Multiple Regression Part 1

STAT 441/541 Statistical Methods II

Objectives of Part 1

- Decide on the dependent variable and independent variables
- Decide on the form of the multiple regression model
- State the multiple regression model and describe all terms in the model
- State the assumptions of multiple regression analysis
- Interpret scatterplot and correlation matrices
- Interpret estimates of partial slope parameters of the multiple regression model
- Use the five-step method for the overall F test of a multiple regression model
- Use the five-step method to test parameters of the multiple regression model

Sections Covered for Part 1

Special Topic: Scatterplot and Correlation Matrices

Chapter 13

- 13.3 Formulating the Model

Chapter 12

- 12.1 Introduction
- 12.3 Estimating Multiple Regression Coefficients
- 12.4 Inferences in Multiple Regression

Chapter 11

- 11.1 Introduction and Abstract of Research Study
  - Transformations

Scatterplot Matrix

- A set of scatterplots for each pair of variables
- Used to examine the relationship between each pair of variables
  - Linear
  - Nonlinear
  - Outliers
- Examine relationship between dependent variable and all independent variables
- Examine relationship between each pair of independent variables

Correlation Coefficients

- A Pearson correlation coefficient for each pair of variables
- Use *p*-values to identify significant correlations
- Examine correlations between dependent variable and all independent variables
- Examine correlations between each pair of independent variables
- Used along with scatterplots to help build a multiple regression model

Section 13.3 Formulating the Model

- First, decide on the dependent variable and candidate independent variables for the regression equation
- The initial selection of independent variables is critical in constructing a multiple regression model
- Knowledge of the subject matter is important for selecting independent variables and the form of each independent variable
- Transformations of the data may be needed to meet the assumptions of multiple regression
  - Logarithmic transformation of the dependent and/or independent variables
  - Inverse transformation of the dependent variable ($1/y$)

Section 12.1 Introduction

- The multiple regression model relates a dependent variable to a set of k independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + + \beta_k x_k + \varepsilon$$

- The only restriction is that no independent variable is a perfect linear function of any other independent variables
- The parameter β_0 is the y-intercept and is the expected value of y when each $x_i = 0$. Only meaningful if it makes sense to have each $x_i = 0$
- The other parameters $(\beta_1, \cdots \beta_k)$ are partial slope parameters and represent the expected change in y for a unit increase in x_i when all other $x_j's$ are held constant.

Note: Expected value is the same as average value

Examples of Multiple Regression Models

- First-order model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

- Model with an interaction term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- Polynomial Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \varepsilon$$

Assumptions for Multiple Regression

- The model has been properly specified
- The variance of the errors is σ_{ε}^{2} for all observations
- The errors are independent
- The error terms are normally distributed and there are no outliers

Some Limitations of Regression Analysis

- The existence of a relationship does not imply that changes in the independent variables cause changes in the dependent variable (cause and effect)
- Do not use an estimated regression equation for extrapolation outside the range of values for all independent variables

Section 12.3 Estimating Multiple Regression Coefficients

- The multiple regression model relates a dependent variable to a set of quantitative independent variables.
- For a random sample of n measurements, the ith observation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + + \beta_k x_{ik} + \varepsilon_i$$
$$\text{for } i = 1, 2, \ldots, n; \text{ and } n > k,$$

  where n=number of observations and

  k=number of partial slope parameters in the model for the independent variables

- The method of least-squares is used to estimate all coefficients in the model $\beta_0, \beta_1, \ldots, \beta_k$
- Each coefficient refers to the effect of changing an independent variable while all other independent variables stay constant

Model Standard Deviation

- It is important to estimate the model standard deviation σ_ε or variance σ_ε^2
- Residuals, e_i, are used to estimate σ_ε
- The estimate of the model standard deviation is denoted as s_ε and is given as "Residual standard error" in the R output summary

Residual: $e_i = y_i - \hat{y}_i = Observed\ - Predicted$

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(k+1)}}$$

Section 12.4 Inferences in Multiple Regression

- Inferences about any of the parameters in the general linear model are the same as for the simple linear regression model
- The coefficient of determination, R^2, is the proportion of the variation in the dependent variable, y, that is explained by the model relating y to $x_1, x_2, \ldots, x_k$. However, we will use Adjusted R^2.
- Multicollinearity is present when the independent variables are themselves highly correlated

Overall Model Test

- Hypotheses
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
  - H_a: At least one $\beta_j \neq 0$
- Test Statistic: F with k numerator df and $n - k - 1$ denominator df (df = degrees of freedom)
- Use *p*-value from output and compare to α
- Check assumptions and draw conclusions

According to the null hypothesis, none of the independent variables included in the model have any predictive value

If the null hypothesis is rejected, there is good evidence of some degree of predictive value somewhere among the independent variables

Hypothesis Test for an Individual Coefficient $\beta_j = 0$

- Hypotheses
  - Case 1: H_0: $\beta_j \leq 0$ versus H_a: $\beta_j > 0$
  - Case 2: H_0: $\beta_j \geq 0$ versus H_a: $\beta_j < 0$
  - Case 3: H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$
- Test Statistic: t value from software output
- Compare *p*-value from output to significance level α
  - Reject the null hypothesis H_0 if *p*-value $\leq \alpha$
    (If *p*-value is low, H_0 must go)
  - Fail to reject the null hypothesis H_0 if *p*-value > α
    (If *p*-value is high, with H_0 we must comply)
- Check assumptions and draw conclusions

Hypothesis Test for $\beta_j = 0$ (continued)

- The null hypothesis does not assert that the independent variable x_j has no predictive value by itself
- It asserts that it has no additional predictive value over and above that contributed by the other independent variables in the model
- When two or more independent variables are highly correlated among themselves, it often happens that no x_j can be shown to have unique predictive value, even though the x's together have been shown to be useful

Section 11.1 Introduction and Abstract of Research Study (Transformations Only)

- For multiple regression, it may be necessary to transform the independent variables, dependent variable, or both
- The text provides several graphs and “Steps for choosing a transformation”
- The regression analysis is then performed on the transformed variables(s) as long as the assumptions are met for the analysis using the transformed variable(s)

Transformations of Data

- First, remember that multiple regression can fit very complex relationships between the dependent and independent variables
- Transformations of data are most often used to meet the assumptions of multiple regression when the original data is used to build a model and the assumptions are not met
- Finding a good transformation often requires trial and error. Our text notes two key features to look for in a scatterplot:
  - Is the relation nonlinear
  - Is there a pattern of increasing variability along the y axis as x increases (or decreases)

Steps for choosing a transformation

1. If the plot indicates a relation that is increasing but at a decreasing rate and if variability around the curve is roughly constant, transform x using a square root, logarithm, or inverse transformation.
2. If the plot indicates a relation that is increasing at an increasing rate and if variability is roughly constant, try using both x and x^2 as predictors. Because this method uses two variables, the multiple regression methods of the next two chapters are needed.
3. If the plot indicates a relation that increases to a maximum and then decreases and if variability around the curve is roughly constant, again try using both x and x^2 as predictors.
4. If the plot indicates a relation that is increasing at a decreasing rate and if variability around the curve increases as the predicted y-value increases, try using y^2 as the dependent variable.
5. If the plot indicates a relation that is increasing at an increasing rate and if variability around the curve increases as the predicted y-value increases, try using $\ln(y)$ as the dependent variable. It sometimes may also be helpful to use $\ln(x)$ as the independent variable. Note that a change in a natural logarithm corresponds quite closely to a percentage change in the original variable. Thus, the slope of a transformed variable can be interpreted quite well as a percentage change.

See out text on page 562 for corresponding plots