ST 352 | Lab Assignment 5 - Guide

Brian Cervantes Alvarez

2024-11-01

Honor Code Reminder:
Complete lab assignments individually!

Objective

In this lab, you'll explore interaction terms in regression models and apply model selection using the `brainhead` dataset. You'll examine gender-based differences, test the significance of interaction terms, and assess multiple regression conditions through transformations.

Part I: Interaction Term and Gender Differences (12 points)

Problem 1: Scatterplot with Regression Lines by Gender (2 points)

Create a scatterplot that includes:

- Different symbols for males and females
- Separate regression lines for each gender
- A legend explaining each symbol and line

Steps

1. Use `plot()` and `points()` to create scatterplots for males and females.
2. Use `abline()` to add regression lines for each gender.

Code Example

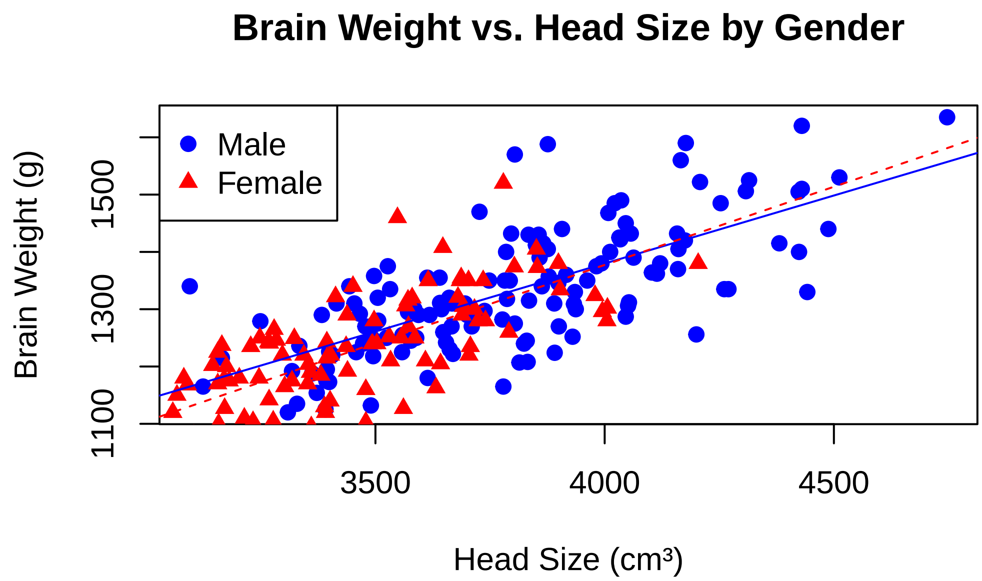
```
# Read brainhead data
brainheadData <- read.table("https://raw.githubusercontent.com/bcervantesalvarez/MS-Statisti

# Separate data by gender
maleData <- subset(brainheadData, gender == "male")
femaleData <- subset(brainheadData, gender == "female")

# Plot
plot(maleData$headsize, maleData$brainwt,
     pch = 19, col = "blue",
     xlab = "Head Size (cm³)",
     ylab = "Brain Weight (g)",
     main = "Brain Weight vs. Head Size by Gender")
points(femaleData$headsize, femaleData$brainwt, pch = 17, col = "red")

# Fit separate regression lines
maleModel <- lm(brainwt ~ headsize, data = maleData)
femaleModel <- lm(brainwt ~ headsize, data = femaleData)

abline(maleModel, col = "blue", lty = 1)
abline(femaleModel, col = "red", lty = 2)
legend("topleft",
       legend = c("Male", "Female"),
       pch = c(19, 17),
       col = c("blue", "red"))
```



Problem 2: Need for an Interaction Term? (2 points)

Using the scatterplot, discuss if an interaction term is necessary.

- **Guiding Questions**: Are the regression lines parallel, or do they have different slopes? If they're similar, an interaction term may not add much value.

Problem 3: Hypothesis Test for Interaction Term (4 points)

Perform a hypothesis test to evaluate the significance of the interaction term between `headsize` and `gender`.

- **a.** State null and alternative hypotheses.
  - **Guiding Questions**: What does it mean for the interaction term to be significant? Does `headsize` affect `brainwt` differently for each gender?
- **b.** Report t-statistic, degrees of freedom, and p-value.

Code Example

```r
# Fit model with interaction
interactionModel <- lm(brainwt ~ headsize * gender, data = brainheadData)
summary(interactionModel)
```

```
Call:
lm(formula = brainwt ~ headsize * gender, data = brainheadData)

Residuals:
     Min       1Q   Median       3Q      Max 
-171.215  -47.721    0.182   47.768  237.690 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         286.08702   83.10127   3.443 0.000683 ***
headsize              0.27280    0.02421  11.269  < 2e-16 ***
gendermale          144.21567  110.44279   1.306 0.192910    
headsize:gendermale  -0.03544    0.03082  -1.150 0.251403    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.13 on 233 degrees of freedom
Multiple R-squared:  0.6453,    Adjusted R-squared:  0.6408 
F-statistic: 141.3 on 3 and 233 DF,  p-value: < 2.2e-16
```

- **c.** Draw a conclusion based on the p-value.
  - **Prompt**: Is the p-value for the interaction term low enough (e.g., below 0.05) to suggest that head size's effect on brain weight differs by gender?

Problem 4: Least-Squares Regression Equation (2 points)

Write the regression equation, including the interaction term.

- **Guiding Questions**: How do we interpret each term? What do `headsize` and `gender` contribute to the equation?

Problem 5: Coefficient Interpretation (2 points)

Interpret the coefficient of the interaction term in the context of gender differences in brain weight.

- **Guiding Questions**: Does the interaction term suggest that head size's impact on brain weight changes depending on gender? How much?

Part II: Multiple Linear Regression Conditions Using the `fish.txt` Dataset (12 points)

Problem 6: Correlation and Scatterplot Matrix (3 points)

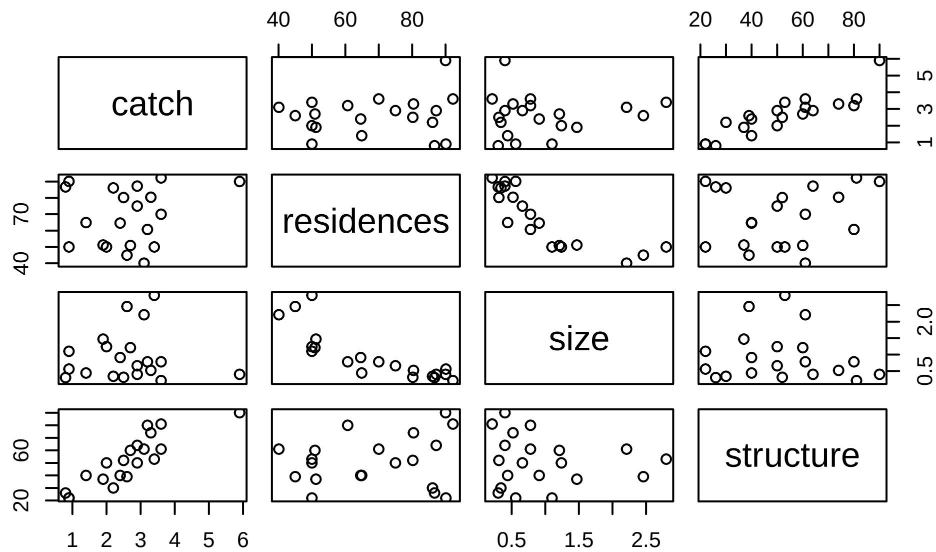
Construct a scatterplot matrix and correlation matrix for the variables in `fish.txt`:

- `catch`: Seasonal bass catch (thousands)
- `residences`: Number of lakeshore residences per square mile
- `size`: Lake size (square miles)
- `structure`: Structure index (measure of lakebed structure)

Code Example

```r
# Load fish data
fishData <- read.table("https://raw.githubusercontent.com/bcervantesalvarez/MS-Statistics/re

# Scatterplot matrix
pairs(fishData[, c("catch", "residences", "size", "structure")])
```




```r
# Correlation matrix
cor(fishData[, c("catch", "residences", "size", "structure")])
```

```
               catch residences       size  structure
catch      1.0000000  0.1491201  0.0428054  0.8753489
residences 0.1491201  1.0000000 -0.8286589  0.1639370
size       0.0428054 -0.8286589  1.0000000 -0.1142250
structure  0.8753489  0.1639370 -0.1142250  1.0000000
```

- **Guiding Questions**: Are there any variables with high correlations? If so, which pairs? Could these correlations suggest multicollinearity in a regression model?

Problem 7: Model Comparison (3 points)

Compare different models using transformations of `catch`, `residences`, `size`, and `structure` to determine which model best satisfies linearity, constant variance, and normality conditions. Create four models to compare:

1. Original scale
2. Log-transformed response variable (`catch`)
3. Log-transformed predictor variables (`residences`, `size`, `structure`)
4. Log-transformation of all variables (except `access`)

Code Example

```r
# Log transformations
fishData$logCatch <- log(fishData$catch)
fishData$logResidences <- log(fishData$residences)
fishData$logSize <- log(fishData$size)
fishData$logStructure <- log(fishData$structure)

# Models
originalModel <- lm(catch ~ residences + size + structure + access, data = fishData)
logCatchModel <- lm(logCatch ~ residences + size + structure + access, data = fishData)
logPredictorsModel <- lm(catch ~ logResidences + logSize + logStructure + access, data = fisl
logAllModel <- lm(logCatch ~ logResidences + logSize + logStructure + access, data = fishData

# Summaries for comparison
summary(originalModel)
```

```
Call:
lm(formula = catch ~ residences + size + structure + access,
    data = fishData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85859 -0.14400 -0.04054  0.21234  0.72653

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.784001   0.815706  -3.413  0.00385 **
residences   0.026794   0.009141   2.931  0.01032 *
size         0.503508   0.220767   2.281  0.03760 *
structure    0.051129   0.004542  11.258 1.03e-08 ***
access       0.742933   0.202128   3.676  0.00225 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3895 on 15 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8906
F-statistic: 39.65 on 4 and 15 DF,  p-value: 8.296e-08
```

```r
summary(logCatchModel)
```

```

Call:
lm(formula = logCatch ~ residences + size + structure + access, 
    data = fishData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31029 -0.12029  0.00497  0.14028  0.31527 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.130217   0.446701  -2.530   0.0231 *  
residences   0.006585   0.005006   1.315   0.2081    
size         0.194501   0.120898   1.609   0.1285    
structure    0.022620   0.002487   9.095 1.72e-07 ***
access       0.291726   0.110690   2.636   0.0187 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2133 on 15 degrees of freedom
Multiple R-squared:  0.8663,	Adjusted R-squared:  0.8306 
F-statistic: 24.29 on 4 and 15 DF,  p-value: 2.096e-06
```

```r
summary(logPredictorsModel)
```

```

Call:
lm(formula = catch ~ logResidences + logSize + logStructure + 
    access, data = fishData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90022 -0.15491 -0.02307  0.16859  1.12108 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -16.5388     4.1208  -4.013  0.00113 **
logResidences   2.3007     0.9901   2.324  0.03460 * 
logSize         0.5740     0.3884   1.478  0.16013   
```

```
logStructure    2.4000     0.2519   9.527 9.42e-08 ***
access          0.6913     0.2568   2.691  0.01674 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4615 on 15 degrees of freedom
Multiple R-squared:  0.8787,    Adjusted R-squared:  0.8463
F-statistic: 27.16 on 4 and 15 DF,  p-value: 1.023e-06
```

```
summary(logAllModel)
```

```
Call:
lm(formula = logCatch ~ logResidences + logSize + logStructure +
    access, data = fishData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30922 -0.05861 -0.00630  0.06515  0.33572

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.35568    1.54448  -4.115 0.000918 ***
logResidences  0.67181    0.37110   1.810 0.090324 .
logSize        0.22036    0.14557   1.514 0.150856
logStructure   1.11068    0.09442  11.763 5.67e-09 ***
access         0.27524    0.09626   2.859 0.011941 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.173 on 15 degrees of freedom
Multiple R-squared:  0.9121,    Adjusted R-squared:  0.8886
F-statistic: 38.89 on 4 and 15 DF,  p-value: 9.465e-08
```

- **Guiding Questions**: Compare the residual plots, constant variance, and normality of each model's residuals. Which model best meets these assumptions? Explain why this model is preferable.

Problem 8: Least-Squares Regression Equation for Chosen Model (2 points)

Write the least-squares regression equation for the model that best satisfies the assumptions, defining each term.

- **Guiding Questions**: For your chosen model, what are the predictors (`residences`, `size`, `structure`, `access`)? What does each term in the equation represent?

$$\log(\text{Catch}) = -6.356 + 0.672 \cdot \log(X_1) + 0.220 \cdot \log(X_2) + 1.111 \cdot \log(X_3) + 0.275 \cdot X_4$$

Here's a breakdown of the equation components:

- **Intercept**: -6.356
- $log(X_1)$ = **Effect of log(Residences)**: 0.672
- $log(X_2)$ = **Effect of log(Size)**: 0.220
- $log(X_3)$ = **Effect of log(Structure)**: 1.111
- X_4 = **Effect of Access**: 0.275

This model indicates that each predictor has a multiplicative effect on the catch when transformed logarithmically, with `Structure` and `Access` showing the most substantial contributions to `Catch` in this log-transformed model.

Problem 9: Predicting Seasonal Bass Catch (2 points)

Use your final model to predict the seasonal bass catch for a lake with:

- 1.5 square miles in size
- 75 lakeshore residences per square mile
- Structure index of 50
- Public access (represented by `1`)

Code Example

```r
# Model from model selection
finalModel <- lm(formula = logCatch ~ logResidences + logSize + logStructure + access, data =
summary(finalModel)
```

```
Call:
lm(formula = logCatch ~ logResidences + logSize + logStructure + 
    access, data = fishData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30922 -0.05861 -0.00630  0.06515  0.33572 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.35568    1.54448  -4.115 0.000918 ***
logResidences  0.67181    0.37110   1.810 0.090324 .
logSize        0.22036    0.14557   1.514 0.150856
logStructure   1.11068    0.09442  11.763 5.67e-09 ***
access         0.27524    0.09626   2.859 0.011941 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.173 on 15 degrees of freedom
Multiple R-squared:  0.9121,    Adjusted R-squared:  0.8886
F-statistic: 38.89 on 4 and 15 DF,  p-value: 9.465e-08
```

```r
# Define new data point with transformed predictors
newLake <- data.frame(
  logResidences = log(75),
  logSize = log(1.5),
  logStructure = log(50),
  access = 1
)

# Predict using the chosen model
predictedLogCatch <- predict(finalModel, newdata = newLake)

# Convert the log prediction back to the original scale if needed
predictedCatch <- exp(predictedLogCatch)
predictedCatch
```

```
       1
3.505907
```

- **Guiding Questions**: Interpret the prediction in the context of the model. If a log transformation was used, remember to back-transform the predicted value.

Problem 10: Interpreting the Coefficient of `access` (2 points)

Interpret the coefficient of `access` in your final model, explaining its impact on predicted seasonal bass catch.

- **Guiding Questions**: How does public access (1 vs. 0) affect predicted bass catch? What does the sign of the `access` coefficient indicate about lakes with public access compared to those without?
