CHAPTER 2

Background knowledge

Obesity is a chronic disease that is increasing in prevalence in adults, adolescents, and children, and is now considered to be a global epidemic. Obesity is associated with a significant increase in mortality and with risk of many disorders, including diabetes mellitus, hypertension, dyslipidemia, heart disease, stroke, sleep apnea, cancer, and others. This chapter is a review about obesity and biomarkers related to body composition. Moreover, statistical methods and the literature of previous studies which are related with this study will help to have more understanding about methodology and the results of this study.

2.1 Basic knowledge about obesity

2.1.1 History of obesity (History and indication of obesity)

In the mid-1990s, WHO responded to the growing obesity epidemic throughout the world by conducting an expert consultation on obesity, or the International Obesity Task Force (IOTF) in June 1997 in Geneva and the report was published in 1998 (11).

Obesity develops when energy intake exceeds energy expenditure. Although the number of fat cells can increase throughout life, individuals with adult-onset obesity in general exhibit increased adipocyte size, whereas individuals with early-onset obesity have both adipocyte hypertrophy and hyperplasia. Fat distribution also plays an important role in metabolic risk since increased intra-abdominal/visceral fat promotes a high risk of metabolic disease, whereas increased subcutaneous fat in the thighs and hips exerts little or no risk.

The past two decades have revealed the role of factors controlling food intake and energy expenditure in body weight regulation and on the transcriptional control and cell biology underlying conversion of pre-adipocytes to adipocytes. Little is known, however, about the developmental origins of adipose tissue; the control of brown versus white pre-adipocyte commitment; the control of the relative amounts and functional heterogeneity among white fat cells in different depots; and the exact pathways and intermediates between the embryonic stem cell and the mature fat cell (12).

2.1.2 Causes of obesity

The balance between calorie intake and energy expenditure determines a person's weight. If a person eats more calories than he or she burns (metabolizes), the person gains weight (the body will store the excess energy as fat). If a person eats fewer calories than he or she metabolizes, he or she will lose weight. Therefore, the most common causes of obesity are overeating and physical inactivity. Ultimately, body weight is the result of genetics, metabolism, environment, behavior, and culture (13).

1) Genetics. A person is more likely to develop obesity if one or both parents are obese. Genetics also affect hormones involved in fat regulation. For example, one genetic cause of obesity is leptin deficiency. Leptin is a hormone produced in fat cells and also in the placenta. Leptin controls weight by signaling the brain to eat less when body fat stores are too high. If, for some reason, the body cannot produce enough leptin or leptin cannot signal the brain to eat less, this control is lost, and obesity occurs. The role of leptin replacement as a treatment for obesity is currently being explored.
2) Overeating. Overeating leads to weight gain, especially if the diet is high in fat. Foods high in fat or sugar (for example, fast food, fried food, and sweets) have high energy density (foods that have a lot of calories in a small amount of food). Epidemiologic studies have shown that diets high in fat contribute to weight gain.

3) A diet high in simple carbohydrates. The role of carbohydrates in weight gain is not clear. Carbohydrates increase blood glucose levels, which in turn stimulate insulin release by the pancreas, and insulin promotes the growth of fat tissue and can cause weight gain. Some scientists believe that simple carbohydrates (sugars, fructose, desserts, soft drinks, beer, wine, etc.) contribute to weight gain because they are more rapidly absorbed into the bloodstream than complex carbohydrates (pasta, brown rice, grains, vegetables, raw fruits, etc.) and thus cause a more pronounced insulin release after meals than complex carbohydrates. This higher insulin release, some scientists believe, contributes to weight gain.

4) Frequency of eating. The relationship between frequency of eating (how often you eat) and weight is somewhat controversial. There are many reports of overweight people eating less often than people with normal weight. Scientists have observed that people who eat small meals four or five times daily, have lower cholesterol levels and lower and/or more stable blood sugar levels than people who eat less frequently (two or three large meals daily). One possible explanation is that small frequent meals produce stable insulin levels, whereas large meals cause large spikes of insulin after meals.

5) Slow metabolism. Women have less muscle than men. Muscle burns (metabolizes) more calories than other tissue (which includes fat). As a result, women have a slower metabolism than men, and hence, have a tendency to put on more weight than men, and weight loss is more difficult for women. As we age, we tend to lose muscle and our metabolism slows; therefore, we tend to gain weight as we get older particularly if we do not reduce our daily caloric intake.

6) Physical inactivity. Sedentary people burn fewer calories than people who are active. The National Health and Nutrition Examination Survey (NHANES) showed that physical inactivity was strongly correlated with weight gain in both gender.

7) Medications. Medications associated with weight gain include certain antidepressants (medications used in treating depression), anticonvulsants (medications used in controlling seizures such ascarbamazepine [Tegretol, Tegretol XR , Equetro, Carbatrol] and valproate [Depacon, Depakene], diabetes medications (medications used in lowering blood sugar such as insulin, sulfonylureas, and thiazolidinediones), certain hormones such as oral contraceptives and most corticosteroids such as prednisone. Weight gain may also be seen with some high blood pressure medications and antihistamines. The reason for the weight gain with the medications differs for each medication and should be discussed the physician rather than discontinuing the medication, as this could have serious effects.

8) Psychological factors. For some people, emotions influence eating habits. Many people eat excessively in response to emotions such as boredom, sadness, stress, or anger. While most overweight people have no more psychological disturbances than normal weight people, about 30% of the people who seek treatment for serious weight problems have difficulties with binge eating.

9) Diseases such as hypothyroidism, insulin resistance, polycystic ovary syndrome, and Cushing's syndrome are also contributors to obesity.

10) Ethnicity. Ethnicity factors may influence the age of onset and the rapidity of weight gain. African-American women and Hispanic women tend to experience weight gain earlier in life than Caucasians and Asians, and age-adjusted obesity rates are higher in these groups. Non-Hispanic black men and Hispanic men have a higher obesity rate then non-Hispanic white men, but the difference in prevalence is significantly less than in women.

11) Childhood weight. A person's weight during childhood, the teenage years, and early adulthood may also influence the development of adult obesity. For example, being mildly overweight in the early 20s was linked to a substantial incidence of obesity by age 35; being overweight during older childhood is highly predictive of adult obesity, especially

if a parent is also obese; being overweight during the teenage years is even a greater predictor of adult obesity.

12) Hormones. Women tend to gain weight especially during certain events such as pregnancy, menopause, and in some cases, with the use of oral contraceptives. However, with the availability of the lower-dose estrogen pills, weight gain has not been as great a risk.

2.1.3 Current obesity situation

Worldwide

The epidemic of obesity is now recognized as one of the most important public health problems facing the world today. Tragically, adult obesity is more common globally than under-nutrition. There are around 475 million obese adults with over twice that number overweight that means around 1.5 billion adults are too fat. Over 200 million school-age children are overweight, making this generation the first predicted to have a shorter lifespan than their parents (12).

Worldwide, at least 2.8 million people die each year as a result of being overweight or obese, and an estimated 35.8 million (2.3%) of global DALYs are caused by overweight or obesity. Overweight and obesity lead to adverse metabolic effects on blood pressure, cholesterol, triglycerides and insulin resistance. Risks of coronary heart disease, ischemic stroke and type 2 diabetes mellitus increase steadily with increasing body mass index (BMI), a measure of weight relative to height. Raised body mass index also increases the risk of cancer of the breast, colon, prostate, endometrium, kidney and gall bladder. Mortality rates increase with increasing degrees of overweight, as measured by body mass index. To achieve optimum health, the median body mass index for an adult population should be in the range of 21 to 23 kg/m^2, while the goal for individuals should be to maintain body mass index in the range 18.5 to 24.9 kg/m^2. There is increased risk of co-morbidities for body mass index 25.0 to 29.9, and moderate to severe risk of co-morbidities for body mass index greater than 30.

In 2008, 35% of adults aged 20+ were overweight (BMI ≥ 25 kg/m^2) (34% men and 35% of women). The worldwide prevalence of obesity has nearly doubled between 1980 and 2008. In 2008, 10% of men and 14% of women in the world were obese (BMI ≥30 kg/m^2), compared with 5% for men and 8% for women in 1980. An estimated 205 million men and 297 million women over the age of 20 were obese a total of more than half a billion adults worldwide.

The prevalence of overweight and obesity were highest in the WHO Regions of the Americas (62% for overweight in both gender, and 26% for obesity) and lowest in the WHO Region for South East Asia (14% overweight in both gender and 3% for obesity). In the WHO Regions for Europe Eastern Mediterranean, the Americas over 50% of women were overweight. For all three of these regions, roughly half of overweight women are obese (23% in Europe, 24% in the Eastern Mediterranean, 29% in the Americas). In all WHO regions women were more likely to be obese than men. In the WHO regions for Africa, Eastern Mediterranean and South East Asia, women had roughly doubled the obesity prevalence of men.

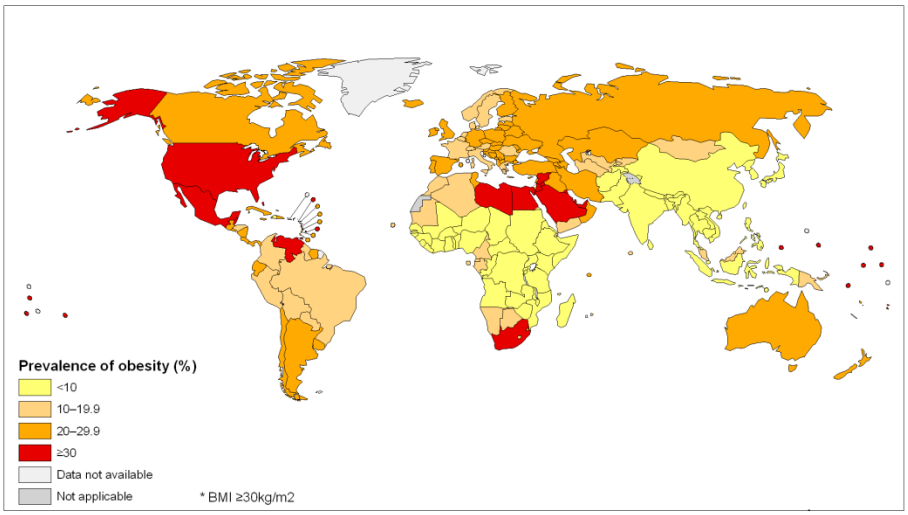

Figure 2.1 Worldwide prevalence of obesity adults (> 20 years) in 2008
Source: Public Health Information and Geographic Information Systems (GIS)
World Health Organization

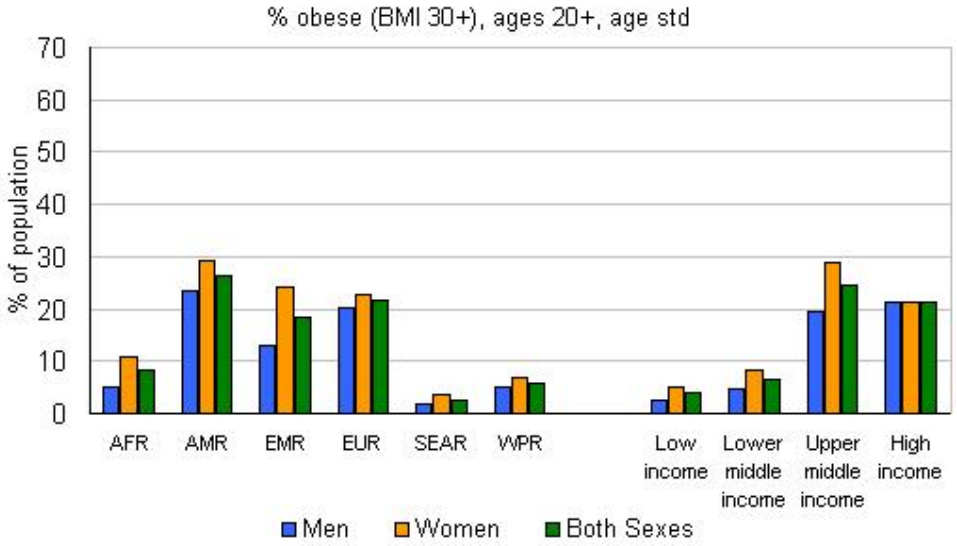

Figure 2.2 The prevalence of obesity classified by gender and income level of countries in 2008
Source: http://www.who.int/gho/ncd/risk_factors/obesity_text/en/

The prevalence of raised body mass index increases with income level of countries up to upper middle income levels. The prevalence of overweight in high income and upper middle income countries was more than double that of low and lower middle income countries. For obesity, the difference more than triples from 7% obesity in both gender in lower middle income countries to 24% in upper middle income countries. Women's obesity was significantly higher than men's, with the exception of high income countries where it was similar. In low and lower middle income countries, obesity among women was approximately double that among men.

Thailand

The incidence of obesity in Thailand is already significantly higher than in most other countries in the region, and worse is yet to come. The potential magnitude of the problem has been recognized by Thai health experts, and some small-scale or experimental remedial programs have been initiated. Those who wield the power to effect the necessary change may, however, be much slower to comprehend the significance of this burgeoning obesity crisis. Even when they do, they will invariably struggle to develop and implement an appropriate response. Failure to act quickly and decisively in addressing

this issue will incur substantial social and economic costs for the Thai community.

Obesity is a huge problem in many countries around the world, and Thailand ranks in the top five Asia-Pacific nations in this regard (14). In the period 2005-2007, obesity rates in Thailand increased from 10 million in 2005 to 17 million in 2007. Since then, and despite further research and some small-scale treatment programs, the incidence of obesity has only accelerated. Furthermore, these increases are now occurring across many demographic groups, and in both urban and rural areas.

With respect to childhood obesity, statistics from Thailand's Ministry of Public Health paint a troubling picture. In the past five years, the percentage of obese preschoolers rose from 5.8 per cent to 7.9 per cent, whilst in school-age children the obesity rate went from 5.8 per cent up to 6.7 per cent. These figures represent obesity growth rates of 36 per cent in pre-school age and 15 per cent in school age. Among Thailand's young adults (those in the 20 to 29 age range), the obesity rate is over the same period increased by 36 per cent among men and 47 per cent for women (15).

Figure 2.3 Overweight prevalence (%) in Southeast Asia for adults of both gender (BMI of >25kg/m^2)
Source: WHO Non-Communicable Diseases Country Profiles, 2011

The prevalence of obesity in Thailand, as in other countries with fast growing economy, has been increasing in an alarming rate. In Thailand, a series of National Health Examination Surveys were conducted consecutively in 1991, 1997 and 2004 where each survey was a national representative cross-

sectional survey. The present study reports on evidence of secular trend of obesity in Thailand using data from National Health Examination Survey I–III (16,17). These surveys were mainly targeted on adult population; however, the second survey also included a sample of children population. The data on measurements of weight and height provide an opportunity to examine the trends of obesity in Thai population.

All the surveys used standardized measurement of anthropometry at time of the survey period. Overall, age-adjusted mean body mass index (BMI) in Thai adults aged 18 years increased from 22.0 kg/m^2 in 1991 to 22.7 kg/m^2 in 1997 and 23.2 kg/m^2 in 2004. This suggests that the distribution of weight and BMI of the entire population is shifting to the right.

The prevalence rates of obesity were determined using BMI cut-off points at ≥25 kg/m^2 (18) and ≥30 kg kg/m^2 (19). The prevalence of obesity with BMI ≥25 kg/m^2 in adults increased dramatically from 18.2% in 1991 to 24.1% in 1997 and 28.1% in 2004. For those with BMI ≥30 kg/m^2, the prevalence increased from 3.5% to 5.8% and 6.9% in the corresponding years (17). Overall, the prevalence was higher in women than in men especially in the middle-age group. The prevalence of obesity with BMI ≥25 kg/m^2 increased in both gender, from 13.0% in men and 23.2% in women in 1991 to 18.6% and 29.5% in 1997 and 22.4% and 34.3% in 2004 respectively.

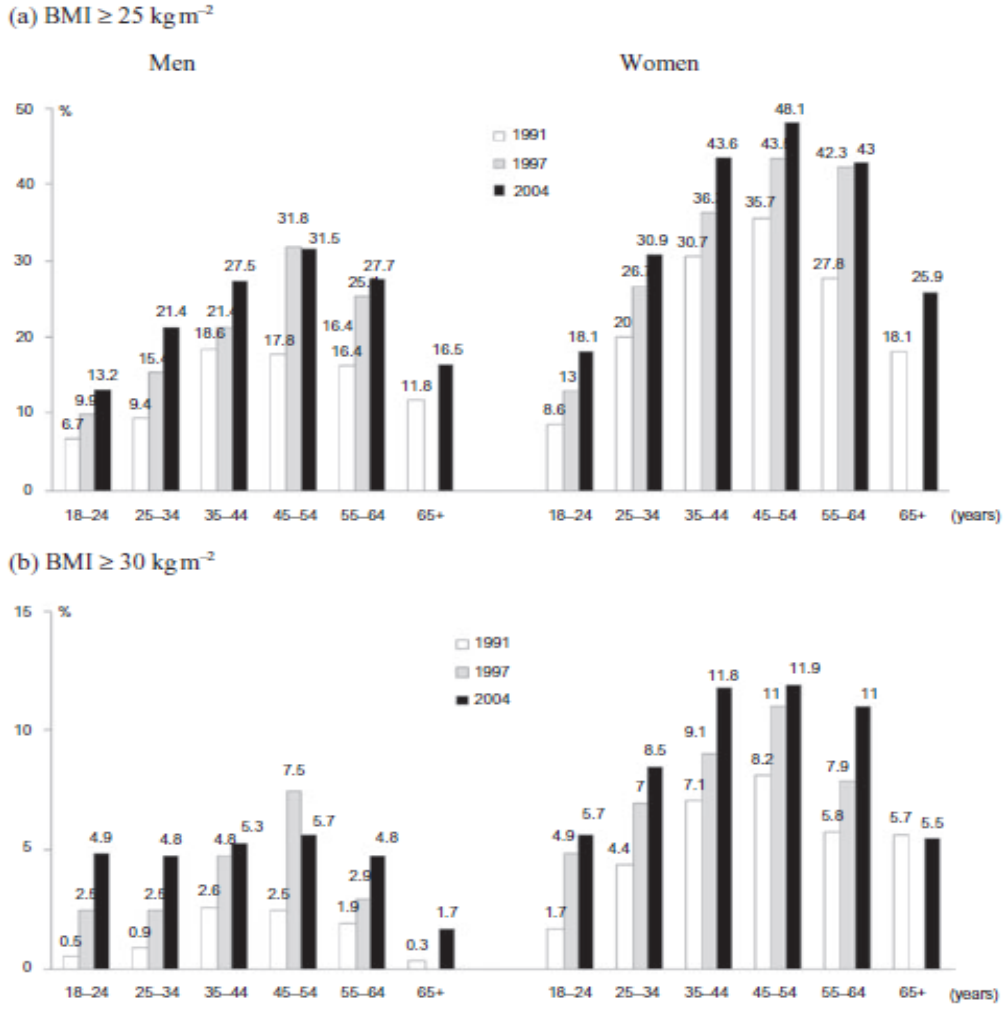

Figure 2.4 Trends of age-standardized prevalence of obesity with body mass index (BMI) ≥25 kg/m^2 and ≥30 kg/m^2 among Thai population
Source: Thai National Health Examination Survey in 1991, 1997 and 2004

Figure 2.4 shows trends in obesity over the last two decades, the prevalence has been increasing in all age groups. The highest prevalence rates of obesity with BMI ≥25 and ≥30 kg/m^2 were in the 45–54 years age group in both men and women. More importantly, the largest increase for BMI ≥25 kg/m^2 was in the 18–34 age group with almost double increases in men and women.

2.1.4 Assessment of obesity

Obesity represents a state of excess storage of body fat. Although similar, the term overweight is puristically defined as an excess of body weight for height. Normal, healthy men have a body fat percentage of 15-20%, while normal, healthy women have a percentage of approximately 25-30% (20). Body fat

can be measured in several ways, with each body fat assessment method having pros and cons.

1) Body-fat percentage (BF%)

The BF% of a human or other living being is the total mass of fat divided by total body mass; body fat includes essential body fat and storage body fat. Essential body fat is necessary to maintain life and reproductive functions. The percentage of essential body fat for women is greater than that for men, due to the demands of childbearing and other hormonal functions. The percentage of essential fat is 3–5% in men, and 8–12% in women (21). Storage body fat consists of fat accumulation in adipose tissue, part of which protects internal organs in the chest and abdomen. The minimum recommended total BF% exceeds the essential fat percentage value reported above. There are many methods to determine BF%, such as measurement with bioelectrical impedance analysis or through the use of dual energy X-ray absorptiometry or skin fold caliper.

Measure of body-fat percentage

- Bioelectrical impedance analysis (BIA)

  The measurement of the BIA depends on the differences in electrical conductivity of fat free mass and fat. The technique measures the impedance of an electrical current passed between two electrodes (typically 800 μA; 50 kHz). For single frequency BIA, two electrodes are generally located on the right ankle and the right wrist of an individual. The impedance is related to volume of a conductor (the human body) and the square of the length of the conductor a distance which is a function of the height of the subject. BIA analysis most closely estimates body water, from which fat free mass is then estimated, on the assumption that the latter contains about 73% water. Fat mass can then be derived as the difference between body weight and fat free mass (22). Because SF-BIA is not valid under conditions of significantly

altered hydration (23), therefore, before BIA, all volunteers were prepared with the following pre-test guidelines. (1) no alcohol consumption within 24 hours. (2) no exercise, caffeine or food within four hours prior to taking the test, and (3) Drinking two to four glasses of water two hours before examination. During the examination, two pairs of sensor electro-cardiograph (ECG) pads were placed on the patient, one on the right wrist and hand and the other on the right foot and ankle. At least 75% of the electrode should be in contact with the patient's skin.

In the new BIA method, multi-frequency measurements have been developed. This method allows the estimation of both total and extracellular body fluid compartments. These estimations have advantages in certain disease conditions involving disturbances in water distribution such as congestive heart disease, renal disease and malnutrition (24,25).

The errors of BIA including the measurement of height, weight, resistance, the criterion reference method used, and errors from the prediction equation which the performance depended on the selection and number of independent variables (26). However, the great advantages of BIA are safety and convenience, and the equipment is portable and relatively inexpensive. In the future, impedance spectrum analysis derived from multi-frequency BIA may be increasingly used to distinguish differences in body water, body composition among individuals as well as specific parts of the body such as muscle and adipose tissue mass in limbs (27).

Figure 2.5 Bioelectrical impedance analysis (BIA)
Source: http://www.builtlean.com/2010/07/13/5-ways-to-measure-body-fat-percentage/

- Dual energy X-ray absorptiometry (DEXA)

  DEXA is now the primary technique for the assessment of the bone mineral content of the axial skeleton but it is also used for determining the relative proportions of the fat free mass, body fat and bone in subjects by whole body scanning. DEXA scanners use a dual energy X ray source that generates X rays at 40 KeV and 70-100 KeV; these pass through the subject. The relative absorption at these two energies is measured to give two estimates of body composition along the beam path using a two compartment model. In bone free regions of the body, the attenuation provides an estimate of the relative proportions of fat and lean tissues. In the other regions, the attenuation provides a measure of the proportions of bone and soft tissues. To provide estimates of the overall relative proportions of the three components – the fat free mass, body fat and bone – the assumption is made that the soft tissue overlaying bone has the same fat to muscle ratio as that in immediately adjacent non-bone regions[4]. However, the computing algorithms used to partition the soft tissue between the body fat and the fat free mass are critically important in assessing body composition and have been shown to

vary significantly with the manufacturer of the equipment. Such algorithms should take into account the different fat distributions in men and women and also the differences generated by overall increases in adiposity. At present, Fan beam technologies, replacing earlier pencil beam techniques, are resulting in a much shorter scan time, lower X ray doses and improved geometrical resolution as well as have a high precision with accurate results (28).

Figure 2.6 Dual energy X-ray absorptiometry (DEXA)
Source: http://www.builtlean.com/2010/07/13/5-ways-to-measure-body-fat-percentage/

- Skin fold caliper

The skinfold estimation methods are based on a skinfold test, also known as a pinch test, whereby a pinch of skin is precisely measured by calipers at several standardized points on the body to determine the subcutaneous fat layer thickness (29,30). These measurements are converted to an estimated body fat percentage by an equation. Some formulas require as few as three measurements, others as many as seven. The accuracy of these estimates is more dependent on a person's unique body fat distribution than on the number of sites measured. As well, it is of utmost importance to test in a precise location with a fixed pressure. Although it may not give an accurate reading of real body fat percentage, it is a reliable measure of body composition

change over a period of time, provided the test is carried out by the same person with the same technique.

The skinfold thickness measurements provide an estimate of the size of the subcutaneous fat deposit which provides an estimate of the total body fat mass. Such estimations are based on two assumptions. First, the thickness of the subcutaneous adipose tissue reflects a constant proportion of the total body fat. Second, the skin fold sites selected for measurement, either single site or combination might represent the average thickness of the entire subcutaneous adipose tissue. However, neither of these is true. In fact, the relationship between subcutaneous and internal fat is nonlinear and varies with body weight and age. In addition, variations in the distribution of subcutaneous fat occur with gender, race or ethnicity and age (31). The following sites of skinfold thickness measurements are commonly used (32,33).

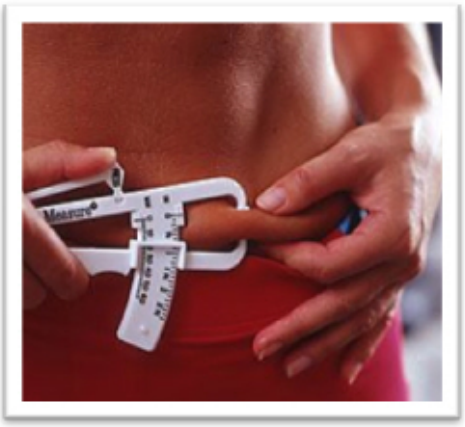

Figure 2.7 Skin fold caliper
Source: http://www.builtlean.com/2010/07/13/5-ways-to-measure-body-fat-percentage/

2) Body mass index (BMI)

BMI is a person's weight in kilograms divided by the square of height in meters. BMI does not measure body fat directly, but research has shown that BMI is moderately correlated with more direct measures of body fat obtained from skinfold thickness measurements, bioelectrical impedance, densitometry (underwater weighing), dual energy x-ray

absorptiometry (DXA) and other methods. Furthermore, BMI appears to be as strongly correlated with various metabolic and disease outcome as are these more direct measures of body fatness. In general, BMI is an inexpensive and easy-to-perform method of screening for weight category, for example underweight, normal or healthy weight, overweight, and obesity.

The BMI cut points recommended from the 1998 WHO consultation on obesity were the first such cut-off points at the international level. Although they have been generally accepted, a number of countries and regions have questioned the relevance of the public health cut-off points to their respective situations. This has been particularly so in the Asia and Pacific regions. It has been amply demonstrated that Asians in general, although not consisting of a homogeneous population, have a higher body-fat percentage at a given BMI than Caucasians. They also have a higher waist-to-hip ratio than Caucasians and a more centralized distribution of body fat. Perhaps of most concern, morbidity and mortality among Asians are occurring in people with lower BMIs and smaller waist circumference. On the other hand, Pacific Islanders tend to be larger and more muscular, with less body fat at higher BMI levels[11].

Although BMI is a useful measurement across populations, it is increasingly apparent that BMI has significant limitations in the assessment of the individual as it does not take into account the distribution of body fat. BMI measurement does not provide any information regarding where body fat is stored (34).

3) Height weight difference index (HWDI)

Body mass index (BMI) is the most common index for assessing weight status of adults, at both individual and population levels. However, calculating BMI without an instrument is quite difficult and time consuming. Thus, the new index Height Weight Difference Index (HWDI) by Sakda Pruenglampoo, et al (3,6) was developed HWDI was calculated using the formula: height (cm) - weight (kg). The researchers

found that the figures of HWDI can be used for predicting underweight, normal weight, overweight and obesity. Nutritional status of the subjects assessed by HWDI were compared with those assessed by BMI. Then the percentages of sensitivity and specificity were calculated. The kappa statistic was used to measure agreement between the assessment of nutritional status by HWDI and by BMI. It may be inferred that HWDI might not be suitable index for screening thin adults from those who have normal nutritional status. However, the study findings suggested that HWDI could be used as a simple and effective index for screening overweight and obesity in adults.

2.1.5 Consequences of obesity

Obesity is an important cause of morbidity, disability and premature death[1]. The health consequences of obesity are many and varied, ranging from an increased risk of premature death to several non-fatal but debilitating complaints that can have a marked effect on the quality of life. It is a major risk factor for:

1) Cardiovascular disease

Obesity predisposes an individual to a number of cardiovascular risk factors including hypertension, raised cholesterol and impaired glucose tolerance. However, long term prospective data now suggest that obesity is also important as an independent risk factor for CHD related morbidity and mortality (35). The Framingham Heart Study ranked body weight as the third most important predictor of CHD among men, after age and dyslipidaemia. Similarly, in women, a large scale prospective study in USA found a positive correlation between BMI and the risk of developing CHD. Weight gain substantially increased this risk (36).

2) Hypertension and stroke

The association between hypertension and obesity is well documented. Both systolic and diastolic blood pressure increase with BMI, and the obese are at higher risk of developing hypertension than lean

individuals. Community-wide surveys in USA (NHANES II) show that the prevalence of hypotension in overweight adults in those aged 20-44 years is 5.6 times greater than that in those aged 45-74 years old, which in turn is twice as high as that for non-overweight adults. The risk of developing hypertension with the duration of obesity, especially in women, and weight reduction leads to fall in blood pressure (19).

3) Cancer

A number of studies have found a positive association between overweight and the incidence of cancer, particularly of hormone dependent and gastrointestinal cancers. Greater risks of endometrial, ovarian, cervical and postmenopausal breast cancer have been documented for obese women, while there is some evidence for an increased risk of prostate cancer among obese men. The increased incidence of these cancers in the obese is greater in those with excess abdominal fat and is thought to be a direct consequence of hormonal change (37). The incidence of gastrointestinal cancers, such as colorectal and gallbladder cancer, has also been reported to be positively associated with body weight or obesity in some but not all studies. And renal cell cancer has consistently been associated with overweight and obesity, especially in women (38).

4) Diabetes mellitus

A positive association between obesity and the risk of developing Non-insulin dependent diabetes mellitus (NIDDM) has been repeatedly observed in both cross-sectional (39), and prospective studies (40). The consistency of the association across population despite difference measures of fatness and criteria for diagnosing NIDDM reflects the strength of the relationship. The risk of NIDDM increases continuously with BMI and decreases with weight loss. Analysis of data from two prospective studies illustrates the impact of overweight and obesity on NIDDM; about 64% of men and 74% of women cases of NIDDM could theoretically have been prevented if no one had a BMI over 25. Detailed analyses of the relationship between obesity and NIDDM have

identified certain characteristics of obese persons that further increase the risk of developing this condition, even after controlling for age, smoking and family history of NIDDM. These include obesity during childhood and adolescence, progressive weight gain from 18 years and intra-abdominal fat accumulation (41). Lack of physical activity and an unhealthy diet, both of which are associated with lifestyle in industrialized countries, also import modifiable risk factors for overweight and obesity. The prevalence of NIDDM is 2-4 fold higher in the less physical activity individuals compared with the most physical active (42).

5) Gallbladder disease

Obesity is a risk factor for gallstones in all age group and, in both men and women, gallstones occur three to four times in obese compared with non-obese individuals and the risk is even greater when excess fat is located around the abdomen. The relative risk of gallstones increases with BMI, and data from the Nurse's Health Study suggest that even moderate overweight may increase the risk (19).

6) Pulmonary diseases

Obesity impairs respiratory function and structure, leading to physiological and pathophysiological impairments. The work of breathing is increased in obesity, mainly as a result of extreme stiffness of the thoracic cage consequent on the accumulation of adipose tissue in and around the ribs, abdomen, diaphragm. Hypoxemia is common, partly because the low relaxation volume causes ventilation to occur at volumes below the closing volume, and is exacerbated when lying down because of the reduced functional residual capacity (19).

7) Disability

In 1990, Rissanen et al showed that obese Finnish adults suffered more often than normal – work disability due to cardiovascular and musculoskeletal disease. A study of obese Swedes showed that obesity accounted for 10% of productivity loss due to sick leave or work disability and that, in particular, disability associated with waist

circumference. In addition, symptoms of osteoarthritis are more severe in heavier patients (19).

8) Mortality

Most studies report relationships between BMI and mortality. BMI comprised both fat mass and fat free mass, both affecting the risk or mortality independently (43) and in opposite directions. Waist circumference is a better alternative than BMI for identifying elderly men with an increased risk of mortality (44). There is an almost linear relationship between BMI and death. The longer the duration of obesity, the higher the risk. Severe obesity is associated with a 12-fold increase in mortality in 25-35 years old compared with lean individuals (19).

9) Reduced life expectancy

Some studies have calculated the number of reduced years of life expectancy caused by obesity. The Framingham study calculated that obesity (BMI$\geq$ 30 kg/m^2) at the age of 40 years was related to a loss of 6-7 years of life. Fontaine et al. calculated that a BMI$\geq$ 33 kg/m^2 from age 40 years was related to a loss of 2-3 years. The studies used different calculation methods and were based on different cohorts (44).

2.2 Statistical methods

2.2.1 Correlation and agreement

Many statistical analyses are conducted to study the relationship between two continuous or ordinal scale variables within a group of patients. Often several quantitative variables are measured on each member of a sample. In considering a pair of such variables, it is frequently of interest to establish if there is a relationship between the two; i.e. to see if they are correlated.

The type of correlation can be categorized by considering what happens to the other variable as one variable increases:

- Positive correlation – the other variable has a tendency to also increase.

- Negative correlation – the other variable has a tendency to decrease.
- No correlation – the other variable does not tend to either increase or decrease.

The starting point of any such analysis should thus be the construction and subsequent examination of a scatterplot. Examples of negative and positive correlation are as follows.

1) Scatter plot

A scatter plot is a simple tool for identifying relationship between two variables X and Y. It is one of the seven basic tools for quality control which is useful when examining dependency and non-linear relationships for a set of data (45). Different scatter plots are illustrated in Figure 2.8 and the linear relationship can be quantified by using Pearson's correlation under certain assumptions.

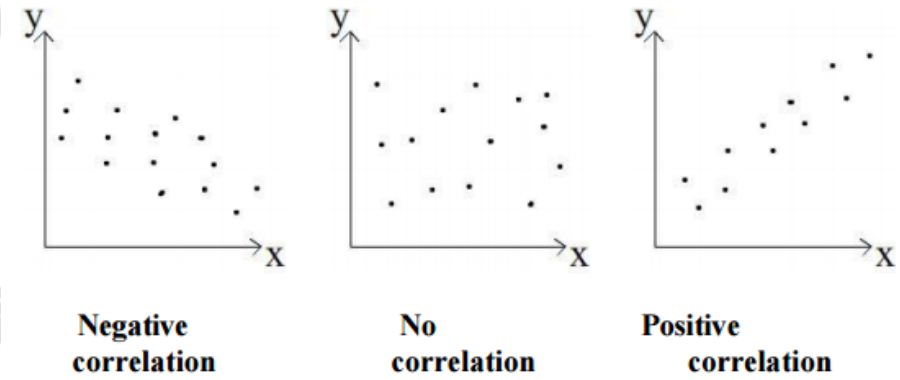

Figure 2.8 Scatter plots of different dependencies

2) Pearson's correlation coefficient

The covariance of two random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

To standardize it, is by dividing it by the standard deviation of each variable involved. This results in a coefficient called Pearson's correlation coefficient, which is the most widely known measure of

dependency since it can be easily calculated by definition $\rho X, Y$ for data population and equation r for sample data, where $\overline{X}$ and $\overline{Y}$ are the averages of X and Y variable, respectively.

$$\rho X, Y = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Pearson's correlation coefficient is a measure of linear dependence between two variables X and Y. The coefficient ρ has a range between $-1 \leq \rho \leq +1$ for the true population. Perfect positive or negative linear coefficient equals to ±1 which corresponds to data sample point lying exactly on a line. Pearson's correlation has the following properties and makes these assumptions about the variables X and Y (45,46).

Assumptions

The calculation of Pearson's correlation coefficient and subsequent significance testing of it requires the following data assumptions:

- Interval or ratio level
- Linearly related
- Bivariate normally distributed.

In practice the last assumption is checked by requiring both variables to be individually normally distributed (which is a by-product consequence of bivariate normality). Pragmatically Pearson's correlation coefficient is sensitive to skewed distributions and outliers, thus it is satisfactory without having these conditions.

2.2.2 Cohen's kappa statistic

Cohen's kappa statistic, is a measure of agreement between categorical variables X and Y. Kappa also can be used to assess the agreement between

alternative methods of categorical assessment when new techniques are under study.

Kappa is calculated from the observed and expected frequencies on the diagonal of a square contingency table. Suppose that there are n subjects on whom X and Y are measured, and suppose that there are g distinct categorical outcomes for both X and Y. Let f_{ij} denote the frequency of the number of subjects with the i*th* categorical response for variable X and the j*th* categorical response for variable Y (47,48).

Then the frequencies can be arranged in the following g × g table:

	Y=1	**Y=2**	**…**	**Y=g**
X=1	f_{11}	f_{12}	…	f_{1g}
X=2	f_{21}	f_{22}	…	f_{2g}
|	|	|	…	|
|	|	|	…	|
X=g	f_{g1}	f_{g2}	…	f_{gg}

The observed proportional agreement between X and Y is defined as:

$$P_0 = \frac{1}{n}\sum_{i=1}^{g} f_{ii}$$

and the expected agreement by chance is:

$$P_e = \frac{1}{n^2}\sum_{i=1}^{g} f_{i+}f_{+i}$$

where f_{i+} is the total for the i*th* row and f_{+i} is the total for the i*th* column. The kappa statistic is:

$$\widehat{K} = \frac{P_0 - P_e}{1 - P_e}$$

Cohen's kappa statistic is an estimate of the population coefficient:

$$K = \frac{\Pr[X = Y] - \Pr[X = Y|X \text{ and } Y \text{ independent}]}{1 - \Pr[X = Y|X \text{ and } Y \text{ independent}]}$$

Kappa is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, ie, potential systematic disagreement between the observers.

A more complete list of how Kappa might be interpreted (48) is given in the following table:

Agreement Level	K
Almost perfect	>0.80
Substantial	0.61-0.80
Moderate	0.41-0.60
Fair	0.21-0.40
Slight	0.00-0.20
Poor	<0.00

2.2.3 Diagnostic test

From a technological and procedural perspective, the diagnostic test for the classification can be relatively simple or complex. From a procedural standpoint, the test may only involve one step which results in one of only two outcomes, positive or negative, or it may involve a vast sequence of procedures that may result in one of an entire spectrum of possible classifications.

The implementation of a diagnostic test should be preconditioned on the practicality and benefit of such a test toward the classification or prediction of the diseased condition. The key criteria that should be considered before implementing a diagnostic test can be adapted from Wilson and Jungner (1968), Cole and Morrison (1980) and Obuchowski et al. (2001), who discuss criteria for useful screening programs which share similar considerations to

the application of diagnostic tests in general. The criteria pertain to the disease (first, second and third criterion), the treatment for the disease (fourth criterion) and to the test itself (fifth and sixth criterion). Firstly, the disease should be serious or potentially so as to merit its use for diagnosis to potentially improve the longevity or quality of life of the subjects. Secondly, the disease should be relatively prevalent in the target population so as to have a potential benefit from testing subjects. Thirdly, the purpose of diagnosing the disease is so that it can be treated, so the disease should be treatable. Fourthly, there must exist an effective treatment to be beneficial for those who test positive. The fifth and sixth criteria pertain to the medical test itself. The fifth criterion is that the test procedure should ideally cause no harm to the individual. However, all tests have more or less negative impact, whether it is financial, physical or emotional discomfort or damage. In practicality, these costs should be reasonably in context and the information from an accurate diagnosis should create potential benefits to be gained by the population or individual being tested. The sixth and final criterion is the accuracy of the test which is discussed in more detail in the next section.

1) Diagnostic accuracy

An accurate test is one that correctly classifies its test population according to the disease or non-disease condition. Inaccurate tests cause those with actual disease to be misclassified as non-diseased, also known as a “false negative”. Conversely, they cause those with no actual disease to be misclassified as diseased, also known as a “false positive”. False negative errors leave diseased subjects untreated. False positive errors open subjects to being subjected to unnecessary procedures and emotional stress. Both false negatives and false positives may also create disillusionment and distrust within the general subjects towards the medical and diagnostic testing community as a whole, potentially making data collection more difficult, biased and costly. Obviously, such errors must be kept to a minimum. As such, the diagnostic accuracy of a test is of utmost importance and must be

thoroughly assessed and understood before such a test can be used in practice.

In order to effectively implement and assess a diagnostic test, the test population the test itself and the resulting observations for many factors which may influence the analysis of the accuracy by applying statistical methodologies must be thoroughly evaluated. The population taking the tests are not influenced by knowledge of their true disease classifications or that the test itself is not influenced by knowledge of the same which could alter the accuracy of the diagnostic test. The persons administering and assessing the results of the test should also be blind to the population's true disease classifications so as not to influence the test results. These situations are more common when assessing more subjective factors of a study.

Many other factors can affect the performance of a diagnostic test for the purpose of detecting disease. These include biased test populations that are not representative of diseased subjects in the general population, inadequate clinical samples that may affect the results of the test, a condition of a repeat testing that results in a positive diseased status which may be counted as tested once rather than twice, the time it takes between when the test is administered and when the results are assessed, patient related factors (demographics, health habits, truthfulness), tester related factors (training, experience), environmental factors (available resources, treatment options, integrity of reporting), etc.

In some cases, statistical methodologies may be enhanced and improved to generate significantly more accurate classification predictions. In other cases, a procedurally simpler statistical methodology may prove to be relatively more efficient than other methodologies, without sacrificing accuracy, especially for computation-heavy studies or for cases in which time is of the essence.

2) Sensitivity and specificity

Sensitivity and specificity are two basic measures of diagnostic accuracy. The two definitions using the following contingency table, Table 1.1 can be illustrated. Firstly, the true condition status by the indicator variable T is denoted, where

$$T = \begin{cases} 1 \text{ with condition} \\ 0 \text{ without condition} \end{cases}$$

The result of the diagnostic test is denoted by the indicator variable X. Test results indicating the condition's presence are called positive, denoted as X = 1, whereas those indicating the condition's absence are called negative, denoted as X = 0, where

$$X = \begin{cases} 1 \text{ postive test results} \\ 0 \text{ negative test results} \end{cases}$$

Table 2.1 illustrates a basic count table specifying the different numbers under different categories. The total numbers with and without the condition are n_1 and n_0, respectively. The total numbers with the condition whose test result is positive and negative are p_1 and p_0, respectively. The total numbers without the condition whose test result is positive and negative are a_1 and a_0, respectively. The total number in the study is N, where $N = p_1 + p_0 + a_1 + a_0$.

Table 2.1 A basic count table

True condition status	Test results		Total
	Positive (X=1)	Negative (X=0)	
Present (T=1)	p_1	p_0	n_1
Absent (T=0)	a_1	a_0	n_0
Total	m_1	m_0	N

The sensitivity (Se) is the test's ability to detect the condition when the condition is present. The sensitivity is the probability that the test result

is positive(X = 1), given the presence of the condition (T = 1), written as

$$\mathrm{Se} = \mathrm{P}(\mathrm{X} = 1|\mathrm{T} = 1)$$

In Table 2.1, among n1 numbers with the condition, p_1 test positive. So, S e = p_1/n_1.

The specificity (Sp) is the test's ability to exclude the condition without the condition. It is the probability that the test result is negative(X = 0), given the absence of the condition (T = 0), written as

$$\mathrm{Sp} = \mathrm{P}(\mathrm{X} = 0|\mathrm{T} = 0)$$

In Table 2.1, among n_0 numbers with the condition, a_0 test positive. Thus, Sp = a_0/n_0.

The data by probabilities, as shown in Table 2.2 are summarized. The consequences associated with the test results are also considered. The test can have two types of errors. One is false positive errors and another one is false negative errors. The true positive fractions (TPF) and false positive fractions (FPF) are defined as follows:

$$\text{false positive fraction} = \mathrm{FPF} = \mathrm{P}(\mathrm{X} = 1|\mathrm{T} = 0)$$

$$\text{true positive fraction} = \mathrm{TPF} = \mathrm{P}(\mathrm{X} = 1|\mathrm{T} = 1)$$

False negative fraction (FNF) is 1-TPF. True negative fraction (TNF) is 1-FPF. The following table illustrates the relationship between them by probabilities.

Table 2.2 Probability table

True condition status	Test results		Total
	Positive (X=1)	Negative (X=0)	
Present (T=1)	Se = p_1/n_1	FNF = p_0/n_1	1.0
Absent (T=0)	FPF = a_1/n_0	Sp = a_0/n_0	1.0

In this usage, sensitivity is known as the TPF and specificity is known as TNF. Under various applications, the terminology for TPF and FPF is often different. In biomedical research, the “sensitivity” (TPF) and “specificity” (1-FPF) are often descriptors of test performance. In engineering and audiology, the terminologies “hit rate” (TPF) and “false alarm rate” (FPF) are often used. In statistical hypothesis testing, the terms ‘significance level’ (FPF) and “statistical power” (TPF) are often used.

3) The Receiver Operating Characteristic curve (ROC curve) (49)

An ROC curve is a plot of the sensitivity of a test which is plotted on the y axis versus the test’s FPF which is plotted on the x axis. Different decision thresholds can generate different points on the graph. Line segments are often used to connect the points from different possible decision thresholds, forming an empirical ROC curve. The diagonal line is called a chance diagonal.

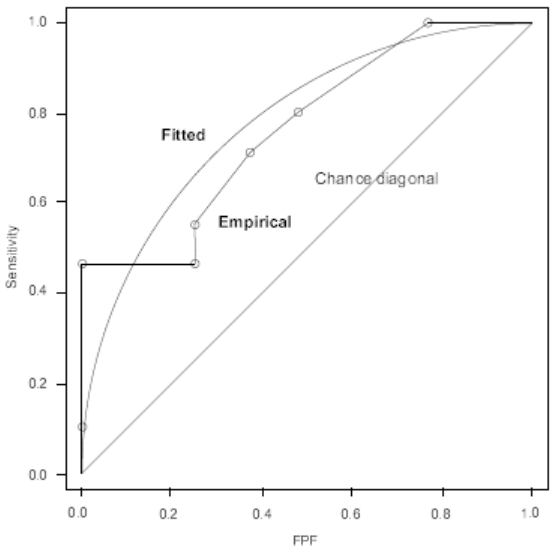

Figure 2.9 An example of an ROC curve

Figure 2.9 illustrates an example of an ROC curve. In this figure, each circle on the empirical ROC curve represents a (FPF, Se) point corresponding to a particular decision threshold. There are seven decision thresholds which provide (FPF, Se) points in addition to the

two points, (0,0) and (1,1). Line segments connect all the points generated from the seven possible decision thresholds and then form empirical ROC curve. It is also convenient to connect all the possible points using a smooth curve which is called a fitted ROC curve, illustrated in Figure 2.9.

Tests are usually ordinal in nature. For example, the clinical symptoms in medical research are often classified as severe, moderate, mild and not present. But it is often convenient to use a statistical model to fit the test results. Now the continuous ROC curves are discussed. A threshold r to define a binary test from the continuous test result X is used as

$$\text{positive if } X \geq r$$

$$\text{negative if } X < r$$

The corresponding true positive fraction at the threshold r TPF(r) and false positive fraction at the threshold r FPF(r) are defined as

$$TPF(r) = P(X \geq r | T = 1)$$

$$FPF(r) = P(X \geq r | T = 0)$$

The ROC plot has many advantages compared to other measures of accuracy. An ROC curve can visually represent the data's accuracy. The scales of the ROC curve plot are two basic measures of accuracy which can be easily read from the plot. The ROC curve includes all the possible decision thresholds so that there is no requirement to select a particular decision threshold. Because sensitivity and specificity are independent of prevalence, the ROC curve is independent of prevalence as well. The ROC curve is also independent of the scale of the test results. That is, the ROC curve does not vary to any monotonic (e.g., linear, logarithmic) transformations of the test results, which is a useful property. Another advantage of the ROC curve is that it can provide a direct and visual comparison of two or more tests on a single set of

scales. It is possible to compare different tests at all decision thresholds by constructing the ROC curves (49-55).

The Area Under the Receiver Operating Characteristic curve (AUROC) summarizes the entire location of the ROC curve rather than depending on a specific operating point. The AUROC is an effective and combined measure of sensitivity and specificity that describes the inherent validity of diagnostic tests (49).

4) The Youden's index

The Youden's index (J), is the difference between the true positive rate and the false positive rate. Maximizing this index allows to find, from the ROC curve, an optimal cut-off point independently from the prevalence. According to its definition and as illustrated on Figure 2.10, J is the vertical distance between the ROC curve and the first bisector (or chance line). If F(x) is the function describing the ROC curve, with x = 1-specificity, it can be written as

$$J(x) = F(x)-x$$

When J is maximal, J'(x) = 0, where J' is the derivative of J.

$$J'(x) = F'(x) -1,$$

where F' is the derivative of F.

Hence, when J is maximal, F'(x) = 1, meaning that the tangent to the ROC curve is parallel to the first bisector (slope = 1). It implies that, around this point, a gain (or a loss) in specificity results in a loss (or a gain) of the same amplitude in sensitivity (56).

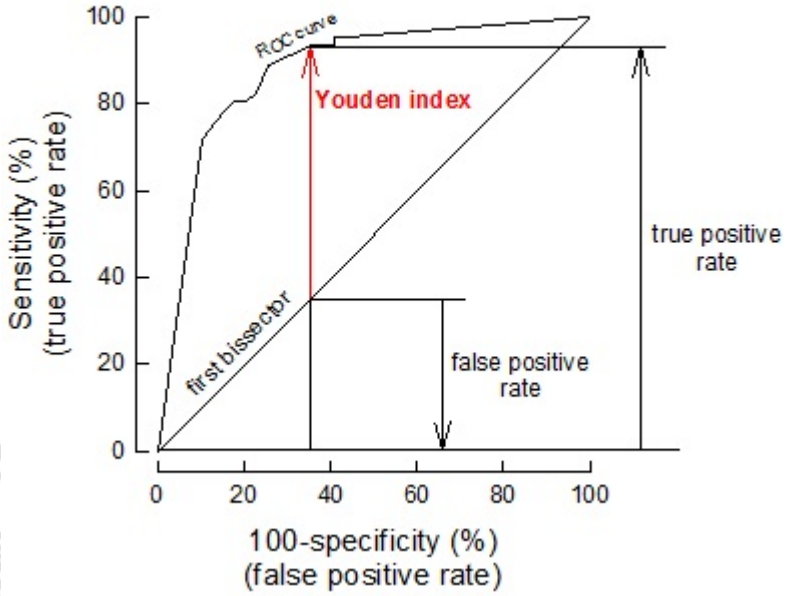

Figure 2.10 Receiver Operating Characteristic curve. Solid red: ROC curve; Dashed line: Chance level; Vertical line (J) maximum value of Youden's index for the ROC curve

Source: Michils A, et al. Exhaled nitric oxide as a marker of asthma control in smoking patients. The European respiratory journal. 2009;33(6):1295-301.

2.2.4 Regression analysis

Regression is a statistical technique to determine the linear relationship between two or more variables.

1) Linear regression (57)

Linear regression is primarily used for prediction and causal inference. In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The magnitude and direction of that relation are given by the slope parameter (β_1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (β_0). An error term (ε) captures the amount of variation not predicted by the

slope and intercept terms. The regression coefficient (R^2) shows how well the values fit the data.

Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. For example, a regression with shoe size as an independent variable and foot size as a dependent variable would show a very high regression coefficient and highly significant parameter estimates, but it cannot be concluded that higher shoe size causes higher foot size. All that the mathematics can tell us is whether or not they are correlated, and if so, by how much.

It is important to recognize that regression analysis is fundamentally different from ascertaining the correlations among different variables. Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail (57).

Simple linear regression model is a special case of multiple linear regression model which involves in more than one predictor variable. i.e., multiple linear regression model describes the relation of response variable y with a number of predictor variables, say X_1, X_2, ..., X_p. The model can be formulated as the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Selection process for multiple regression

The basis of a multiple linear regression is to assess whether one continuous dependent variable can be predicted from a set of independent (or predictor) variables. Or in other words, how much variance in a continuous dependent variable is explained by a set of predictors. Certain regression selection approaches are helpful in testing predictors, thereby increasing the efficiency of analysis.

- Entry method is simultaneous (the enter method); all independent variables are entered into the equation at the same time. This is an appropriate analysis when dealing with a small set of predictors and when the researcher does not know which independent variables will create the best prediction equation. Each predictor is assessed as though it were entered after all the other independent variables were entered, and assessed by what it offers to the prediction of the dependent variable that is different from the predictions offered by the other variables entered into the model.
- Forward selection begins with an empty equation. Predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable. Variables of greater theoretical importance are entered first. Once in the equation, the variable remains there.
- Backward elimination (or backward deletion) is the reverse process. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.
- Stepwise selection is considered a variation of the previous two methods. Stepwise selection involves analysis at each step to determine the contribution of the predictor variable entered previously in the equation. In this way it is possible to understand the contribution of the previous variables now that another variable has been added. Variables can be retained or deleted based on their statistical contribution.

Least squares estimation of model parameters

In practice, the parameters β_0 and β_1 are unknown and must be estimated. One widely used criterion is to minimize the error sum of squares:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \rightarrow \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

This is done by calculus, by taking the partial derivatives of Q with respect to β_0 and β_1 and setting each equation to 0. The values of β_0 and β_1 that set these equations to 0 are the **least squares estimates** and are labeled b_0 and b_1.

First, take the partial derivatives of Q with respect to β_0 and β_1:

$$\frac{\partial Q}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-1)$$

$$\frac{\partial Q}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-X_i)$$

Next, set these 2 equations to 0, replacing β_0 and β_1 with b_0 and b_1 since these are the values that minimize the error sum of squares:

$$-2 \sum_{i=1}^{n} (Y_i - b_0 + b_1 X_i) = 0 \rightarrow \sum_{i=1}^{n} Y_i = n b_0 + b_1 \sum_{i=1}^{n} X_i$$

$$-2 \sum_{i=1}^{n} (Y_i - b_0 + b_1 X_i) X_i = 0 \rightarrow \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

These two equations are referred to as the **normal equations** (although, note that nothing is said YET, about normally distributed data).

Solving these two equations yields:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{(X_i - \bar{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \sum_{i=1}^{n} \left[\frac{1}{n} - \bar{X} k_i \right] Y_i = \sum_{i=1}^{n} l_i Y_i$$

where k_i and l_i constants, and Y_i is a random variable with mean and variance given above:

$$k_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$l_i = \frac{1}{n} - \overline{X}k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

The **fitted** regression **line**, also known as the **prediction equation** is:

$$\hat{Y} = b_o + b_1 X$$

The **fitted values** for the individual observations are obtained by plugging in the corresponding level of the predictor variable (X_i) into the fitted equation. The **residuals** are the vertical distances between the **observed values** (Y_i) and their **fitted values** ($\hat{Y}_i$), and are denoted as e_i.

$$\hat{Y}_i = b_o + b_1 X_i \qquad e_i = Y_i - \hat{Y}_i$$

Properties of the fitted regression line

- $\sum_{i=1}^{n} e_i = 0$ The residuals sum to 0
- $\sum_{i=1}^{n} X_i e_i = 0$ The sum of the weighted (by X) residuals is 0
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ The sum of the weighted (by $\hat{Y}$) residuals is 0
- The regression line goes through the point $(\overline{X}, \overline{Y})$

These can be derived via their definitions and the normal equations.

Estimation of the error variance

Note that for a random variable, its variance is the expected value of the squared deviation from the mean. That is, for a random variable W, with mean μ_W its variance is:

$$\sigma^2\{W\} = E\{(W - \mu_W)^2\}$$

For the simple linear regression model, the errors have mean 0, and variance σ^2. This means that for the actual observed values Y_i, their mean and variance are as follows:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad \sigma^2\{Y_i\} = E\{(Y_i - (\beta_0 + \beta_1 X_i))^2\} = \sigma^2$$

First, the unknown mean $\beta_0 + \beta_1 X_i$ is replaced with its fitted value $\hat{Y}_i = b_o + b_1 X_i$, then the "average" squared distance from the observed values to their fitted values is taken. The sum of squared errors is divided by *n*-2 to obtain an unbiased estimate of σ^2 (recall how you computed a sample variance when sampling from a single population).

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Common notation is to label the numerator as the **error sum of squares (SSE)**.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Also, the estimated variance is referred to as the **error (or residual) mean square (MSE)**.

$$MSE = s^2 = \frac{SSE}{n-2}$$

To obtain an estimate of the standard deviation (which is in the units of the data), the square root of the error mean square is taken. $s = \sqrt{MSE}$.

A shortcut formula for the error sum of squares, which can cause problems due to round-off errors is:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 - b_1 \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Some notation makes life easier when writing out elements of the regression model:

$$SS_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

$$SS_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SS_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Note that most all of the simple linear regression analysis from these quantities, the sample means, and the sample size are obtained.

$$b_1 \frac{SS_{XY}}{XX_{XX}} \qquad SSE = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}}$$

Normal error regression model

If the random errors follow a normal distribution, then the response variable also has a normal distribution, with mean and variance given above. The notation, used for the errors, and the data is:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad Y_i = N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Test statistic

The error sum of squares for the full model will always be less than or equal to the error sum of squares for reduced model, by definition of least squares. The test statistic will be:

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

where df_R, df_F are the error degrees of freedom for the full and reduced models. This method is used throughout course.

For the simple linear regression model, the following quantities are obtained:

$$SSE(F) = SSE \quad \mathrm{df_F} = n - 2 \quad SSE(R) = SSTO \quad \mathrm{df_R} = \mathrm{n} - 1$$

thus the *F*-Statistic for the General Linear Test can be written:

$$\mathrm{F^*} = \frac{\frac{\mathrm{SSE(R)} - \mathrm{SSE(F)}}{\mathrm{df_R} - \mathrm{df_F}}}{\frac{\mathrm{SSE(F)}}{\mathrm{df_F}}} = \frac{\frac{\mathrm{SSTO} - \mathrm{SSE}}{(\mathrm{n}-1)-(\mathrm{n}-2)}}{\frac{\mathrm{SSE}}{\mathrm{n}-2}} = \frac{\frac{\mathrm{SSR}}{1}}{\frac{\mathrm{SSE}}{\mathrm{n}-2}} = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

Thus, for this particular null hypothesis, the general linear test "generalizes" to the *F*-test.

Descriptive measures of association

Along with the slope, *Y*-intercept, and error variance; several other measures are often reported.

Coefficient of determination (r^2)

The coefficient of determination measures the proportion of the variation in *Y* that is "explained" by the regression on *X*. It is computed as the regression sum of squares divided by the total (corrected) sum of squares. Values near 0 imply that the regression model has done little to "explain" variation in *Y*, while values near 1 imply that the model has "explained" a large portion of the variation in *Y*. If all the data fall exactly on the fitted line, r^2=1. The coefficient of determination will lie between 0 and 1.

$$\mathrm{r^2} = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}} \qquad 0 \le \mathrm{r^2} \le 1$$

Adjusted r squared

The use of an adjusted $\mathrm{r^2}$ is an attempt to take account of the phenomenon of the $\mathrm{r^2}$ automatically and spuriously increasing when extra explanatory variables are added to the model. It is a modification

due to Theil of r^2 that adjusts for the number of explanatory terms in a model relative to the number of data points. The adjusted r^2 can be negative, and its value will always be less than or equal to that of r^2. Unlike r^2, the adjusted r^2 increases only when the increase in r^2 (due to the inclusion of a new explanatory variable) is more than one would expect to see by chance. If a set of explanatory variables with a predetermined hierarchy of importance are introduced into a regression one at a time, with the adjusted r^2 computed each time, the level at which adjusted r^2 reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess/unnecessary terms. The adjusted r^2 is defined as

$$r^2_{Adj} = 1 - (1 - r^2)\frac{n-1}{n-p} = r^2 - (1 - r^2)\frac{p-1}{n-p}$$

where p is the total number of explanatory variables in the model (not including the constant term), and n is the sample size.

Adjusted r^2 can also be written as

$$r^2_{Adj} = 1 - \frac{SS_{res} / df_e}{SS_{tot} / df_t}$$

where df_t is the degrees of freedom n– 1 of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom n – p – 1 of the estimate of the underlying population error variance.

Adjusted r^2 does not have the same interpretation as r^2 while r^2 is a measure of fit, adjusted r^2 is instead a comparative measure of suitability of alternative nested sets of explanators. As such, care must be taken in interpreting and reporting this statistic. Adjusted r^2 is particularly useful in the feature selection stage of model building (57).

Coefficient of correlation (*r*)

The coefficient of correlation is a measure of the strength of the linear association between *Y* and *X*. It will always be the same sign as the slope estimate (b_1), but it has several advantages:

- In some applications, a clear dependent and independent variable, cannot be identified. How two variables vary together in a population (peoples heights and weights, closing stock prices of two firms, etc). Unlike the slope estimate, the coefficient of correlation does not depend on which variable is labeled as *Y*, and which is labeled as *X*.
- The slope estimate depends on the units of X and Y, while the correlation coefficient does not.
- The slope estimate has no bound on its range of potential values. The correlation coefficient is bounded by –1 and +1, with higher values (in absolute value) implying stronger linear association (it is not useful in measuring nonlinear association which may exist, however).

$$r = sgn(b_1)\sqrt{r^2} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})(Y_i - \bar{Y})}} = \frac{s_x}{s_y} b_1 \qquad -1 \leq \mathrm{r} \leq 1$$

where sgn(b_1) is the sign (positive or negative) of b_1, and s_x, s_y are the sample standard deviations of *X* and *Y*, respectively.

2) Polynomial regression

Consider fitting polynomial regression equation between independent variable x and dependent variable y. Let this be represented by

$$\hat{\mathrm{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X} + \hat{\beta}_2 \mathrm{X}^2 + \cdots + \hat{\beta}_\mathrm{p} \mathrm{X}^\mathrm{p}$$

In principle this is no different from fitting multiple regression model except that the powers of X play the role of different independent variables. In matrix notation a polynomial regression model written as:

$$E(Y) = X\beta$$

where is 1 x n vector of observations, x is n(p+1) matrix given by

$$X = \left[1\ x_1\ x_1^2 \dots\ x_1^k\right]; \text{where } x_1^t = (x_1^t\ x_2^t \dots\ x_n^t)'\ ; t = 1, \dots, p$$

and is a vector of order (p+1) unknown parameters. 1 is a vector of unities.

Even though the problem of fitting polynomial regression is similar to the one of fitting multiple regression, polynomial regression has special features.

To smooth out fluctuations in the data caused by random or uncontrolled errors, not because it is thought to represent the relationship. If clear cut linear or parabolic relationship is no clear from the scatter of the data, one may draw free hand curve. This method has however, the disadvantage of biased-ness and impossibility of making a valid estimate of the residual variation about the curve.

While fitting the polynomial regression the form of the null hypothesis takes is that polynomial regression being fitted represents certain relationship and secondly, whether terms of higher degree contributes significantly to the relationship. As for example if a regression of degree four is fitted; the first test would be to test the significant of overall regression. If it is not significant, there is no need of further testing. Suppose, the overall regression comes out to be significant than one must test the significance of fourth order regression. If it is significant retain the equation. If it is not significant the process will have to be revised for fitting lower order polynomial to fit polynomial of degree three and so on.

There are, however, exceptions to this procedure. For example, when the null hypothesis specifies regression through origin; then it is correct to test the significance of $\hat{\beta}_0$ before testing other coefficients.

Orthogonal polynomial regression

Suppose an observation (X_I, Y_I), I=1,2, …,n. Where X is a predictor variable and Y, it is desired to fit the following model.

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon$$

In most of the situations the columns of X will not be orthogonal. If at some stage it is intended to include another term. $\beta_p X^{p+1}$ in the model, the changes will occur in all the other co-efficient. In order to simplify the computations, regression variable, but polynomial of increasing degree of X which are un-correlated. These polynomial are known as orthogonal polynomial. The advantages of defining independent variable in such a way are (i) each regression coefficient on each successive polynomial may be calculated independently of the other (ii) the sum of squares for regression attributable to each polynomial is independently calculated and represents the amount by which the regression sum of squares is increased by passage from an equation of lower degree.

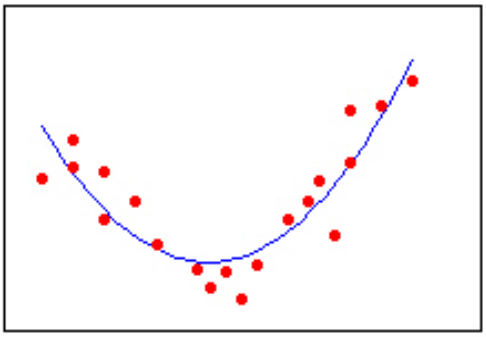

Figure 2.11 Quadratic (second order): $Y = b_0 + b_1 X + b_2 X^2$

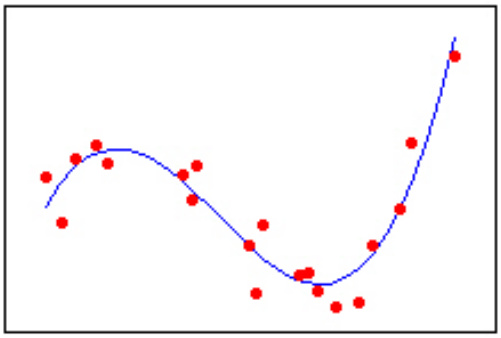

Figure 2.12 Cubic (third order): $Y = b_0 + b_1X + b_2X^2 + b_3X^3$

Another issue in fitting the polynomials in one variables is ill conditioning. An assumption in usual multiple linear regression analysis is that all the independent variables are independent. In polynomial regression model, this assumption is not satisfied. Even if the ill-conditioning is removed by centering, there may exist still high levels of multicollinearity. Such difficulty is overcome by orthogonal polynomials.

The classical cases of orthogonal polynomials of special kinds are due to Legendre, Hermite and Tehebycheff polynomials. These are continuous orthogonal polynomials (where the orthogonality relation involve integrating) whereas in our case, we have discrete orthogonal polynomials (where the orthogonality relation involves summation).

Analysis:

Consider the polynomial model of order is one variable as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \, , i = 1,2, \ldots . , n$$

When writing this model as

$$y = X\beta + \varepsilon$$

the columns of X will not be orthogonal. If we add another terms $\beta_{p+1}x_i^{p+1}$, then the matrix $(X'X)^{-1}$ has to be recomputed and consequently, the lower order parameters $\hat{\beta}_0, \ \hat{\beta}_1, \ldots , \ \hat{\beta}_p$ will also change.

This regression sum of squares does not depend on other parameters in the model. The analysis of variance table in this case is given as follows

Source of variation	Degrees of freedom	Sum of squares	Mean squares
$\hat{\beta}_0$	1	$SS(\hat{\beta}_0)$	-
$\hat{\beta}_1$	1	$SS(\hat{\beta}_1)$	$SS(\hat{\beta}_1)$
$\hat{\beta}_2$	1	$SS(\hat{\beta}_2)$	$SS(\hat{\beta}_2)$
⋮	⋮	⋮	⋮
$\hat{\beta}_p$	1	$SS(\hat{\beta}_p)$	$SS(\hat{\beta}_p)$
Residual	n-p-1	$SS_{res}(p)$(by subtraction)	SS_{res}
Total	n	SS_T	

Notice that:

• We need not to bother for other terms in the model.

• Simply concentrate on the newly added term only.

• No re-computation of $(X'X)^{-1}$ or any other $\hat{\alpha}_j (j \neq p+1)$ is necessary due to orthogonality of polynomials.

• Thus higher order polynomials can be fitted with ease.

• Terminate the process when a suitably fitted model is obtained.

Test of significance:

To test the significance of highest order term, we test the null hypothesis

$$H_0: \hat{\beta}_p = 0$$

This hypothesis is equivalent to $H_0: \beta_p = 0$ in polynomial regression model.

We would use

$$F_0 = \frac{SS_{reg}(\hat{\beta}_p)}{SS_{reg}(p)/(n-p-1)}$$

If order of the model is changed to (p+r), we need to compute only new coefficients. The remaining coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ do not change due to orthogonality

property of polynomials. Thus the sequential fitting of the model is computationally easy.

When X_i are equally spaced, the tables of orthogonal polynomials are available and the orthogonal polynomials can be easily constructed (58).

2.3 Literature reviews

Although body mass index and percentage of body fat are the indicators used in many clinical situations, there are some limitations including complicated calculation of BMI and difficulties of percent body fat measurement. Therefore HWDI, which is easier method to measurement than BMI and percent body fat, was recommended. Many researchers tried to study obesity evaluation, literature review are shown as the below.

Chittawatanarat K, et al. (7) identified variations of BMI and body fat across the age spectrum as well as comparing results between BMI predicted body fat and bioelectrical impedance results on age. Healthy volunteers were recruited. A total of 2,324 volunteers were included in this study. Multivariable linear regression coefficients were calculated. For results, the overall body composition and weight status, average body weight, height, body mass index (BMI), fat mass (FM), fat free mass (FFM), and its derivatives were significantly different among age groups. The coefficient of age altered the percentage fat mass (PFM) differently between younger, middle, and older groups (0.07; P=0.02 vs 0.13; P<0.01 vs 0.26; P<0.01; respectively). All coefficients of age alterations in all FM and FFM derived variables between each age spectrum were tested, demonstrating a significant difference between the younger (<60 years) and older (≥60 years) age groups, except the percentage fat free mass (PFFM) to BMI ratio (difference of PFM and FMI [95% confidence interval]: 17.8 [12.8-22.8], P<0.01; and 4.58 [3.4-5.8], P<0.01; respectively). The comparison between measured PFM and calculated PFM demonstrated a significant difference with increments of age.

Chathuranga R, et al. (59) studied the BMI and percent body fat relationship, in a group of South Asian adults who have a different body composition compared to

presently study ethnic groups. Pearson's correlation coefficient was calculated to see the relationship between BMI and percent body fat in the different age groups. Multiple regression analysis was performed to determine association between age and gender, and polynomial regression examined the linearity of the BMI and percent body fat relationship. The relationships between age and BMI, age and percent body fat were separately assessed. The results were out of 1,114 participants, A significant positive correlation was observed between BMI and percent body fat, in men ($r = 0.75$, $p < 0.01$; SEE = 4.17) and in women ($r = 0.82$, $p < 0.01$; SEE = 3.54) of all ages. Effect of age and gender in the BMI and percent body fat relationship was significant ($p < 0.001$); with more effect from gender. Regression line found to be curvilinear in nature at higher BMI values where women ($p<0.000$) having a better fit of the curve compared to men ($p<0.05$). In both genders, with increase of age, BMI seemed to increase in curvilinear fashion, whereas percent body fat increased in a linear fashion.

Nirav R, et al. (60) studied the effectiveness of precise biomarkers and duel-energe x-ray absorptiometry (DXA) to help diagnose and treat obesity. A cross-sectional study of adults with BMI, DXA, fasting leptin and insulin results were measured from 1998-2009. This study examined concordance and discordance of biomarkers, and Scatter plot of the relationship between BMI, percent body fat and leptin. A Receiver Operating Curve (ROC) analysis was used to identify cut points for BMI to optimize the area under the ROC curve (AUROC), specifically sensitivity and specificity, relative to percent body fat. For results, BMI characterized 26% of the subjects as obese, while DXA indicated that 64% of them were obese. 39% of the subjects were classified as non-obese by BMI, but were found to be obese by DXA. BMI misclassified 25% men and 48% women. Meanwhile, a strong relationship was demonstrated between increased leptin and increased body fat. Finally, new BMI cut-points for defining obesity would increase sensitivity with small tradeoffs in specificity.

Pruenglampoo S, et al. (3) assessed the use of height weight difference index (HWDI), which was named shortly as healthy index (HI), for screening overweight and obesity in adults. These were 2,234 Thai subjects (including men and women),

aged between 20 to 35 years, and enrolled in a community cohort project, Chiang Mai province, Thailand. Pearsons' correlation coefficient was calculated to see the relationship between BMI and HWDI. Linear regression model was used to estimate HWDI and the kappa statistic was used to measure agreement between the assessment of nutritional status by HWDI and by BMI. There was a negative correlation between BMI and HWDI ($r = -0.97$, $P < 0.001$, $n = 2,234$) with linear regression equation: HWDI = 158.69–2.54 * BMI ($P< 0.001$). The study findings suggest that HWDI could be used as a simple and effective index for screening overweight and obesity in adults.

Bedogni G, et al. (61) evaluated the agreement of air displacement plethysmography (ADP) and bioelectrical impedance analysis (BIA) with dual-energy X-ray absorptiometry (DXA) for the assessment of percent fat mass (PFM) in morbidly obese women. Fifty-seven women aged 19-55 years and with a body mass index (BMI) ranging from 37.3 to 55.2 kg/m^2 were studied. Values of PFM were obtained directly from ADP and DXA, whereas for BIA, we estimated fat free mass (FFM) from an equation for morbidly obese subjects and calculated PFM as (weight-FFM)/weight. As a result, this study was the mean (s.d.) difference between ADP and DXA for the assessment of PFM was -2.4% (3.3%) with limits of agreement (LOA) from -8.8% to 4.1%. The mean (s.d.) difference between BIA and DXA for the assessment of %FM was 1.7% (3.3%) with LOA from -4.9% to 8.2%.

Flegal K, et al. (62) investigated the relations between body mass index (BMI), waist circumference (WC), the waist-stature ratio (WSR), and percentage body fat (measured by DXA) in adults in a large nationally representative US population sample from the National Health and Nutrition Examination Survey (NHANES). As a result, WC, WSR, and BMI were significantly more correlated with each other than with percentage body fat ($P < 0.0001$ for all gender-age groups). Percentage body fat tended to be significantly more correlated with WC than with BMI in men but significantly more correlated with BMI than with WC in women ($P < 0.0001$ except in the oldest age group). WSR tended to be slightly more correlated with percentage body fat than WC. Percentile values of BMI, WC, and WSR are shown

that correspond to percentiles of percentage body fat increments of 5 percentage points. More than 90% of the sample could be categorized to within one category of percentage body fat by each measure.

Lazarus R, et al. (63) appraised the screening performance of BMI by using appropriate epidemiologic methods sample of 230 (119 men, 111 women) healthy Australian volunteers aged 4-20 years inclusive. Receiver operating characteristic (ROC) curves were prepared for detecting percentage body fat at or beyond the 85th percentile, using BMI as the screening test. Screening performance was slightly better for girls than for boys, but the differences were not significant. Reasonable true-positive (0.71, 95% CI: 0.53, 0.85) and low false-positive (0.05, 95% CI: 0.02, 0.09) rates were observed at the 85th percentile cut point for BMI. At the 95th percentile cut point for BMI, both true-positive (0.29, 95% CI: 0.15, 0.47) and false-positive (0.01, 95% CI: 0.00, 0.03) rates were lower.

Nair C, et al. (64) studied the relationship between BMI and percent body fat and health risks outcomes (specifically hypertension and type 2 diabetes) in men residents of Luck now city, north India to evaluate the validity of BMI cut-off points for overweight. One thousand one hundred and eleven men volunteer subjects (18-69 years) who participated indifferent programmers organized by the Institute during 2005 to 2008 were included in the study. The proposed cut-off for BMI based on percent body fat was calculated using receiver operating characteristics (ROC) curve analysis. The results, which were forty four percent subjects, showed higher percent body fat (>25%) with BMI range (24 - 24.99 kg/m^2). Sensitivity and specificity at BMI cut-off at 24.5 kg/m^3 were 83.2 and 77.5, respectively. Sensitivity at BMI cut-off >25 kg/m^2 was reduced by 5 percent and specificity was increased by 4.6 percent when comparing to 24.5 cut-off.

Meeuwsen S, et al. (65) studied the effects of age, gender and age-gender interactions on BMI-percent body fat relationships over a wide range of BMI and age. It also aimed to examine controversies regarding linear or curvilinear BMI-percent body fat relationships. Body composition was measured using validated bioimpedance equipment (Bodystat) in a large self-selected sample of 23,627 UK adults aged 18-99 (99% ≤70) years, of which 11,582 were men with a mean BMI

of 26.3±4.7 (sd) kg/ m^2, and 12,044 women, with a mean BMI of 25.7±5.1 kg/m^2. Multiple regression analysis was used. The results, BMI progressively increased with age in women and plateaued between 40 and 70 years in men. At a fixed BMI, body fat mass increased with age (1.9 kg/decade), as did percent fat (1.1-1.4% per decade). The relationship between BMI and percent fat was found to be curvilinear (quadratic) rather than linear, with a weaker association at lower BMI. There was also a small but significant age-gender interaction. The association between BMI and percent body fat is not strong, particularly in the desirable BMI range, is curvilinear rather than linear, and is affected by age.

Macias N, et al. (66) compared the classification accuracy of percent body fat, BMI and waist circumference for the detection of metabolic risk factors in a sample of Mexican adults; optimized cut-offs as well as sensitivity and specificity at commonly used percent body fat and BMI international cut-offs were estimated. Conditional percent body fat means at BMI international cut-offs. The methods this study performed a cross-sectional analysis of data on body composition, anthropometry and metabolic risk factors (high glucose, high triglycerides, low HDL cholesterol and hypertension) from 5,100 Mexican men and women. The association between BMI, waist circumference and percent body fat was evaluated with linear regression models. The percent body fat, BMI and waist circumference optimal cut-offs for the detection of metabolic risk factors were selected at the point where sensitivity was closest to specificity. Areas under the ROC Curve (AUROC) were compared among classifiers using a non-parametric method. The results, after adjustment for waist circumference, a 1% increase in BMI was associated with a percent body fat rise of 0.05 percentage points (p.p.) in men ($P < 0.05$) and 0.25 p.p. in women ($P < 0.001$). At BMI = 25.0 predicted percent body fat was 27.6 ± 0.16 (mean ± SE) in men and 41.2 ± 0.07 in women. Estimated percent body fat cut-offs for detection of metabolic risk factors were close to 30.0 in men and close to 44.0 in women. In men waist circumference had higher AUROC than percent body fat for the classification of all conditions whereas BMI had higher AUROC than percent body fat for the classification of high triglycerides and hypertension. In women BMI and waist circumference had higher AUROC than percent body fat for the classification of all metabolic risk factors. It concluded that the BMI and

waist circumference were more accurate than percent body fat for classifying the studied metabolic disorders. International percent body fat cutoffs had very low specificity and thus produced a high rate of false positives in both gender.

Ho-Pham L, et al. (67) studied the relationship between percent body fat and body mass index (BMI) in the Vietnamese population. This study included 1,217 individuals of Vietnamese (862 women) aged 20 years and older (average age 47 years old) who were randomly selected from the general population in Ho Chi Minh City. Lean mass (LM) and fat mass (FM) were measured by DXA. Percent body fat was derived as FM over body weight. The results, the prevalence of obesity (BMI≥30) was 1.1% and 1.3% for men and women, respectively. The prevalence of overweight and obesity combined (BMI≥25) was ~24% and ~19% in men and women, respectively. Based on the quadratic relationship between BMI and percent body fat, the approximate percent body fat corresponding to the BMI threshold of 30 (obese) were 30.5 in men and 41 in women. Using the criteria of percent body fat >30 in men and percent body fat >40 in women, approximately 15% of men and women were considered obese. This study suggest that body mass index underestimates the prevalence of obesity. It is recommended that PBF >30 in men or PBF >40 in women is used as criteria for the diagnosis of obesity in Vietnamese adults. Using these criteria, 15% of Vietnamese adults in Ho Chi Minh City was considered obese.