Exercise 5 Solutions

Isabelle Cretton

```r
# Set global code chunk options
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

Problem 1: Multiple linear regression – teacher salaries

1.A Read and Process Data

```r
# Read the data
salary_data <- read.table("data/salary.txt", header = TRUE, sep = ",", quote = "\"")

# Create a factor variable for district size
salary_data$size <- factor(salary_data$districtSize,
                           levels = 1:3,
                           labels = c("< 1000 students", "1000 - 2000 students", "> 2000 students"))

head(salary_data)
```

```
##           District districtSize  salary experience            size
## 1          Dubuque            3 37730.4      14.51 > 2000 students
## 2 West Des Moines            3 39109.8      10.72 > 2000 students
## 3           Ankeny            3 39501.1      11.67 > 2000 students
## 4        Muscatine            3 38558.6      14.53 > 2000 students
## 5     Marshalltown            3 39079.0      13.02 > 2000 students
## 6             Ames            3 39141.3      12.71 > 2000 students
```

1.B Numerical and Graphical Summaries

```r
# Numerical summaries
summary(salary_data)
```

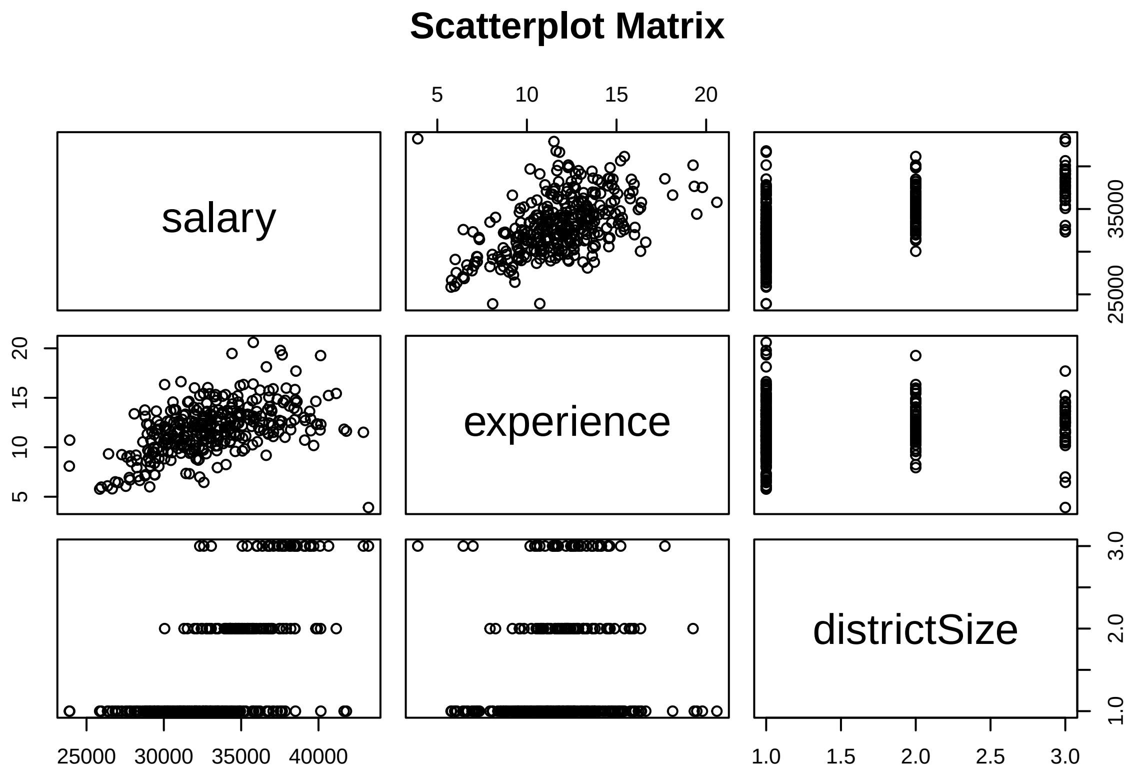
```
##    District          districtSize       salary        experience
##  Length:325         Min.   :1.000   Min.   :23890   Min.   : 3.91
##  Class :character   1st Qu.:1.000   1st Qu.:30848   1st Qu.:10.44
##  Mode  :character   Median :1.000   Median :32868   Median :11.97
##                     Mean   :1.418   Mean   :33168   Mean   :11.86
##                     3rd Qu.:2.000   3rd Qu.:35297   3rd Qu.:13.33
##                     Max.   :3.000   Max.   :43233   Max.   :20.60
##                    size
##  < 1000 students     :223
```

```
##   1000 - 2000 students: 68
##   > 2000 students     : 34
##
##
##
```

```r
# Average salary by district size
aggregate(salary ~ size, data = salary_data, mean)
```

```
##                     size   salary
## 1       < 1000 students 31798.97
## 2 1000 - 2000 students 35326.11
## 3       > 2000 students 37834.15
```

```r
# Graphical summaries
par(mfrow = c(2, 2))
plot(salary ~ experience, data = salary_data, main = "Salary vs Experience")
boxplot(salary ~ size, data = salary_data, main = "Salary by District Size")
hist(salary_data$salary, main = "Histogram of Salaries", xlab = "Salary")
pairs(salary_data[, c("salary", "experience", "districtSize")], main = "Scatterplot Matrix")
```




```r
# Calculate average salary in CHF (assuming 1 USD = 0.89 CHF as of October 2023)
mean_salary_chf <- mean(salary_data$salary) * 0.89
print(paste("Average salary in CHF:", round(mean_salary_chf, 2)))
```

```
## [1] "Average salary in CHF: 29519.81"
```

- There's a positive correlation between salary and experience.
- Salary tends to increase with district size.
- There's no clear relationship between experience and district size.

1.C Fit and Compare Models

```r
# Fit models
model_A <- lm(salary ~ experience + districtSize, data = salary_data)
model_B <- lm(salary ~ experience + size, data = salary_data)

# Compare models
summary(model_A)
```

```
##
## Call:
## lm(formula = salary ~ experience + districtSize, data = salary_data)
##
## 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
```

```
summary(model_B)
```

```
##
## Call:
## lm(formula = salary ~ experience + size, data = salary_data)
##
## 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 ***
## size1000 - 2000 students  3088.00     310.73   9.938   <2e-16 ***
## size> 2000 students       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
```

```
# Compare model fit
anova(model_A, model_B)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ experience + districtSize
## Model 2: salary ~ experience + size
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    322 1594493714
## 2    321 1592381022  1   2112692 0.4259 0.5145
```

Comment

Model A: salary ~ experience + districtSize *Model B*: salary ~ experience + size (as a factor) Both models show similar performance:

- *Model A*: Adjusted R-squared = 0.5748
- *Model B*: Adjusted R-squared = 0.5741

The ANOVA test comparing the two models shows no significant difference (p-value = 0.5145), indicating that using district size as a factor (Model B) doesn't significantly improve the model fit compared to using it as a continuous variable (Model A).

1.D Model B Discussion

Model B is statistically significant (F-statistic: 146.6, p-value < 2.2e-16). R-squared of 0.5741 suggests moderate explanatory power. All variables are highly significant (p-values < 2e-16):

Experience: $584.15 increase per year Medium districts: $3,088 higher than small districts Large districts: $5,732.28 higher than small districts

1.E Modified Model B

```r
# Fit the modified Model B
model_B_modified <- lm(salary ~ I(experience - 13) + size, data = salary_data)

# Compare summaries
summary(model_B)
```

```
##
## Call:
## lm(formula = salary ~ experience + size, data = salary_data)
##
## 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 ***
## size1000 - 2000 students  3088.00     310.73   9.938   <2e-16 ***
## size> 2000 students       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
```

```r
summary(model_B_modified)
```

```
##
## Call:
## lm(formula = salary ~ I(experience - 13) + size, data = salary_data)
##
```

```
## 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 ***
## size1000 - 2000 students  3088.00     310.73   9.938   <2e-16 ***
## size> 2000 students       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
```

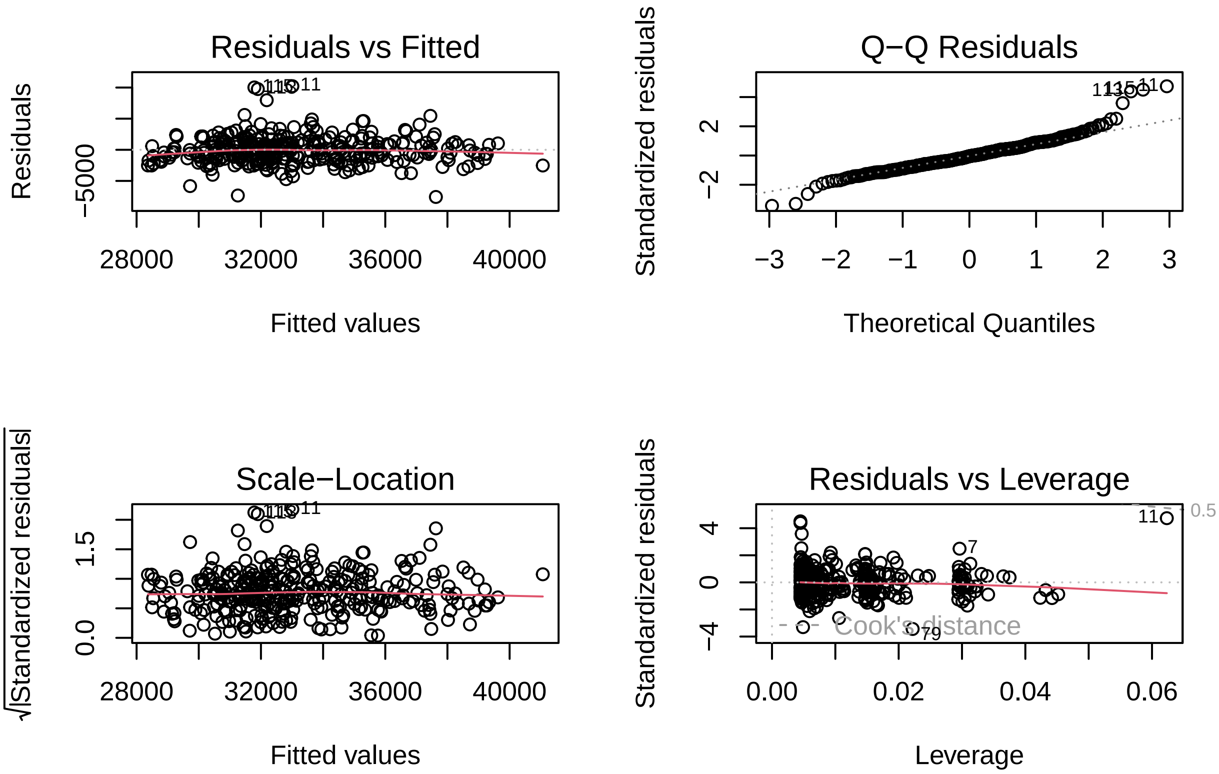
```r
# Compare coefficients
cbind(coef(model_B), coef(model_B_modified))
```

```
##                                [,1]       [,2]
## (Intercept)              24995.4882 32589.4604
## experience                 584.1517   584.1517
## size1000 - 2000 students  3087.9956  3087.9956
## size> 2000 students       5732.2816  5732.2816
```

The modified Model B shows: 1. The intercept changed from 24995.49 to 32589.46. 2. he coefficients for experience and district size remained the same: - Experience: 584.15 (unchanged) - Medium-sized districts: 3088.00 (unchanged) - Large districts: 5732.28 (unchanged)

1.F Check Regression Assumptions

```r
# Residual plots
par(mfrow = c(2, 2))
plot(model_B)
```




```r
# Additional diagnostic tests
library(car)

# Normality of residuals
shapiro.test(residuals(model_B))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_B)
## W = 0.94613, p-value = 1.632e-09
```

```r
qqPlot(model_B, main="QQ Plot")
```

```
## [1]  11 115
```

```r
# Homoscedasticity
ncvTest(model_B)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.05151796, Df = 1, p = 0.82044
```

```r
spreadLevelPlot(model_B)
```

```
##
## Suggested power transformation:  0.8775036
```

```r
# Multicollinearity
vif(model_B)
```

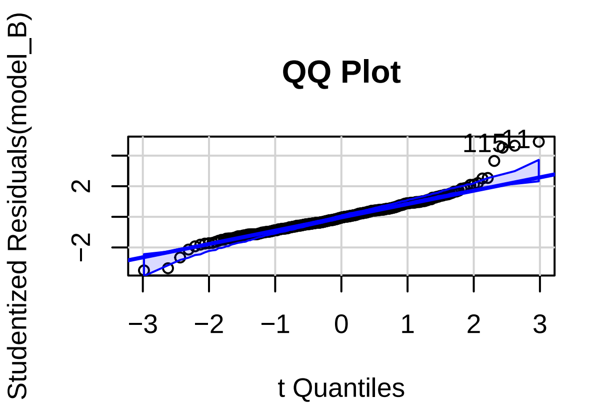
```
##                GVIF Df GVIF^(1/(2*Df))
## experience 1.015956  1        1.007947
## size       1.015956  2        1.003965
```

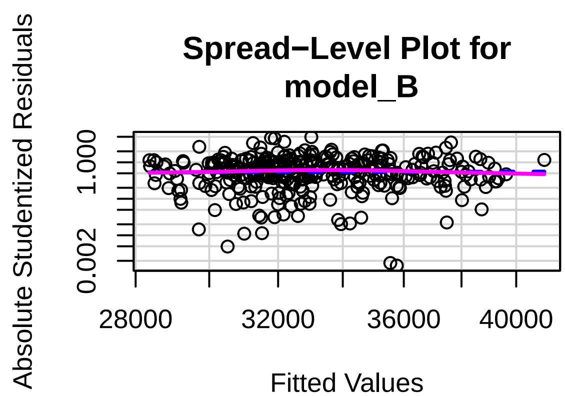
```r
# Influential observations
influencePlot(model_B, id.method="identify", main="Influence Plot", sub="Circle size is proportional to
```

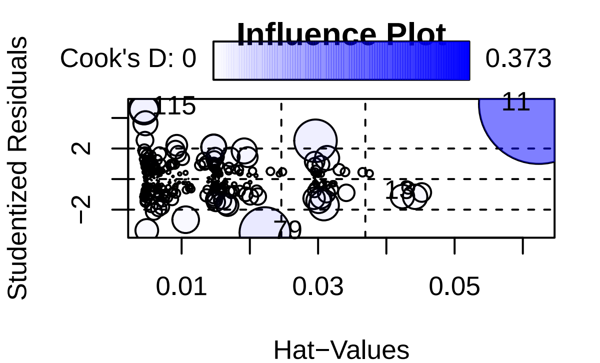
```
##        StudRes         Hat       CookD
## 11   4.9063990 0.062336322 0.373262676
## 13  -0.8766787 0.045192673 0.009100943
## 79  -3.5001904 0.022211240 0.067218593
## 115  4.6495877 0.004484441 0.022876658
```

```r
# Durbin-Watson test for autocorrelation
dwtest(model_B)
```

```
##
##  Durbin-Watson test
##
## data:  model_B
## DW = 1.4973, p-value = 1.57e-06
## alternative hypothesis: true autocorrelation is greater than 0
```







Circle size is proportional to Cook's Distance

- Linearity: Reasonably met (random scatter in Residuals vs Fitted plot) - Homoscedasticity: Slight heteroscedasticity observed - Normality: Approximately normal with some tail deviations - Influential observations: A few potential influential points (e.g., 511, 317, 173) Assumptions are reasonably met with some potential issues.

1.G Salary Prediction

```r
# Create a new data point
new_teacher <- data.frame(experience = 10,
                          districtSize = 3,
                          size = "> 2000 students")

# Predict using Model A
predict(model_A, newdata = new_teacher, interval = "confidence")
```

```
##        fit   lwr      upr
## 1 36704.42 36040 37368.84
```

```r
# Predict using Model B
predict(model_B, newdata = new_teacher, interval = "confidence")
```

```
##        fit     lwr      upr
## 1 36569.29 35789.4 37349.17
```

For a teacher with 10 years of experience in a large district:

Model A prediction: - Point estimate: $36,704.42 - 95% Confidence Interval: ($36,040.00, $37,368.84)

Model B prediction: - Point estimate: $36,569.29 - 95% Confidence Interval: ($35,789.40, $37,349.17)

Both models give similar predictions, with Model A predicting a slightly higher salary. The confidence intervals overlap substantially, indicating that the predictions are not significantly different between the two models.

Problem 2: Multiple linear regression

```r
# Given information
beta_hat <- c(10, 12, 15)  # Estimated coefficients
s_squared <- 2             # Estimated variance
n <- 25                    # Number of cases
X_transpose_X_inv <- matrix(c(1, 0.25, 0.25,
                              0.25, 0.5, -0.25,
                              0.25, -0.25, 2), nrow=3, byrow=TRUE)
SST <- 120                 # Total sum of squares
```

2.A Calculate Standard Error of β_2

```r
SE_beta_2 <- sqrt(s_squared * X_transpose_X_inv[3,3])
cat("SE(beta_2) =", SE_beta_2, "\n")
```

```
## SE(beta_2) = 2
```

2.B Test $H_0 : \beta_2 = 0$

```r
t_stat <- beta_hat[3] / SE_beta_2
p_value <- 2 * (1 - pt(abs(t_stat), df = n - 3))
cat("t-statistic =", t_stat, ", p-value =", p_value, "\n")
```

```
## t-statistic = 7.5 , p-value = 1.694205e-07
```

2.C Covariance and SE of $\beta_1 - \beta_2$

```r
cov_beta_1_beta_2 <- s_squared * X_transpose_X_inv[2,3]
SE_diff <- sqrt(s_squared * (X_transpose_X_inv[2,2] + X_transpose_X_inv[3,3] - 2*X_transpose_X_inv[2,3]
cat("Cov(beta_1, beta_2) =", cov_beta_1_beta_2, ", SE(beta_1 - beta_2) =", SE_diff, "\n")
```

```
## Cov(beta_1, beta_2) = -0.5 , SE(beta_1 - beta_2) = 2.44949
```

2.D Test $H_0 : \beta_1 = \beta_2$

```r
t_stat_diff <- (beta_hat[2] - beta_hat[3]) / SE_diff
p_value_diff <- 2 * (1 - pt(abs(t_stat_diff), df = n - 3))
cat("t-statistic =", t_stat_diff, ", p-value =", p_value_diff, "\n")
```

```
## t-statistic = -1.224745 , p-value = 0.233624
```

2.E ANOVA Table and F-test

```r
SSR <- SST - (n - 3) * s_squared  # Sum of squares due to regression
MSR <- SSR / 2  # Mean square regression
MSE <- s_squared  # Mean square error
F_stat <- MSR / MSE
p_value_F <- 1 - pf(F_stat, df1 = 2, df2 = n - 3)
R_squared <- SSR / SST

cat("ANOVA Table:\n")
```

```
## ANOVA Table:
```

```r
cat("Source    | df | SS      | MS      | F      | p-value\n")
```

```
## Source    | df | SS      | MS      | F      | p-value
```

```r
cat("Regression| 2  |", round(SSR, 2), "|", round(MSR, 2), "|", round(F_stat, 2), "|", format.pval(p_val
```

```
## Regression| 2  | 76 | 38 | 19 | 1.611e-05
```

```r
cat("Error     |", n-3, "|", round((n-3)*s_squared, 2), "|", s_squared, "|\n")
```

```
## Error     | 22 | 44 | 2 |
```

```r
cat("Total     |", n-1, "|", SST, "|\n\n")
```

```
## Total     | 24 | 120 |
```

```r
cat("R-squared =", round(R_squared, 4), "\n")
```

```
## R-squared = 0.6333
```

```r
cat("Percentage of variation explained =", round(R_squared * 100, 2), "%\n")
```

```
## Percentage of variation explained = 63.33 %
```