***Multiple Regression***

Multiple regression is similar to simple regression, except we have multiple independent variables and one dependent variable.

Multiple regression is a powerful analytical tool with many applications in business. For example, it can be used to....

***Screen job applicants:*** Multiple regression can be used to build a mathematical model to identify applicants that are more likely to do well in a particular job

* The model would simultaneously consider multiple factors such as work experience, college major, GPA, and performance on screening tests, to “predict” success
* This model won’t actually predict whether a particular applicant will do well. It will simply describe what characteristics have been associated with success for past applicants who were hired

***Identify successful locations:*** Multiple regression can be used to build a model to estimate/predict which retail locations are most likely to be successful for a sandwich shop

* Variables in the model might be amount of foot traffic, amount of auto traffic, proximity of competitors, demographic characteristics of nearby residents, local economic conditions, etc.
* Again, this model won’t actually predict success, but is a guide describing what’s been successful in the past that helps the company make guesses for the future

***Factors associated with video streaming service use:*** Multiple regression can be used to identify the characteristics of people who are heavy users of video streaming services

* Variables in the model might include age, hours of paid employment per week, availability of alternative entertainment options nearby, marital or dating status, and whether the person has children living at home
* This model will indicate which characteristics are most strongly associated with video streaming service use at the time the data were collected; if the environment changes, the model would no longer be valid

**Kinds of variables we can include in multiple regression analysis**

The dependent variable (the thing we’re trying to “predict”) must be

• interval level of measurement, or

• ratio level of measurement

The independent variables must be

• binary variables,

• interval level of measurement, or

• ratio level of measurement

**Reviewing levels of measurement**

***Nominal variables*** are composed of categories that have no order to them:

the state a business is located in

a student’s major in college

location of a pharmacy (a free standing store, in a strip mall, or inside a grocery store)

***Binary variables*** are a type of nominal variable that has two categories:

yes/no

college graduate/not a college graduate

has coding experience/doesn’t have coding experience

mall store location/free-standing store location

offers delivery/doesn’t offer delivery

***Ordinal variables*** have ordered categories:

age categories (18-25, 26-35, 36-45, 46-60, 60+)

movie ratings (G, PG, PG-13, R, NC-17)

clothing sizes (small, medium, large)

***Interval variables*** are scaled variables with equal distances between scale points:

Job performance evaluation, where 1 = far below expectations and 10 = outstanding performance

Taste evaluation scale, where 1 = tastes terrible and 7 = tastes delicious

A variable like the income measure below, which has equal distances between categories

––– less than $25,000

––– $25,000 - $49,999

––– $50,000 - $74,999

––– $75,000 - $99,999

––– $100,000 - $124,999

––– $125,000 - $149,999

***Ratio variables*** have specific quantities, such as

number of employees

number of years of work experience

points on a test

price of a pair of jeans

Only binary, interval, and ratio level variables can be used in regression analysis

**Identifying suitable independent variables**

Prime Fit is a chain of 72 fitness centers on the west coast. An analyst at the company is developing a model to identify the characteristics of gym locations that are most effective in retaining members.

The dependent variable will be the percentage of members who renew their membership after their first 12-month contract.

For independent variables, below is a list of the variables the analyst has data for. Which ones would be suitable to include in the model and which would not? (In other words, which ones have the proper level of measurement for inclusion?)

Potential independent variables

1. size of the gym in square feet
2. number of personal trainers employed by the gym
3. number of miles to the closest competitor gym
4. gym identification number
5. whether the gym is free-standing or located within a strip mall
6. score on most recent inspection report (graded 1 through 10)
7. the state in which the gym is located

Which of the above variables should not be included in the regression model because of level of measurement?

**Multiple regression example: Identifying frequent visitors to fast food restaurants**

A major fast food franchise wants to identify the factors related to frequency of fast food restaurant visits by consumers. An analyst used regression to develop the model below.

The dependent variable is the number of fast food restaurant visits a person makes per month.

The independent variables are

• number of hours working outside the home per day

• monthly income (in thousands of dollars; $000)

• presence of children at home (binary variable; 0 = children not present, 1 = children present)

Here is the regression model:

# of visits = 2.1 + .68 (# of hours worked) - .11 (monthly income) +. 51 (if children present)

**Estimates from the model:**

For a person who works 6 hours a day outside the home, makes $3,500 per month, and has children, the model would estimate the following:

2.10 + .68 (6) -.11 (3.5) +. 51 (1) = 6.37 visits per month

Here’s the estimate for a person with the same work schedule and income, but with no children:

2.10 + .68 (6) -.11 (3.5) +. 51 (0) = 5.86 visits per month

And here’s the estimate for a person who doesn’t work outside the home, has a monthly income of $6,000, and has children:

2.10 + .68 (0) -.11 (6.0) +. 51 (1) = 2.01 visits per month

**Interpreting the model coefficients:**

Here is the model again:

# of visits = 2.10 + .68 (# of hours worked) - .11 (monthly income) +. 51 (if children present)

2.10 – Theoretically, this number represents how many visits a person in the sample made to fast food restaurants if they worked outside the home 0 hours per week, had 0 monthly income, and had no children in the home. Because the data didn’t include any people with 0 income, this situation is outside the relevant range of the model and not of practical interest.

.68 – This is the coefficient for number of hours worked outside the home per week. For each additional hour worked outside the home, the number of fast food visits per month by people in the sample increased by .68 (on average).

-.11 – This is the coefficient for monthly income (in thousands of dollars). For each additional $1,000 in income, the number of fast food visits per month by people in the sample decreased by .11 visits (on average).

.51 – This is the coefficient for the binary variable representing the presence of children. The number of fast food visits per month was .51 visits higher (on average) for people in the sample who had children in the household.

This is a very simple example model. In a real application, the model would be much more complex and include many more independent variables, including additional demographic variables, information about alternative food sources available to the consumer, and attitudes about nutrition and convenience, and more.

**Model fit and usefulness**

R2 for this model is .14, meaning that 14 percent of the variation in fast food visits is accounted for by this model. This isn’t a very good fit. A model that included more information like that described in the preceding paragraph would have a better fit. Models that have poor fit are less useful than models with good fit.

**Another example: Characteristics associated with purchase volume of sales accounts**

Midwest Veterinary is a wholesaler that provides equipment and supplies to veterinary offices. Sales management is interested in finding a model that would identify the factors associated with the size of monthly purchases by customers. Multiple regression was used to build the model below.

The dependent variable is prior month purchases by each of the current customers (in dollars)

There are three independent variables:

* veterinary practice size (number of employees)
* how satisfied the customer is with service provided by Midwest Veterinary on a 10-point rating scale
* whether the practice is located in a rural or suburban area (1 = rural, 0 = suburban)

Here’s the multiple regression model that was developed:

Purchases = 708.81 + 329.96 (# of employees) + 449.35 (satisfaction rating) – 317.39 (if a rural location)

**Estimates from the model:**

Among current customers, a veterinary practice with 7 employees that rated their satisfaction with Midwest Vet as a 5 and has a rural location, the model estimates their monthly purchases, on average,

to be:

708.81 + 329.96 (7) + 449.35 (5) - 317.39 (1) = $4,947.88 monthly purchases

Here’s the estimate for a customer with 15 employees, a satisfaction assessment of 9, and in a suburban location:

708.81 + 329.96 (15) + 449.35 (9) - 317.39 (0) = $9,702.35 monthly purchases

And here’s the estimate for a customer with the same situation as the preceding customer but in a rural location:

708.81 + 329.96 (15) + 449.35 (9) - 317.39 (1) = $9,384.96 monthly purchases

**Interpreting the model coefficients:**

Here’s the model again:

Purchases = 708.81 + 329.96 (# of employees) + 449.35 (satisfaction rating) – 317.39 (if a rural location)

708.81 – Theoretically, this number represents purchase amount by a veterinary practice that has 0 employees, a satisfaction rating of 0, and is in a suburban location. This situation is outside the relevant range (the data, for obvious reasons, didn’t include any veterinary practices with 0 employees) and therefore is not of practical interest.

329.96 – This is the coefficient for number of employees. For each additional employee in a veterinary practice, the dollar value of purchases per month increased $329.96, on average, among Midwest Vet’s customers.

449.35 – This is the coefficient for satisfaction rating. For each additional one-point increase in their satisfaction rating, the dollar value of purchases increased $449.35, on average, among Midwest Vet’s customers.

-317.39 – This is the coefficient for the binary variable representing the customer’s location (rural or suburban). Monthly purchases were $317.39 lower, on average, for customers whose practice was in a rural location.

**Model fit**

R2 for this model is .57, meaning that 57 percent of the variation in monthly purchases is accounted for by this model. 43 percent of the variation in purchase size is explained by other variables not in the model.

**Summing up—what you need to know about correlation and regression for this course**

Below are the essential things you should remember about correlation and regression

Correlation

1. what a scatter plot looks like and what it represents
2. what a nonlinear relationship looks like on a scatter plot
3. how to interpret correlation coefficients when there is a positive association, negative association, or no association between two variables
4. why correlation does not indicate causation

Simple regression

1. what the dependent and independent variables are in a regression model
2. what a regression line is, and what it means that a regression line is “best fitting”
3. what the intercept and slope coefficient in a regression model represent
4. how to interpret the slope coefficient in a regression model
5. understand that regression models don’t predict anything, but instead describe something that has already happened
6. what the requirements are for a valid regression model
7. why the relevant range for a model is important, and how it is determined
8. what “error” is in a regression model, and how you can tell whether a model fits the actual data well

Multiple regression

1. what kind of variables (based on level of measurement) are eligible for inclusion in a regression model
2. how to create estimates from a multiple regression model
3. how to interpret multiple regression slope coefficients
4. interpreting intercepts can be a little tricky sometimes—information about intercept interpretation has been provided, but it isn’t always easy to tell when it is or is not meaningful to interpret the intercept, so you won’t be tested on this
5. how to interpret R2 and “percentage of variation explained” means

These class sessions have given you an introduction to regression analysis and shown you a few of the ways it can be used by companies.

Actually conducting regression analysis requires additional training to avoid making analysis and interpretation errors.

Students who enjoyed the material on correlation and regression should consider getting a marketing analytics certificate. In that coursework, you’ll learn the technical details of running regression analysis plus many other ways data can be analyzed to help companies make marketing decisions.