ML CLASS – 27-JAN-2017 - REGRESSION

The objective of statistics is to make inferences about a population based on information contained in a sample.

Populations are characterized by numerical descriptive measures called *parameters*.

Typical population parameters are the mean m, the median *M*, the standard deviation σ, and a proportion π.

Most inferential problems can be formulated as an inference about one or more parameters of a population.

Population Mean Score = μ.

Methods for making inferences about parameters fall into one of two categories:

1) Either we will **estimate** the value of the population parameter of interest or
2) We will **test a hypothesis** about the value of the parameter.

These two methods of statistical inference—estimation and hypothesis testing—involve different procedures, and, more important, they answer two different questions about the parameter.

In estimating a population parameter, we are answering the question, "What is the value of the population parameter?"

In testing a hypothesis, we are seeking an answer to the question, "Does the population parameter satisfy a specified condition?" For example, "$\mu > 20$" or "$\pi < .3$."

LINEAR REGRESSION

The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques.

We refer to this type of modeling as regression analysis.

A regression model provides the user with a functional relationship between the response variable and explanatory variables that allows the user to determine which of the explanatory variables have an effect on the response.

The regression model allows the user to explore what happens to the response variable for specified changes in the explanatory variables.

For example, financial officers must predict future cash flows based on specified values of interest rates, raw material costs, salary increases, and so on.

When designing new training programs for employees, a company would want to study the relationship between employee efficiency and explanatory variables such as the results from employment tests, experience on similar jobs, educational background, and previous training.

The basic idea of regression analysis is to obtain a model for the functional relationship between a **response variable** (often referred to as the dependent variable) and one or more **explanatory variables** (often referred to as the independent variables).

Regression Model Applications:

1. The model ***provides a description of the major features of the data set***. In some cases, a subset of the explanatory variables will not affect the response variable and hence the researcher will not have to measure or control any of these variables in future studies. This may result in significant savings in future studies or experiments.

2. ***The equation relating the response variable to the explanatory variables produced from the regression analysis provides estimates of the response variable for values of the explanatory not observed in the study***.

For example, ***a clinical trial is designed to study the response of a subject to various dose levels of a new drug***. Because of time and budgetary constraints, only a limited number of dose levels are used in the study. The regression equation will provide estimates of the subjects' response for dose levels not included in the study. The accuracy of these estimates will depend heavily on how well the final model fits the observed data.

3. In business applications, the prediction of future sales of a product is crucial to production planning. ***If the data provide a model that has a good fit in relating current sales to sales in previous months, prediction of sales in future months is possible***. However, a crucial element in the accuracy of these predictions is that the business conditions during which model building data were collected remains fairly stable over the months for which the predictions are desired.

4. In some applications of regression analysis, ***the researcher is seeking a model which can accurately estimate the values of a variable that is difficult or expensive to measure using explanatory variables that are inexpensive to measure and obtain.*** If such a model is obtained, then in future applications it is possible to avoid having to obtain the values of the expensive variable by measuring the values of the inexpensive variables and using the regression equation to estimate the value of the expensive variable.

For example, **a physical fitness center wants to determine the physical well-being of its new clients**. Maximal oxygen uptake is recognized as the single best measure of cardiorespiratory fitness but its measurement is expensive. Therefore, the director of the fitness center would want a model that provides accurate estimates of maximal oxygen uptake using easily measured variables such as weight, age, heart rate after 1-mile walk, time needed to walk 1 mile, and so on.

We can distinguish between **PREDICTION** (***reference to future values***) and **EXPLANATION** (***reference to current or past values***). Because of the virtues of hindsight, explanation is easier than prediction. However, it is often clearer to use the term *prediction* to include both cases.

For prediction (or explanation) to make much sense, there must be some connection between **THE VARIABLE WE'RE PREDICTING (the Dependent Variable)** and **THE VARIABLE WE'RE USING TO MAKE THE PREDICTION (The Independent Variable)**.

No doubt, if you tried long enough, you could find 30 common stocks whose price changes over a year have been accurately predicted by the won–lost percentage of the 30 major league baseball teams on the fourth of July. However, such a prediction is absurd because there is no connection between the two variables.

Prediction requires a **unit of association;** there should be an entity that relates the two variables.

With time-series data, the unit of association may simply be time. The variables may be measured at the same time period or, for genuine prediction, the independent variable may be measured at a time period before the dependent variable. For cross-sectional data, an economic or physical entity should connect the variables. If we are trying to predict the change in market share of various soft drinks, we should consider the promotional activity for those drinks, not the advertising for various brands of spaghetti sauce. The need for a unit of association seems obvious, but many predictions are made for situations in which no such unit is evident.

<u>Simple Linear Regression Analysis</u>, in which there $\hat{\beta}$ is a single independent variable and the equation for predicting a dependent variable *y* is a linear function of a given independent variable *x*. Suppose, for example, that the director of a county highway department wants to predict the cost of a resurfacing contract that is up for bids. We could reasonably predict the costs to be a function of the road miles to be resurfaced. A reasonable first attempt is to use a linear production function.

Let *y* = total cost of a project in thousands of dollars, *x* = number of miles to be resurfaced, and $\hat{y}$ = the predicted cost, also in thousands of dollars. A prediction equation (for example) is a linear equation.

The constant term, such as the 2.0, is the **intercept** term and is interpreted as the predicted value of *y* when *x* = 0. In the road resurfacing example, we may interpret the intercept as the fixed cost of beginning the project. The coefficient of *x*, such as the 3.0, is the **slope** of the line, the predicted change in *y* when there is a one-unit change in *x*.

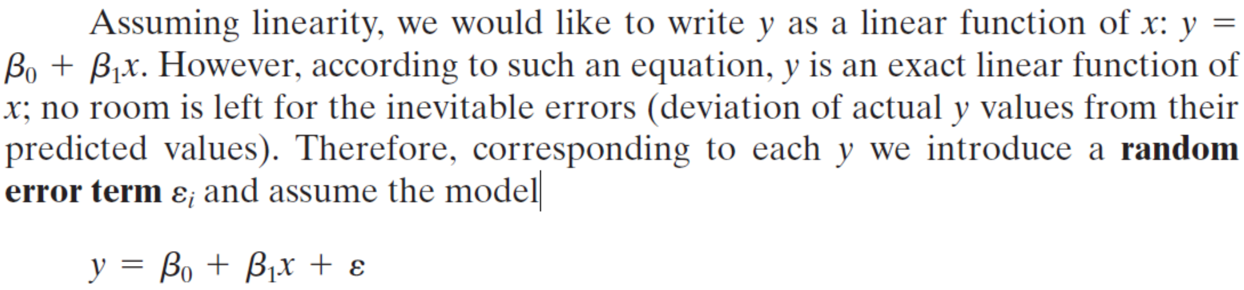
In the road resurfacing example, if two projects differed by 1 mile in length, we would predict that the longer project cost 3 (thousand dollars) more than the shorter one. In general, we write the prediction equation as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope.

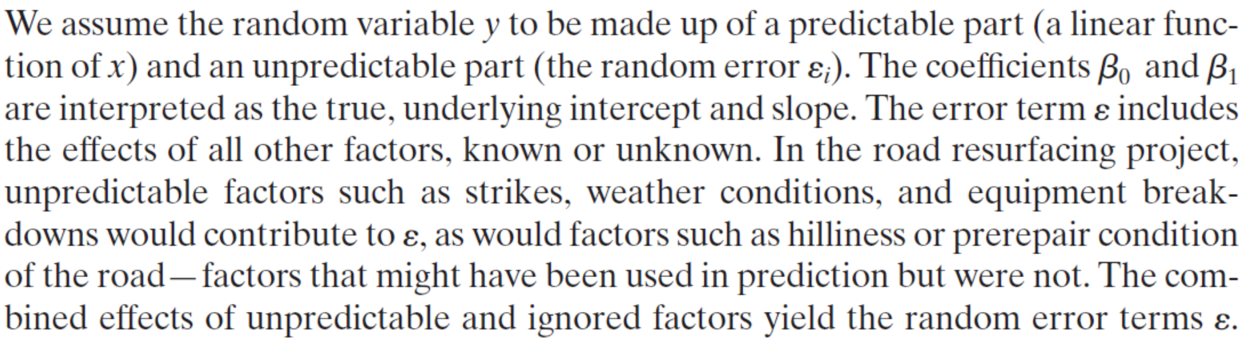
The basic idea of simple linear regression is to use data to fit a prediction line that relates a dependent variable *y* and a single independent variable *x*. The first assumption in simple regression is that the relation is, in fact, linear. According to the **assumption of linearity,** the slope of the equation does not change as *x* changes. In the road resurfacing example, we would assume that there were no (substantial) economies or diseconomies from projects of longer mileage. There is little point in using simple linear regression unless the linearity assumption makes sense (at least roughly).

Linearity is not always a reasonable assumption, on its face. For example, if we tried to predict *y* = number of drivers that are aware of a car dealer's midsummer sale using *x* = number of repetitions of the dealer's radio commercial, the assumption of linearity means that the first broadcast of the commercial leads to no greater an increase in aware drivers than the thousand-and-first. (You've heard commercials like that.) We strongly doubt that such an assumption is valid over a wide range of *x* values. It makes far more sense to us that the effect of repetition would diminish as the number of repetitions got larger, so a straight-line prediction wouldn't work well.



Assuming linearity, we would like to write y as a linear function of x: $y = \beta_0 + \beta_1 x$. However, according to such an equation, y is an exact linear function of x; no room is left for the inevitable errors (deviation of actual y values from their predicted values). Therefore, corresponding to each y we introduce a **random error term** ε_i and assume the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$



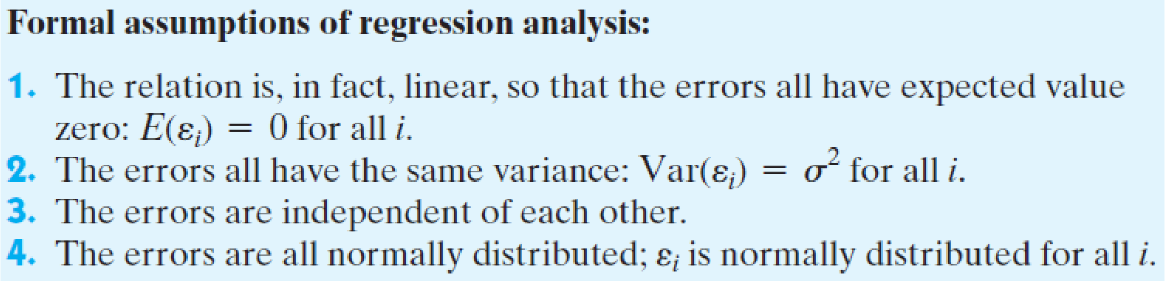
We assume the random variable y to be made up of a predictable part (a linear function of x) and an unpredictable part (the random error ε_i). The coefficients β_0 and β_1 are interpreted as the true, underlying intercept and slope. The error term ε includes the effects of all other factors, known or unknown. In the road resurfacing project, unpredictable factors such as strikes, weather conditions, and equipment breakdowns would contribute to ε, as would factors such as hilliness or prerepair condition of the road—factors that might have been used in prediction but were not. The combined effects of unpredictable and ignored factors yield the random error terms ε.

For example, one way to predict the gas mileage of various new cars (the dependent variable) based on their curb weight (the independent variable) would be to assign each car to a different driver, say, for a 1-month period. What unpredictable and ignored factors might contribute to prediction error? Unpredictable (random) factors in this study would include the driving habits and skills of the drivers, the type of driving done (city versus highway), and the number of stoplights encountered.

Factors that would be ignored in a regression analysis of mileage and weight would include engine size and type of transmission (manual versus automatic).

In regression studies, the values of the independent variable (the x_i values) are usually taken as predetermined constants, so the only source of randomness is the *i* terms. Although most economic and business applications have fixed x_i values, this is not always the case. For example, suppose that *xi* is the score of an applicant on an aptitude test and y_i is the productivity of the applicant. If the data are based on a random sample of applicants, x_i (as well as y_i) is a random variable. The question of fixed versus random in regard to *x* is not crucial for regression studies. If the x_is are random, we can simply regard all probability statements as conditional on the observed x_is.

When we assume that the x_is are constants, the only random portion of the model for y_i is the randomerror term ε_i. We make the following formal assumptions.



Formal assumptions of regression analysis:

1. The relation is, in fact, linear, so that the errors all have expected value zero: $E(\varepsilon_i) = 0$ for all i.
2. The errors all have the same variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i.
3. The errors are independent of each other.
4. The errors are all normally distributed; ε_i is normally distributed for all i.

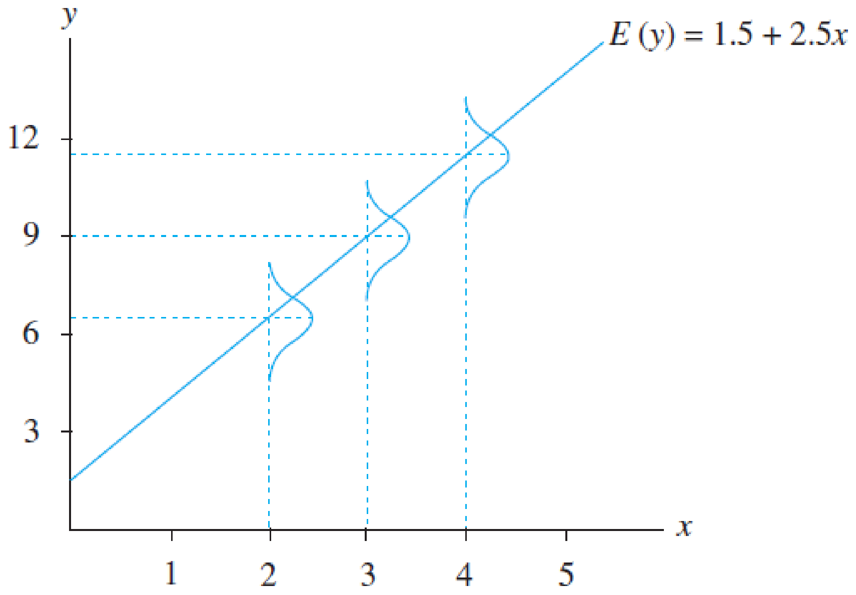
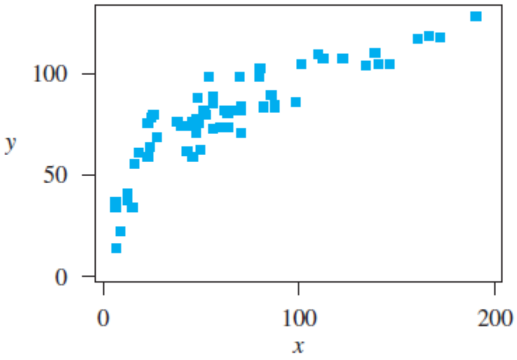


Figure: Theoretical distribution if y in Regression

These assumptions are illustrated in Figure above. The actual values of the dependent variable are distributed normally, with mean values falling on the regression line and the same standard deviation at all values of the independent variable. The only assumption not shown in the figure is independence from one measurement to another.

These are the formal assumptions, made in order to derive the significance tests and prediction methods that follow. We can begin to check these assumptions by looking at a **scatterplot** of the data. This is simply a plot of each (*x*, *y*) point, with the independent variable value on the horizontal axis, and the dependent variable value measured on the vertical axis. Look to see whether the points basically fall around a straight line or whether there is a definite curve in the pattern. Also look to see whether there are any evident outliers falling far from the general pattern of the data. A scatterplot is shown below.



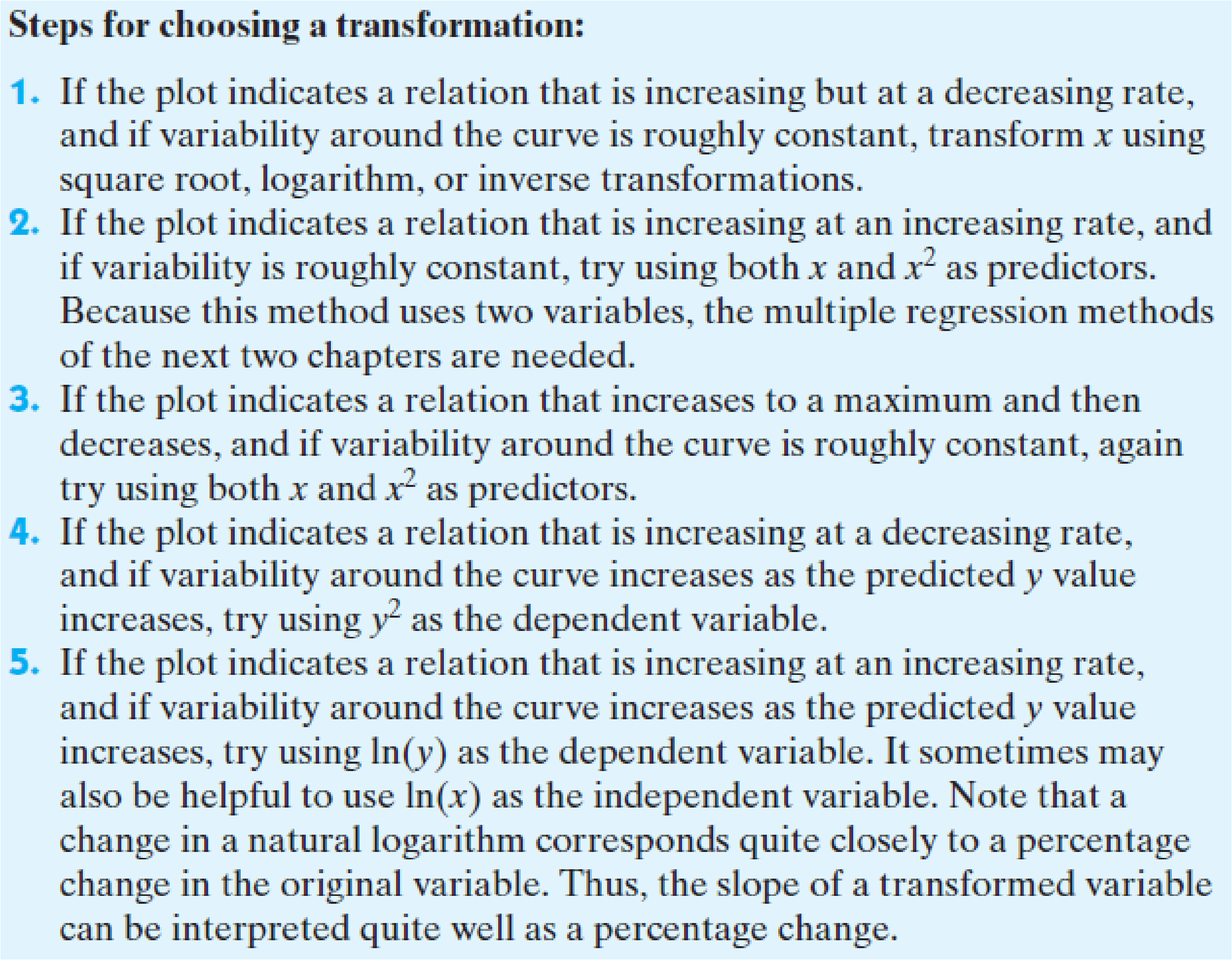
Many economic relations are not linear. For example, any diminishing returns pattern will tend to yield a relation that increases, but at a decreasing rate.

If the scatterplot does not appear linear, by itself or when fitted with a LOWESS curve, it can often be “straightened out” by a **transformation** of either the independent variable or the dependent variable.

A good statistical computer package or a spreadsheet program will compute such functions as the square root of each value of a variable. The transformed variable should be thought of as simply another variable.

We can ***try several transformations of the independent variable to find a scatterplot in which the points more nearly fall along a straight line***.

Three ***common transformations are square root, natural logarithm, and inverse (one divided by the variable)***.



Steps for choosing a transformation:

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