Class Notes – Day 6 **Saturday, February 17, 2018**

1. Simple Regression

For any bivariate data, where the value of y depends on x (x is an independent variable, and y is the dependent variable, the value it takes depends on the value of x), y can be expressed as a function of x and a constant - β, with an error function - ε.

In general, a linear regression model takes the following form:

$$y_t = f(x_t, \beta_t) + \epsilon$$

where:

y: dependent variable (criterion variable)

x: independent variable (predictor variable)

β: constants describing the functional relationship in the population

ε: overall error component.

Assumptions for this form:

$\varepsilon \sim N(0, \sigma^2)$ (The error component is normally distributed, with mean 0 and variance σ^2).

$Cov(\varepsilon_t, \varepsilon_{t+1}) = 0$ (The errors at different points are completely random, and do not depend on the previous value - ε_{t+1} does not depend on ε_t).

For generating data from the above form, we take only observable parameters that impact the data into consideration.

1.1 Linear Regression

This has been covered in earlier classes, previous class notes may be referred.

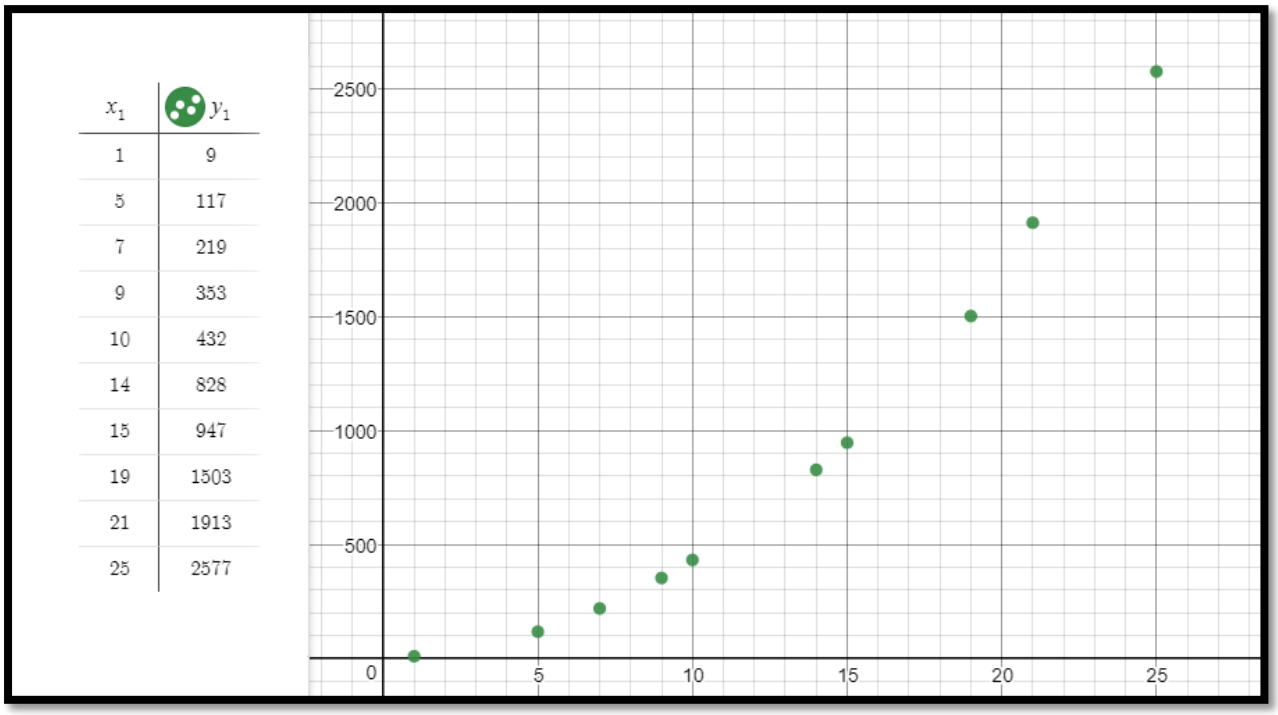
1.2 Polynomial Regression

We encounter a situation of polynomial regression, if the data does not exhibit a linear relation. We need to find the best model that fits the data "curve". By simply observing the scatter plot and doing a bit of analysis and finding the slope at strategic points, we can have a tentative picture of the curve, and probably its behavior on extrapolation too.

As an example, we take the following data:

x	y
1	9
5	117
7	219
9	353
10	432
14	828

15	947
19	1503
21	1913
25	2577



From the scatter plot, we can see that the data is clearly not linear. It exhibits a curve similar to a parabolic curve oriented on y-axis. We deduce that x is of a higher degree.

Now, the question is, what is the degree of x? It might be 2, 3, 4 and so on. From the slope of the curve when it is approaching the origin, we observe that it tends to 0, which is not the case when x = 3. Hence, we try to fit the curve to a quadratic function.

A simple quadratic equation is of the form:

$$y = \beta_1 x^2 + \beta_2 x + \beta_0$$

The best estimate of β_t for the generic relation $y_t = f(x_t, \beta_t) + \epsilon$ is obtained by minimizing the square error:

$$e^2 = (y - f(x))^2$$

For all points,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Using the general quadratic equation in the above equation:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_1 x_i^2 + \beta_2 x_i + \beta_0))^2$$

For minimizing this value, we partially differentiate the above equation with each of the constants β_0, β_1 and β_2. On partial differentiation, we get the slope of the curve and since slope is 0 at its maxima and minima, we equate the partial derivatives to 0 to obtain the minima.

$$\frac{\partial \sum_{i=1}^{n} (e_i^2)}{\partial \beta_1} = 2 \sum_{i=1}^{n} [y_i - (\beta_1 x_i^2 + \beta_2 x_i + \beta_0)](-x_i^2) = 0$$

$$\frac{\partial \sum_{i=1}^{n} (e_i^2)}{\partial \beta_2} = 2 \sum_{i=1}^{n} [y_i - (\beta_1 x_i^2 + \beta_2 x_i + \beta_0)](-x_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} (e_i^2)}{\partial \beta_0} = 2 \sum_{i=1}^{n} [y_i - (\beta_1 x_i^2 + \beta_2 x_i + \beta_0)](-1) = 0$$

Dividing both sides by 0 and simplifying further, we get the following 3 equations

$$\sum_{i=1}^{n} x_i^2 y_i = \beta_1 \sum_{i=1}^{n} x_i^4 + \beta_2 \sum_{i=1}^{n} x_i^3 + \beta_0 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i = \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i + n\beta_0$$

In the above three equations, after calculating and substituting the values of summations of x and y, we'll have three unknowns *viz.* β_0, β_1 and β_2, solving the equations simultaneously would fetch these values too. Let us solve the above example for an instance.

S. no.	x	y	xy	x^2	x^3	x^4	x^2y
1	1	9	9	1	1	1	9
2	5	117	585	25	125	625	2925
3	7	219	1533	49	343	2401	10731
4	9	353	3177	81	729	6561	28593
5	10	432	4320	100	1000	10000	43200
6	14	828	11592	196	2744	38416	162288
7	15	947	14205	225	3375	50625	213075
8	19	1503	28557	361	6859	130321	542583
9	21	1913	40173	441	9261	194481	843633
10	25	2577	64425	625	15625	390625	1610625
sum	126	8898	168576	2104	40062	824056	3457662

After getting the values of summations from the above table, we substitute in the above three equations and get the following three equations with three unknown variables:

$$3457662 = 824056\beta_1 + 40062\beta_2 + 2104\beta_0$$
$$168576 = 40062\beta_1 + 2104\beta_2 + 126\beta_0$$
$$8898 = 2104\beta_1 + 126\beta_2 + 10\beta_0$$

When we solve these equations simultaneously, we get the values for the betas.

$$\beta_1 = 3.91, \beta_2 = 6.86, \beta_0 = -20.08$$

Multiple Regression

Multiple regression makes use of multiple observed values to predict a dependent variable.

Observed variables are variables for which we have measurements in our dataset, whereas unobserved (latent) variables are variables for which we don't.

To offer a contrived example, suppose our dataset includes strongly correlated variables “people closing their bank accounts” and “people adapting mobile banking”. We might suspect that there's an unobserved variable (age), acting as a common cause, driving the correlation.

Apart from this, we also need to take into consideration how much impact an observed variable has on the dependent variable. An observed variable having infinitesimal effect on the dependent variable should be avoided from being taken into consideration, based on the below factors:

Measurement effort.

The error propagated by the observed variable – The variable itself might have very less effect on the variable being predicted, but since it is being included in the analysis, its error component also gets added up, resulting in a greater error component. Do we need to include a component which is of less use, but causes more harm?

If we include unnecessary variables, because they are present in the dataset, the accuracy of our prediction becomes questionable.

Now, the question is, how do we come to know whether the variable has impact on the variable being predicted? Is the impact significant? Here’s where ANOVA (Analysis of variables) comes into picture. ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance.

In the above example, account closure (the dependent variable y) can be depicted as a function of two independent variables mobile banking usage (x_1) and age (x_2). The resulting function can be written as:

$$y = ax_1 + bx_2 + c$$

In general, for y to be dependent on x, there should be a cause-effect relationship between x and y. We should be able to say that x is the cause of y, and y is the effect of change in x. In the above example, we can say that age can be the cause of account closure in banks, since oldsters tend to do so, while youngsters are busy saving money and finding different sources of investment. Whereas, we cannot say that increase in mobile banking technology leads to account closures. In fact, advancement in technology attracts more young customers, enthusiastic to stay updated with technology. We might even question if at all bank account closure depends on mobile banking usage. If we think logically, young people tend to use more technology as compared to old people, and since old people tend to leave their affiliation with banks, we presume that account closure is because of mobile banking usage, but we miss the internal cause effect relationship of age and mobile banking usage. Since these both are strongly related, we might be under a misconception that the dependent variable is related the latter.

Partial Correlation coefficient:

In probability theory and statistics, partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. If we are interested in finding whether or to what extent there is a numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another, confounding, variable that is numerically related to both variables of interest (variables of interest: age, account closure. Variable numerically related to age: mobile banking usage). This misleading information can be avoided by controlling for the confounding variable, which is done by computing the partial correlation coefficient.

Consider a dependent variable which depends on three variables x_1, x_2 and x_3.

$$y = ax_1 + bx_2 + cx_3 + d$$

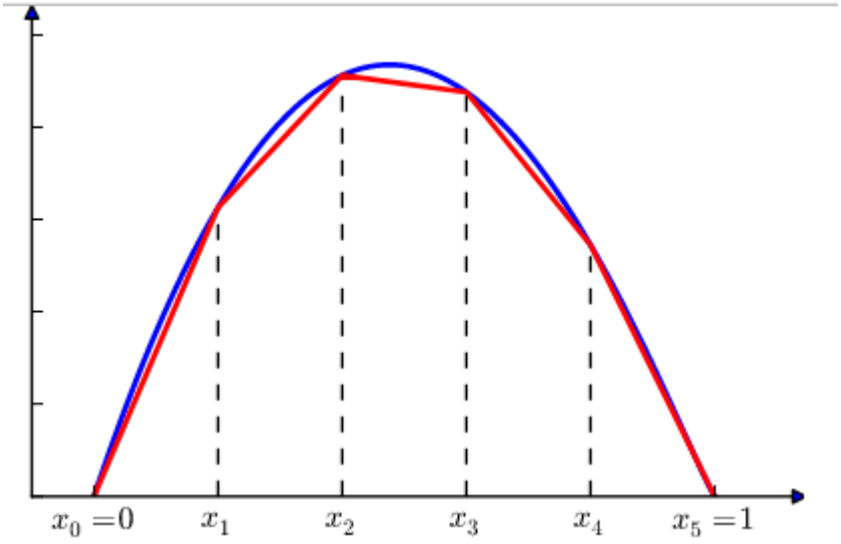
Partial correlation coefficient of x_1 and y is calculated, while ignoring (nullifying) the effect of x_2 and x_3 on y, and similarly while calculating the partial correlation coefficient for x_2 and x_3, the effect of other two variables is ignored.

Piecewise linear function:

A piecewise linear function is a real – valued function, whose graph is composed of straight line sections, as in the figure below.

Limitations:

- Each straight line segment should be connected to the adjacent segment, *ie.* The graph should be continuous.
- Extrapolation of the segments could be misleading, since the slope might vary.



Notes documented by: P G Mahadevan.