A Regression Analysis on BMI – Cholesterol and the Environmental Tobacco Smoke Exposure

Akash Kumar Meher[1]

Ispat Autonomous College, Rourkela, India

Soumya Ranjan Kabi[2]

Ispat Autonomous College, Rourkela

Abstract:

The basics of linear and nonlinear regression analysis will be reviewed along with the statistical methods that are used with regression, such as confidence intervals, prediction intervals, and hypothesis tests. Although the emphasis will be on traditional topics, with standard techniques applied primarily to data sets, the course will also touch upon important developments of the past several years. This paper provides an introduction to the construction and validation of regression models from a practical point of view. Scientific and engineering examples such as BMI, total cholesterol and the Controversy over Environmental Tobacco Smoke Exposure are used to show the steps in the model-building process and to give an intuitive understanding of regression algorithms and the associated hypothesis tests and statistical intervals. Both linear and nonlinear regression is covered, with special attention to calibration and outlier resistant regression. A computer-based approach is used for calculations to minimize the number of equations and formulas used.

Key words: regression analysis, hypothesis, regression model, data set, BMI – cholesterol, environmental tobacco smoke exposure.

1. Introduction:

The introduction to associations between two quantitative variables was done by using correlation and regression. Some of the complexity of the formulas disappears when these techniques are described in terms of standardized versions of the variables. This simplified approach also leads to a more intuitive understanding of correlation and regression. More specifically, the following facts about correlation and regression are simply expressed.

1.1 Objectives of the Study:

i. To illustrating the relationship between BMI and total cholesterol.

ii. To study on the Controversy over Environmental Tobacco Smoke Exposure

1.2 Research Methodology:

The study is based on critical evaluation and analysis of basically Secondary Data. The primary sources include customers. A study is undertaken in the sampled regions to see its impact for which a detailed questionnaire is prepared to collect relevant information from the primary source for the guidance of the researchers. With the assistance of the questionnaire, thorough considerations were made with the certain foundations of primary data to comprehend their views, thinking and attitude which would aid to provide the researchers useful recommendations, if any. The questionnaire is processed with the help of statistical tools like tabulations, grouping, percentages, averages, testing of hypothesis etc. Questionnaire is used mainly to analyze the opinion of the respondents.

1.3 Thrust Area:

The study was done by using Primary and Secondary sources of investigation. Around 450 samples of questionnaire were surveyed and by the enumerator to obtain the answer from the respondents.

2. Regression Analysis:

Regression analysis is a widely used technique which is useful for evaluating multiple independent variables. As a result, it is particularly useful for assess and adjusting for confounding. It can also be used to assess the presence of effect modification.

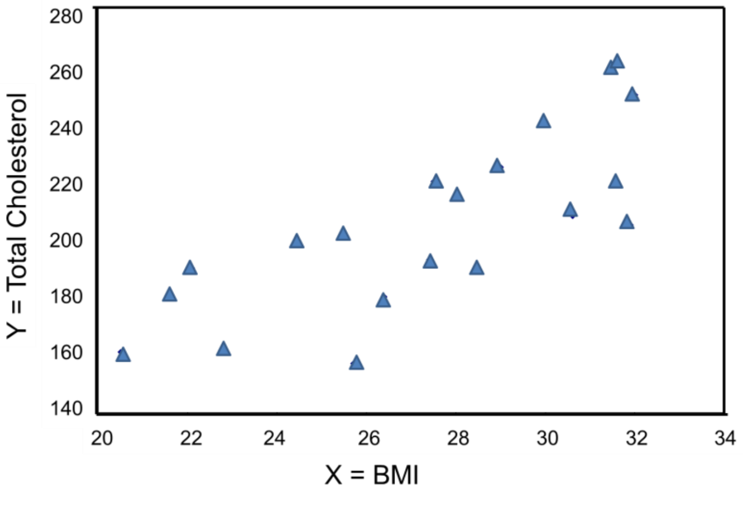
2.1 Simple Linear Regression:

Suppose we want to assess the association between total cholesterol and body mass index (BMI) in which total cholesterol is the dependent variable, and BMI is the independent variable. In regression analysis, the dependent variable is denoted Y and the independent variable are denoted X. So, in this case, Y=total cholesterol and X=BMI.

When there is a single continuous dependent variable and a single independent variable, the analysis is called a **simple linear regression analysis**. This analysis assumes that there is a linear association between the two variables. (If a different relationship is hypothesized, such as a curvilinear or exponential relationship, alternative regression analyses are performed.)

The figure below is a scatter diagram illustrating the relationship between BMI and total cholesterol. Each point represents the (X, Y) pair, in this case, BMI and the corresponding total cholesterol measured in each participant. Note that the independent variable is on the horizontal axis and the dependent variable on the vertical axis.

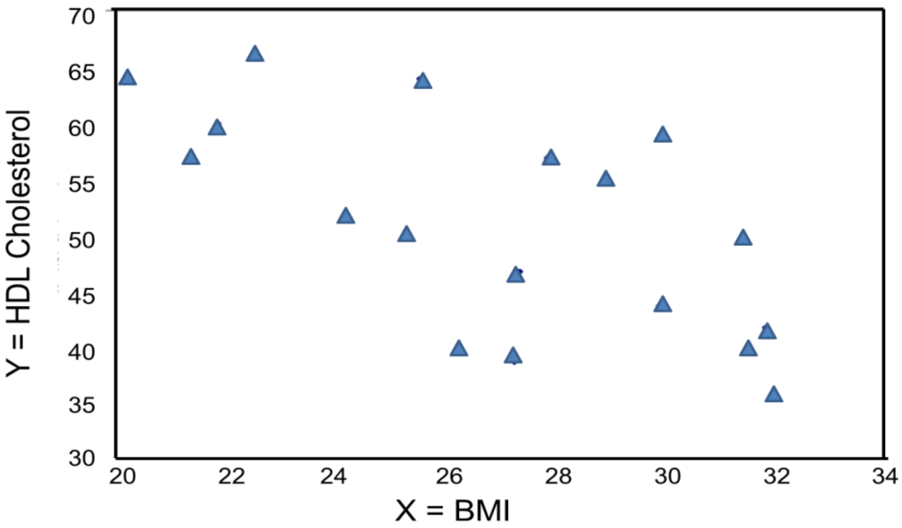
2.2 BMI and Total Cholesterol



The graph shows that there is a positive or direct association between BMI and total cholesterol; participants with lower BMI are more likely to have lower total cholesterol levels and participants with higher BMI are more likely to have higher total cholesterol levels. In contrast, suppose we examine the association between BMI and HDL cholesterol.

In contrast, the graph below depicts the relationship between BMI and HDL **HDL cholesterol** in the same sample of n=20 participants.

2.3 BMI and HDL Cholesterol:



This graph shows a negative or inverse association between BMI and HDL cholesterol, i.e., those with lower BMI are more likely to have higher HDL cholesterol levels and those with higher BMI are more likely to have lower HDL cholesterol levels.

For either of these relationships we could use simple linear regression analysis to estimate the equation of the line that best describes the association between the independent variable and the dependent variable. The simple linear regression equation is as follows:

$\hat{Y} = b_0 + b_1 X$, where

$\hat{Y}$ is the predicted or expected value of the outcome, **X** is the predictor , **b_0** is the estimated Y-intercept, and **b_1** is the estimated slope. The Y-intercept and slope are estimated from the sample data so as to minimize the sum of the squared differences between the observed and the predicted values of the outcome, i.e., the estimates minimize:

$$\Sigma(Y - \hat{Y})^2$$

These differences between observed and predicted values of the outcome are called **residuals**. The estimates of the Y-intercept and slope minimize the sum of the squared residuals, and are called the **least squares estimates**.

Residuals

Conceptually, if the values of X provided a perfect prediction of Y than the sum of the squared differences between observed and predicted values of Y would be 0. That would mean that variability in Y could be completely explained by differences in X. However, if the differences between observed and predicted values are not 0, then we are unable to entirely account for differences in Y based on X, then there are residual errors in the prediction. The residual error could result from inaccurate measurements of X or Y, or there could be other variables besides X that affect the value of Y.

Based on the observed data, the best estimate of a linear relationship will be obtained from an equation for the line that minimizes the differences between observed and predicted values of the outcome. The **Y-intercept** of this line is the value of the dependent variable (Y) when the independent variable (X) is zero. The **slope** of the line is the change in the dependent variable (Y) relative to a one unit change in the independent variable (X). The least squares estimates of the y-intercept and slopes are computed as follows:

$$b_1 = r\frac{s_y}{s_x} \quad \text{and} \quad b_0 = \overline{Y} - b_1\overline{X},$$

Where

- r is the sample correlation coefficient,
- the sample means are $\overline{X}$ and $\overline{Y}$
- And Sx and Sy are the standard deviations of the independent variable x and the dependent variable y, respectively.

2.4 BMI and Total Cholesterol:

The least squares estimates of the regression coefficients, b $_0$ and b_1, describing the relationship between BMI and total cholesterol are b_0 = 28.07 and b_1=6.49. These are computed as follows:

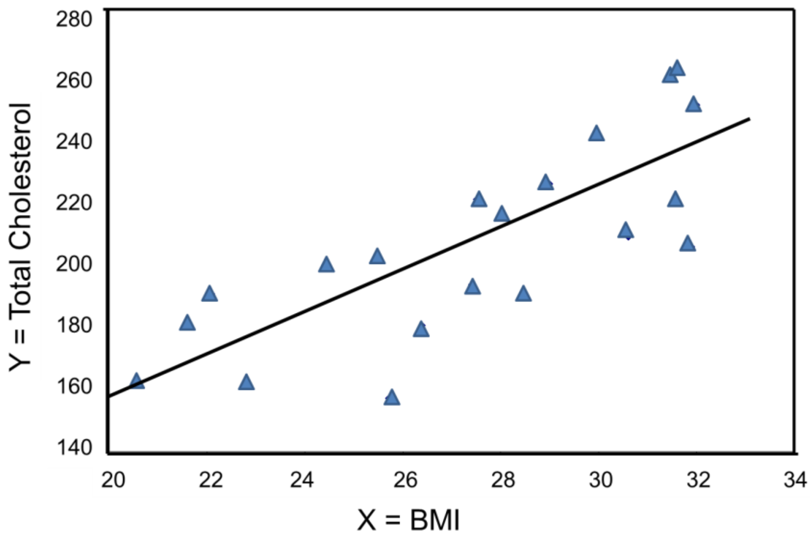
$$b_1 = r\frac{s_y}{s_x} = 0.78\frac{30.8}{3.7} = 6.49 \quad \text{and} \quad b_0 = \overline{Y} - b_1\overline{X} = 205.9 - 6.49(27.4) = 28.07.$$

The estimate of the Y-intercept (b_0 = 28.07) represents the estimated total cholesterol level when BMI is zero. Because a BMI of zero is meaningless, the Y-intercept is not informative. The estimate of the slope (b_1 = 6.49) represents the change in total cholesterol relative to a one unit change in BMI. For example, if we compare two participants whose BMIs differ by 1 unit, we would expect their total cholesterols to differ by approximately 6.49 units (with the person with the higher BMI having the higher total cholesterol).

The equation of the regression line is as follows:

$$\hat{Y} = 28.07 + 6.49\ \text{BMI}$$

The graph below shows the estimated regression line superimposed on the scatter diagram.



The regression equation can be used to estimate a participant's total cholesterol as a function of his/her BMI. For example, suppose a participant has a BMI of 25. We would estimate their total cholesterol to be 28.07 + 6.49(25) = 190.32. The equation can also be used to estimate total cholesterol for other values of BMI. However, the equation should only be used to estimate cholesterol levels for persons who are BMIs are in the range of the data used to generate the regression equation. In our sample, BMI ranges from 20 to 32, thus the equation should only be used to generate estimates of total cholesterol for persons with BMI in that range.

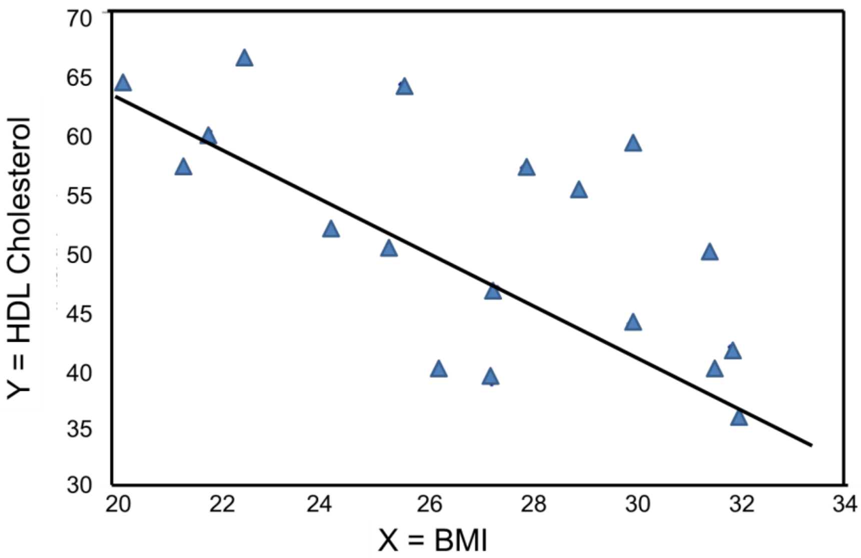
There are statistical tests that can be performed to assess whether the estimated regression coefficients (b_0 and b_1) are statistically significantly different from zero. The test of most interest is usually H_0: $b_1=0$ versus H_1: $b_1\neq 0$, where b_1 is the population slope. If the population slope is significantly different from zero, we conclude that there is a statistically significant association between the independent and dependent variables.

2.5 BMI and HDL Cholesterol:

The least squares estimates of the regression coefficients, b_0 and b_1, describing the relationship between BMI and HDL cholesterol are as follows: b_0 = 111.77 and b_1 = -2.35. These are computed as follows:

$$b_1 = r\frac{s_y}{s_x} = -0.72\frac{12.1}{3.7} = -2.35 \quad \text{and} \quad b_0 = \overline{Y} - b_1\overline{X} = 47.4 - (-2.35)(274) = 111.79.$$

Again, the Y-intercept in uninformative because a BMI of zero is meaningless. The estimate of the slope (b_1 = -2.35) represents the change in HDL cholesterol relative to a one unit change in BMI. If we compare two participants whose BMIs differ by 1 unit, we would expect their HDL cholesterols to differ by approximately 2.35 units (with the person with the higher BMI having the lower HDL cholesterol. The figure below shows the regression line superimposed on the scatter diagram for BMI and HDL cholesterol.



Linear regression analysis rests on the assumption that the dependent variable is continuous and that the distribution of the dependent variable (Y) at each value of the independent variable (X) is approximately normally distributed. Note, however, that the independent variable can be continuous (e.g., BMI) or can be dichotomous (see below).

2.6 Comparing Mean HDL Levels With Regression Analysis:

We previously considered data from a clinical trial that evaluated the efficacy of a new drug to increase HDL cholesterol (see page 4 of this module). We compared the mean HDL levels between treatment groups using a two independent samples t test. Note, however, that regression analysis can also be used to compare mean HDL levels between treatments.

HDL cholesterol is the continuous dependent variable and treatment (new drug versus placebo) is the independent variable. A simple linear regression equation is estimated as follows:

$\hat{Y} = 39.21 + 0.95\ X$ where

$\hat{Y}$ is the estimated HDL level and X is a dichotomous variable (also called an indicator variable, i.e., indicating whether the active treatment was given or not). In this example, X is coded as 1 for participants who received the new drug and as 0 for participants who received the placebo.

The estimate of the Y-intercept is b_0=39.21. The Y-intercept is the value of Y (HDL cholesterol) when X is zero. In this example, X=0 indicates the placebo group. Thus, the **Y-intercept is exactly equal to the mean HDL level in the placebo group**. The slope is b_1=0.95. The slope represents the change in Y (HDL cholesterol) relative to a one unit change in X. A one unit change in X represents a difference in treatment assignment (placebo versus new drug). The **slope represents the difference in mean HDL levels between the treatment groups.** Dichotomous (or indicator) variables are usually coded as 0 or 1, where 0 is assigned to participants who do not have a particular risk factor, exposure or characteristic and 1 is assigned to participants who have the particular risk factor, exposure or characteristic. In a later section we will present **multiple logistic regression analysis** which applies in situations where the outcome is dichotomous (e.g., incident CVD).

3. The Controversy Over Environmental Tobacco Smoke Exposure

There is convincing evidence that active smoking is a *cause* of lung cancer and heart disease. Many studies done in a wide variety of circumstances have consistently demonstrated a strong association and also indicate that the risk of lung cancer and cardiovascular disease (i.e.., heart attacks) increases in a dose-related way. These studies have led to the conclusion that active smoking is causally related to lung cancer and cardiovascular disease. Studies in active smokers have had the advantage that the lifetime exposure to tobacco smoke can be quantified with reasonable accuracy, since the unit dose is consistent (one cigarette) and the habitual nature of tobacco smoking makes it possible for most smokers to provide a reasonable estimate of their total lifetime exposure quantified in terms of cigarettes per day or packs per day. Frequently, average daily exposure (cigarettes or packs) is combined with duration of use in years in order to quantify exposure as "pack-years".

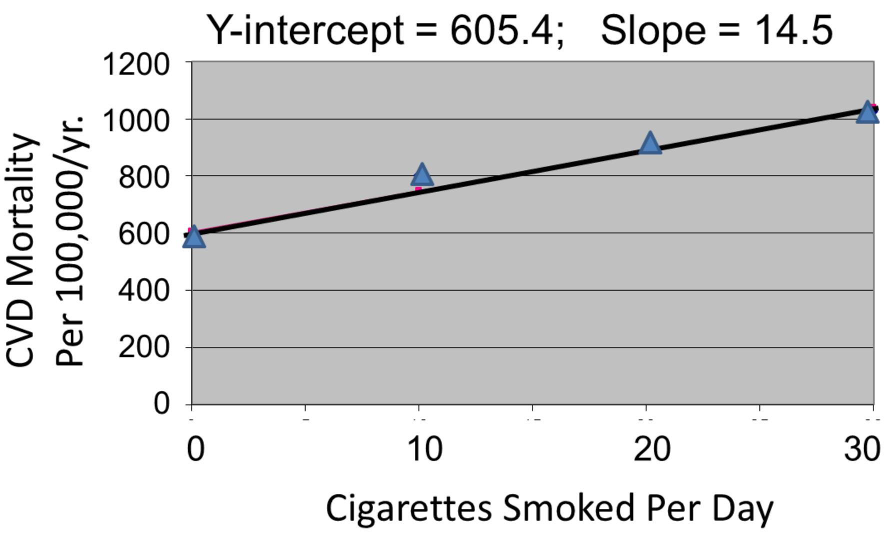
It has been much more difficult to establish whether environmental tobacco smoke (ETS) exposure is causally related to chronic diseases like heart disease and lung cancer, because the total lifetime exposure dosage is lower, and it is much more difficult to accurately estimate total lifetime exposure. In addition, quantifying these risks is also complicated because of confounding factors. For example, ETS exposure is usually classified based on parental or spousal smoking, but these studies are unable to quantify other environmental exposures to tobacco smoke, and inability to quantify and adjust for other environmental exposures such as air pollution makes it difficult to demonstrate an association even if one existed. As a result, there continues to be controversy over the risk imposed by environmental tobacco smoke (ETS).

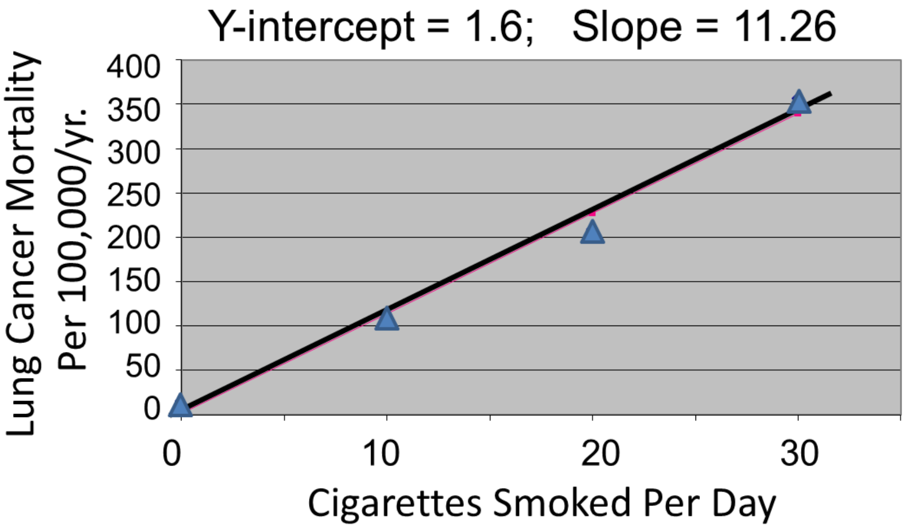
Correlation analysis provides a useful tool for thinking about this controversy. Consider data from the below. This reported the annual mortality for a variety of disease at four levels of cigarette smoking per day: Never smoked, 1-14/day, 15-24/day, and 25+/day. In order to perform a correlation analysis, I rounded the exposure levels to 0, 10, 20, and 30 respectively.

Cigarettes Smoked Per Day	**CVD Mortality/100,000 men/yr.**	**Lung Cancer Mortality/100,000 men/yr.**
0	572	14
10 (actually 1-14)	802	105
20 (actually 15-24)	892	208
30 (actually >24)	1025	355

The figures below show the two estimated regression lines superimposed on the scatter diagram. The correlation with amount of smoking was strong for both CVD mortality (r= 0.98) and for lung cancer (r = 0.99). Note also that the Y-intercept is a meaningful number here; it represents the predicted annual death rate from these diseases in individuals who never smoked. The Y-intercept for prediction of CVD is slightly higher than the observed rate in never smokers, while the Y-intercept for lung cancer is lower than the observed rate in never smokers.

The linearity of these relationships suggests that there is an incremental risk with each additional cigarette smoked per day, and the additional risk is estimated by the slopes. This perhaps helps us think about the consequences of ETS exposure. For example, the risk of lung cancer in never smokers is quite low, but there is a finite risk; various reports suggest a risk of 10-15 lung cancers/100,000 per year. If an individual who never smoked actively was exposed to the equivalent of one cigarette's smoke in the form of ETS, then the regression suggests that their risk would increase by 11.26 lung cancer deaths per 100,000 per year. However, the risk is clearly dose-related. Therefore, if a non-smoker was employed by a tavern with heavy levels of ETS, the risk might be substantially greater.





Finally, it should be noted that some findings suggest that the association between smoking and heart disease is non-linear at the very lowest exposure levels, meaning that non-smokers have a disproportionate increase in risk when exposed to ETS due to an increase in platelet aggregation.

4. Multiple Linear Regression Analysis:

Multiple linear regression analysis is an extension of simple linear regression analysis, used to assess the association between two or more independent variables and a single continuous dependent variable. The multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p ,$$

where $\hat{Y}$ is the predicted or expected value of the dependent variable, X_1 through X_p are p distinct independent or predictor variables, b_0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, and b_1 through b_p are the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. In the multiple regression situation, b_1, for example, is the change in Y relative to a one unit change in X_1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed). Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

4.1 Controlling for Confounding With Multiple Linear Regression:

Multiple regression analysis is also used to assess whether confounding exists. Since multiple linear regression analysis allows us to estimate the association between a given independent variable and the outcome holding all other variables constant, it provides a way of adjusting for (or accounting for) potentially confounding variables that have been included in the model.

Suppose we have a risk factor or an exposure variable, which we denote X_1 (e.g., X_1=obesity or X_1=treatment), and an outcome or dependent variable which we denote Y. We can estimate a simple linear regression equation relating the risk factor (the independent variable) to the dependent variable as follows:

$$\hat{Y} = b_0 + b_1X_1$$

where b_1 is the estimated regression coefficient that quantifies the association between the risk factor and the outcome.

Suppose we now want to **assess whether a third variable (e.g., age) is a confounder**. We denote the potential confounder X_2, and then estimate a multiple linear regression equation as follows:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 .$$

In the multiple linear regression equation, b_1 is the estimated regression coefficient that quantifies the association between the risk factor X_1 and the outcome, adjusted for X_2 (b_2 is the estimated regression coefficient that quantifies the association between the potential confounder and the outcome). As noted earlier, some investigators assess confounding by assessing how much the regression coefficient associated with the risk factor (i.e., the measure of association) changes after adjusting for the potential confounder. In this case, we compare b_1 from the simple linear regression model to b_1 from the multiple linear regression models. As a rule of thumb, if the regression coefficient from the simple linear regression model changes by more than 10%, then X_2 is said to be a confounder.

Once a variable is identified as a confounder, we can then use multiple linear regression analysis to estimate the association between the risk factor and the outcome adjusting for that confounder. The test of significance of the regression coefficient associated with the risk factor can be used to assess whether the association between the risk factor is statistically significant after accounting for one or more confounding variables. This is also illustrated below.

Example - The Association between BMI and Systolic Blood Pressure

Suppose we want to assess the association between BMI and systolic blood pressure using data collected in the seventh examination of the Framingham Offspring Study. A total of n=3,539 participants attended the exam, and their mean systolic blood pressure was 127.3 with a standard deviation of 19.0. The mean BMI in the sample was 28.2 with a standard deviation of 5.3.

A simple linear regression analysis reveals the following:

Independent Variable	Regression Coefficient	T	P-value
Intercept	108.28	62.61	0.0001
BMI	0.67	11.06	0.0001

The simple linear regression model is:

$$\hat{Y} = 108.28 + 0.67\ (\text{BMI})$$

where

$\hat{Y}$ is the predicted of expected systolic blood pressure. The regression coefficient associated with BMI is 0.67 suggesting that each one unit increase in BMI is associated with a 0.67 unit increase in systolic blood pressure. The association between BMI and systolic blood pressure is also statistically significant (p=0.0001).

Suppose we now want to assess whether age (a continuous variable, measured in years), male gender (yes/no), and treatment for hypertension (yes/no) are potential confounders, and if so, appropriately account for these using multiple linear regression analysis. For analytic purposes, treatment for hypertension is coded as 1=yes and 0=no. Gender is coded as 1=male and 0=female. A multiple regression analysis reveals the following:

Independent Variable	Regression Coefficient	T	P-value
Intercept	68.15	26.33	0.0001
BMI	0.58	10.30	0.0001
Age	0.65	20.22	0.0001
Male gender	0.94	1.58	0.1133
Treatment for hypertension	6.44	9.74	0.0001

The multiple regression model is:

$\hat{Y}$ = 68.15 + 0.58 (BMI) + 0.65 (Age) + 0.94 (Male gender) + 6.44 (Treatment for hypertension).

Notice that the association between BMI and systolic blood pressure is smaller (0.58 versus 0.67) after adjustment for age, gender and treatment for hypertension. BMI remains statistically significantly associated with systolic blood pressure (p=0.0001), but the magnitude of the association is lower after adjustment. The regression coefficient decreases by 13%.

[Actually, doesn't it decrease by 15.5%. In this case the true "beginning value" was 0.58, and confounding caused it to appear to be 0.67. So the actual % change = 0.09/0.58 = 15.5%.]

Using the informal rule (i.e., a change in the coefficient in either direction by 10% or more), we meet the criteria for confounding. Thus, part of the association between BMI and systolic blood pressure is explained by age, gender and treatment for hypertension.

This also suggests a useful way of identifying confounding. Typically, we try to establish the association between a primary risk factor and a given outcome after adjusting for one or more other risk factors. One useful strategy is to use multiple regression models to examine the association between the primary risk factor and the outcome before and after including possible confounding factors. If the inclusion of a possible confounding variable in the model causes the association between the primary risk factor and the outcome to change by 10% or more, then the additional variable is a confounder.

4.2 Relative Importance of the Independent Variables

Assessing only the p-values suggests that these three independent variables are equally statistically significant. The magnitude of the t statistics provides a means to judge relative importance of the independent variables. In this example, age is the most significant independent variable, followed by BMI, treatment for hypertension and then male gender. In fact, male gender does not reach statistical significance (p=0.1133) in the multiple regression model.

Some investigators argue that regardless of whether an important variable such as gender reaches statistical significance it should be retained in the model. Other investigators only retain variables that are statistically significant.

This is yet another example of the complexity involved in multivariable modeling. The multiple regression models produces an estimate of the association between BMI and systolic blood pressure that accounts for differences in systolic blood pressure due to age, gender and treatment for hypertension.

A one unit increase in BMI is associated with a 0.58 unit increase in systolic blood pressure holding age, gender and treatment for hypertension constant. Each additional year of age is associated with a 0.65 unit increase in systolic blood pressure, holding BMI, gender and treatment for hypertension constant.

Men have higher systolic blood pressures, by approximately 0.94 units, holding BMI, age and treatment for hypertension constant and persons on treatment for hypertension have higher systolic blood pressures, by approximately 6.44 units, holding BMI, age and gender constant. The multiple regression equation can be used to estimate systolic blood pressures as a function of a participant's BMI, age, gender and treatment for hypertension status. For example, we can estimate the blood pressure of a 50 year old male, with a BMI of 25 who is not on treatment for hypertension as follows:

$$\hat{Y} = 68.15 + 0.58\ (25) + 0.65\ (50) + 0.94\ (1) + 6.44\ (0) = 116.09$$

We can estimate the blood pressure of a 50 year old female, with a BMI of 25 who is on treatment for hypertension as follows:

$$\hat{Y} = 68.15 + 0.58\ (25) + 0.65\ (50) + 0.94\ (0) + 6.44\ (1) = 121.59.$$

4.3 Evaluating Effect Modification with Multiple Linear Regression

On page 4 of this module we considered data from a clinical trial designed to evaluate the efficacy of a new drug to increase HDL cholesterol. One hundred patients enrolled in the study and were randomized to receive either the new drug or a placebo. The investigators were at first disappointed to find very little difference in the mean HDL cholesterol levels of treated and untreated subjects.

	Sample Size	Mean HDL	Standard Deviation of HDL
New Drug	50	40.16	4.46
Placebo	50	39.21	3.91

However, when they analyzed the data separately in men and women, they found evidence of an effect in men, but not in women. We noted that when the magnitude of association differs at different levels of another variable (in this case gender), it suggests that effect modification is present.

WOMEN	**Sample Size**	**Mean HDL**	**Standard Deviation of HDL**
New Drug	40	38.88	3.97
Placebo	41	39.24	4.21
MEN			
New Drug	10	45.25	1.89
Placebo	9	39.06	2.22

Multiple regression analysis can be used to assess effect modification. This is done by estimating a multiple regression equation relating the outcome of interest (Y) to independent variables representing the treatment assignment, sex and the product of the two (called the **treatment by sex interaction variable**). For the analysis, we let T = the treatment assignment (1=new drug and 0=placebo), M = male gender (1=yes, 0=no) and TM, i.e., T * M or T x M, the product of treatment and male gender. In this case, the multiple regression analysis revealed the following:

Independent Variable	Regression Coefficient	T	P-value
Intercept	39.24	65.89	0.0001
T (Treatment)	-0.36	-0.43	0.6711
M (Male Gender)	-0.18	-0.13	0.8991
TM (Treatment x Male Gender)	6.55	3.37	0.0011

The multiple regression model is:

$$\hat{Y} = 39.24\ 0.36\ T - 0.18\ M + 6.55\ TM$$

The details of the test are not shown here, but note in the table above that in this model, the regression coefficient associated with the interaction term, b_3, is statistically significant (i.e., H_0: $b_3 = 0$ versus H_1: $b_3 \neq 0$). The fact that this is statistically significant indicates that the association between treatment and outcome differs by sex.

The model shown above can be used to estimate the mean HDL levels for men and women who are assigned to the new medication and to the placebo. In order to use the model to generate these estimates, we must recall the coding scheme (i.e., T = 1 indicates new drug, T=0 indicates placebo, M=1 indicates male sex and M=0 indicates female sex).

The expected or predicted HDL for men (M=1) assigned to the new drug (T=1) can be estimated as follows:

$$\hat{Y} = 39.24 -0.36\ (1) -0.18\ (1) + 6.55\ (1)(1) = 45.25$$

The expected HDL for men (M=1) assigned to the placebo (T=0) is:

$$\hat{Y} = 39.24 -0.36\ (0) -0.18\ (1) + 6.55\ (0)(1) = 39.06$$

Similarly, the expected HDL for women (M=0) assigned to the new drug (T=1) is:

$$\hat{Y} = 39.24 -0.36\ (1) -0.18\ (0) + 6.55\ (1)(0) = 38.88$$

The expected HDL for women (M=0) assigned to the placebo (T=0) is:

$$\hat{Y} = 39.24 -0.36\ (0) -0.18\ (0) + 6.55\ (0)(0) = 39.24$$

Notice that the expected HDL levels for men and women on the new drug and on placebo are identical to the means shown the table summarizing the stratified analysis. Because there is effect modification, separate simple linear regression models are estimated to assess the treatment effect in men and women:

MEN	Regression Coefficient	T	P-value
Intercept	39.08	57.09	0.0001
T (Treatment)	6.19	6.56	0.0001
WOMEN	Regression Coefficient	T	P-value
Intercept	39.24	61.36	0.0001
T (Treatment)	-0.36	-0.40	0.6927

The regression models are:

In Men: $\hat{Y} = 39.06 + 6.19\ T$

In Women: $\hat{Y} = 39.24\ 0.36\ T$

In men, the regression coefficient associated with treatment (b_1=6.19) is statistically significant (details not shown), but in women, the regression coefficient associated with treatment (b_1= -0.36) is not statistically significant (details not shown).

Multiple linear regression analysis is a widely applied technique. In this section we showed here how it can be used to assess and account for confounding and to assess effect modification. The techniques we described can be extended to adjust for several confounders simultaneously and to investigate more complex effect modification (e.g., three-way statistical interactions).

There is an important distinction between confounding and effect modification. Confounding is a distortion of an estimated association caused by an unequal distribution of another risk factor. When there is confounding, we would like to account for it (or adjust for it) in order to estimate the association without distortion. In contrast, effect modification is a biological phenomenon in which the magnitude of association is differs at different levels of another factor, e.g., a drug that has an effect on men, but not in women. In the example, present above it would be in inappropriate to pool the results in men and women. Instead, the goal should be to describe effect modification and report the different effects separately.

There are many other applications of multiple regression analysis. A popular application is to assess the relationships between several predictor variables simultaneously, and a single, continuous outcome. For example, it may be of interest to determine which predictors, in a relatively large set of candidate predictors, are most important or most strongly associated with an outcome. It is always important in statistical analysis, particularly in the multivariable arena, that statistical modeling is guided by biologically plausible associations.

Conclusion

Multivariable methods are computationally complex and generally require the use of a statistical computing package. Multivariable methods can be used to assess and adjust for confounding, to determine whether there is effect modification, or to assess the relationships of several exposure or risk factors on an outcome simultaneously. Multivariable analyses are complex, and should always be planned to reflect biologically plausible relationships. While it is relatively easy to consider an additional variable in a multiple linear or multiple logistic regression model, only variables that are clinically meaningful should be included.

It is important to remember that multivariable models can only adjust or account for differences in confounding variables that were measured in the study. In addition, multivariable models should only be used to account for confounding when there is some overlap in the distribution of the confounder each of the risk factor groups.

Stratified analyses are very informative, but if the samples in specific strata are too small, the analyses may lack precision. In planning studies, investigators must pay careful attention to potential effect modifiers. If there is a suspicion that an association between an exposure and risk factor is different in specific groups, then the study must be designed to ensure sufficient numbers of participants in each of those groups. Sample size formulas must be used to determine the numbers of subjects required in each stratum to ensure adequate precision or power in the analysis.

References

[1] Kleinbaum D, Kupper LL, Muller KE. *Applied Regression Analysis and Other Multivariable Methods*. PWS-Kent, Boston, MA, 2nd edition, 1988.

[2] Jewell NP. *Statistics for Epidemiology*. New York, NY: Chapman and Hall/CRC. 2004.

[3] Hosmer D, Lemeshow S. Applied Logistic Regression. New York: NY: John Wiley & Sons, Inc. 1989.

[4] Anderson M, Wilson PW, Odell PM, Kannell WB. An updated coronary risk profile: A statement for health professionals. *Circulation*. 1991; 83: 356-362.

[5] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998; 97: 1837-1847.

[6] SAS version 9.1© 2002-2003 by SAS Insitute, Inc., Cary, NC.

[7] Aschengrau A, Seage GR. *Essentials of Epidemiology for Public Health*. Sudbury, MA: Jones and Bartlett Publishers, Inc., 2006.

[8] Goldberg JI, Borgen PI. Breast cancer susceptibility testing: Past, present and future. *Expert Review of Anticancer Therapy*. 2006; 6(8): 1205-1214.

[9] Meigs JB, Hu FB, Rifai N, Manson J. Biomarkers of endothelial dysfunction and risk of Type 2 Diabetes Mellitus. Journal of the American Medical Association. 2004; 291: 1978-1986.