STA2005S Regression Assignment

Questions

Introduction

This assignment consists of two parts:

1. **Part One: Analysis** (44 marks) — Analyse a dataset on household electricity consumption.
2. **Part Two: Simulation** (16 marks) — Simulate statistical power for regression models.

In addition, there are 10 marks for **presentation** across both parts, focusing on clarity, structure, and visualisation quality. Details at the end of this document.

As such, the total marks available are **70**.

Part One: Analysis

Introduction

Background

Household electricity demand is a key driver of load on South Africa's power grid. Understanding what influences daily household consumption can support demand management, tariff design, and energy-efficiency interventions.

A study of **150** households across the Cape Town metro recorded daily observations with the following variables:

- *consumption_kwh*: Daily household electricity consumption (kWh) *(Response Variable)*.
- *outside_temperature*: Daily average outdoor temperature (°C).
- *humidity*: Daily average relative humidity (%).
- *wind_speed*: Daily average wind speed (m/s).
- *household_size*: Number of residents.
- *appliance_index*: Composite index (0–100) of appliance use intensity.
- *energy_efficiency*: Household efficiency rating ("Poor", "Average", "Good", "Excellent").
- *solar_installed*: Whether the household has solar PV ("Yes", "No").
- *day_of_week*: Day of the week ("Monday" to "Sunday").
- *holiday*: Whether the day is a public holiday ("Yes", "No").

Objective

Analyse the relationships between these variables and `consumption_kwh` to inform demand-side management and efficiency policy.

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_energy_use", package = "DataTidyRodoSTA2005S")
```

If you cannot install the package (i.e. the `remotes::install_github()` commands fails) due to an error related to your GitHub token, then it is possible that your GitHub token has expired. A quick fix for doing this in `R` is available at this link. Your mileage may vary.

Questions (Total: 70 marks)

Part One: Analysis (44 marks)

Section 1: Introduction (3 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.
3. *Nature of analysis* (1 mark)
   - State and justify whether the analysis goal is primarily predictive or explanatory.

Section 2: Data Exploration (13 marks)

1. *Density Plots* (3 marks)
   - Plot a histogram of `consumption_kwh` (`freq = FALSE`, if using base `R`) with an overlaid normal density. Comment on its shape.
   - Plot the density of `consumption_kwh`, stratified by `solar_installed`, with overlaid normal densities. Comment on the shape, and consider why the densities changes from the unstratified case.
2. *Pairwise Plots* (2 marks)
   - Create pairwise scatterplots for all continuous variables, including `consumption_kwh`.
3. *Categorical Variable Plots* (2 marks)
   - Plot `consumption_kwh` against each categorical variable.
   - Ensure ordinal variables (e.g., `energy_efficiency`, `day_of_week`) are correctly ordered.
4. *Categorical Relationships* (2 marks)

- Tabulate relationships between a) energy efficiency and solar usage, and b) days of the week and holiday.

5. *Comments* (4 marks)

    - Observed Relationships (2 marks): identify any explanatory variables with apparent relationships with `consumption_kwh`.
    - Potential Collinearity (1 mark): identify any potential collinearity among explanatory variables.
    - Outliers (1 mark): note any outlying observations.

Section 3: Simple Linear Regression (11 marks)

For this section, you do **not** need to comment on findings; demonstrate calculations from first principles. By first principles, we mean that you should perform all matrix calculations manually, without using built-in functions like `lm()`.

1. *Model Fitting* (8 marks)

    - From first principles (manual calculation), fit a simple linear regression of `consumption_kwh` on `outside_temperature`.
    - Reproduce the `summary()` output, working from the coefficients table down (ignore call and residuals sections).

2. *Simultaneous Hypothesis Test* (3 marks)

    - From first principles, perform a simultaneous hypothesis test for the effect of `energy_efficiency` on `consumption_kwh`.

Section 4: Multiple Linear Regression (12 marks)

1. *Fit Model* (3 marks)

    - Fit a multiple linear regression including **all** explanatory variables.
    - Include an interaction term between `outside_temperature` and `humidity`.
    - Construct and display a table of coefficients with 95% confidence intervals and p-values.

2. *Hypothesis Testing* (4 marks)

    - Test whether the following significantly affect `consumption_kwh`:
        - Outside Temperature (1 mark)
        - Humidity (1 mark)
        - Any categorical variables with more than one level (2 marks)

- Hint: use the `anova()` function to compare nested models.

3. *Interpretation* (5 marks)

   - Interpret coefficients of variables that are statistically significant at the 5% level, focusing on:
     - Statistical significance (p-values)
     - Effect sizes (magnitude and direction)
     - Confidence intervals

Section 5: Conclusion (5 marks)

1. *Summary* (2 marks)

   - Synthesise key findings from the multiple regression analysis.

2. *Recommendations* (2 marks)

   - Discuss practical implications for demand-side management and efficiency policy.

3. *Future Research* (1 mark)

   - Suggest areas for further data collection or modelling.

Part Two: Simulation of Power (16 marks)

Introduction

Statistical power is the probability of correctly rejecting a false null hypothesis for a given effect size.

In this part, you will simulate statistical power for testing the effect of a predictor variable in regression models under two different scenarios:

1. **Scenario A:** The true relationship is linear, errors are i.i.d. normal (baseline). You will vary the effect size β_1 in a linear model for a fixed sample size.
2. **Scenario B:** The true relationship is *exponential* in x, and you compare power between a mis-specified linear fit and the correctly specified exponential fit across varying sample sizes for a fixed effect size.

General Simulation set-up

- In both scenarios:
  - Use a **fixed** x vector in each simulation for a given scenario.
  - Set $\beta_0 = 12$.
  - Target marginal error variance $\sigma^2 = 6.25$ (i.e., SD = 2.5).
  - Run **1 000** simulations per effect size and scenario with a fixed seed (e.g., `set.seed(123)`).
- For each simulation (for each scenario):
  1. Generate y values from the **true model** for the scenario.
  2. Fit the specified model(s) to the simulated data.
  3. Test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ (two-sided).
  4. Record a rejection (1) or not (0).
- **Estimated power** = proportion of rejections over the 1 000 simulations.

Scenarios

Scenario A — Baseline Normal Errors (6 marks)

Use the code below to set up the **x** vector fixed across all effect sizes:

```r
set.seed(25)
x_vec <- runif(25, 0, 1)  # fixed design points
```

- True model:

$$Y = \beta_0 + \beta_1 X_1 + e,\ e \overset{i.i.d.}{\sim} N(\mu = 0, \sigma^2 = 6.25).$$

- Consider effect sizes $\beta_1 \in \{0.5, 1, 2, 5, 10\}$.
- For each effect size, fit the correct linear model and estimate power.

Marks:

- Implementation of simulation (4 marks).
- Tabulate results and comment briefly on the power trend across β_1 (2 marks).

Scenario B — Non-linear Truth: Exponential Relationship (7 marks)

- True model:

$$Y = \beta_0 + 0.01 \times \exp(X_1) + e, \quad e \overset{i.i.d.}{\sim} N(\mu = 0, \sigma^2 = 6.25).$$

- Consider sample sizes $n \in \{5, 10, 20, 50\}$.
- For each sample size:

  1. Generate a fixed $\mathbf{x}$ vector of size n using `seq(0, 7, length.out = n)`.

  - For example, for $n = 5$, $\mathbf{x} = (0, 1.75, 3.5, 5.25, 7)$, whereas for $n = 10$, $\mathbf{x} = (0, 0.777\ldots, 1.555\ldots, \ldots, 6.222\ldots, 7)$.

  2. Fit a **linear model** $Y = \beta_0 + \beta_1 x + e$ (mis-specified).
  3. Fit the **correct exponential model** $Y = \beta_0 + \beta_1 \exp(x) + e$.
  4. Compute and compare power for each model form.

Marks:

- Implementation for both fits (5 marks).
- Plot results and comment on them (2 marks).

Conclusion & design insight (3 marks)

- Summarise how power changes with effect size in both scenarios.
- Identify which scenario yields the **lowest** power and explain why.

Presentation (10 marks)

You start with 10 marks for presentation. Marks will be deducted for violations specified below.

Display of R Code:

- Including R code is optional but recommended for clarity.
- If you include code:
  - Do not display boilerplate (e.g., loading packages, reading data).
  - Include only code relevant to calculations or model fitting.
  - Ensure 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 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 colour choices
    - Inappropriate graph types
    - Unnecessary 3D effects
- *Tables*:
  - Missing 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.