ST 352 | Lab Assignment 4 - Guide

Multiple Regression Analysis and Model Selection

Brian Cervantes Alvarez

2024-11-01

Reminder of the honor code:
Lab assignments are to be completed individually!

Objective

In this lab assignment, you will perform multiple regression analysis and model selection using the `bodyfat` dataset. You will explore the relationships between various explanatory variables and the response variable (percent body fat), assess regression assumptions, conduct hypothesis tests, and apply model selection techniques to build an optimal predictive model.

Data Description

The `bodyfat` dataset contains the following variables:

- `fat`: Percent body fat
- `age`: Age in years
- `weight`: Weight in pounds
- `height`: Height in inches
- `chest`: Chest circumference in centimeters

Problem 1: Identifying Highly Correlated Explanatory Variables

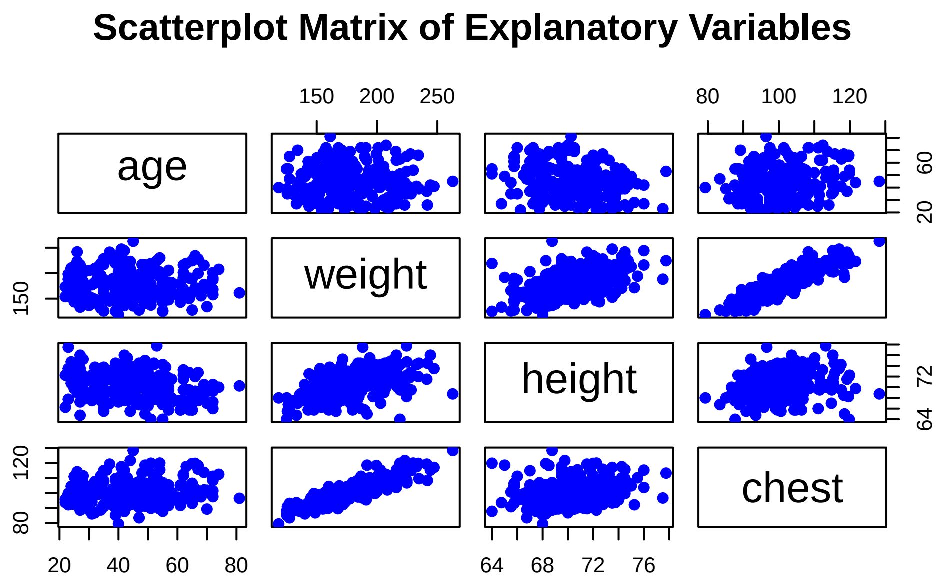
Based on both the scatterplot matrix of the explanatory variables and the correlation matrix, which two explanatory variables are "highly correlated"? Explain. You must use and refer to both the scatterplot matrix and the correlation matrix in your explanation of why these two variables are considered highly correlated. (Include the scatterplot matrix here. You do not have to include the correlation matrix.)

Note: We exclude explanatory variables from the correlation matrix to focus on relationships among predictors and avoid redundancy. Including them can lead to multicollinearity, which complicates interpreting each predictor's unique effect on the outcome.

```r
# Read the bodyfat data
bodyfatData <- read.table("https://raw.githubusercontent.com/bcervantesalvarez/MS-Statistics/

# Select explanatory variables (age, weight, height, chest)
explanatoryVars <- bodyfatData[c("age", "weight", "height", "chest")]

# Create scatterplot matrix
pairs(explanatoryVars,
      main = "Scatterplot Matrix of Explanatory Variables",
      pch = 19,
      col = "blue")
```



Problem 2: Variable Selection Based on Correlation

In the Lab 4 Notes, a strategy was shown to help with deciding which highly correlated explanatory variable to remove. From that strategy, which of the two highly correlated explanatory variables mentioned in the lab notes would be removed? Explain why that variable would be removed.

Here's the strategy!

```r
# Fit a model with weight & chest since they are highly correlated
model <- lm(fat ~ weight + chest, data = bodyfatData)
summary(model)
```

```
Call:
lm(formula = fat ~ weight + chest, data = bodyfatData)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.9701  -4.1056  -0.2213   3.7342  14.6604 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -53.99451    5.93133  -9.103  < 2e-16 ***
weight       -0.01074    0.03071  -0.350    0.727    
chest         0.74446    0.10182   7.312 3.66e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.94 on 247 degrees of freedom
Multiple R-squared:  0.4912,    Adjusted R-squared:  0.4871 
F-statistic: 119.2 on 2 and 247 DF,  p-value: < 2.2e-16
```

Which predictor is significant? Ideally, we should remove any insignificant predictors—though this doesn't always apply in every dataset. Here, it's safe to remove the weight parameter since our goal is to explain body fat percentage, and weight is already a contributing factor.

Problem 3: Assessing Regression Assumptions

Include the graphical displays to assess the linearity, constant variation, and normality conditions (scatterplot matrix that includes the response variable, residual plot, and normal probability plot of the residuals).

```r
# Select response and remaining explanatory variables after removing weight
responseVar <- bodyfatData$fat
selectedExplanatory <- bodyfatData[c("age", "height", "chest")]

# Fit the initial multiple regression model
initialModel <- lm(fat ~ age + height + chest, data = bodyfatData)
summary(initialModel)
```

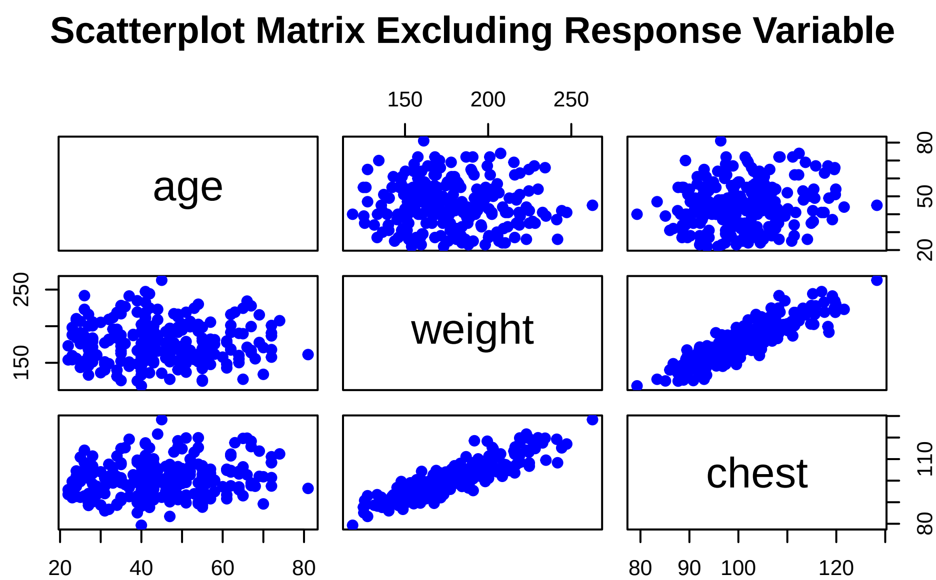
```
Call:
lm(formula = fat ~ age + height + chest, data = bodyfatData)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.2711  -4.3709  -0.4543   3.6338  14.4474 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.56283   10.43100  -2.163 0.031501 *  
age           0.08304    0.03016   2.753 0.006338 ** 
height       -0.49966    0.14713  -3.396 0.000797 ***
chest         0.72514    0.04654  15.581  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.65 on 246 degrees of freedom
Multiple R-squared:  0.5415,	Adjusted R-squared:  0.5359 
F-statistic: 96.85 on 3 and 246 DF,  p-value: < 2.2e-16
```

```r
# Scatterplot matrix excluding the response variable
pairs(bodyfatData[c("age", "weight", "chest")],
      main = "Scatterplot Matrix Excluding Response Variable",
      pch = 19,
      col = "blue")
```




```
# Residual plot


# Normal probability plot of residuals
```

Problem 4: Assessing Linearity

Using the appropriate graph, discuss whether or not the linearity condition is satisfied. Make sure to reference which plot you are using to assess this condition.

Problem 5: Assessing Constant Variation

Using the appropriate graph, discuss whether or not the constant variation condition is satisfied. Make sure to reference which plot you are using to assess this condition.

Problem 6: Assessing Normality

Using the appropriate graph, discuss whether or not the normality condition is satisfied. Make sure to reference which plot you are using to assess this condition.

Problem 7: F-test for Overall Significance

Perform an F-test to determine if at least one explanatory variable is helpful in predicting the response.

a.

State the null and alternative hypotheses in words.

- **Null Hypothesis** H_0: None of the explanatory variables (`age`, `height`, `chest`) are associated with the response variable (`fat`). In other words, all regression coefficients are equal to zero.

  All the $\beta_1 = \beta_2 = \beta_3 = 0$

- **Alternative Hypothesis** H_a: At least one of the explanatory variables (`age`, `height`, `chest`) is associated with the response variable (`fat`). In other words, at least one regression coefficient is not equal to zero.

  At least one $\beta_1, \beta_2, \beta_3 \neq 0$

b.

Using the regression output, report the F-statistic with degrees of freedom and the p-value. (No calculations are necessary! Include the regression output from R – please put the output below your answer to this question.)

```r
# Perform the F-test by summarizing the initial model
# Initial model
initialModel <- lm(fat ~ age + height + chest, data = bodyfatData)
summary(initialModel)
```

```

Call:
lm(formula = fat ~ age + height + chest, data = bodyfatData)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.2711  -4.3709  -0.4543   3.6338  14.4474 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -22.56283   10.43100  -2.163 0.031501 *
age           0.08304    0.03016   2.753 0.006338 **
height       -0.49966    0.14713  -3.396 0.000797 ***
chest         0.72514    0.04654  15.581  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.65 on 246 degrees of freedom
Multiple R-squared:  0.5415,	Adjusted R-squared:  0.5359
F-statistic: 96.85 on 3 and 246 DF,  p-value: < 2.2e-16
```

c.

State a conclusion from the F-test in the context of the problem.

Problem 8: T-tests for Individual Explanatory Variables

For each explanatory variable,

a.

Report the t-statistic with degrees of freedom and the p-value for the t-test (testing if that explanatory variable explains the response variable after accounting for the effects of the other variables).

b.

State a conclusion in the context of the problem based on the p-value from the t-test. (You should have three separate sentences, one for the conclusion for each explanatory variable. Much of each sentence may contain the same wording, but you still need to write three separate sentences with three separate conclusions.)

Problem 9: Backwards Selection Process

Suppose a backwards selection process was performed. Would any of the explanatory variables drop out? Why or why not? If so, which one would drop out first? Why?

Problem 10: Final Model After Backwards Selection

If needed, perform a backwards selection process to obtain a model with only significant explanatory variables. Use the model with only significant explanatory variables to answer these questions:

This is not needed in our case. Why? Well, I ran it to demonstrate that it's going to keep the same model as before!

```
# Perform backwards selection using step from base R
# Initial model
initialModel <- lm(fat ~ age + height + chest, data = bodyfatData)
# Perform backwards selection
backwardModel <- step(initialModel, direction = "backward", trace = FALSE)
# Summary of the final model after backward selection
summary(backwardModel)
```

```
Call:
lm(formula = fat ~ age + height + chest, data = bodyfatData)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.2711  -4.3709  -0.4543   3.6338  14.4474 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.56283   10.43100  -2.163 0.031501 *  
age           0.08304    0.03016   2.753 0.006338 ** 
height       -0.49966    0.14713  -3.396 0.000797 ***
chest         0.72514    0.04654  15.581  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.65 on 246 degrees of freedom
Multiple R-squared:  0.5415,    Adjusted R-squared:  0.5359 
F-statistic: 96.85 on 3 and 246 DF,  p-value: < 2.2e-16
```

a.

Include the regression output of your final model.

```r
# Summary of the final model after backward selection
summary(backwardModel)
```

```
Call:
lm(formula = fat ~ age + height + chest, data = bodyfatData)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.2711  -4.3709  -0.4543   3.6338  14.4474 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.56283   10.43100  -2.163 0.031501 *  
age           0.08304    0.03016   2.753 0.006338 ** 
height       -0.49966    0.14713  -3.396 0.000797 ***
chest         0.72514    0.04654  15.581  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.65 on 246 degrees of freedom
Multiple R-squared:  0.5415,	Adjusted R-squared:  0.5359 
F-statistic: 96.85 on 3 and 246 DF,  p-value: < 2.2e-16
```

Final Model Regression Output:

b.

Write the least-squares regression equation. Define the terms in the equation (i.e., what the x's and $\hat{y}$ represent in the context of the problem).

The least-squares regression equation outline:

$$fat_i = \beta_0 + \beta_1 \times \text{age}_i + \beta_2 \times \text{height}_i + \beta_3 \times \text{chest}_i + \epsilon_i$$

Where:

- fat is the predicted percent body fat.
- **age** is the age of the individual in years.
- **height** is the height of the individual in inches.
- **chest** is the chest circumference in centimeters.

c.

Interpret the coefficient of Age in the context of the problem.

d.

Predict percent body fat for a 23-year-old who is 73 inches tall, and has a chest circumference of 120 cm. Use R to find this predicted value.

```
# Define the new data point
newData <- data.frame(
  age = 23,
  height = 73,    # Height in inches
  chest = 120      # Chest circumference in centimeters
)
# Predict percent body fat
predictedBodyFat <- predict(backwardModel, newdata = newData)
predictedBodyFat
```

```
       1
29.88957
```

Predicted Percent Body Fat:

e.

Report and interpret a 95% prediction interval for the person in question 10d.

```
# Calculate the 95% prediction interval
predictionInterval <- predict(backwardModel, newdata = newData,
                              interval = "prediction", level = 0.95)
predictionInterval
```

```
       fit      lwr      upr
1 29.88957 18.50715 41.27199
```