf + Prof, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + yrs.service + sex1 + disc +
##     AssocProf + Prof, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65248 -13211  -1775  10384  99592
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    80372.9     4372.3  18.382  < 2e-16 ***
## yrs.since.phd    535.1      241.0   2.220  0.02698 *
## yrs.service     -489.5      211.9  -2.310  0.02143 *
## sex1            4783.5     3858.7   1.240  0.21584
## disc          -14417.6     2342.9  -6.154 1.88e-09 ***
## AssocProf      12907.6     4145.3   3.114  0.00198 **
## Prof           45066.0     4237.5  10.635  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22540 on 390 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4463
## F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

All of the predictors' p-values have values less than 0.05 except for sex. From this data, we can interpret that yrs.since.phd, yrs.service, discipline, and rank matters in predicting salaries.

f. Interpret the coefficients for rank1, rank2, and disc.

Based on the coefficients in rank1, rank2, and disc. We can see that on average if the person is a professor, he or she will make about 45066.0 more than the base pay. Associate Professors on average make 12907.60 more than the base pay. Discipline A on average makes 14417.6 less than discipline B.

g. Interpret the coefficients of the numerical predictors.

For the numerical predictors yrs.since.phd, we can see that the longer the time the person has acquired their phd, they tend to make on average 535.1 dollars more than the base pay. For yrs.service, we can see that longer services tend to make 489.5 dollars less.

h. The coefficient for the years of service indicates a decrease in salary as the number of years of service increases. Does it make sense for the slope of this variable to be negative? What do you think caused this scenario?

The negative slope of this variable seems to be balanced out by the years since phd. This could be caused by the scenario of staying stagnant in a job pays you less over time than changing jobs.

i. Based on the p-value from the coefficients table, is there a difference in salary between male and female professors? Why or why not?

Based on the p-value from the coefficients table, there seems to be lack of evidence to prove that there is a significant difference between male and female professors since the p-value of sex1 is less than 0.05.

j. Perform stepwise regression using the full model in part e as input, with $\alpha_{in} = 0.15$ and $\alpha_{out} = 0.15$. What are the predictors in the final model?

```
ols_step_both_p(model, pent=0.15)
```

```
##
##                                Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step    Variable     Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ----------------------------------------------------------------------------------------------
##    1      Prof       addition       0.379       0.377    51.2730    9135.5524    23903.1032
##    2      disc       addition       0.428       0.425    18.1280    9104.8349    22967.2623
##    3    AssocProf    addition       0.445       0.441     7.9340    9094.8231    22651.1854
## ----------------------------------------------------------------------------------------------
```

The predictors in the final model are Prof, AssocProf, and disc. ### k. Perform best subset regression using the full model in part e as input. Specify the predictors in the best model selected using these criteria: r^2_{adj}, r^2_{PRED}, C_p, AIC, and SBC.

```
ols_step_best_subset(model)
```

```
##                       Best Subsets Regression
## --------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------
##      1         Prof
##      2         disc Prof
```

```
##      3         disc AssocProf Prof
##      4         sex1 disc AssocProf Prof
##      5         yrs.since.phd yrs.service disc AssocProf Prof
##      6         yrs.since.phd yrs.service sex1 disc AssocProf Prof
## -----------------------------------------------------------------
##
##                                                                  Subsets Regression Summary
## ------------------------------------------------------------------------------------------------
##                          Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)         AIC         SBIC          SBC
## ------------------------------------------------------------------------------------------------
##   1        0.3788      0.3772      0.3736    51.2731    9135.5524    8008.4670    9147.5042    226829
##   2        0.4279      0.4250      0.4203    18.1278    9104.8349    7978.0181    9120.7707    209416
##   3        0.4450      0.4407      0.4372     7.9344    9094.8231    7968.1872    9114.7428    203693
##   4        0.4469      0.4412       0.438     8.5681    9095.4542    7968.8532    9119.3578    203511
##   5        0.4525      0.4455      0.4344     6.5368    9093.3875    7966.9174    9121.2751    201954
##   6        0.4547      0.4463      0.4355     7.0000    9093.8262    7967.4397    9125.6977    201678
## ------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Based on R^2 adj, model 6 is the better fit because its R^2 adj is the highest. Based on R^2 pred, model 4 is the better fit because its R^2 pred is the highest. Based on c(p), model 5 is the better fit because its c(p) is the lowest. Based on AIC, model 5 is the better fit because its AIC is the lowest. Based on SBC, model 3 is the better fit because its SBC is the lowest.

l. Which model would you recommend to use and why?

I would recommend to use model 5 because it was proven to be a better fit twice based on c(p) and AIC. We also had already determined that sex was not a significant variable in the dataset in predicting salaries while the other variables were significant.

m. Based on the model you selected, construct a 95% prediction interval for a female associate professor in the applied department with 8 years of service and 10 years after PhD. Note: Use the values only for the predictors in your final model.

sex1 = 0 yrs.service = 8 yrs.since.phd = 10

```
model5 = lm(salary ~ yrs.since.phd + yrs.service + disc + AssocProf + Prof, data=data)
new = data.frame(sex1 = c(1), disc = c(0), yrs.service = c(8), yrs.since.phd = c(10), AssocProf = c(1),
predict(model5, new)
```

```
##        1
## 98738.27
```

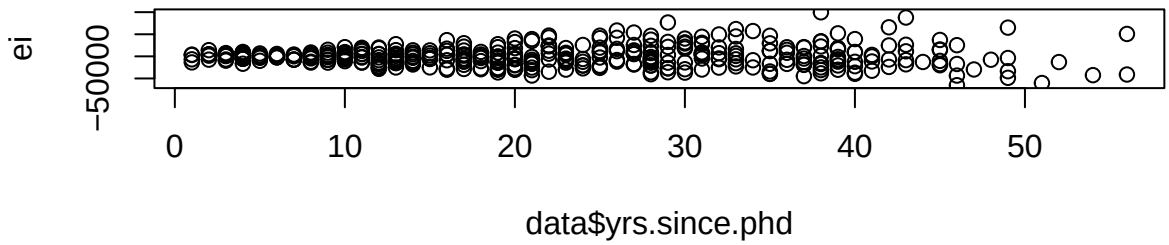
```
predict(model5, new, interval = "prediction")
```

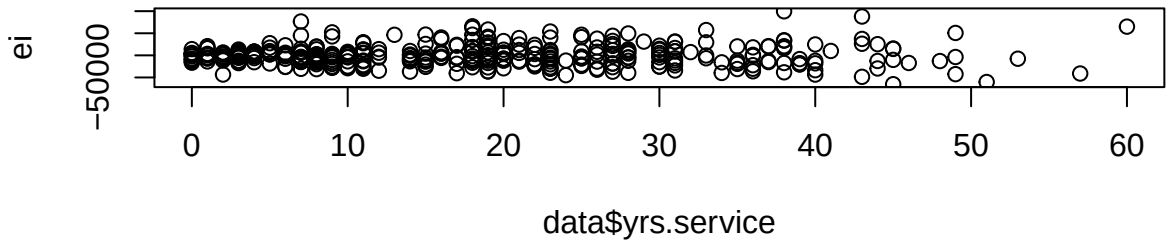
```
##        fit      lwr      upr
## 1 98738.27 54001.04 143475.5
```

n. Perform diagnostic checking on your final model. Identify if there's any outliers or influential observations in the dataset using the model you selected.

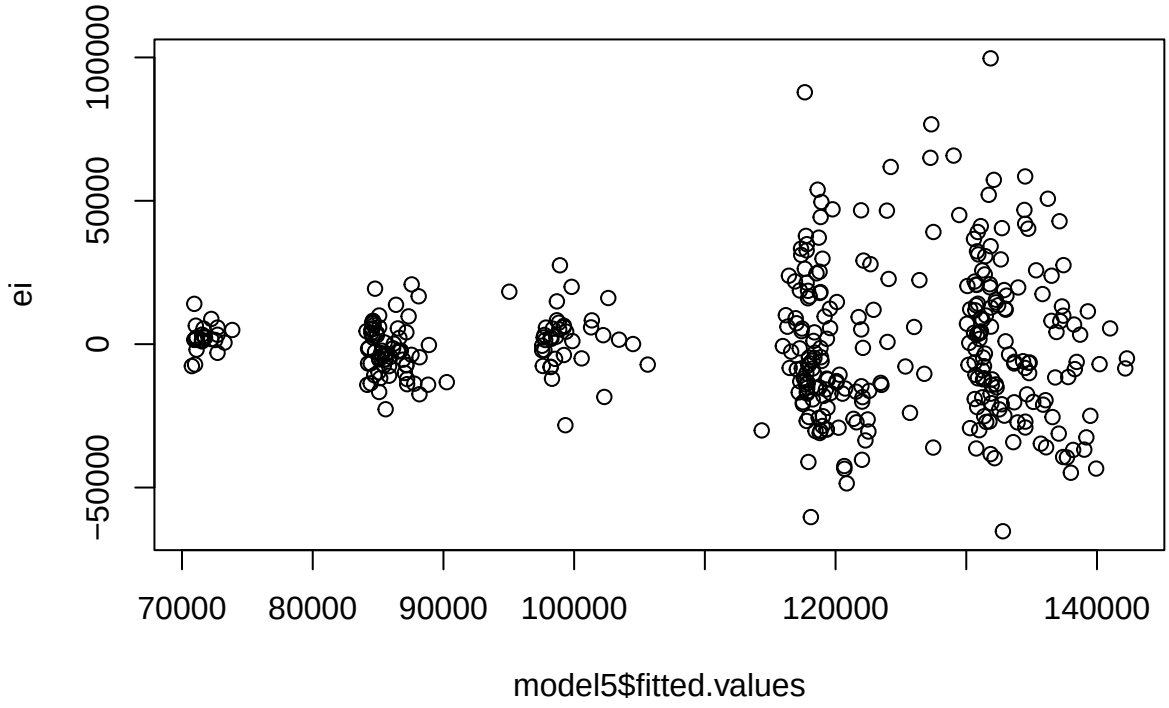
```
ei = model5$residuals

par(mfrow= c(2,1))
plot(data$yrs.since.phd, ei)
plot(data$yrs.service, ei)
```

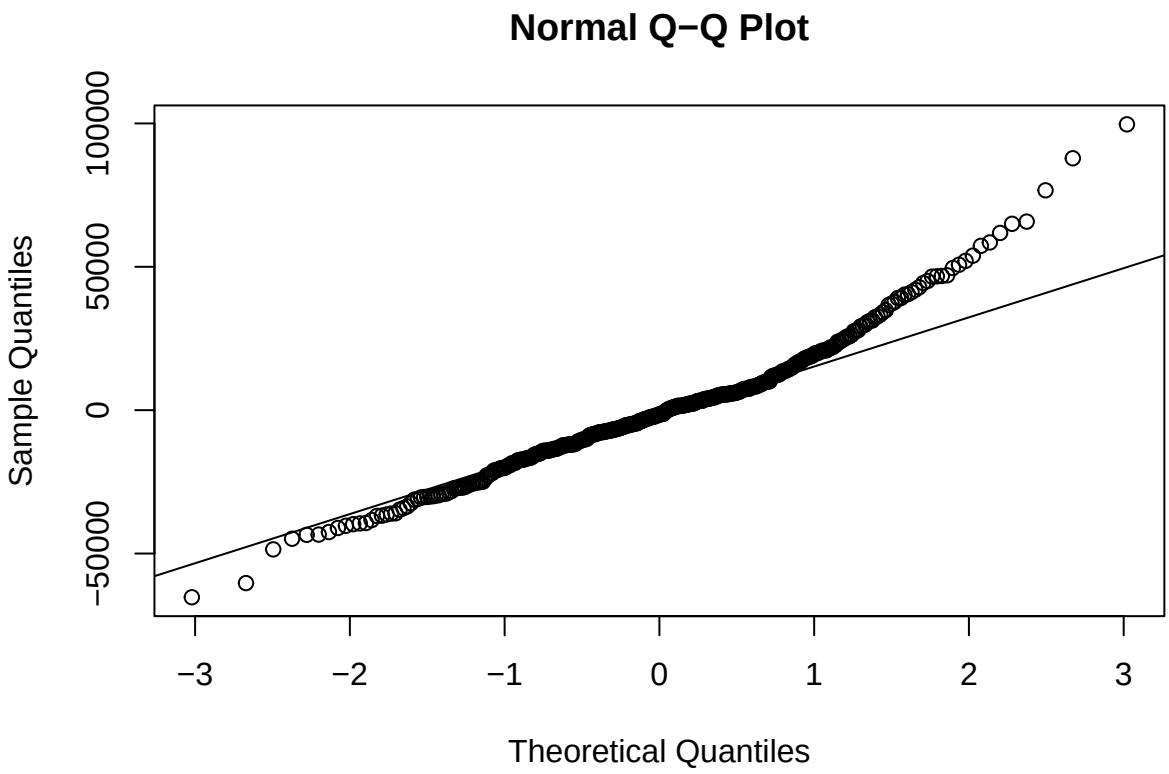





```
par(mfrow = c(1,1))
plot(model5$fitted.values, ei)
```



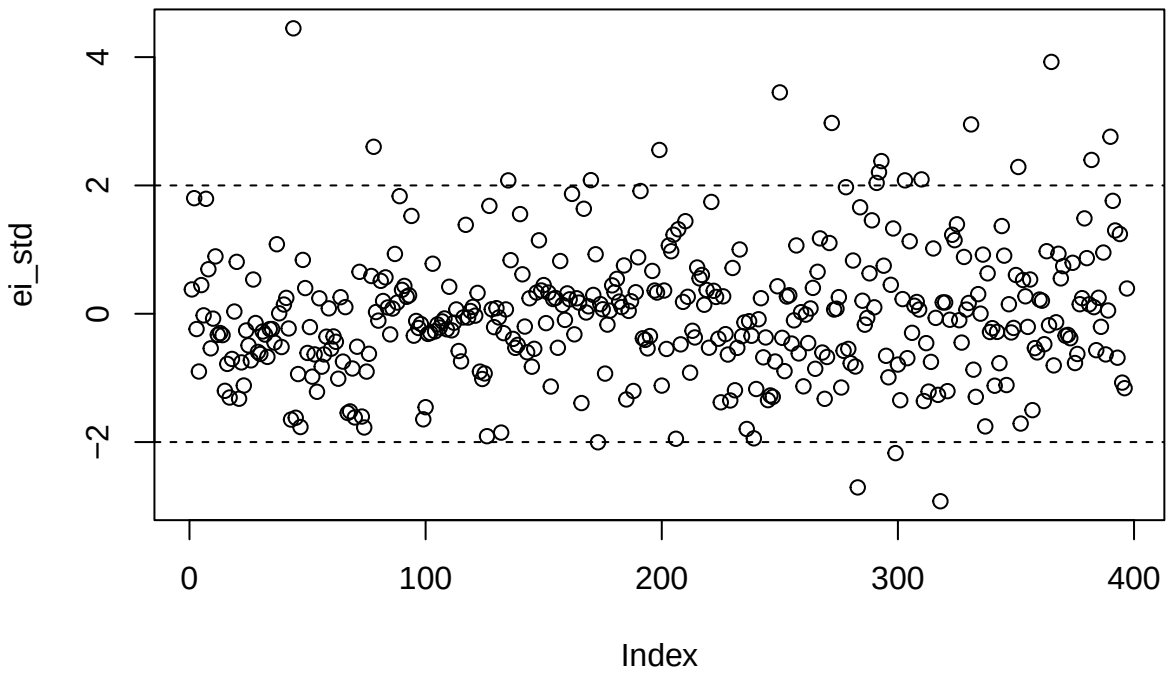

```
qqnorm(ei)
qqline(ei)
```



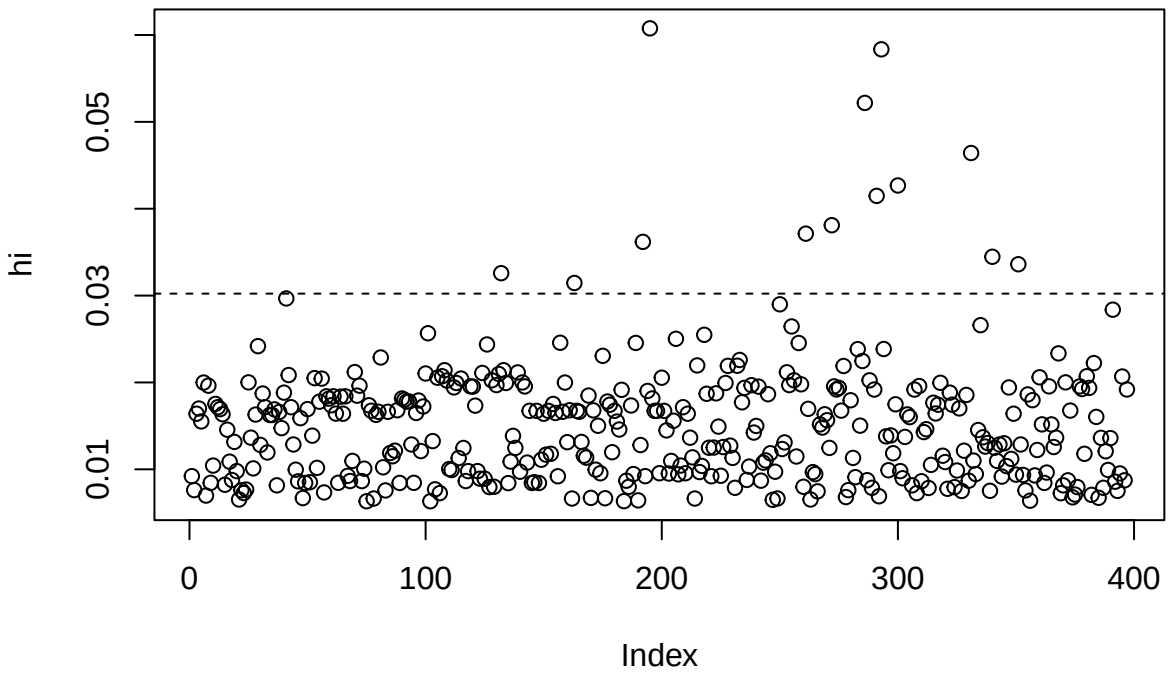

```
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.96734, p-value = 9.546e-08
```

```
ei_std = rstandard(model5)
plot(ei_std)
abline(h=c(-2,2), lty=2)
```




```
hi = hatvalues(model5)
bm = 2*(5+1)/397
plot(hi)
abline(h=bm, lty=2)
```



Based on our plots, linearity seems to be satisfied for each numerical variable, constant variance seems to be satisfied, and based on the shapiro test, our p-value is below 0.05 so it is not normality is not satisfied. There also seems to be many outliers and influential values.