DATA 605 - Discussion 12

Omer Ozeren

Table of Contents

## Objective

Using R, build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

I am using salary data that include observations on six variables for 52 tenure-track professors in a small college. The original data also can be found in “<http://data.princeton.edu/wws509/datasets/>”

The variables are:

* sx: Sex, coded 1 for female and 0 for male
* rk: Rank, coded 1 for assistant professor, 2 for associate professor, and 3 for full professor
* yr: Number of years in current rank
* dg: Highest degree, coded 1 if doctorate, 0 if masters
* yd: Number of years since highest degree was earned
* sl: Academic year salary, in dollars.

## Data Import

# Data import  
library(foreign)  
salary <- read.dta("http://data.princeton.edu/wws509/datasets/salary.dta")  
# Summary of Salary Data  
summary(salary)

## sx rk yr dg   
## Male :38 Assistant:18 Min. : 0.000 Min. :0.0000   
## Female:14 Associate:14 1st Qu.: 3.000 1st Qu.:0.0000   
## Full :20 Median : 7.000 Median :1.0000   
## Mean : 7.481 Mean :0.6538   
## 3rd Qu.:11.000 3rd Qu.:1.0000   
## Max. :25.000 Max. :1.0000   
## yd sl   
## Min. : 1.00 Min. :15000   
## 1st Qu.: 6.75 1st Qu.:18247   
## Median :15.50 Median :23719   
## Mean :16.12 Mean :23798   
## 3rd Qu.:23.25 3rd Qu.:27259   
## Max. :35.00 Max. :38045

# Data sample  
knitr::kable(head(salary))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sx | rk | yr | dg | yd | sl |
| Male | Full | 25 | 1 | 35 | 36350 |
| Male | Full | 13 | 1 | 22 | 35350 |
| Male | Full | 10 | 1 | 23 | 28200 |
| Female | Full | 7 | 1 | 27 | 26775 |
| Male | Full | 19 | 0 | 30 | 33696 |
| Male | Full | 16 | 1 | 21 | 28516 |

## Data Engineering : Sex (sx) will be **dichotomous** variable.

Convert *sex* and *rank* into numerical representation.

salary$sx <- as.character(salary$sx)  
salary$sx[salary$sx == "Male"] <- 0  
salary$sx[salary$sx == "Female"] <- 1  
salary$sx <- as.integer(salary$sx)  
salary$rk <- as.character(salary$rk)  
salary$rk[salary$rk == "Assistant"] <- 1  
salary$rk[salary$rk == "Associate"] <- 2  
salary$rk[salary$rk == "Full"] <- 3  
salary$rk <- as.integer(salary$rk)

## Iniitial Model

# Quadratic variable  
rk2 <- salary$rk^2  
sx\_yd <- salary$sx \* salary$yd  
  
# Initial model  
salary\_lm <- lm(sl ~ sx + rk + rk2 + yr + dg + yd + sx\_yd, data=salary)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ sx + rk + rk2 + yr + dg + yd + sx\_yd, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3827.5 -1180.3 -288.7 844.7 8709.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13045.79 3302.80 3.950 0.000279 \*\*\*  
## sx 127.83 1359.23 0.094 0.925502   
## rk 4230.31 3530.16 1.198 0.237202   
## rk2 340.06 845.99 0.402 0.689655   
## yr 523.83 105.21 4.979 1.03e-05 \*\*\*  
## dg -1514.35 1024.89 -1.478 0.146645   
## yd -174.42 90.99 -1.917 0.061751 .   
## sx\_yd 80.00 76.74 1.042 0.302882   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2396 on 44 degrees of freedom  
## Multiple R-squared: 0.8585, Adjusted R-squared: 0.836   
## F-statistic: 38.15 on 7 and 44 DF, p-value: < 2.2e-16

Perform **backwards elimination** - removing one variable (the one with highest p-value) at a time. Removing *sex*.

# Version 2  
salary\_lm <- update(salary\_lm, .~. -sx)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ rk + rk2 + yr + dg + yd + sx\_yd, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3822.3 -1186.7 -284.7 851.5 8710.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13159.28 3040.37 4.328 8.28e-05 \*\*\*  
## rk 4142.04 3365.41 1.231 0.2248   
## rk2 358.78 813.13 0.441 0.6612   
## yr 523.94 104.04 5.036 8.16e-06 \*\*\*  
## dg -1506.10 1009.82 -1.491 0.1428   
## yd -175.51 89.24 -1.967 0.0554 .   
## sx\_yd 85.29 51.63 1.652 0.1055   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2370 on 45 degrees of freedom  
## Multiple R-squared: 0.8585, Adjusted R-squared: 0.8396   
## F-statistic: 45.51 on 6 and 45 DF, p-value: < 2.2e-16

Removing *square of rank*.

# Version 3  
salary\_lm <- update(salary\_lm, .~. -rk2)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ rk + yr + dg + yd + sx\_yd, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3635.0 -1330.9 -218.3 615.3 8730.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11898.01 1026.60 11.590 3.04e-15 \*\*\*  
## rk 5598.87 645.84 8.669 3.11e-11 \*\*\*  
## yr 531.62 101.67 5.229 4.06e-06 \*\*\*  
## dg -1411.88 978.31 -1.443 0.1557   
## yd -180.92 87.61 -2.065 0.0446 \*   
## sx\_yd 88.38 50.70 1.743 0.0880 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2349 on 46 degrees of freedom  
## Multiple R-squared: 0.8579, Adjusted R-squared: 0.8424   
## F-statistic: 55.54 on 5 and 46 DF, p-value: < 2.2e-16

Removing *highest degree*.

# Version 4  
salary\_lm <- update(salary\_lm, .~. -dg)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ rk + yr + yd + sx\_yd, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3545.3 -1585.0 -432.7 884.0 8520.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11087.98 869.42 12.753 < 2e-16 \*\*\*  
## rk 5090.45 547.49 9.298 3.18e-12 \*\*\*  
## yr 480.09 96.28 4.986 8.81e-06 \*\*\*  
## yd -94.26 64.53 -1.461 0.151   
## sx\_yd 66.10 48.85 1.353 0.182   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2376 on 47 degrees of freedom  
## Multiple R-squared: 0.8515, Adjusted R-squared: 0.8388   
## F-statistic: 67.35 on 4 and 47 DF, p-value: < 2.2e-16

Removing *interaction between sex and number of years since highest degree was earned*.

# Version 5  
salary\_lm <- update(salary\_lm, .~. -sx\_yd)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ rk + yr + yd, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3329.7 -1135.6 -377.9 801.5 9576.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11282.90 864.79 13.047 < 2e-16 \*\*\*  
## rk 4973.64 545.30 9.121 4.71e-12 \*\*\*  
## yr 405.67 79.71 5.089 5.94e-06 \*\*\*  
## yd -40.86 51.50 -0.794 0.431   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2396 on 48 degrees of freedom  
## Multiple R-squared: 0.8457, Adjusted R-squared: 0.836   
## F-statistic: 87.68 on 3 and 48 DF, p-value: < 2.2e-16

Removing *number of years since highest degree was earned*.

# Version 6  
salary\_lm <- update(salary\_lm, .~. -yd)  
summary(salary\_lm)

##   
## Call:  
## lm(formula = sl ~ rk + yr, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3339.4 -1451.0 -323.3 821.3 9502.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11336.67 858.87 13.200 < 2e-16 \*\*\*  
## rk 4731.26 450.01 10.514 3.72e-14 \*\*\*  
## yr 376.50 70.46 5.344 2.36e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2387 on 49 degrees of freedom  
## Multiple R-squared: 0.8436, Adjusted R-squared: 0.8373   
## F-statistic: 132.2 on 2 and 49 DF, p-value: < 2.2e-16

## Summary of Model Results

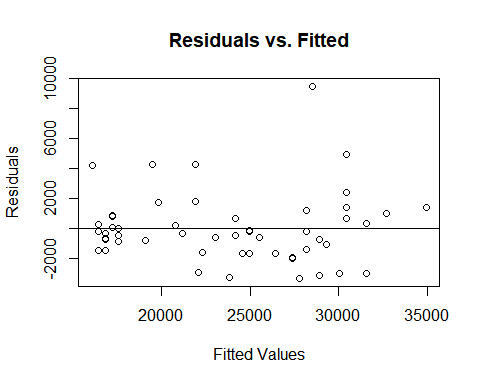
The final model has two variables - *rank* and *number of years in current rank* - that can be used to predict the target variable.

Two coefficients imply that for every increase in rank the salary increases by $4,731.26 and with every year in the current rank the salary increases by $376.50.

Based on the Residuals vs. Fitted plot below there are some outliers in the data, but overall variability is fairly consistent. Based on the Q-Q plot, distribution of residuals is close to normal.

Based on value, the model explains 84.36% of variability in the data.

plot(salary\_lm$fitted.values, salary\_lm$residuals, xlab="Fitted Values", ylab="Residuals", main="Residuals vs. Fitted")  
abline(h=0)



qqnorm(salary\_lm$residuals)  
qqline(salary\_lm$residuals)

