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

Meredith Sliger

Contents

Analysis of covariance: Salaries for Professors 1
- Load the dataset 1
- Exploratory Data Analysis 2
- Model Fitting 9

Non-linear Regression: An Simulated Example 11

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. was collected as part of the ongoing effort of the college's administration to monitor salary differences between male and female faculty members.

Load the dataset

Code:

```
library(carData)
```

```
## Warning: package 'carData' was built under R version 4.4.2
```

```
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 variables in this data set

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

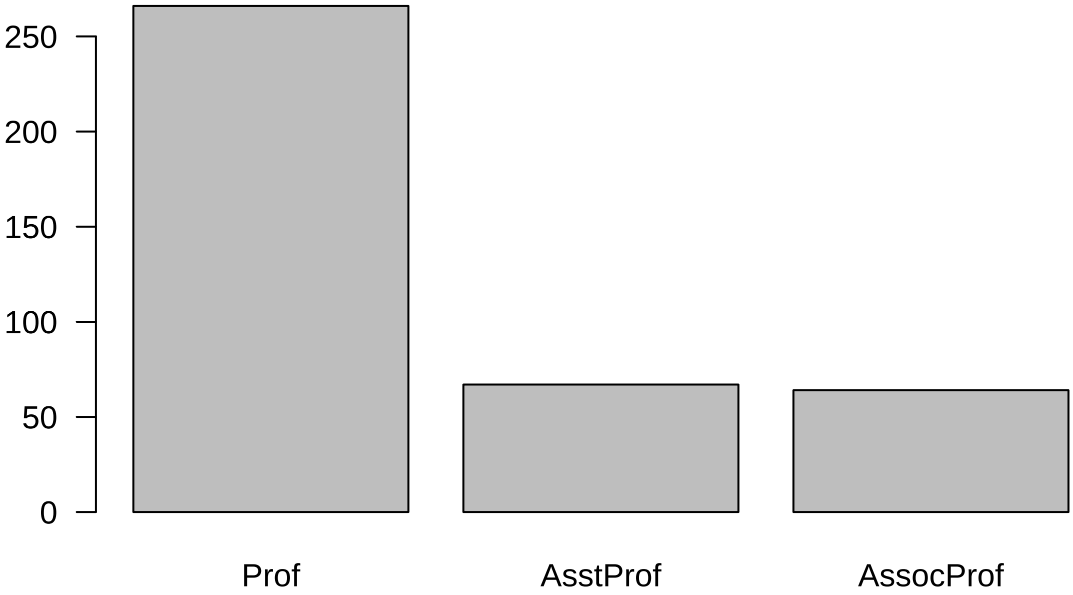
2. Summarize each variable numerically and graphically, briefly describe your findings

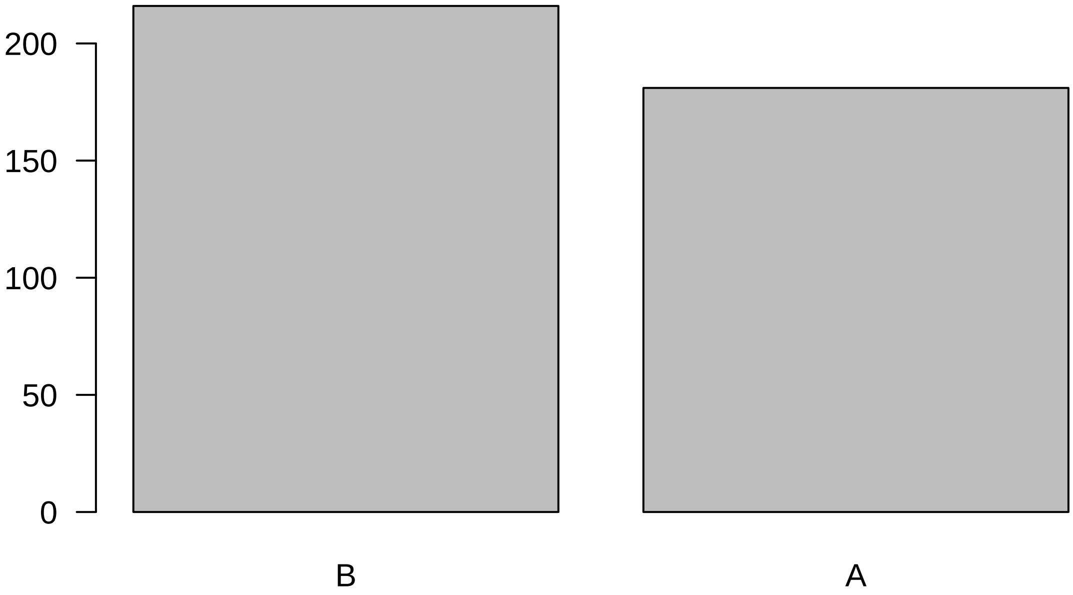
Code:

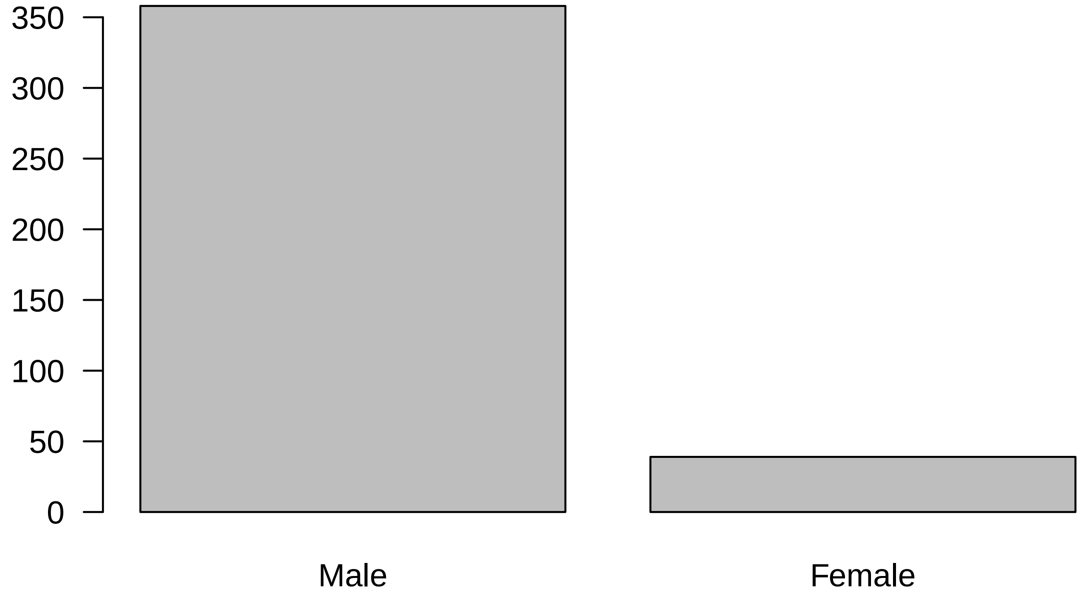
```
summary(Salaries)
```

```
##         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
```

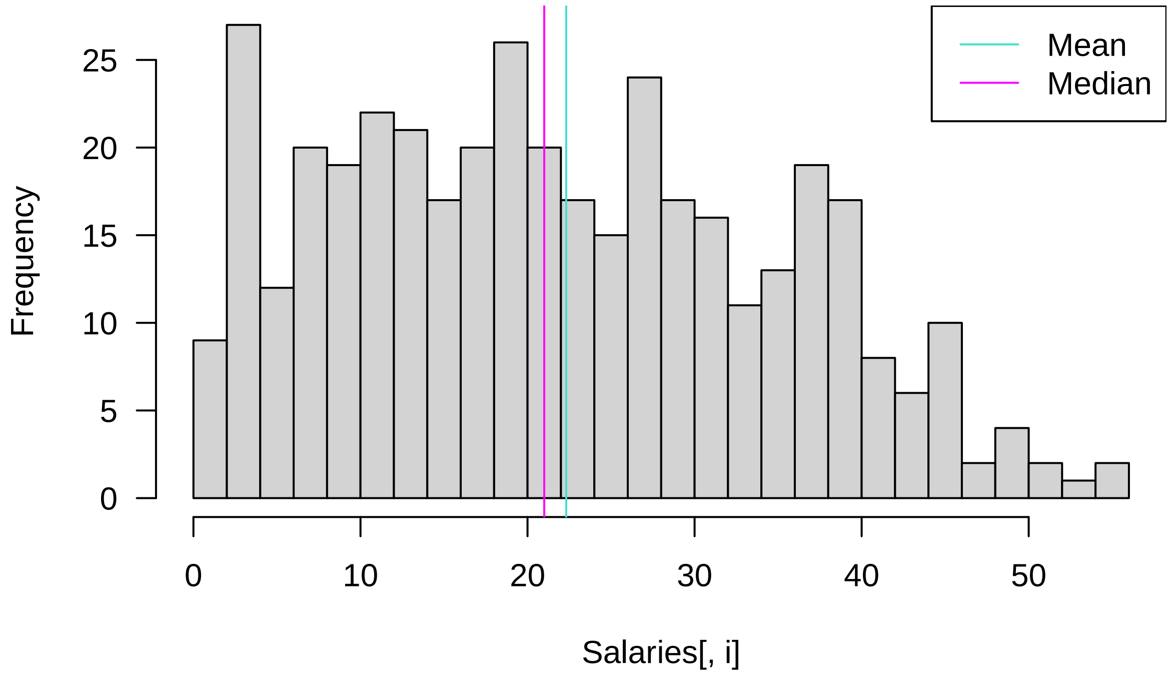
```
catVars <- c("rank", "discipline", "sex")
numVars <- c("yrs.since.phd", "yrs.service", "salary")

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

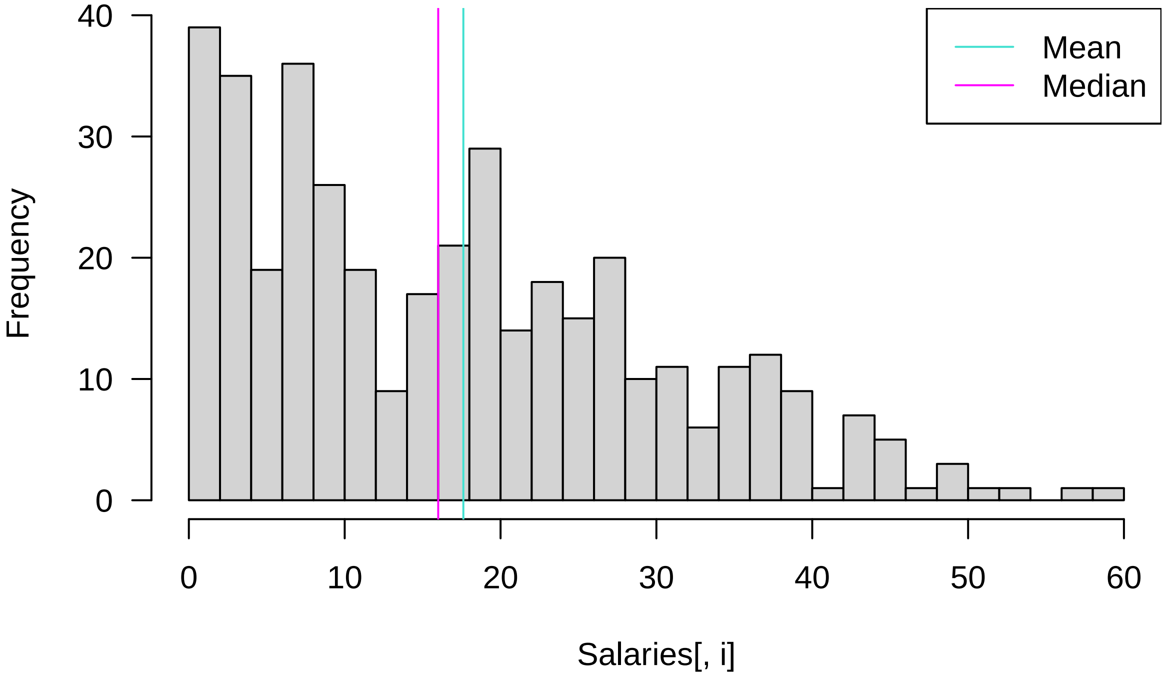

250
200
150
100
50
0
Prof
AsstProf
AssocProf

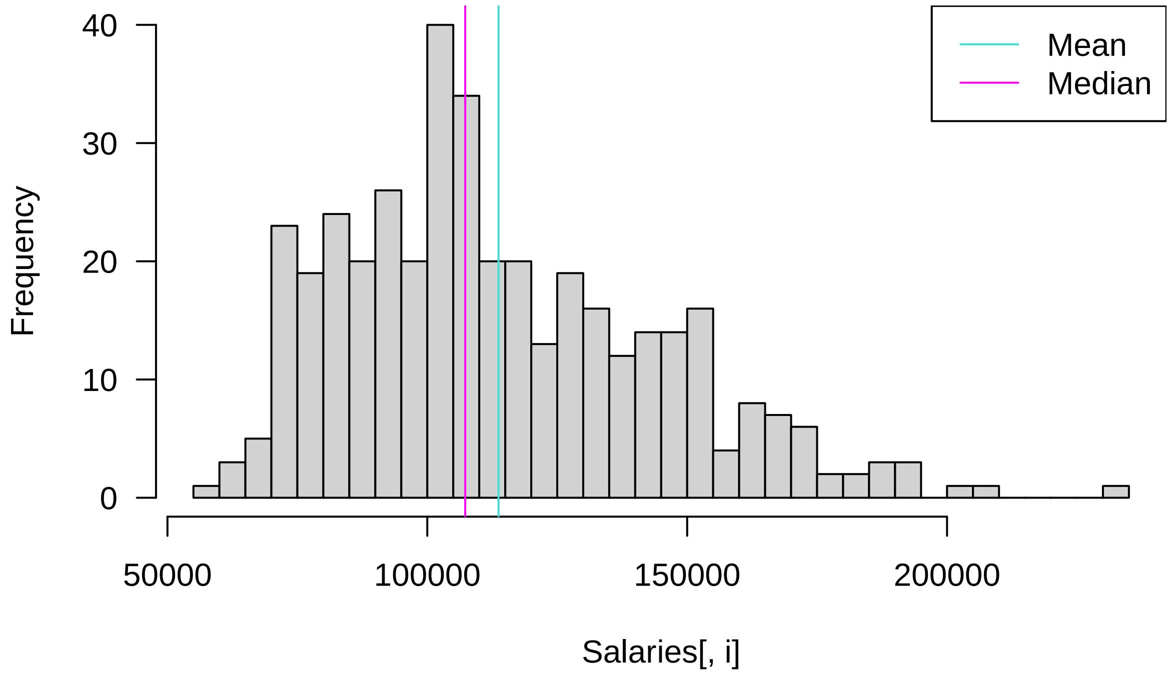

200
150
100
50
0
B
A



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


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


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

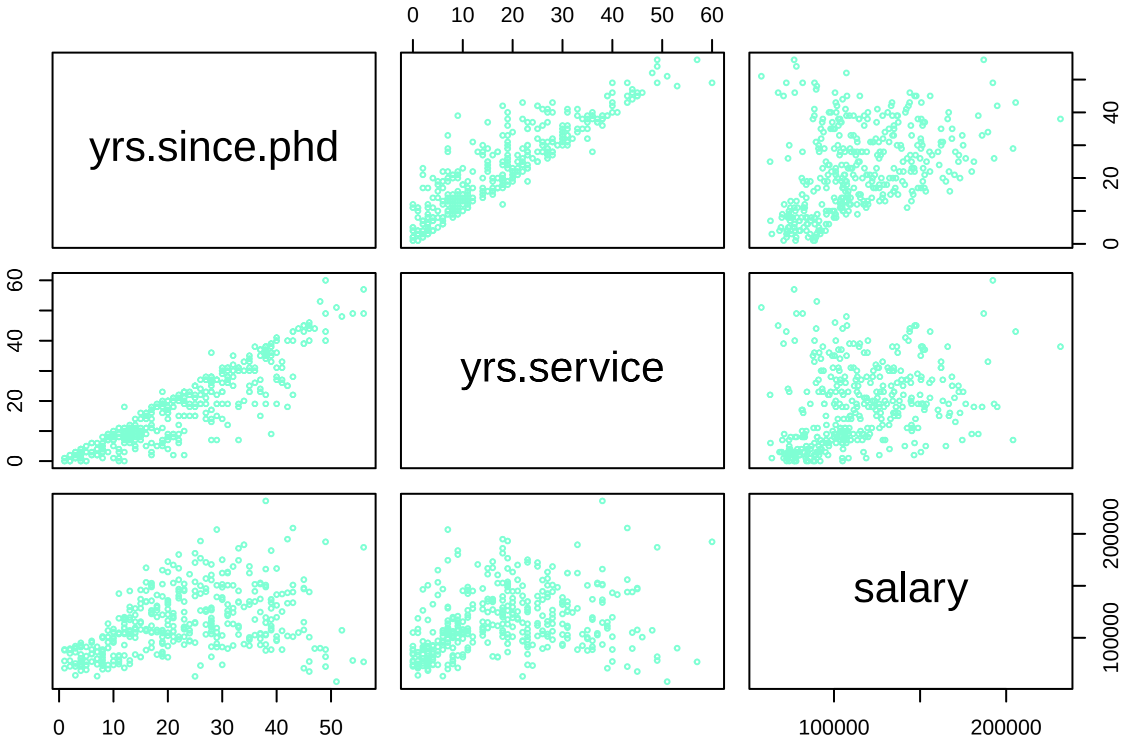


Answer: The distributions of both years of service and years since earning a PhD are right-skewed. Similarly, salary follows a right-skewed pattern but appears closer to symmetric. The majority of faculty members are professors, making up approximately two-thirds of the dataset, while associate and assistant professors together account for the remaining one-third. The distribution across disciplines is fairly balanced. However, there is a notable disparity in gender representation, with significantly more male professors than female professors.

3. Create a scatterplot matrix and briefly describe your findings

Code:

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



```
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: As anticipated, there is a strong positive relationship between years of service and years since earning a PhD, which makes sense given that faculty members typically begin their careers soon after completing their doctoral degrees. Additionally, salary tends to increase moderately with years since PhD, suggesting that experience and time in academia contribute to higher earnings. However, the relationship between salary and years of service is weaker, indicating that factors beyond just service time, such as rank, discipline, and external funding opportunities, may play a more significant role in salary determination.

Model Fitting

4. Fit a multiple linear regression model (MLR) with `yrs.since.phd`, `discipline`, `rank`, and `sex` as 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: $\widehat{\texttt{salary}} = 67884.32 + 61.01 \times \texttt{yrs.since.phd}$

Female, Assistant Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 13937.47) + 61.01 \times \texttt{yrs.since.phd}$

Female, Associate Professor, Theoretical: $\widehat{\texttt{salary}} = (67884.32 + 13104.15) + 61.01 \times \texttt{yrs.since.phd}$

Female, Associate Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 13104.15) + 61.01 \times \texttt{yrs.since.phd}$

Female, Professor, Theoretical: $\widehat{\texttt{salary}} = (67884.32 + 46032.55) + 61.01 \times \texttt{yrs.since.phd}$

Female, Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 46032.55) + 61.01 \times \texttt{yrs.since.phd}$

Male, Assistant Professor, Theoretical: $\widehat{\texttt{salary}} = (67884.32 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

Male, Assistant Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

Male, Associate Professor, Theoretical: $\widehat{\texttt{salary}} = (67884.32 + 13104.15 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

Male, Associate Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 13937.47 + 13104.15 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

Male, Professor, Theoretical:$\widehat{\texttt{salary}} = (67884.32 + 46032.55 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

Male, Professor, Applied: $\widehat{\texttt{salary}} = (67884.32 + 46032.55 + 4349.37) + 61.01 \times \texttt{yrs.since.phd}$

5. State the model assumptions in the previous regression model

Answer: The multiple linear regression model relies on several key assumptions. It assumes a linear relationship between salary and years since earning a PhD, meaning the effect of experience on salary remains consistent across all 12 categories. Additionally, the model assumes that the random error terms are normally distributed, ensuring that residuals are symmetrically distributed around zero. Another assumption is constant variance (homoscedasticity), meaning that the spread of residuals does not change across different levels of the predictors. Lastly, the model assumes independence of errors, indicating that salary observations are not correlated beyond what is explained by the model's predictors.

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: $\texttt{sal}\hat{\texttt{a}}\texttt{ry} = 68155.8 + 1574.6 \times \texttt{yrs.since.phd}$

Female, Applied: $\texttt{sal}\hat{\texttt{a}}\texttt{ry} = (68155.8 + 6386.7) + (1574.6 + 403.9) \times \texttt{yrs.since.phd}$

Male, Theoretical: $\texttt{sal}\hat{\texttt{a}}\texttt{ry} = (68155.8 + 19608.8) + (1574.6 - 728.9) \times \texttt{yrs.since.phd}$

Male, Applied: $\texttt{sal}\hat{\texttt{a}}\texttt{ry} = (68155.8 + 1906.8 + 6386.7) + (1574.6 + 403.9 - 728.9) \times \texttt{yrs.since.phd}$

Non-linear Regression: An Simulated Example

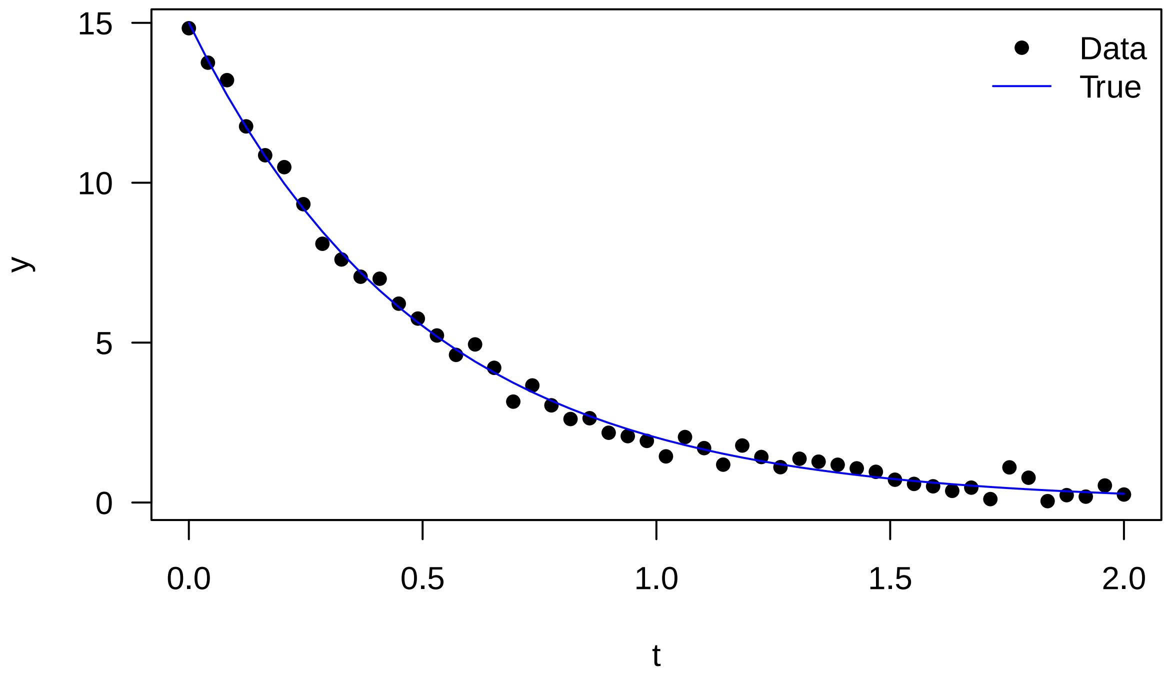
Suppose the response y depends on the predictor t in the following form:

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

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, and the true α, β, and σ^2 are 15, 2 and 0.16, 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 the **nls** function to obtain nonlinear least-squares estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$. To use **nls**, 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:

```
NLFit <- nls(y ~ alpha * exp(-beta * t), start = list(alpha = 5, beta = 1))
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: 5
## Achieved convergence tolerance: 3.747e-07
```

Answer: $\hat{\alpha} = 15.0962245$, $\hat{\beta} = 2.0113177$, and $\hat{\sigma} = 0.2795961$

8. Write down the fitted equation and the estimated variance $\hat{\sigma}^2$

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

9. Apply the natural log transformation to the simulated response, then fit a simple linear regression. Back-transform to obtain the fit on the original scale

Code:

```
logTrimFit <- lm(log(y) ~ t)
summary(logTrimFit)
```

```
##
## 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: $\hat{\beta} = 2.1332$ ($SE(\hat{\beta}) = 0.1062$) and $\hat{\alpha} = \exp(2.7798) = 16.1158$

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

Answer: The nonlinear regression method is preferred because it directly estimates the parameters $\hat{\alpha}$ and $\hat{\beta}$ in their original exponential form, resulting in more accurate estimates closer to the true values. Unlike the log-transformed linear regression, which assumes normally distributed errors after transformation and may introduce bias when back-transforming, the nonlinear model maintains the correct error structure. Additionally, it provides a more intuitive interpretation, as it expresses the relationship naturally as an exponential decay function without requiring exponentiation. While the linear model offers a reasonable fit, the linear approach ensures more precise parameter estimation and avoids potential distortions from transformations.