**Question 1**

In Question 1, we use a data set, “Rent\_Data.txt”, with details about apartments. The data includes rent of the apartments as well as other corresponding variable data such as Rent (in $), Distance from Airport (in Miles), Distance to Downtown (in Miles), and Distance to University (in Miles).

1. Open the ‘Rent\_Data.txt’ data in R-studio

Import DatasetFrom Text (base) Choose ‘Rent\_Data.txt’Hit Import

1. Create a pairwise correlation plot and correlation matrix.

# R-code: gpairs() (run line 17 on the R-file) and cor() (run line 18 on the R-file)

[Insert R-output]

1. Using scatterplots and correlation matrix, briefly describe the relationship between the response variable (Rent) and three predictors using 4 elements such as (1) form, (2) direction, (3) strength, and (4) outliers.

[Write your answer here ]

1. Describe histograms of all 4 variables based on (1) number of peaks, (2) skewness (symmetric vs. asymmetric), and (3) outliers.

[Write your answer here]

1. Fit **THREE** simple regression models with’ Rent’ as the response variable and ‘Distance.from.Airport’, ‘Distance.to…Downtown’, and ‘Distance.to.University’ as a predictor for each model. Use the summary() function to print the results.

[Insert R-output]

1. Summarize the regression results using a table below.

\*Using the form coefficient (p-value); 3.14 (0.0012) if the coefficient is 3.14 and its p-value is 0.0012.

[Write your answer in the table below]

|  |  |  |  |
| --- | --- | --- | --- |
| Response variable  = Rent | Model 1  (Distance.from.Airport) | Model 2  (Distance.to…Downtown) | Model3 (Distance.to.University) |
| Intercept |  |  |  |
| Estimate |  |  |  |
| Residual Standard Error |  |  |  |
|  |  |  |  |
| F-statistics |  |  |  |

1. Using ‘confint()’ function, obtain 95% confidence intervals of the coefficient of each model.

# R-code: confint() (run line 31 on the R-file, modify the line 31 for model 2 and 3)

[Insert R-output]

[Write your answer in the table below]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictors** | **Coefficient ()** | **Confidence interval**  **(Lower CI, Upper CI)** | **Test Hypothesis** | |
| **Is ‘0’ inside of CI** | **Decision** |
| Distance.from.Airport |  |  |  |  |
| Distance.to…Downtown |  |  |  |  |
| Distance.to.University |  |  |  |  |

1. When we build the regression model to predict outcome variable, it is very important to check the assumptions. Without verifying that the model has satisfied the assumptions, the results from the model may be misleading.

Discuss about the assumptions we need to consider after fitting the mode to verify the model (Which assumptions we need to check, how to check the assumptions, when the assumptions are violated, and how can we resolve the violations).

[Write your answer here]

1. Check the assumptions of the models using plot() and gvlma(). Is there evidence of outliers or high leverage observations in the models? If so, please inspect the outliers.

# R-code: plot(), influencePlot(), gvlma() (run lines 38-45 on the R-file, modify the codes for the model 2 and 3)

[Insert R-output]

[Write your answer in the table below]

|  |  |  |
| --- | --- | --- |
| **Assumptions** | **Residual plot** | **Normal Q-Q plot** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

[Write additional comments here]

1. Discuss about the strength and the weakness of the models in Question 1. And, briefly discuss the way(s) to overcome the weakness of the models.

[Write your answer here]

**Question 2**

The dataset ‘teengamb’ in the ‘faraway’ package concerns a study of teenage gambling in Britain. This data contains 47 observations with 5 variables (please see the below for the data frame).

|  |  |
| --- | --- |
| Variable | Descriptions |
| sex | 0=male, 1=female |
| status | Socioeconomic status score based on parents’ occupation |
| income | Income in pounds per week |
| verbal | Verbal score in words out of 12 correctly defined |
| gamble (response) | Expenditure on gambling in pounds per year |

1. Create a pairwise correlation plot and correlation matrix. **Are there relationships between the predictors? Explain.**

# R-code: gpairs() (run line 58 on the R-file) and cor() (run line 59 on the R-file)

[Insert R-output]

[Write your answer here ]

1. Fit a multiple regression model to predict ‘gamble’. Write an estimated regression model using the coefficients on the summary output and interpret the significant coefficients.

# R-code: lm() (run line 65 on the R-file) and summary() (run line 66 on the R-file)

[Insert R-output]

[Write your answer here ]

1. Plot the side-by-side boxplot of ‘gamble’ for different gender groups. Does that suggest that male and female students have different gambling behavior?

# R-code: plot() (run line 70 on the R-file)

[Insert R-output]

[Write your answer here ]

1. Considering your answer of part 3), update the model in part 2 with interaction term (sex\*income). Using the model summary, side-by-side box plot (in part 3), and interaction graph, test the significance of interaction effect between the two significant predictors, sex and income in the form of

.

# R-code: lm() (run line 73 on the R-file), and summary() (run line 74 on the R-file)

[Insert R-output]

Chart, line chart

Description automatically generated

[Write your answer here]

**Question 3**

Wage data in the ‘ISLR’ package includes the 3000 observations on the following 11 variables.

|  |  |
| --- | --- |
| Variable | Descriptions |
| year | Year that wage information was recorded |
| age | Age of worker |
| maritl | A factor with levels 1=Never married, 2=Married, 3=Widowed, 4=Divorces, and 5=Separated indicating marital status |
| race | A factor with levels 1=White, 2=Black, 3=Asian, and 4=Other indicating race |
| education | A factor with levels 1=<HS Grad, 2=HS Grad, 3=Some College, 4=College Grad, and 5=Advanced Degree indicating education level |
| region | Region of the country(mid-atlantic only) |
| jobclass | A factor with levels 1=Industrial and 2=Information indicating type of job |
| health | A factor with levels 1<=Good and 2>=Very Good indicating health level of worker |
| Health\_ins | A factor with levels 1=Yes and 2=No indicating whether worker has health incurance |
| logwage | Log of workers wage |
| wage | Workers raw wage |

**h**

**g**

1. Summarize the Wage dataset using dim(), str(), and summary().

# R-code: dim(), str(), and summary() (run line 85-88)

[Insert R-output]

[Write your answer here]

1. Create a scatterplot using age and wage. Describe the scatterplot in terms of form, direction, strength, and outlier(s).

# R-code: plot() (run line 92 on the R-file)

[Insert R-output]

[Write your answer here]

1. Fit a linear regression model that explains the wage according to the age. Does age significantly explain the wage? Evaluate the model using summary statistics.

# R-code: lm() (run line 95-96 on the R-file)

[Insert R-output ]

[Write your answer here ]

1. Check the assumptions. Which assumptions are violated? Justify your answer.

# R-code: run lines 99-103

[Insert R-output ]

[Write your answer here ]

1. Considering your answer on part 4), please discuss about the two most commonly violated assumptions of the regression model; which assumptions are frequently violated? How can we relax those assumptions?

[Write your answer here ]

1. Fit polynomial models with different degrees of power. Using the summary tables, plots (for assumptions), global test, and ANOVA table, choose the best model to explain the wage.

Model 1: From part 3.

Model 2:

Model 3:

Model 4:

Model 5:

# R-code: run lines 105-126

[Insert R-output]

[Summarize the model fitting ]

[Check assumptions ]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Assumptions | Models | | | | |
| Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Linearity |  |  |  |  |  |
| Normality |  |  |  |  |  |
| Independent |  |  |  |  |  |
| Constant variance |  |  |  |  |  |

Additional comments:

[Interpret the results on ANOVA test ]

[Write your answer here: What is the best model? Why?]