# CHAPTER 3

# Multiple Regression

## 3.5 Correlated Predictors

## 3.6 Testing Subsets of Predictors

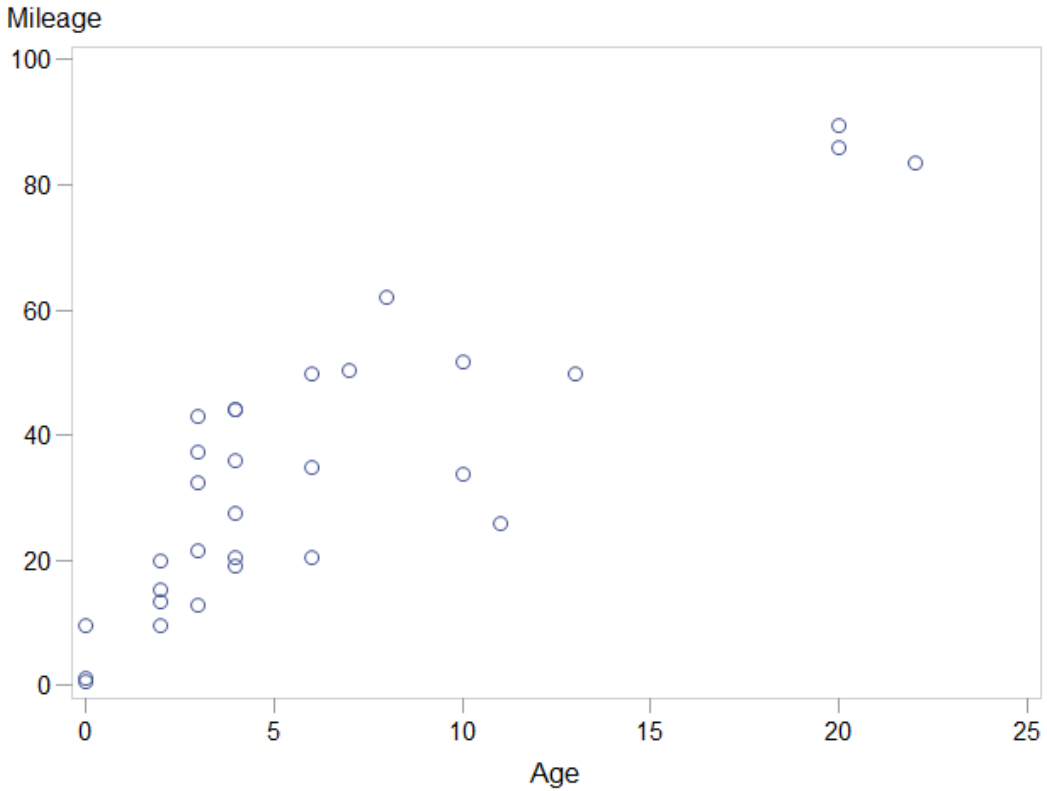
# 3.5 Correlated Predictors:

When fitting multiple regression model, we often encounter predictor variables that are correlated with one another. We should not surprised because this is not a necessarily a bad thing, but it can lead to difficulty in fitting and interpreting the mode.

# Example 1: *(Porsche prices)*

For dataset *Porscheprices.csv*.

1. Graph the scatterplot of the age vs. the mileage?
2. Calculate the correlation coefficient for part (1)?

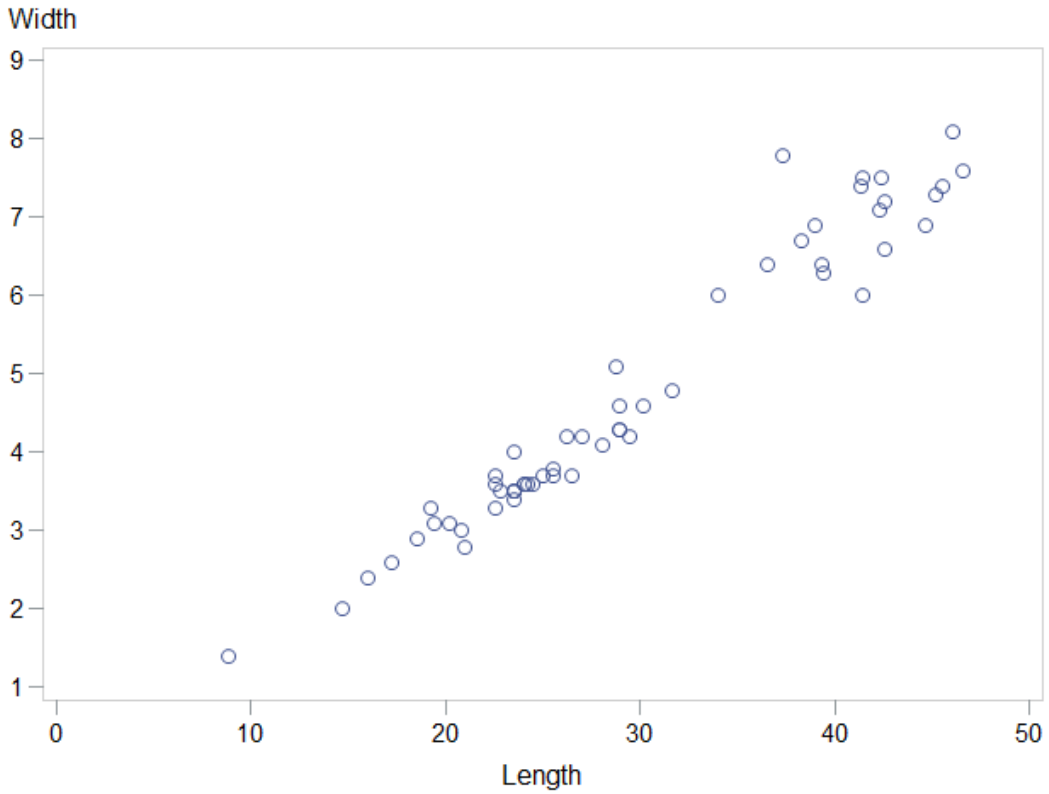



| Pearson Correlation Coefficients, N = 30<br>Prob > \|r\| under H0: Rho=0 | |
|---|---|
| | Age |
| Mileage | 0.86313 |
| | <.0001 |

# Example 2: *(Perch weight)*

For dataset *Perch.csv*.

1. Graph the scatterplot of the length vs. the width?
2. Calculate the correlation coefficient for part (1)?




| Pearson Correlation Coefficients, N = 56<br>Prob > \|r\| under H0: Rho=0 | |
|---|---|
| | Length |
| Width | 0.97511 |
| | <.0001 |

# Multicollinearity:

We say that a set of predictors exhibits **multicollinearity** when one or more of the predictors is strongly correlated with some combination of the other predictors in the set.

# Detecting Multicollinearity:

How do we know when **multicollinearity** might be an issue with a set of predictors?

- The quick check is to examine the pairwise correlation between the predictors.
- The common procedure is **Variance Inflation Factor (VIF)**.

# Variance Inflation Factor (VIF):

For any predictor $X_i$ in a model, the **Variance Inflation Factor (VIF)** is computed as

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where $R_i^2$ is the coefficient of multiple determination for a model to predict $X_i$ using other predictors in the model.

As rough rule, we suspect multicolinearity with predictors for which the $VIF > 5$, which is equivalent to $R_i^2 > 80\%$.

# Example 3: *(Porsche prices)*

For dataset *Porscheprices.csv*. Run multiple regression analysis and find the VIF for each predictor. Check the multicollinearity.

```
proc reg data = PorschePrice;
model price = mileage age / vif;
run;
```

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
|---|---|---|---|---|---|---|
| Intercept | 1 | 70.91916 | 2.48352 | 28.56 | <.0001 | 0 |
| Mileage | 1 | -0.56134 | 0.11407 | -4.92 | <.0001 | 3.92150 |
| Age | 1 | -0.13023 | 0.45684 | -0.29 | 0.7778 | 3.92150 |

The VIF values for the age and the mileage are less than 5, so which indicates that the age is not strongly related to the mileage.

# Example 4: *(Perch weight)*

For dataset *Perch.csv*. Check the multicollinearity.

```
proc reg data = Perch;
model Weight = Length Width / VIF;
run;
```

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | 1 | -578.75777 | 43.66725 | -13.25 | <.0001 | 0 |
| Length | 1 | 14.30738 | 5.65880 | 2.53 | 0.0145 | 20.33948 |
| Width | 1 | 113.49966 | 30.26474 | 3.75 | 0.0004 | 20.33948 |

The **VIF** values for the Length and Age are greater than 5, so which indicates that there a strong relationship between the Length and Width.

# What should we do if we detect multicollinearity in a set of predictors?

1. Drop some predictors.
2. Combine some predictors. (create a new variable with some formula)
3. Discount the individual coefficients and t-tests. (read pp. 134)

# 3.6 Testing Subset of Predictors:

- The F-test (ANOVA) tests the effectiveness of all of the predictors in the model as a group.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$vs \qquad H_1: at\ least\ one\ \beta_i \neq 0$$

- The individual t-tests for regression predictors allow us to check the importance of terms in the model one at time.

$$H_0: \beta_i = 0 \quad vs \quad H_1: \beta_i \neq 0$$

- Could we test only a subset of the predictors?

# Nested F-test:

To test a subset of predictors in a multiple regression model,

$$H_0: \beta_i = 0 \ \ (for\ all\ predictors\ in\ the\ subset)$$

$$H_1: \beta_i \neq 0\ (at\ least\ one\ predictor\ in\ the\ subset)$$

Let the full model denote one with all $k$ predictors and the reduced model be the nested model obtained by dropping the predictors that are being tested.

The test statistics is

$$F = \frac{(SSModel_{full} - SSModel_{reduced})/\#\ predictors\ tested}{SSE_{full}/(n-k-1)}$$

The p-value is computed from an F-distribution with numerator degrees of freedom equal to the number of predictors being tested and denominator degrees of freedom equal to the error degrees of freedom for the full model.

# Nested Model:

Nested means that one model is a subset of another.

## Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2 + \beta_3 x_1 x_2 + \epsilon \quad (1)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon \quad (2)$$

Model 1 is nested in (within) Model 2 because model 2 contains all parameters in model 1.

# Example 5: *(Perch weight)*

In lab 7, the complete second-order model (full model) was:

$$\widehat{Weight} = 156.3 - 25.0\,Length + 20.9\,Width + 1.6\,Length^2 + 34.4\,Width^2 - 9.8\,Length.Width$$

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 156.34862 | 61.41521 | 2.55 | 0.0140 |
| Length | 1 | -25.00073 | 14.27289 | -1.75 | 0.0860 |
| Width | 1 | 20.97715 | 82.58767 | 0.25 | 0.8005 |
| Length2 | 1 | 1.57194 | 0.72439 | 2.17 | 0.0348 |
| Width2 | 1 | 34.40585 | 18.74550 | 1.84 | 0.0724 |
| LengthXWidth | 1 | -9.77630 | 7.14548 | -1.37 | 0.1774 |

# Example 5: *(Perch weight)*

Let's test whether adding the two quadratic terms ($Length^2$ and $Width^2$) actual provides a substantial improvement over the interaction model.

We need to test this hypothesis

$$H_0: \beta_3 = \beta_4 = 0 \quad vs \quad H_1: \beta_3 \neq 0 \ or \ \beta_4 \neq 0$$

```
proc reg data = Perch1;
model Weight = Length Width L2 W2 LW;
test L2, W2;
run;
```

Test 1 Results for Dependent Variable Weight

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 2 | 4382.28185 | 2.36 | 0.1052 |
| Denominator | 50 | 1860.00201 | | |

Since $p - value > 0.05$, so we fail reject $H_0$ which means that the quadratic terms are not helpful to this model.

# Reading Assignment

Read section 3.5, 3.6, and 3.8