Sahiti Kovvuri R Homework #1

STA 4155

R Homework for Chapter 4

**Question 1: Produce a scatterplot between the Cost of Living Index and EACH of the other index variables. As a result, there should be 4 scatterplots in total. Examine the relationship shown in each scatterplot in terms of its form, strength and direction**

**Q1 Answer:** The relationship between the **Cost of Living Index** and the **Rent Index** has a Moderate Positive Linear Relationship

**R code:** plot(Cost.of. Living.Index ~ Rent.Index, data = cost)

**R output:** Chart, scatter chart

Description automatically generated

**Q1 Answer:** The relationship between the **Cost of Living Index** and the **Groceries Index** has a Strong Positive Linear Relationship

**R code:** plot(Cost.of.Living.Index ~ Groceries.Index, data = cost)

**R output:** Chart, scatter chart

Description automatically generated

**Q1 Answer:** The relationship between the **Cost of Living Index** and the **Restaurants Index** has a Strong Positive Linear Relationship

**R code:** plot(Cost.of.Living.Index ~ Restaurant.Price.Index, data = cost)

**R output:**Chart, scatter chart

Description automatically generated

**Q1 Answer:** The relationship between the **Cost of Living Index** and the **Local Purchasing Power Index** has a Weak Positive Linear Relationship

**R code:** plot(Cost.of.Living.Index ~ Local.Purchasing.Power.Index, data = cost)

**R output:** Chart, scatter chart

Description automatically generated

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 2: Compute the correlation coefficients for all the scatterplots obtained above**

**Q2 Answer:** The correlation coefficient between the Cost of Living Index the **Rent Index is** 0.7722926

**R code:** cor(cost$Cost.of.Living.Index,cost$Rent.Index) #correlation of the COL index and rent

**R output:** [1] 0.7722926

**Q2 Answer:** The correlation coefficient between the Cost of Living Index the **Groceries Index is** 0.9538616

**R code:** cor(cost$Cost.of.Living.Index,cost$Groceries.Index) #correlation of the COL index and groceries

**R output:** [1] 0.9538616

**Q2 Answer:** The correlation coefficient between the Cost of Living Index the **Restaurants Index is** 0.9493554

**R code:** cor(cost$Cost.of.Living.Index,cost$Restaurant.Price.Index) #correlation of the COL index and restaurants

**R output:** [1] 0.9493554

**Q2 Answer:** The correlation coefficient between the Cost of Living Index the **Local Purchasing Power Index is** 0.525902

**R code:** cor(cost$Cost.of.Living.Index,cost$Local.Purchasing.Power.Index) #correlation of the COL index and LPP

**R output:** [1] 0.525902

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 3: Verify the conditions for EACH correlation coefficient computed above**

**Q3 Answer:**

Rent Index

* For the Rent Index, the scatter plot follows a linear pattern that shows that **linearity condition is met.**
* For the Rent Index, the scatter plot follows the Rent Index (integer) and the Cost of Living (integer) which refers to the city which is qualitative. The **quantitative condition is met**
* For the Rent Index, the scatter plot can be observed to have an outlier which can distort the data, the **outlier assumption is not met**

Groceries Index

* For the Groceries Index, the scatter plot follows a linear pattern that shows that **linearity condition is met.**
* For the Groceries Index, the scatter plot follows the Rent Index (integer) and the Cost of Living (integer) which refers to the city which is qualitative. The **quantitative condition is met**
* For the Groceries Index, the scatter plot can be observed to not have an outlier which can distort the data, the **outlier assumption is met**

Restaurant Price Index

* For the Restaurant Price Index, the scatter plot follows a linear pattern that shows that **linearity condition is met.**
* For the Restaurant Price Index, the scatter plot follows the Rent Index (integer) and the Cost of Living (integer) which refers to the city which is qualitative. The **quantitative condition is met**
* For the Restaurant Price Index, the scatter plot can be observed to not have an outlier which can distort the data, the **outlier assumption is met**

Local Purchasing Power Index

* For the Local Purchasing Power Index, the scatter plot follows a sparsely non linear pattern that shows that **linearity condition is not met.**
* For the Local Purchasing Power Index, the scatter plot follows the Rent Index (integer) and the Cost of Living (integer) which refers to the city which is qualitative. The **quantitative condition is met**
* For the Local Purchasing Power Index, the scatter plot can be observed to have an outlier which can distort the data, the **outlier assumption is not met**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 4: Fit a linear regression model between the Cost of Living Index and each of the other index variables. As a result, there should be 4 regression models in total. Interpret the resulting estimated slope in each model.**

**Q4 Answer:** The estimated slope in this model is 1.024624

**R code**: mod1 <- lm(Cost.of.Living.Index ~ Rent.Index, data = cost)

coef (mod1)

**R output**: (Intercept) Rent.Index

45.232600 1.024624

**Q4 Answer:** The estimated slope in this model is 0.9529463

**R code**: mod2 <- lm(Cost.of.Living.Index ~ Groceries.Index, data = cost)

coef (mod2)

**R output**: (Intercept) Groceries.Index

9.2178364 0.9529463

**Q4 Answer:** The estimated slope in this model is 0.8033304

**R code**: mod3 <- lm(Cost.of.Living.Index ~ Restaurant.Price.Index, data = cost)

coef (mod3)

**R output**: (Intercept) Restaurant.Price.Index

24.6635984 0.8033304

**Q4 Answer:** The estimated slope in this model is 0.3761637

**R code**: mod4 <- lm(Cost.of.Living.Index ~ Local.Purchasing.Power.Index, data = cost)

coef (mod4)

**R output**: (Intercept) Local.Purchasing.Power.Index

48.9974246 0.3761637

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 5: Based on the correlation coefficients and the regression models obtained above, which item would be the best predictor of overall cost in these cities? Which would be the worst? Explain.**

**Q5 Answer:** The better predictor of overall cost in these cities is the Groceries Index because its model (Model 2 has a high r (which tells us the correlation) as well as a high R-squared (0.909852), where as the worse model to judge the overall cost in these cities is the Local Purchasing Power Index because of its low r and R-Squared Value (0.2765729)

**Q5 Answer:** The r squared value of model 1 is 0.5964358

**R code**: summary(mod1)$r.squared

**R output**: [1] 0.5964358

**Q5 Answer:** The r squared value of model 2 is 0.909852

**R code**: summary(mod2)$r.squared

**R output**: [1] 0.909852

**Q5 Answer:** The r squared value of model 3 is 0.9012757

**R code**: summary(mod3)$r.squared

**R output**: [1] 0.9012757

**Q5 Answer:** The r squared value of model 4 is 0.2765729

**R code**: summary(mod4)$r.squared

**R output**: [1] 0.2765729

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 6: Find the cost of living as predicted by Groceries Index and its residual for Beijing, China. (Hint: Find row index of Beijing in the dataset, and then use that index to extract the corresponding fitted value and residual from the regression result.)**

**Q6 Answer:** The Cost of Living as predicted by the Groceries Index (mod2) is 88.85556

**R code**: predval = which(cost$City == 'Beijing, China')

mod2$fitted.values[predval]

**R output**: 172

88.85556

**Q6 Answer:** The residual of the Cost of Living as predicted by the Groceries Index (mod2) is -11.66556

**R code**: mod2$residuals[predval]

**R output**: 172

-11.66556

**Appendix:**

cost <- read.table('Cost\_Of\_Living\_Index.txt', sep = '\t', header = TRUE)

#1 Produce a scatterplot between the Cost of Living Index and EACH of the other index variables. As a result, there should be 4 scatterplots in total. Examine the relationship shown in each scatterplot in terms of its form, strength and directiion

plot(Cost.of.Living.Index ~ Rent.Index, data = cost)

#moderate positive linear relationship

plot(Cost.of.Living.Index ~ Groceries.Index, data = cost)

#Strong positive linear relationship

plot(Cost.of.Living.Index ~ Restaurant.Price.Index, data = cost)

#Strong Positive linear relationship

plot(Cost.of.Living.Index ~ Local.Purchasing.Power.Index, data = cost)

#Weak linear relationship

#2 Compute the correlation coefficients for all the scatterplots obtained above

cor(cost$Cost.of.Living.Index,cost$Rent.Index) #correlation of the COL index and rent

cor(cost$Cost.of.Living.Index,cost$Groceries.Index) #correlation of the COL index and groceries

cor(cost$Cost.of.Living.Index,cost$Restaurant.Price.Index) #correlation of the COL index and restaurants

cor(cost$Cost.of.Living.Index,cost$Local.Purchasing.Power.Index) #correlation of the COL index and LPP

#4 Fit a linear regression model between the Cost of Living Index and each of the other index variables. As a result, there should be 4 regression models in total. Interpret the resulting estimated slope in each model

mod1 <- lm(Cost.of.Living.Index ~ Rent.Index, data = cost)

coef (mod1)

mod2 <- lm(Cost.of.Living.Index ~ Groceries.Index, data = cost)

coef (mod2)

mod3 <- lm(Cost.of.Living.Index ~ Restaurant.Price.Index, data = cost)

coef (mod3)

mod4 <- lm(Cost.of.Living.Index ~ Local.Purchasing.Power.Index, data = cost)

coef (mod4)

#5 Based on the correlation coefficients and the regression models obtained above, which item would be the best predictor of overall cost in these cities? Which would be the worst? Explain.

summary(mod1)$r.squared

summary(mod2)$r.squared

summary(mod3)$r.squared

summary(mod4)$r.squared

#The better predictor of overall cost in these cities is the Groceries Index because its model (Model 2 has a high r (which tells us the correlation) as well as a high R-squared (0.909852), where as the worse model to judge the overall cost in these cities is the Local Purchasing Power Index because of its low r and R-Squared Value (0.2765729)

#6 Find the cost of living as predicted by Groceries Index and its residual for Beijing, China. (Hint: Find row index of Beijing in the dataset, and then use that index to extract the corresponding fitted value and residual from the regression result.)

#corresponding fitted value

predval = which(cost$City == 'Beijing, China')

mod2$fitted.values[predval]

#residual from the regression result

mod2$residuals[predval]