06_homework_multiple_regression

Bill Perry

```
# Load required packages
library(patchwork)    # For combining plots
library(broom)        # For tidy statistical output
library(car)          # For regression diagnostics
```

```
Loading required package: carData
```

```
library(lmtest)       # For assumption testing
```

```
Loading required package: zoo
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
library(corrplot)     # For correlation matrices
```

```
corrplot 0.95 loaded
```

```
library(GGally)       # For pairs plots
```

```
Loading required package: ggplot2
```

```
library(tinytable)    # For simple tables
library(tidyverse)    # For data manipulation and visualization
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ lubridate 1.9.4     ✔ tibble    3.3.0
✔ purrr     1.1.0     ✔ tidyr     1.3.1
```

```
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
✖ dplyr::recode() masks car::recode()
✖ purrr::some()   masks car::some()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
# Set ggplot theme
theme_set(theme_light())
```

Assignment Overview

This homework assignment will test your understanding of **multiple regression analysis** using real lake water quality data from New Zealand. You will analyze the relationships between chlorophyll-a concentrations and nutrient levels in New Zealand lakes, building on the research findings from Abell et al. (2010) on nitrogen and phosphorus limitation in New Zealand freshwater systems. You will apply the statistical concepts from Lecture 11 to understand how multiple predictors influence water quality.

Learning Objectives

By completing this assignment, you will be able to:

1. **Understand multiple regression concepts and applications**
2. **Perform correlation analyses between multiple variables**
3. **Conduct multiple linear regression**
4. **Test statistical assumptions for multiple regression**
5. **Interpret multicollinearity and its effects**
6. **Create publication-quality figures**

Data Description

The dataset `matched_lakes_data.csv` contains water quality measurements from 111 New Zealand lakes. This dataset represents a carefully matched subset of lakes with complete measurements for all variables. Key variables include:

- `chl_a_mg_m3`: Chlorophyll-a concentration (mg/m³) - **Response variable**
- `tp_mg_m3`: Total phosphorus concentration (mg/m³) - **Predictor variable**
- `tn_mg_m3`: Total nitrogen concentration (mg/m³) - **Predictor variable**
- `lake_id`: Unique identifier for each lake

> **ℹ Important Background**
>
> This dataset builds on the work of Abell et al. (2010), who found that New Zealand lakes show different nutrient limitation patterns compared to European lakes. Chlorophyll-a is a key indicator of algal biomass and water quality in lakes. High chlorophyll concentrations often indicate eutrophication (nutrient enrichment) which can lead to algal blooms and poor water quality. In New Zealand freshwater systems, both phosphorus and nitrogen can act as limiting nutrients, making this an ideal system for studying multiple regression relationships.

Part 1: Data Loading and Initial Exploration

1.1 Load and Examine the Data

```
# Load the New Zealand lake data - abel_et_al_lakes_data.csv
```

1.2 Initial Data Cleaning and Preparation

```
# Check for missing values in key variables
```

Part 2: Descriptive Statistics and Data Exploration

2.1 Summary Statistics for Key Variables

! Question 1: Summary Statistics

Task: Calculate and interpret summary statistics for the three main variables: chlorophyll-a, total phosphorus, and total nitrogen.

Instructions: - Use appropriate R functions to calculate mean, median, standard deviation, and quartiles - Comment on the distribution shape (symmetric, skewed) based on these statistics - Identify any potential outliers or unusual values - Consider the ecological meaning of these values in the context of New Zealand lakes

```
# Calculate comprehensive summary statistics for the three main variables
```

```
# Check for skewness by comparing mean and median or make a graph
```

Interpretation:

2.2 Graphical Data Exploration

! Question 2: Data Visualization

Task: Create appropriate plots to visualize the distribution of each variable and relationships between variables.

Instructions: - Create histograms for each continuous variable - Create a pairs plot or correlation matrix plot - Comment on distributions and potential relationships - Consider whether transformations might be needed

```
# Create histograms for each continuous variable
```

```
# Create pairs plot for the three main variables
```

Interpretation:

Part 3: Correlation Analysis

3.1 Correlation Analysis Setup

! Question 3: Correlation Analysis Setup

Task: Before conducting the correlation analysis, clearly state:

1. **Type of analysis:** What type of statistical test are you using and why?
2. **Hypotheses:** State your null and alternative hypotheses for each pair
3. **Assumptions:** List the assumptions for this test
4. **Variables:** Identify which variables you're analyzing and their measurement scales
5. **Ecological predictions:** What relationships do you expect based on New Zealand lake ecology?

Type of Analysis: This is a Pearson correlation analysis to quantify the strength and direction of linear relationships between chlorophyll-a and the two nutrient variables (phosphorus and nitrogen) in New Zealand lakes. Correlation analysis is appropriate as an initial step before multiple regression to understand the bivariate relationships and assess potential multicollinearity between predictors.

Hypotheses: -

- H_0 for chl_a vs tp: $\rho = 0$ (There is no linear correlation between chlorophyll-a and total phosphorus in New Zealand lakes)
- H_1 for chl_a vs tp: $\rho \neq 0$ (There is a significant linear correlation between chlorophyll-a and total phosphorus)
- H_0 for chl_a vs tn: $\rho = 0$ (There is no linear correlation between chlorophyll-a and total nitrogen in New Zealand lakes)
- H_1 for chl_a vs tn: $\rho \neq 0$ (There is a significant linear correlation between chlorophyll-a and total nitrogen)
- H_0 for tp vs tn: $\rho = 0$ (There is no linear correlation between total phosphorus and total nitrogen)
- H_1 for tp vs tn: $\rho \neq 0$ (There is a significant linear correlation between total phosphorus and total nitrogen)

Assumptions:

1. Random sampling from the population of New Zealand lakes
2. Bivariate normality (both variables normally distributed)
3. Linear relationship between the variables
4. Independence of observations (each lake represents an independent sample)

Variables:

- Variable 1: Chlorophyll-a concentration (chl_a_mg_m3) - continuous numerical variable (mg/m^3)
- Variable 2: Total phosphorus concentration (tp_mg_m3) - continuous numerical variable (mg/m^3)
- Variable 3: Total nitrogen concentration (tn_mg_m3) - continuous numerical variable (mg/m^3)

3.2 Perform Correlation Analysis

```
# Calculate correlation matrix for the three variables
```

```
# Check normality assumptions for correlation
```

3.3 Interpret Correlation Results

! Question 4: Correlation Interpretation

Task: Interpret your correlation results by addressing:

1. **Strength and direction:** How strong are the relationships and in what direction?
2. **Statistical significance:** Which correlations are statistically significant?
3. **Ecological significance:** What do these relationships mean for New Zealand lake ecology?
4. **Multicollinearity concerns:** Are the predictors correlated with each other?

Part 4: Multiple Regression Analysis

4.1 Multiple Regression Analysis Setup

! Question 5: Multiple Regression Analysis Setup

Task: For the multiple regression of chlorophyll on phosphorus and nitrogen, clearly state:

1. **Type of analysis:** What type of statistical analysis are you using?
2. **Model equation:** Write out your multiple regression model
3. **Hypotheses:** State your null and alternative hypotheses
4. **Variables:** Clearly identify your predictors and response variables
5. **Biological rationale:** Why use multiple regression for this New Zealand lake question?

Type of Analysis: This is multiple linear regression analysis to model chlorophyll-a concentration as a function of both total phosphorus and total nitrogen concentrations simultaneously in New Zealand lakes. This approach allows us to assess the independent effects of each nutrient while controlling for the other.

Model Equation: chl_a_mg_m3 = β_0 + β_1(tp_mg_m3) + β_2(tn_mg_m3) + ε

Where:

- - chl_a_mg_m3 = chlorophyll-a concentration (response variable)
- - β_0 = intercept (baseline chlorophyll when both nutrients = 0)
- - β_1 = partial slope for phosphorus (change in chlorophyll per unit phosphorus, holding nitrogen constant)
- - β_2 = partial slope for nitrogen (change in chlorophyll per unit nitrogen, holding phosphorus constant)
- - ε = random error term

Hypotheses:

4.2 Perform Multiple Regression Analysis

```
# Fit the multiple linear model (untransformed)
```

```
# Create ANOVA table
```

```
# Check for multicollinearity using VIF
# vif_mr_nz_model <- vif(multiple_reg_nz_model)
# vif_mr_nz_model
```

```
# # Get confidence intervals for parameters
# confint_mr_nz_model <- confint(multiple_reg_nz_model)
# confint_mr_nz_model
```

Interpretation of Model Output:

- **Overall Model:**
- **R-squared:**
- **Individual Predictors and significance:**
  - - Total phosphorus: $\beta_1 = 0.321$ ($t = 49.0$, $p < 0.001$)
  - - Total nitrogen: $\beta_2 =$
- **Multicollinearity:** VIF values of 4.1 for both predictors indicate moderate multicollinearity (VIF > 3 but < 10), which is expected given the strong correlation between nutrients but not severe enough to compromise the analysis.

4.3 Test Regression Assumptions

```
# # Create diagnostic plots
# par(mfrow = c(2, 2))
# plot(multiple_reg_nz_model)
# par(mfrow = c(1, 1))
```

```
# Formal tests for assumptions
# Shapiro-Wilk test for normality of residuals
```

4.4 Transformation of data??

> **i** Data Transformation
>
> Given the assumption violations and the power-law nature of ecological relationships, we'll apply log transformation to all variables to improve normality and homoscedasticity.

```
# Create log-transformed variables
```

```
# Create diagnostic plots for log-transformed model
```

```
# Formal tests for log model assumptions
```

4.5 Model Comparison

Compare models using AIC

```
# AIC(multiple_reg_nz_model)
# AIC(log_multiple_reg_nz_model)
```

4.6 Interpret Final Multiple Regression Results

! Question 6: Multiple Regression Interpretation

Task: Provide a complete interpretation of your final (log-transformed) multiple regression results:

1. **Overall model:** Is the overall model statistically significant?
2. **Individual predictors:** Which predictors are significant and what are their effects?
3. **Parameter estimates:** What are the slope estimates and their meanings?
4. **Model fit:** How much variance is explained (R^2)?
5. **Assumptions:** Were the assumptions met? Any concerns?
6. **Multicollinearity:** Is multicollinearity a problem?
7. **Ecological interpretation:** What does this mean for New Zealand lake management?

Final Model Interpretation:

Overall Model:

Model Equation: log(chlorophyll) = ???? + ??? × log(phosphorus) + ??? × log(nitrogen)

Individual Predictors:

- - Log total phosphorus:
- - Log total nitrogen:

Parameter Interpretation: The coefficients represent elasticities: a 1% increase in phosphorus leads to a ????% increase in chlorophyll, while a 1% increase in nitrogen leads to a ???% increase in chlorophyll, holding the other nutrient constant.

what does the above mean?

Model Fit: The model explains ???% of variance in log-chlorophyll

Ecological Interpretation: Phosphorus has approximately ??? times stronger effect than nitrogen, but both nutrients are highly significant predictors. And what should be controlled

---

Part 6: Publication-Quality Figure and Write-up

6.1 Create Publication-Quality Figure

! Question 8: Publication Figure

Task: Create a publication-quality figure that effectively displays your multiple regression results.

Requirements: - Show the relationship between chlorophyll and both predictors - Consider using multiple panels or 3D visualization - Include model information and fit statistics - Use appropriate colors and themes - Include informative title and axis labels - Consider the target audience (New Zealand water quality managers)

Submission Guidelines

What to turn in -

1. a quarto markdown file and dataframe. Note that your code should be able to run with what you turn in.
2. a self-contained word and html file showing the code and output
3. annotations in the quarto file that shows or tells what is being done in the r code chunks describing what you are trying to do - credit will be given even if it does not work as long as you detail what you are doing. As we start to move into more statistics you will be expected to interpret the results.

Points

- summary stats - 1 points
- exploratory graphs - 1 point
- Hypotheses - 2 points
- Assumptions - 2 points
- Model and Test Results - 4
- interpretation - 3 points
- Assumption Tests - 4 points
- Final figure - 3 points