Day5 exercise solutions

Ali Movasati, Isabelle Cretton, Tristan Koning

Oct. 14th, 2024

```
# Set global code chunk options
knitr::opts_chunk$set(warning = FALSE)
```

```
# load required libraries
library(skimr)
library(ggplot2)
library(ggpubr)
library(magrittr)
library(dplyr)
library(tibble)


# define functions
`%notin%` <- Negate(`%in%`)
```

Problem 1

```
# read in the data

salary <- read.table(file = "/Users/alimos313/Documents/studies/phd/university/courses/stat-modelling/S
```

1.A)

```
salary %<>% mutate(size = factor(districtSize))
```

1.B)

- Numerical summary

```
# summary of dataset
skim(salary)
```

Table 1: Data summary

Name	salary
Number of rows	325
Number of columns	5
Column type frequency:	
character	1
factor	1

numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
District	0	1	3	24	0	325	0

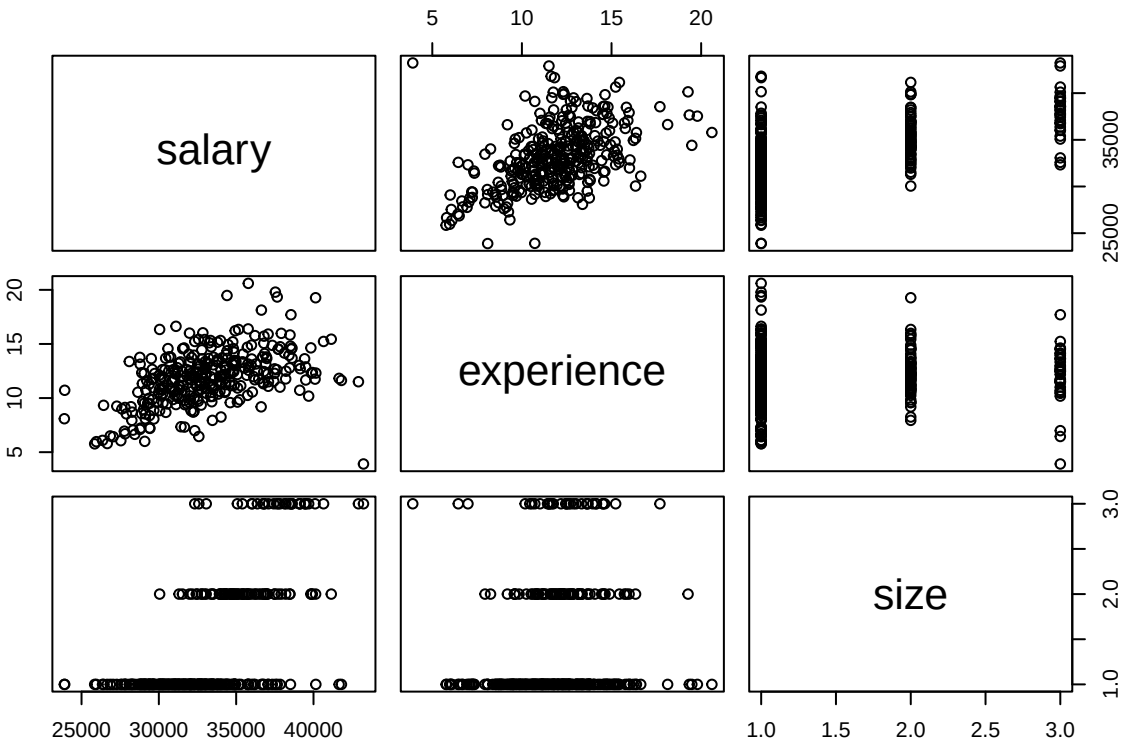
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
size	0	1	FALSE	3	1: 223, 2: 68, 3: 34

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
districtSize	0	1	1.42	0.67	1.00	1.00	1.00	2.00	3.0	
salary	0	1	33168.33	3412.77	23889.50	30847.70	32867.50	35296.70	43232.6	
experience	0	1	11.86	2.55	3.91	10.44	11.97	13.33	20.6	

```
# Graphical summeries
pairs(salary[c("salary", "experience", "size")])
```

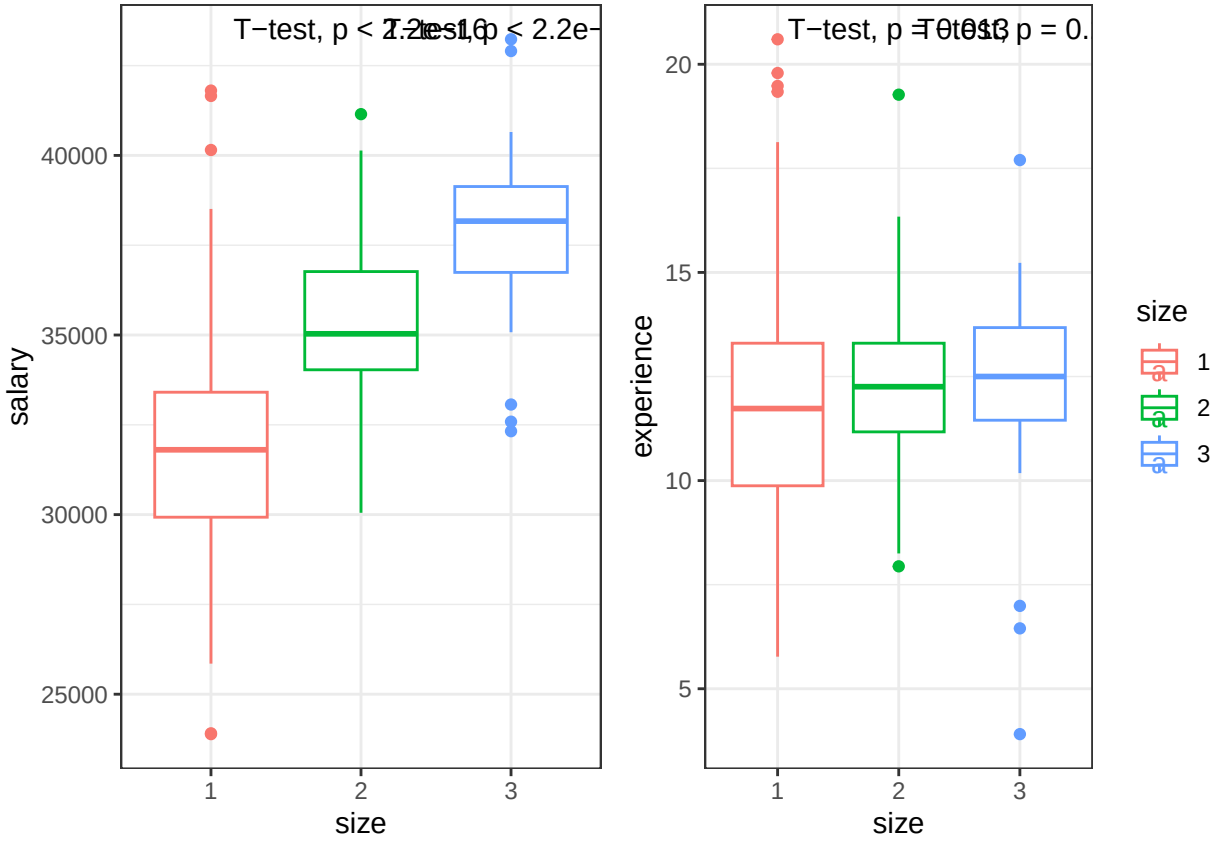


```r
## boxplots


bxp1 <- salary %>%
    ggplot(aes(x = size, y = salary, color = size)) +
    geom_boxplot() +
    stat_compare_means(ref.group = "1", method = "t.test") +
    theme_bw() +
    theme(legend.position = "none")

bxp2 <- salary %>%
    ggplot(aes(x = size, y = experience, color = size)) +
    geom_boxplot() +
    stat_compare_means(ref.group = "1", method = "t.test") +
    theme_bw()

# Arrange the plots side by side
grid.arrange(bxp1, bxp2, ncol = 2)
```




```r
hist(salary$salary, main = "Histogram of Salaries", xlab = "Salary")
```


```
# fit models

model_a <- lm(salary ~ 1 + experience + districtSize, data = salary)
model_b <- lm(salary ~ 1 + experience + size, data = salary)

## model A

summary(model_a)
```

```
##
## Call:
## lm(formula = salary ~ 1 + experience + districtSize, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7446.0 -1307.9  -180.7  1142.7 10099.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22065.55     622.14   35.47   <2e-16 ***
## experience     586.42      48.79   12.02   <2e-16 ***
## districtSize  2924.90     184.47   15.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2225 on 322 degrees of freedom
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5748
## F-statistic:   220 on 2 and 322 DF,  p-value: < 2.2e-16
```

```
## model A

summary(model_b)
```

```
##
## Call:
## lm(formula = salary ~ 1 + experience + size, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7577.1 -1283.9  -108.9  1141.3 10220.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24995.49     589.35  42.412   <2e-16 ***
## experience    584.15      48.96  11.932   <2e-16 ***
## size2        3088.00     310.73   9.938   <2e-16 ***
## size3        5732.28     410.84  13.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 321 degrees of freedom
## Multiple R-squared:  0.578,  Adjusted R-squared:  0.5741
## F-statistic: 146.6 on 3 and 321 DF,  p-value: < 2.2e-16
```

« comments »

The two models are similar in terms of their goodness-of-fit values. Model A has one degree of freedom lower as it treats districtSize as one independent variable, while model B treats factorized size variable as 2.

```
AIC(model_a, model_b)
```

```
##         df      AIC
## model_a  4 5937.259
## model_b  5 5938.828
```

```
BIC(model_a, model_b)
```

```
##         df      BIC
## model_a  4 5952.394
## model_b  5 5957.747
```

Looking at the AIC and BIC, model A slightly outperforms model B, which is surprising given that variables districtSize and size in theory represent exactly the same information. Considering the summary of both models, we can see that the linear model takes each factor of variable size in consideration with their own estimate (Intercept, size2, size3), while model A with the variable districtSize only uses one estimate which is multiplied with the districtSize. Therefore, we may conclude that model B slightly overfits the values because it has more variables to estimate on, and thats why we see a slight difference in performance between the models.

1.D)

```
# save model summary in an object
summary_b <- summary(model_b)
```

« comments »

The adjusted coefficient of determination for model B is 0.5740808. Therefore, 57.4080767% of variability in the response variable salary can be explained by the proposed linear regression model.

Given p-values < 0.05, it indicates that all predictors has a significant effect on the dependent variable. Therefore, it is not advisable to drop any of the predictor variables.

The parameter estimate for variable "experience" is 584.1517062. That means for 1 year additional experience, given the district is the same, the salary will be increase by 584.1517062.

The parameter estimate for variable "district 2" is 3087.9955923. That means for the same level of experience, teachers in district size 2, earn 3087.9955923 USD more than teachers in the reference district of size 1.

The parameter estimate for variable "district 3" is 5732.2815637. That means for the same level of experience, teachers in district size 3, earn 5732.2815637 USD more than teachers in the reference district of size 1.

1.E)

```
model_b_transformed <- lm(salary ~ I(experience - 13) + size, data = salary)

summary(model_b_transformed)
```

```
##
## Call:
## lm(formula = salary ~ I(experience - 13) + size, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7577.1 -1283.9  -108.9  1141.3 10220.8
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32589.46     163.20 199.691   <2e-16 ***
## I(experience - 13)   584.15      48.96  11.932   <2e-16 ***
## size2               3088.00     310.73   9.938   <2e-16 ***
## size3               5732.28     410.84  13.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 321 degrees of freedom
## Multiple R-squared:  0.578,  Adjusted R-squared:  0.5741
## F-statistic: 146.6 on 3 and 321 DF,  p-value: < 2.2e-16
```

« comments »

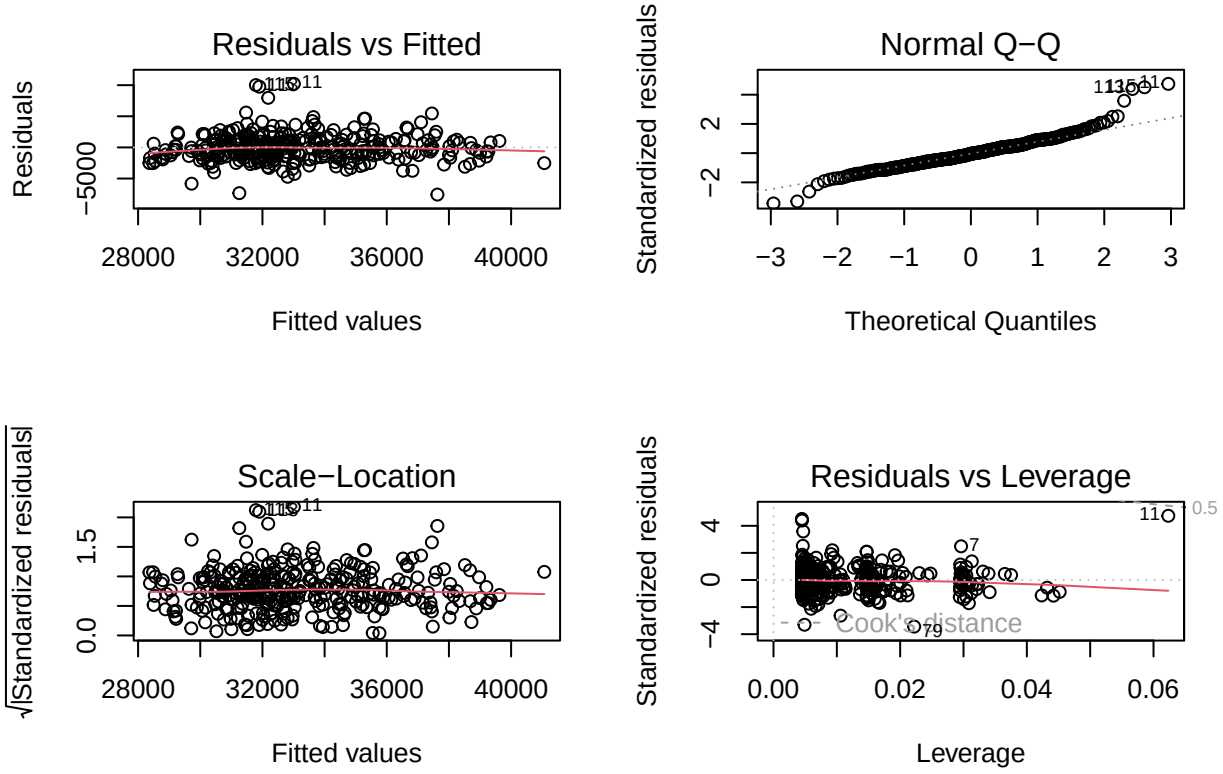
When we modify the model by using "experience - 13" instead of experience, the interpretation of the coefficient for intercept and experience will change, but the overall fit and statistical properties of the model (likeR^2, adjusted R^2, p-values) will remain unchanged.

The coefficient for I(experience - 13) will now represent the change in salary for each year of experience compared to 13 years, and the intercept will reflect the average salary when the experience is 13 years, rather than 0 years.

Intercept: In the original model, the intercept represents the expected salary when experience = 0. This can be less meaningful if no teachers in the dataset have exactly 0 years of experience. By shifting the experience to center it around 13 years, the intercept will now represent the expected salary for a teacher with 13 years of experience.

1.F)

```
# Plot diagnostic plots
par(mfrow = c(2, 2))  # Arrange 4 diagnostic plots
plot(model_b)
```



« comments »

- based on the residuals vs fitted plot we can confirm linear relationship between the response and predictor variables
- based on Q-Q Plot for Residuals we can confirm normality of error term of our linear regression model
- based on scale-location plot we can confirm homoscedasticity of the residuals across all levels of the independent variables

1.G)

```
example_data_a <- data.frame(experience = c(10), districtSize = c(3))
example_data_b <- data.frame(experience = c(10), size = c("3"))
```

« comments »

According to model A the predicted salary will be 3.6704423×10^4 and according to model B 3.6569287×10^4

Problem 2

2.A)

we have the following formule:

SE(B^j) = sqrt(s**2.(XTX)^-1jj)

Therefore, we need the third element of the diagonal of the given matrix to compute SE which is ‘r sqrt()

```
s2 <- 2

mat <- matrix(c(1,0.25,0.25,0.25,0.5,-0.25, 0.25, -0.25, 2), byrow = 3, ncol = 3)
```

Therefore, we need the third element of the diagonal of the given matrix to compute SE which is 2

2.B)

To test the hypothesis that 2=0 we will use a t-test for the regression coefficient ^2.

Hypotheses: Null Hypothesis (0): 2=0 (there is no effect of the predictor 2) Alternative Hypothesis ( ): 2 0 (there is a significant effect of 2).

Test Statistic:

The test statistic for the hypothesis test is calculated as:

 = 2/ (2)

```
b2 <- 15
se_b2 <- sqrt(s2*mat[3,3])
t_statistic <- b2/se_b2

# degree of freedom

df <- 25 - 3


# get p.value

2 * pt(-7.5, df = 22)
```

```
## [1] 1.694205e-07
```

since the p-value is smaller than 0.05, we reject the null hypothesis, concluding that 2 0 and therefore 2 has a significant effect on y.

2.C)

cov(1, 2) is the (2,3) of covariance matrix times s^2, therefore it is -0.5

```
# formula for getting the SE of (B1-B2)

se_deduction <- sqrt(mat[2,2]*s2+mat[3,3]*s2-2*mat[2,3]*s2)
```

The standard error of ^1 − ^2 is 2.4494897

2.D)

To test the hypothesis that 1= 2 we will use a t-test as follow:

Hypotheses: Null Hypothesis (0): 1= 2 Alternative Hypothesis ( ): 1 2

Test Statistic:

The test statistic for the hypothesis test is calculated as:

=(1- 2)/ (1- 2)

```
b1 <- 12
b2 <- 15
t_statistic <- (b1-b2)/se_deduction

2 * pt(-abs(t_statistic), df = 22)
```

```
## [1] 0.233624
```

The p-value is greater than 0.05, therefore we fail to reject the null hypothesis and conclude that there is no significant difference between ^1 and ^2

2.E)

```
sst <- 120
sse <- s2*df

ssr <- sst - sse
```

SSR or the percentage of variation in y that is explained by the model is 63.3333333%

```
msr <- ssr/2
mse <- sse/22

f_statistic <- msr/mse

pf(f_statistic, df1 = 2, df2 = 22, lower.tail = FALSE)
```

```
## [1] 1.610593e-05
```

The p-value is lower than 0.05 therefore the H0 (1= 2=0) can be rejected. This means that the model with both predictors significantly explains the variation in the response variable y.