MULTIPLE REGRESSION ANALYSIS

Seth Howells
Concepts of Statistics II
Week# 4 Assignment – Multiple Regression
08/02/20

TABLE OF CONTENTS

OVERVIEW

Multiple Regression analysis, a regression model that examines the relationship between one dependent variable (criterion) and two or more independent variables (predictor), is widely used in business as a statistical technique. Most businesses use this technique in efforts to forecast, or predict, a single dependent variable based on the selection of two or more predictor variables.

EQUATION OVERVIEW

Multiple Regression analysis is interested in $\hat{y}$, which is comprised of the following:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 +_{\dots} b_px_p + \varepsilon$$

for *i* in *n* observations:

b_0: y-intercept (constant), the mean
b_1: regression coefficients, for each one unit increase/decrease, y will increase/decrease
x_1: explanatory variable
ε: random error
together, b_1x_1 represents the linear effect of X_1
**please note: exponential variables, such as x^2, will create inflection points of curvilinear relationships.*

The equation derives from the slope formula (*y = mx + b*) and extends the linear regression equations to include any two or more variables. Consequently, multiple regression attempts to create a slope that minimizes the distance (sum of the squared residuals) from the aggregate data points to the fitted regression line as shown in *FIG 4.1 (Least Squares Method).* **Residuals *(e or ε)*** is the difference between the observed value and the predicted value. Interpretation of R^2, the proportion of residuals above or below the regression line, is below:

- $R^2 = \frac{variance\ explained\ by\ the\ model}{total\ variance}$, represented by a percent. That is, R^2 represents the percent of variation in the dependent variable explained by the independent variable.
- R^2 *Assumptions*: generally, a high R^2 value indicates a better fit for the modal because a large portion of the total variance is explained by the independent variables

Least Squares Method examines the data points' distance to the regression line. This application is mostly related to data fitting. The reason the difference is squared is because the analysis simply focuses on the difference (distance) regardless of the actual positivity or negativity of the difference.

- *Assumptions*: the lower the sum of square values, the better fit because the difference indicates a close proximity to the regression line.
- *FIG 4.1* graphically shows how this method treats the observed value in relation to the predicted (regression line) value.

Least Squares Method: $\Sigma e_{i^2} = \Sigma (Y_i - \hat{Y}_i)^2$

CATEGORICAL VARIABLES IN MULTIPLE REGRESSION

Multiple Regression analysis falls under the Dependence Methods of the multivariate umbrella and as such requires metric variables. However, a researcher might want to investigate independent variables' affect on the dependent variable in which one of the variables is nominal. Utilizing **dummy variables**, numerical variables created to replace nominal text that typically take a "1" or "0" value, is a solution to this problem. Dummy variables represent the existence or nonexistence of a nominal value. For example, a researcher could convert the observations existence or nonexistence row each corresponding row.

- X_1 = 1, if Republican; X_1 = 0, otherwise.
- X_2 = 1, if Democrat; X_2 = 0, otherwise.

INTERPRETATION

There are many items to interpret with multiple regression analysis. Some items aim to identify whether a particular independent value is statistically significant, others to explain the relationship between variance of values, and some items aim to predict the effects of an independent variable's rate of change toward the dependent variable. Three interpretations that are key to multiple regression analysis are: R^2 vs Adjusted-R^2, Significance Level, and Regression Coefficients, also known as Parameter Estimates.

a. ***R^2 vs Adjusted-R^2*** – percent of variation in the dependent variable explained by the independent variables.
    1. R^2 : to be used for simple regression, 1 dependent and 1 independent variable
    2. Adjusted-R^2: used for multiple regression, 1 dependent and 2 or more independent variables
b. ***Significance Level***: examines p-value
    1. High >0.05 – statistically insignificant, does not greatly affect the dependent variable
        i. Highest p-value should be removed and multiple regression to be performed again
    2. Low <0.05 – statistically significant, affects the dependent variable

c. ***Regression Coefficients (Parameter Estimates)***: dependent rate of change (positive or negative) per 1 unit change of the independent variable

ESTIMATION TECHNIQUES

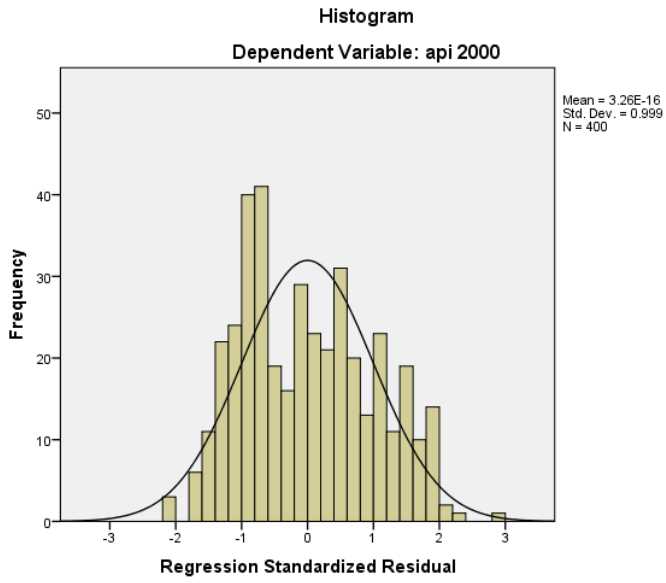
There are different types of selection processes for multiple regression. Two popular processes for multiple regression are *simultaneous* and *stepwise*.

- *Simultaneous*:
  - Standard method, also known as entry method
  - Includes all the independent variables into the regression model at the same time
  - Appropriate to use when there is a small set of predictor variables
- *Stepwise*:
  - Requires analysis at each step in efforts to determine the contribution of the predictor variable
  - Appropriate to use when predictors are grouped based on theoretical reasons (sales vs production efficiency)

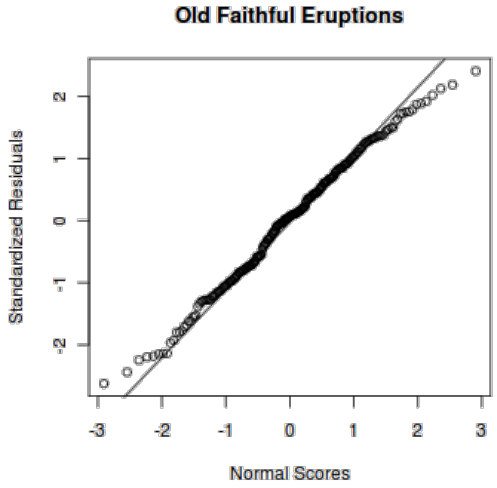
VISUALIZATION

Because multiple regression analysis has many items to examine, difference graphical plots are used to get a clearer, visually appealing understanding of the multivariate technique. More specifically, the visualizations below focus on the random error / residual and its relationship of the independent variables and dependent variable.

- **Histogram of standardized residuals:** determine whether the errors are normally distributed



- **Normal probability plot:** compares the observed standardized residuals to the expected standardized residuals from a normal distribution.



- **Scatter plot of residuals:** compares standardized predicted values of dependent variables to the standardized residuals from the regression equation.

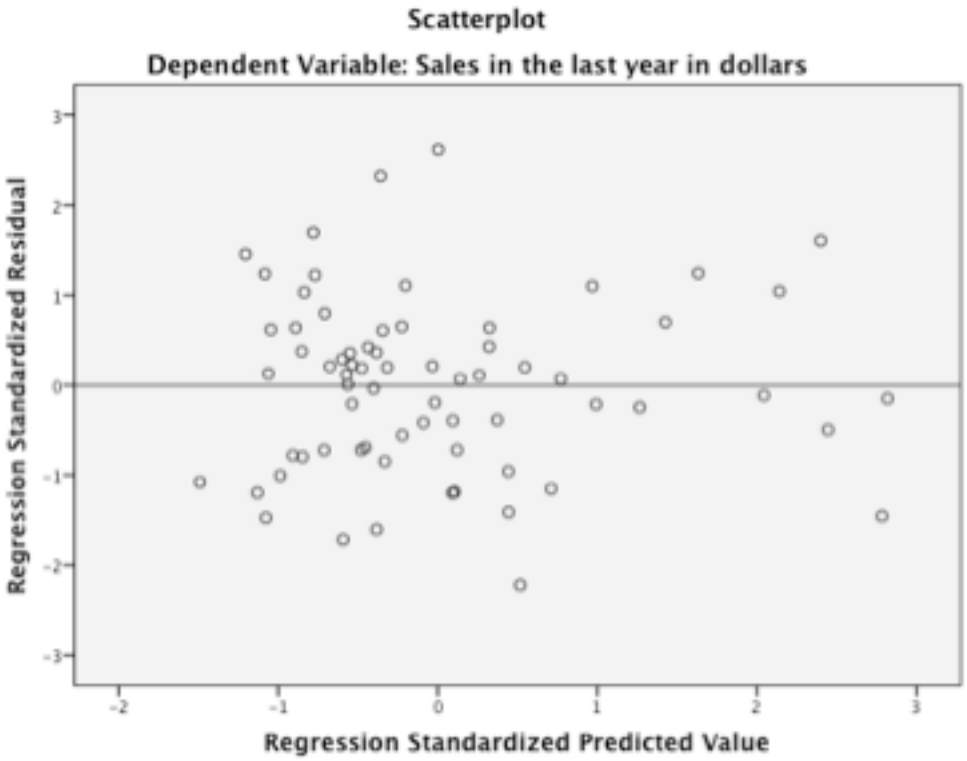


FIG 4.1 – Multiple Regression Least Squares Method

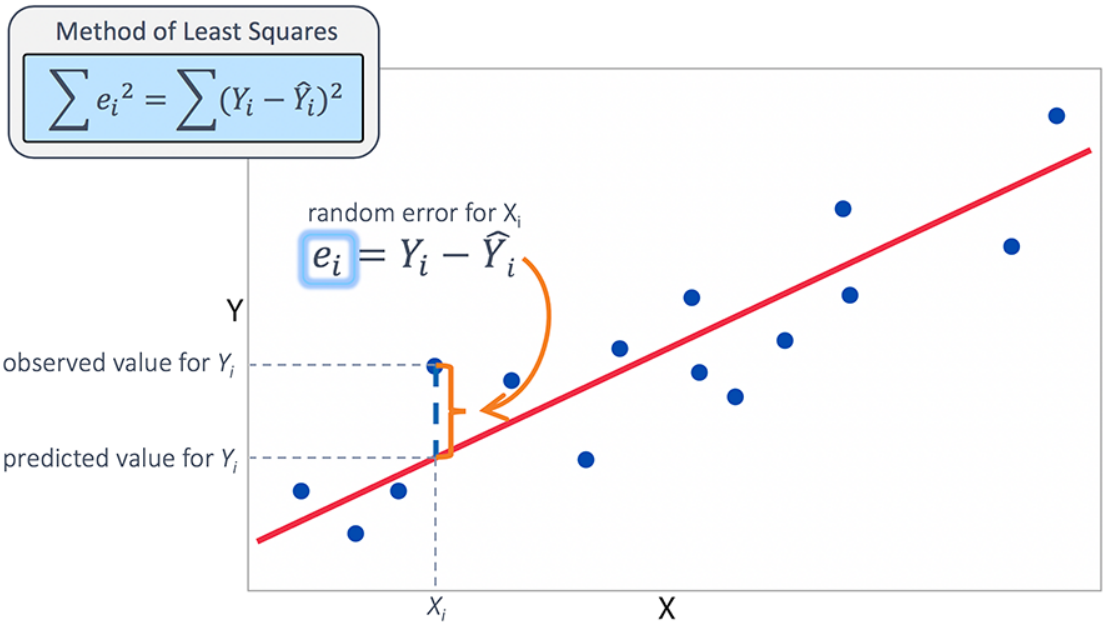


FIG 4.2 – 3-dimensional multiple linear regression scatter plot