STA2005S Regression Assignment

Questions

Background

Air pollution is a pressing environmental and health issue in South Africa's urban centers. Elevated levels of air pollutants like particulate matter (`particulate_matter`) are linked to respiratory diseases and other health problems. Understanding the factors influencing air quality is essential for developing effective policies and interventions.

The South African Urban Air Quality Initiative (SAUAQI) has collected data from 150 monitoring stations across various cities. The dataset includes multiple explanatory variables:

- *particulate_matter*: Fine particulate matter concentration (µg/m^3) *(Response Variable)*.
- *traffic_density*: Number of vehicles per hour in the area.
- *industrial_activity*: Level of industrial activity near the monitoring station ("None", "Low", "Moderate", "High").
- *temperature*: Average ambient temperature (°C).
- *humidity*: Average relative humidity (%).
- *wind_speed*: Average wind speed (m/s).
- *day_of_week*: Day of the week ("Monday" to "Sunday").
- *holiday*: Whether the day is a public holiday ("Yes", "No").
- *urban_greenery*: Percentage of area covered by green spaces.

Objective

Analyze the relationships between these variables and `particulate_matter` levels to inform urban planning and public health strategies.

Data Loading Instructions

```r
#| results: hide
#| warning: false
#| message: false
#| error: false
if (!requireNamespace("remotes", quietly = TRUE)) {
  install.packages("remotes")
}
remotes::install_github("MiguelRodo/DataTidyRodoSTA2005S")
data("data_tidy_air_quality", package = "DataTidyRodoSTA2005S")
head(data_tidy_air_quality)
```

Questions (Total: 65 marks)

Part One: Analysis (40 marks)

Section 1: Introduction (2 marks)

1. *Problem & Unknown* (1 mark):

   - Clearly articulate the problem being addressed.
   - Identify the unknown factors that need investigation.

2. *Analysis Summary* (1 mark):

   - Provide a brief overview of your analytical approach.
   - Outline the expected findings.

Section 2: Data Exploration (12 marks)

1. *Density Plot* (1 mark):

   - Plot the density of `particulate_matter` levels with an overlaid normal distribution density.

2. *Pairwise Plots* (2 marks):

   - Create pairwise scatterplots for all continuous variables, including `particulate_matter`.

3. *Categorical Variable Plots* (2 marks):

   - Plot `particulate_matter` levels against each categorical variable.
   - Ensure that ordinal variables (e.g., `industrial_activity`, `day_of_week`) are properly ordered in your plots.

4. *Categorical Relationships* (2 marks):

   - Tabulate or graph the relationships between all categorical variables.

5. *Comments* (5 marks):

   - Distribution Characteristics (1 mark):

  - Discuss the distribution of `particulate_matter` levels.

- Observed Relationships (2 marks):
  - Describe any observed relationships between `particulate_matter` and other variables.

- Potential Collinearity (1 mark):
  - Identify any potential collinearity among explanatory variables.

Section 3: Simple Linear Regression (11 marks)

For this section, you do not need to comment on your findings. The purpose is to demonstrate ability to perform calculations.

1. *Model Fitting* (8 marks):

   - From first principles (i.e. manual calculation), fit a simple linear regression model of `particulate_matter` against `traffic_density`.
   - Reproduce the output from the `summary` function, working from from the coefficients table down (i.e. ignore the call and residuals sections).

2. *Simultaneous Hypothesis Test* (3 marks):

   - From first principles, perform a simultaneous hypothesis test to assess the effect of `industrial_activity` on `particulate_matter` levels.

Section 4: Multiple Linear Regression (10 marks)

1. *Fit Model* (3 marks):

   - Fit a multiple linear regression model including all explanatory variables.
   - Include an interaction term between `temperature` and `humidity`.
   - Display confidence intervals for individual coefficients (table or plot).

2. *Hypothesis Testing* (2 marks):

   - Perform hypothesis tests to determine if the following variables significantly affect `particulate_matter` levels:
     - Temperature (1 mark)

    - Humidity (1 mark)
    - Categorical Variables with More Than Two Levels (1 mark), such as `industrial_activity` or `day_of_week`.

3. *Interpretation* (5 marks):

- Interpret the coefficients of the significant variables, focusing on:
  - Statistical significance (p-values)
  - Effect sizes (magnitude and direction)
  - Confidence intervals

Section 5: Conclusion (5 marks)

1. *Summary* (2 marks):
   - Synthesize your key findings from the multiple regression analysis.

2. *Recommendations* (2 marks):
   - Discuss the practical implications for urban planning and public health.

3. *Future Research* (1 mark):
   - Suggest areas for further research or data collection.

Part Two: Simulation (15 marks)

Section 6: Type I Error Simulation (15 marks)

Objective

Suppose we fit the model:

$$Y = \beta_0 + \beta_1 \times \text{temperature} + e,$$

where we assume (when estimating regression coefficients and performing inference) that $e \sim \mathcal{N}(0, \sigma^2)$.

We want to simulate the Type I error rate for testing the effect of `temperature` on `particulate_matter` levels under different scenarios violating model assumptions.

Simulation assumptions

- $\beta_0 = 30$
- $\beta_1 = 0$ (under the null hypothesis; no effect of temperature)
- $\sigma^2 = 100$
- Use the `temperature` values from the `data_tidy_air_quality` dataset.
- Keep the `temperature` values constant across simulations.

Instructions

For each scenario:

- *Methodology* (1 mark per scenario):
  - Describe how you will simulate the data under the null hypothesis ($\beta_1 = 0$).
  - Specify how you introduce the violation of the assumption.
- *Simulation* (2 marks per scenario):
  - Set the seed (e.g., `set.seed(123)`) for reproducibility.
  - Run 1,000 simulations.
  - In each simulation:
    * Generate the error term e according to the scenario.
    * Calculate $Y_i = 30 + e$.
    * Fit the model $Y = \beta_0 + \beta_1 \times \text{temperature} + \text{error}$.
    * Perform the hypothesis test for β_1 (test if $\beta_1 = 0$).

- Record whether the null hypothesis is incorrectly rejected.

- *Results* (1 mark per scenario):
  - Report the proportion of simulations where the null hypothesis is incorrectly rejected (Type I error rate).
  - Discuss how the violation of the assumption affects the Type I error rate.

Scenarios

1. *Scenario A* (4 marks):
   - Errors follow a uniform distribution instead of a normal distribution.
   - *Guidance*:
     - If $e \sim U(a, b)$ with $a = -b$, choose a and b such that $\text{Var}(e) = 100$.
2. *Scenario B* (4 marks):
   - Errors exhibit heteroscedasticity.
   - *Guidance*:
     - You will need to simulate the variance of the error term for each observation.
     - Simulate the errors from a distribution where the variance has a mean 100 and a variance of 50.
3. *Scenario C* (4 marks):
   - Errors are autocorrelated with a correlation coefficient of $\rho = 0.3$ (i.e., $\text{Corr}(\varepsilon_i, \varepsilon_{i+1}) = 0.3$).
   - *Guidance*:
     - Use the following function to generate autocorrelated errors:

```r
generate_ar1_errors <- function(n, rho, var_epsilon) {
  # n: number of observations
  # rho: correlation coefficient
  # var_epsilon: desired variance of the errors
  sigma_u_squared <- var_epsilon * (1 - rho^2)
  u <- rnorm(n, mean = 0, sd = sqrt(sigma_u_squared))
  epsilon <- numeric(n)
  epsilon[1] <- u[1]
  for (i in 2:n) {
    epsilon[i] <- rho * epsilon[i - 1] + u[i]
  }
  epsilon
}
```

4. *Conclusion* (3 marks):

    - Summarize your findings across all scenarios.
    - Discuss the implications for the validity of hypothesis testing when model assumptions are violated.

Part Three: Presentation (10 marks)

Display of R Code:

- Including R code is optional but recommended for clarity.
- If you include code:
  - Do not display boilerplate code (e.g., loading packages, reading data).
  - Include code relevant to calculations or model fitting.
  - Ensure your code is well-formatted and commented.

Presentation Guidelines:

You begin the assignment with *10 marks* for presentation. Marks will be deducted for the following:

- *Graphs*:
  - Missing any of the following elements:
    - Figure caption (title below the figure)
    - Axes labels
    - Legend (if applicable)
  - Inappropriate scales or hard-to-read visuals due to:
    - Excessive decimal places
    - Small or illegible text
    - Poor color choices
    - Inappropriate graph types
    - Unnecessary 3D effects
- *Tables*:
  - Missing any of the following elements:
    - Table caption (title above the table)
    - Column headings
  - Hard-to-read tables due to:
    - Crowded text
    - Small or illegible text
    - Excessive decimal places
- *Text*:
  - Lack of clarity or conciseness.

  - Excessive spelling or grammatical errors.

- *Overall Structure*:

  - Illogical flow of sections.
  - Missing clear headings and subheadings.