DSA 8020 R Lab 5: Analysis of covariance and Non-linear Regression

Whitney

February 17, 2021

Contents

Analysis of covariance: Salaries for Professors

The 2008-09 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.

Load the dataset

Code:

```
library(carData)
data(Salaries)
head(Salaries)
```

```
##        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
```

Description of the variables

- `rank`: a factor with levels Assistant Professor ("AsstProf"); Associate Professor ("AssocProf"); Full Professor ("Prof")

- `discipline`: a factor with levels A ("theoretical" departments) or B ("applied" departments)
- `yrs.since.phd`: years since her/his PhD
- `sex`: a factor with levels "Female" and "Male"
- `salary`: nine-month salary, in dollars

Exploratory Data Analysis

1. Identify the numerical variables and categorical variabes in this data set

Answer: Numerical variables: `yrs.since.phd`, `yrs.service`, `salary`. Categorical variabes: `rank`, `discipline`, `sex`

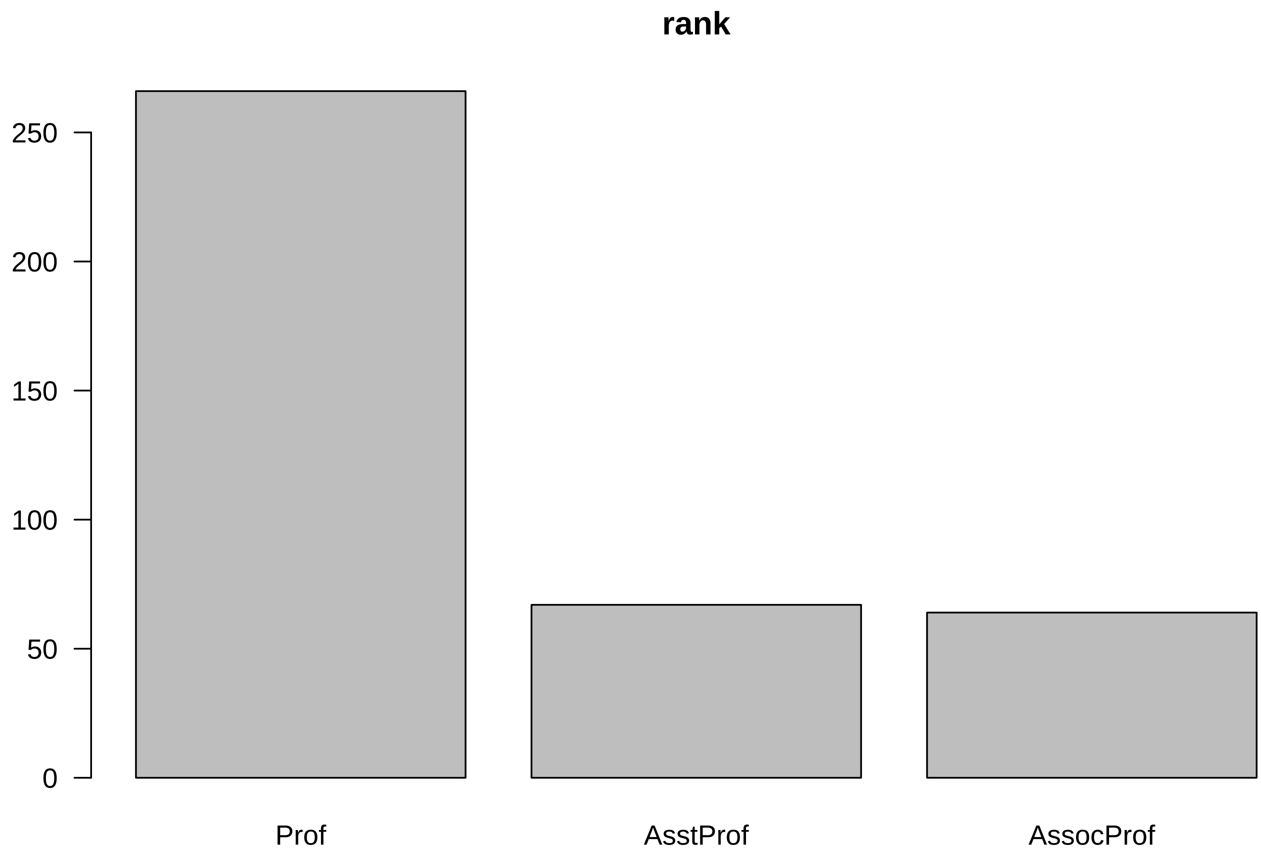
2. Summarize each varaible numerically and graphically, briefly describe your findings

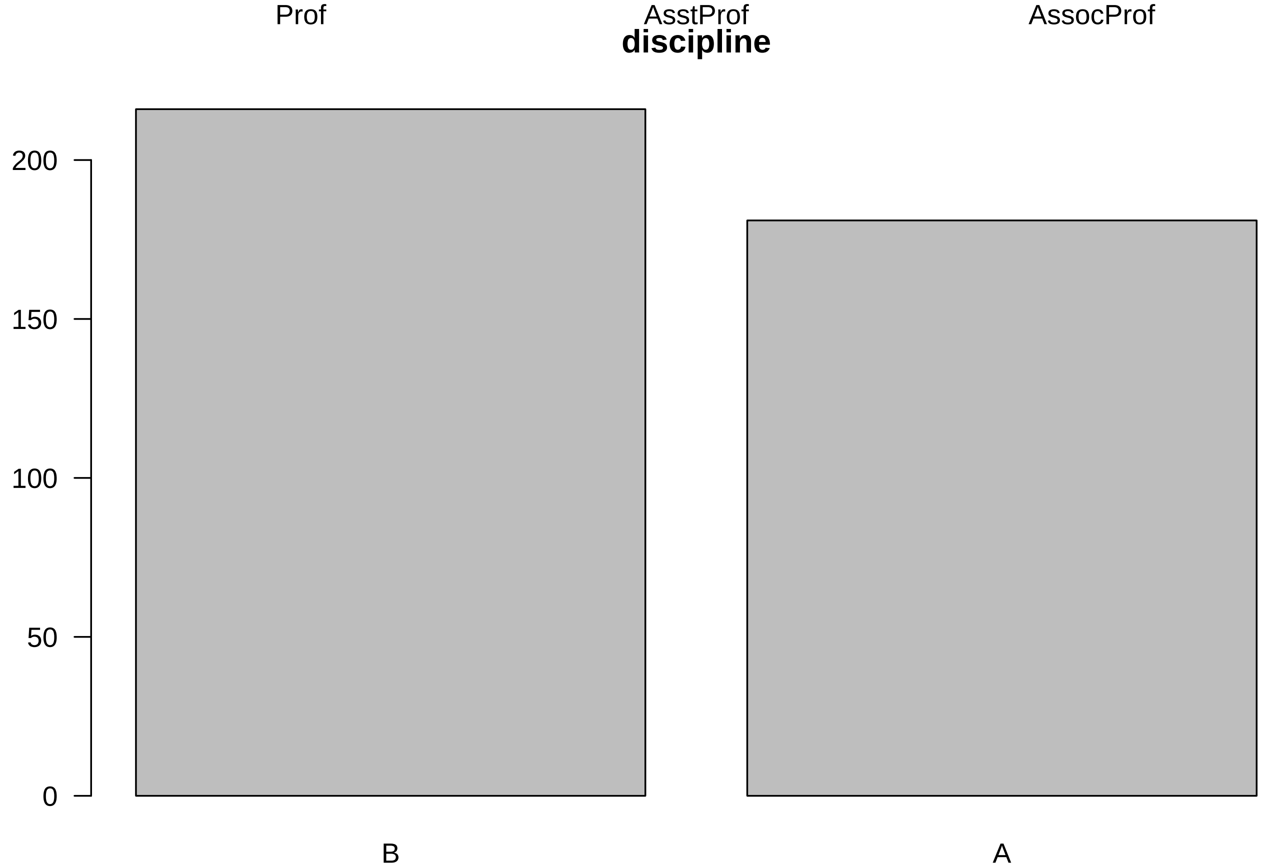
Code:

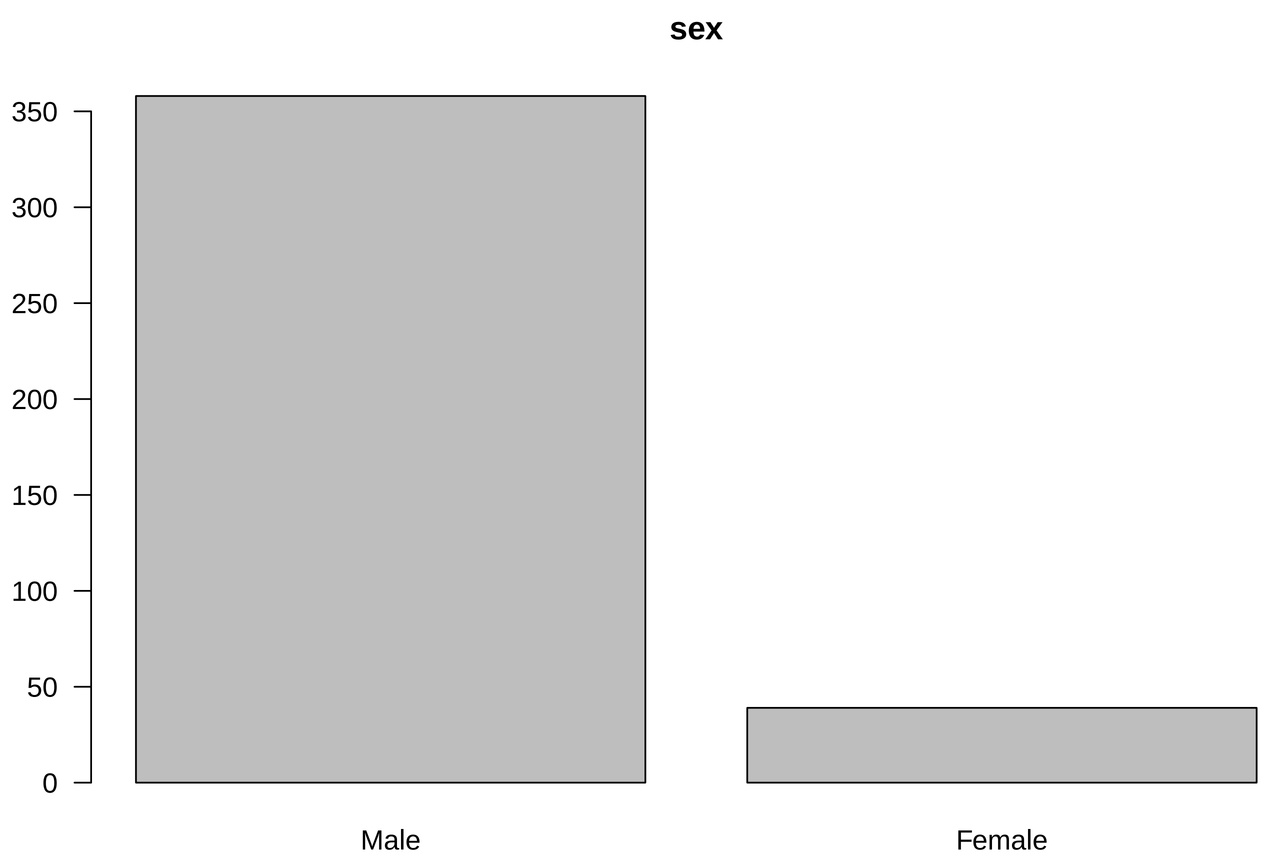
```r
summary(Salaries)
```

```
##        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
```

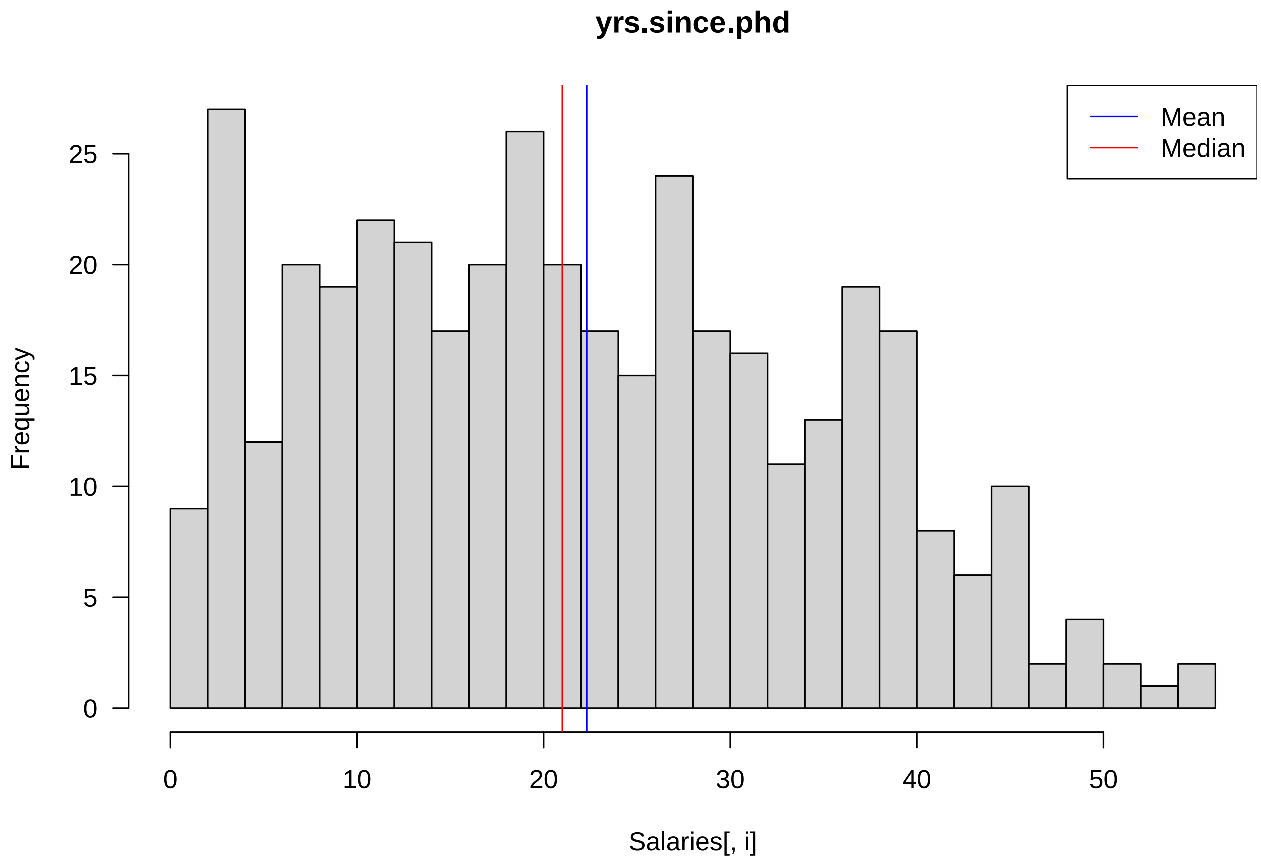
```r
catVars <- which(colnames(Salaries) %in% c("rank", "discipline", "sex"))
numVars <- which(colnames(Salaries) %in% c("yrs.since.phd", "yrs.service", "salary"))


for (i in catVars) barplot(sort(table(Salaries[,i]), decreasing = T), las = 1,
                           main = colnames(Salaries)[i])
```

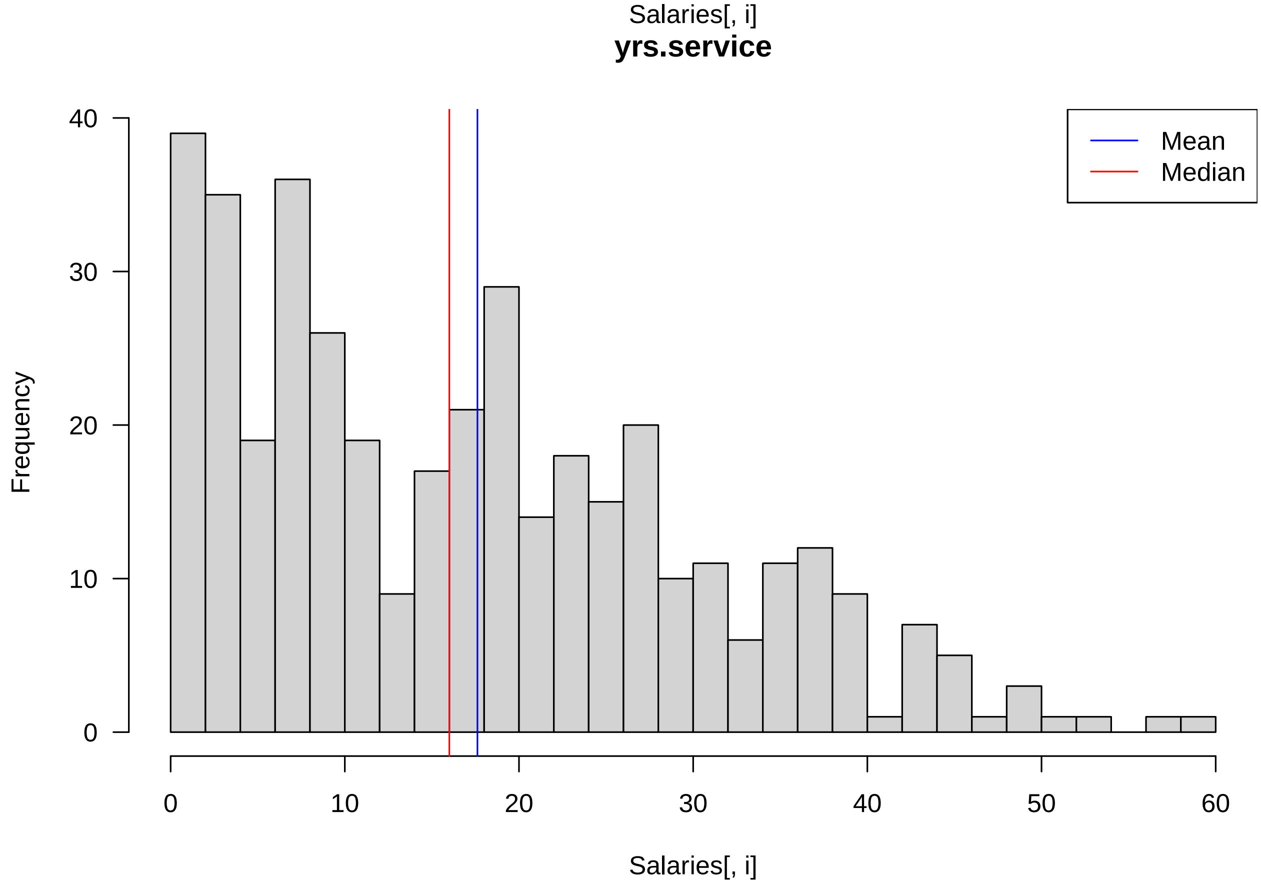

rank
250
200
150
100
50
0
Prof
AsstProf
AssocProf

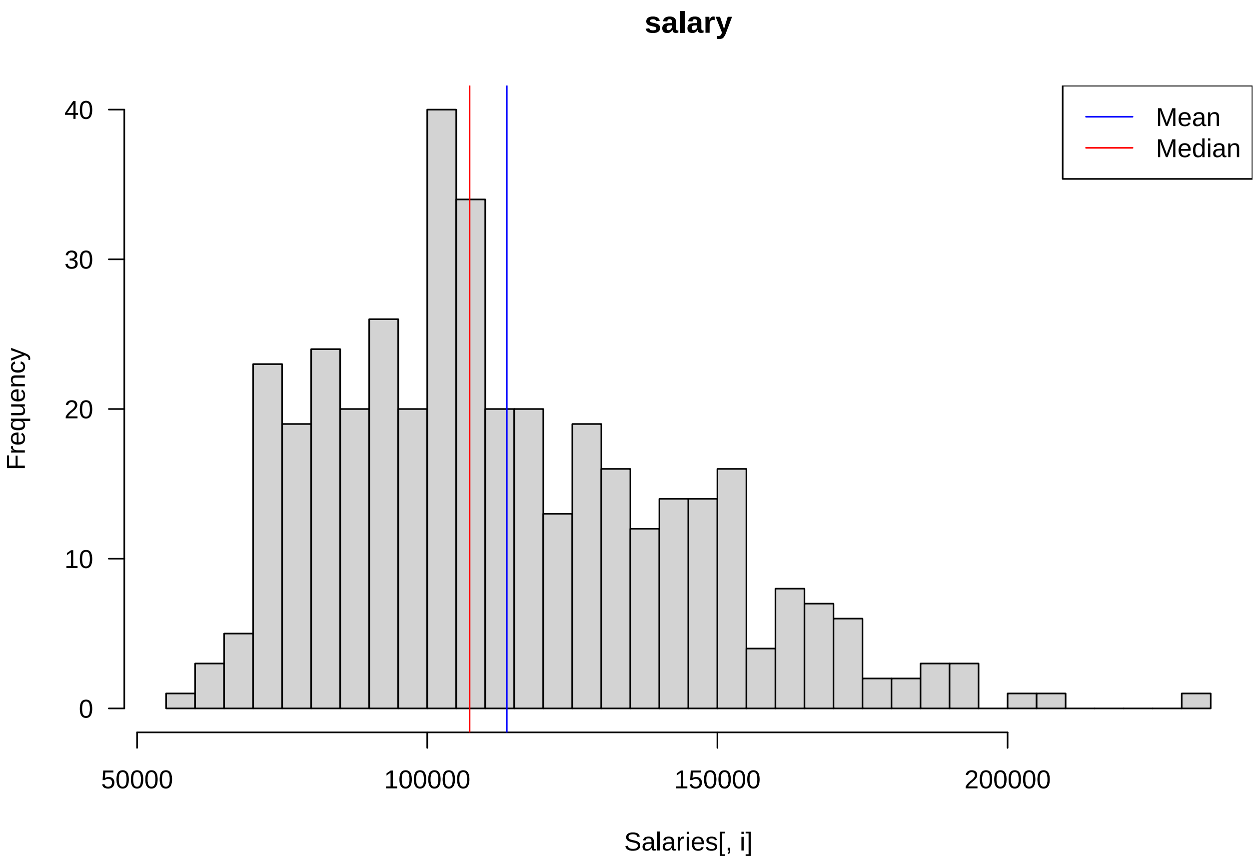

discipline
200
150
100
50
0
B
A




```r
for (i in numVars){
  hist(Salaries[,i], 30, main = colnames(Salaries)[i], las = 1)
  abline(v = mean(Salaries[,i]), col = "blue")
  abline(v = median(Salaries[,i]), col = "red")
  legend("topright", legend = c("Mean", "Median"), lty = 1, col = c("blue", "red"))
}
```


yrs.since.phd
Mean
Median
Frequency
0
5
10
15
20
25
0
10
20
30
40
50
Salaries[, i]


yrs.service
Mean
Median
Frequency
0
10
20
30
40
0
10
20
30
40
50
60
Salaries[, i]

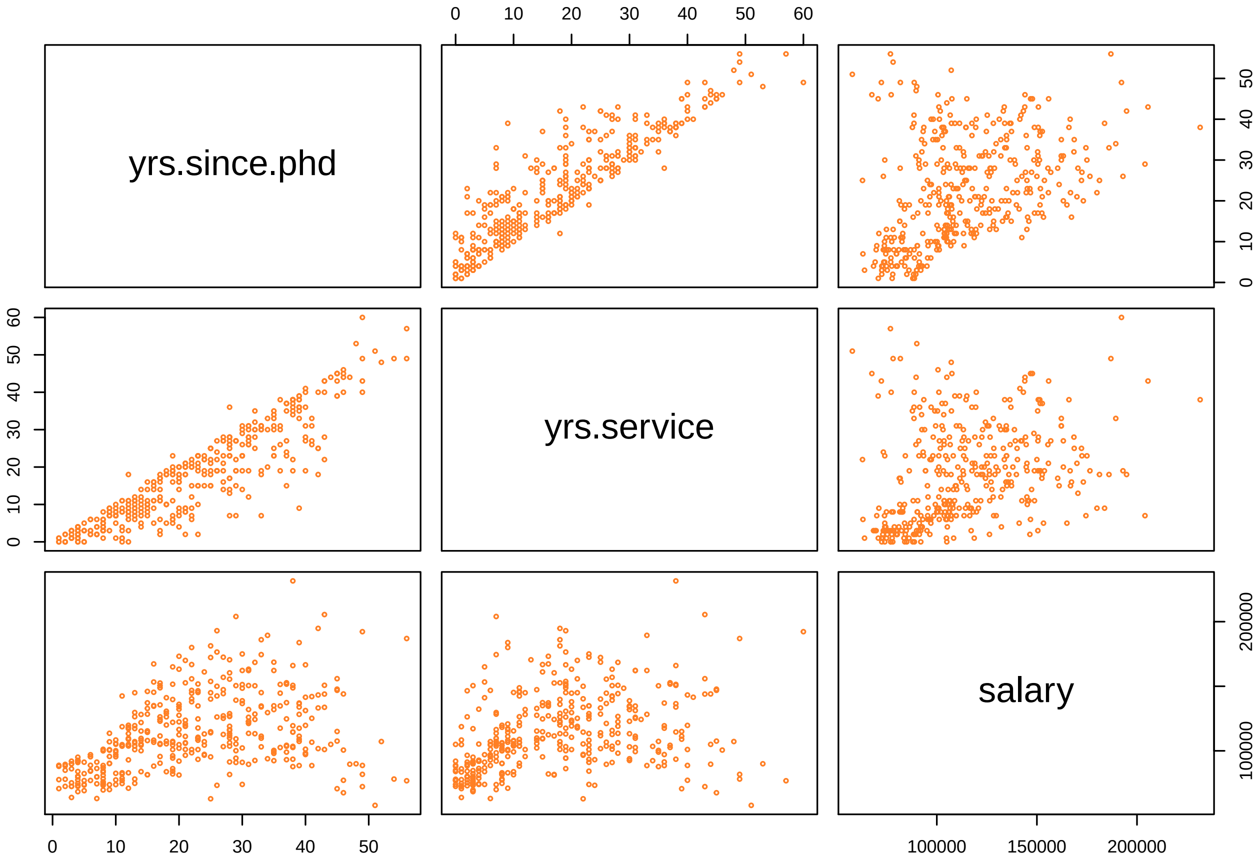


Answer: The years of service and years since PhD are both skewed right. Salary distribution is skewed right but a bit closer to symmertic. There are more professors (~2/3) than associate and assistant professors combined (~1/3). The discaplines are relatively close to equal.There are way more male than female professors.

3. Create a scatterplot matrix and briefly describe your findings

Code:

```
pairs(Salaries[, numVars], cex = 0.5, col = "chocolate1")
```




```r
cor(Salaries[, numVars])
```

```
##               yrs.since.phd yrs.service    salary
## yrs.since.phd     1.0000000   0.9096491 0.4192311
## yrs.service       0.9096491   1.0000000 0.3347447
## salary            0.4192311   0.3347447 1.0000000
```

Answer: There is, as expected, a storng postive linear relationship with `yrs.service` and `yrs.since.phd`. There is a moderate postive linear relationship between `salary` and `yrs.since.phd` and a even weaker postive linear relationship between `salary` and `yrs.service`.

Model Fitting

4. Fit a MLR with `yrs.since.phd`, `discipline`, `rank`, and `sex` as the predictors. Write down the fitted regression equations for each category (e.g., Female, Assistant Professor, theoretical departments). There are 12 categories in total

Code:

```r
model1 <- lm(salary ~ yrs.since.phd + discipline + rank + sex, data = Salaries)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = salary ~ yrs.since.phd + discipline + rank + sex,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -13860  -1549  10716  97023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
## yrs.since.phd    61.01     127.01   0.480  0.63124
## disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
## rankAssocProf 13104.15    4167.31   3.145  0.00179 **
## rankProf      46032.55    4240.12  10.856  < 2e-16 ***
## sexMale        4349.37    3875.39   1.122  0.26242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22660 on 391 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4401
## F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

Answer:

- Female, Assistant Professor, theoretical departments: $\widehat{\texttt{salary}} = 67884.32 + 61.01 \times \texttt{yrs.since.phd}$
- Female, Assistant Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32 + 13937.47) + 61.01 \times \texttt{yrs.since.phd}$
- Female, Associate Professor, theoretical departments: $\widehat{\texttt{salary}} = (67884.32 + 13104.15) + 61.01 \times \texttt{yrs.since.phd}$
- Female, Associate Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32+13937.47+13104.15)+61.01\times \texttt{yrs.since.phd}$
- Female, Professor, theoretical departments: $\widehat{\texttt{salary}} = (67884.32 + 46032.55) + 61.01 \times \texttt{yrs.since.phd}$
- Female, Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 46032.55) + 61.01 \times \texttt{yrs.since.phd}$
- Male, Assistant Professor, theoretical departments: $\widehat{\texttt{salary}} = (67884.32 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$
- Male, Assistant Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$
- Male, Associate Professor, theoretical departments: $\widehat{\texttt{salary}} = (67884.32+13104.15+4349.37)+61.01\times \texttt{yrs.since.phd}$
- Male, Associate Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32+13937.47+13104.15+4349.37)+ 61.01 \times \texttt{yrs.since.phd}$
- Male, Professor, theoretical departments: $\widehat{\texttt{salary}} = (67884.32 + 46032.55 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$
- Male, Professor, applied departments: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 46032.55 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

5. State the model assumptions in the previous regression model

Answer: There is a linear relationship between `salary` and `yrs.since.phd` and the regression is the same across all 12 categories. Also the random error term follows a normal distribution with constant variance and all these random errors are independent to each other.

6. Now fit another MLR with `yrs.since.phd`, `discipline`, `sex` and their interactions. Write down the fitted regression equations for each category

Code:

```r
model2 <- lm(salary ~ yrs.since.phd * discipline +  yrs.since.phd * sex, data = Salaries)
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd * discipline + yrs.since.phd *
##     sex, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84074 -17993  -3246  15708  91709
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               68155.8     8752.1   7.787 6.21e-14 ***
## yrs.since.phd              1574.6      442.8   3.556 0.000423 ***
## disciplineB                6386.7     5493.0   1.163 0.245665
## sexMale                   19608.8     8840.5   2.218 0.027125 *
## yrs.since.phd:disciplineB   403.9      210.9   1.915 0.056195 .
## yrs.since.phd:sexMale      -728.9      450.2  -1.619 0.106257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26300 on 391 degrees of freedom
## Multiple R-squared:  0.2558, Adjusted R-squared:  0.2463
## F-statistic: 26.88 on 5 and 391 DF,  p-value: < 2.2e-16
```

Answer:

- Female, theoretical departments: $\widehat{\texttt{salary}} = 68155.8 + 1574.6 \times \texttt{yrs.since.phd}$
- Female, applied departments: $\widehat{\texttt{salary}} = (68155.8 + 6386.7) + (1574.6 + 403.9) \times \texttt{yrs.since.phd}$
- Male, theoretical departments: $\widehat{\texttt{salary}} = (68155.8 + 19608.8) + (1574.6 - 728.9) \times \texttt{yrs.since.phd}$
- Male, applied departments: $\widehat{\texttt{salary}} = (68155.8 + 19608.8 + 6386.7) + (1574.6 + 403.9 - 728.9) \times \texttt{yrs.since.phd}$

Non-linear Regression: An Simulated Example

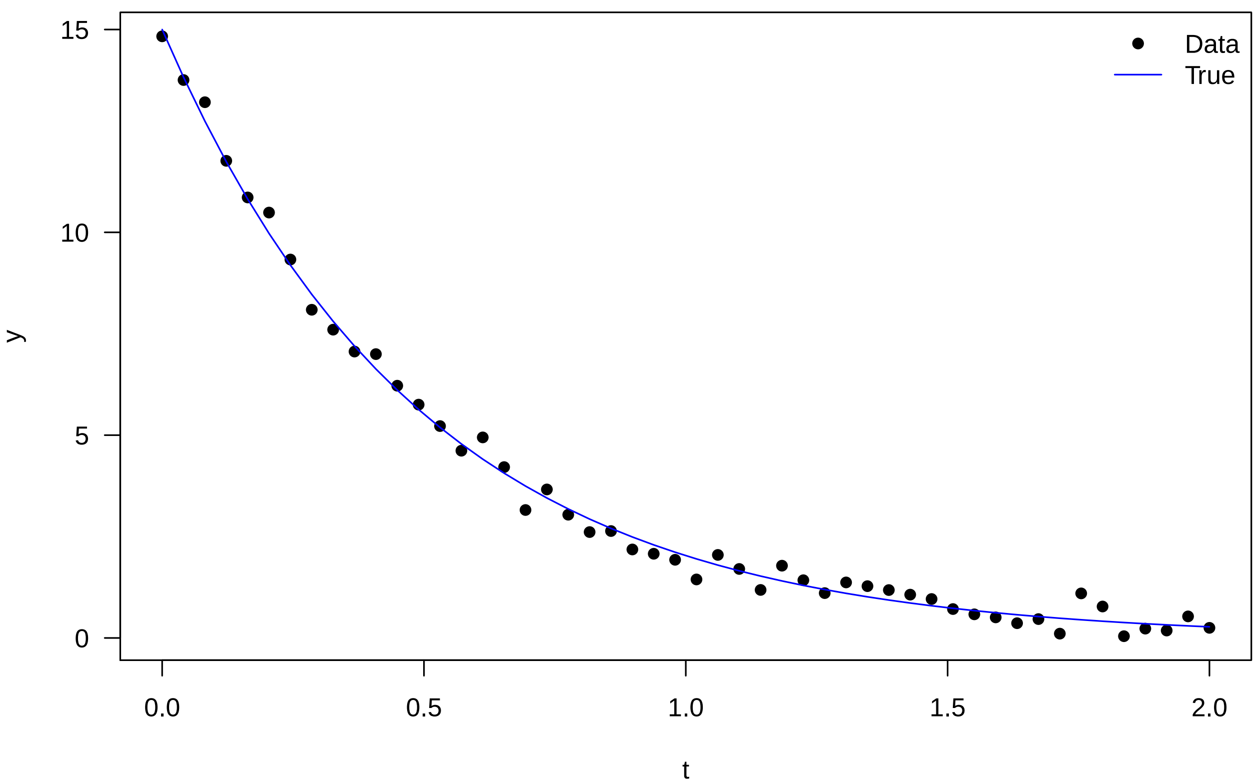
Suppose the response y depends on the predictor t in the following form:

$$y = \alpha \exp(-\beta t) + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, and the true α, β, and σ^2 are 15, 2 and 0.09, respectively. First, let's simulate some data points from this nonlinear model:

Code:

```
alpha = 15; beta = 2; sigma.sq = 0.09
n <- 50
t <- seq(0, 2, len = 50)
set.seed(123)
y <- alpha * exp(-beta * t) + rnorm(n, sd = sqrt(sigma.sq))
data <- data.frame(y = y, t = t)

plot(t, y, las = 1, pch = 16)
lines(t, alpha * exp(-beta * t), type = "l", col = "blue")
legend("topright", legend = c("Data", "True"), pch = c(16, NA), lty = c(NA, 1),
       col = c("black", "blue"), bty = "n")
```



7. Use `nls` function to obtain nonlinear least-squares estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$. In order to use `nls` you would need to provide formula = y ~ alpha * exp(-beta * t), start = list(alpha = alpha_0, beta = beta_0), where alpha_0 and beta_0 are initial guesses of the parameters α and β

Code:

```r
NLFit <- nls(y ~ alpha * exp(-beta * t), start = list(alpha = 8, beta = 3))
summary(NLFit)
```

```
##
## Formula: y ~ alpha * exp(-beta * t)
##
## Parameters:
##       Estimate Std. Error t value Pr(>|t|)
## alpha  15.0962     0.1484  101.72   <2e-16 ***
## beta    2.0113     0.0293   68.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2796 on 48 degrees of freedom
##
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 1.822e-06
```

Answer: $\hat{\alpha} = 15.0962261$, $\hat{\beta} = 2.011318$, and $\hat{\sigma} = 0.2795961$

8. Write down the fitted equation

Answer: $y = 15.0962 \times \exp(-2.0113 \times t)$

9. Apply the natural log transformation to the simulated response and fit a simple linear regression. Back transform to get the fit on the original scale.

Note that $\mathbb{E}(y) = \alpha \exp(-\beta t) \Rightarrow \log(\mathbb{E}(y)) = \log(\alpha) - \beta t$

Code:

```r
logTrlmFit <- lm(log(y) ~ t)
summary(logTrlmFit)
```

```
##
## Call:
## lm(formula = log(y) ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98777 -0.06917 -0.00311  0.10335  1.05859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7798     0.1233   22.55   <2e-16 ***
## t            -2.1332     0.1062  -20.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4423 on 48 degrees of freedom
## Multiple R-squared:  0.8937, Adjusted R-squared:  0.8915
## F-statistic: 403.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

Answer:

We have $\hat{\beta} = 2.1332$ ($SE(\hat{\beta}) = 0.1062$) and $\hat{\alpha} = \exp(2.7798) = 16.1158$

10. Comaring the nonlinear regression method and the linear regression with log-transformed response, which method would you prefer in this example. Explain your answer

Answer:

Will discuss this in today's office hour.