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

your name here

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. 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)
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:

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

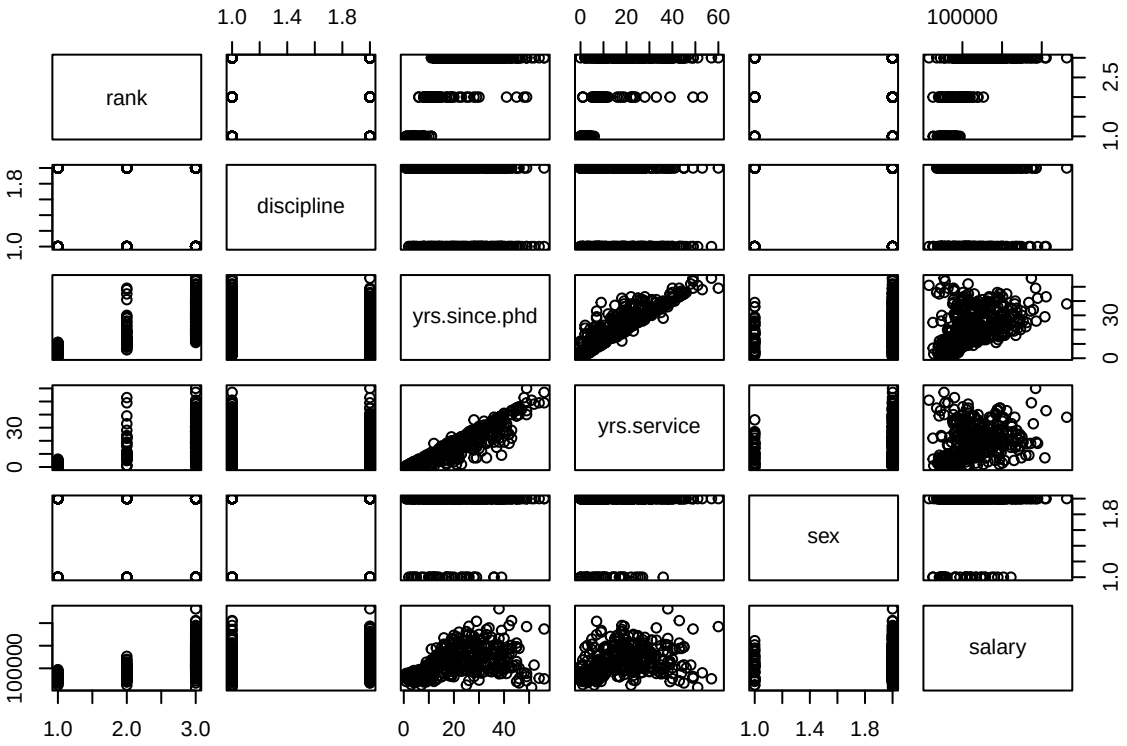
Code:

Answer:

3. Create a scatterplot matrix and briefly describe your findings

Code:

```
pairs(Salaries)
```




```
library(GGally)
```

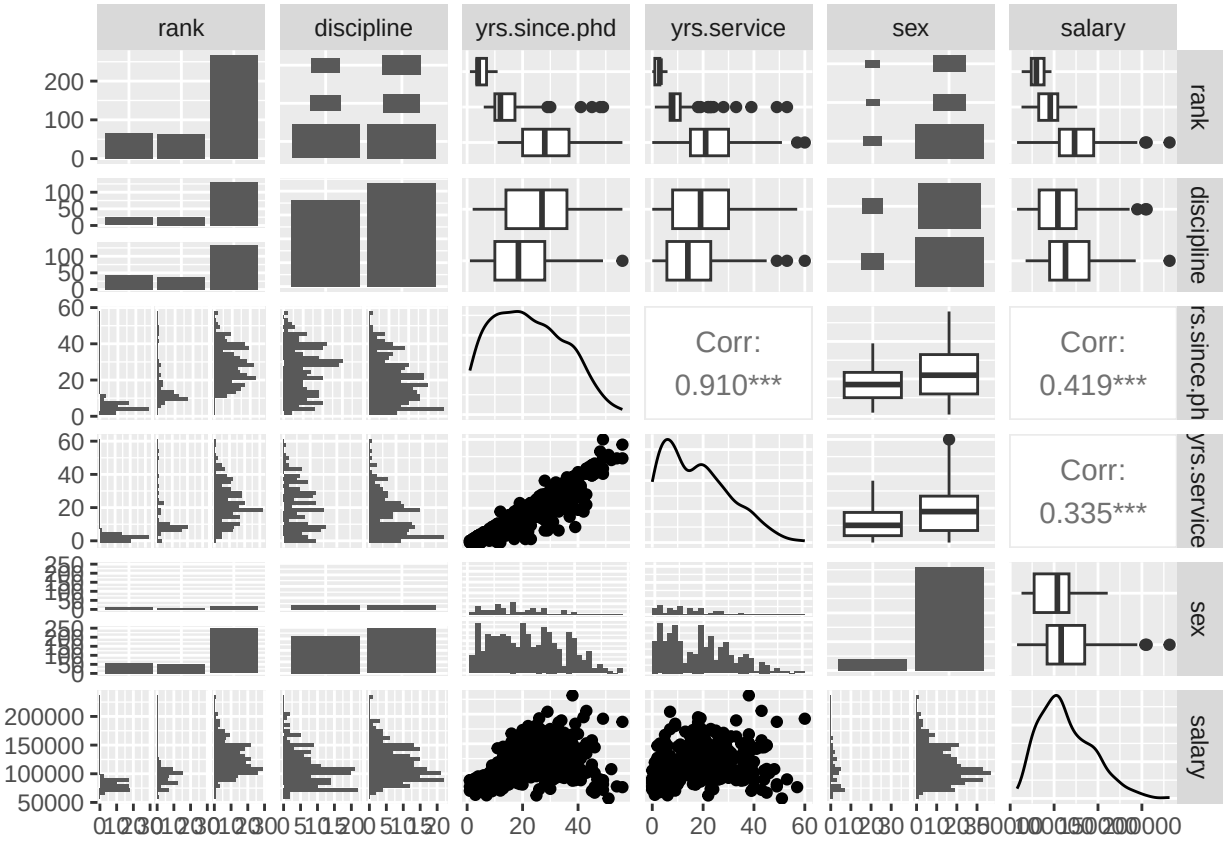
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(Salaries)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Answer:

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:

Answer:

5. State the model assumptions in the previous regression model

Answer:

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

Code:

Answer:

Non-linear Regression: An Simulated Example

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

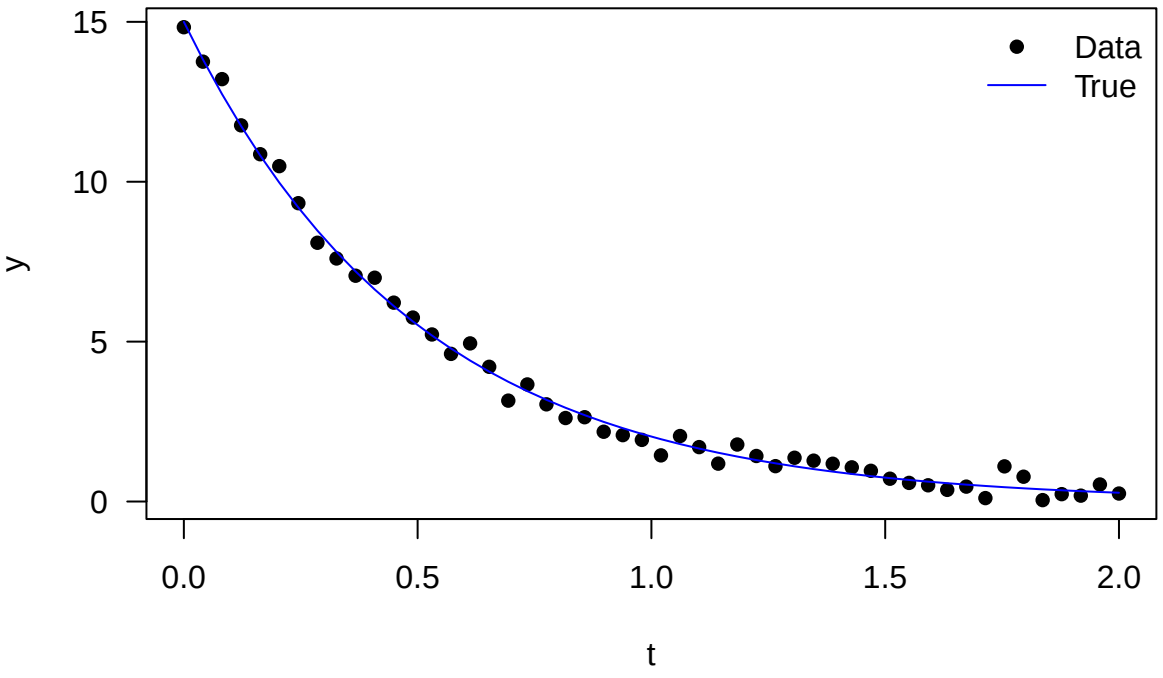
$$y = \alpha \exp(-\beta t) + \epsilon,$$

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 $\sim$ 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:

Answer:

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

Answer:

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:

Answer:

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: