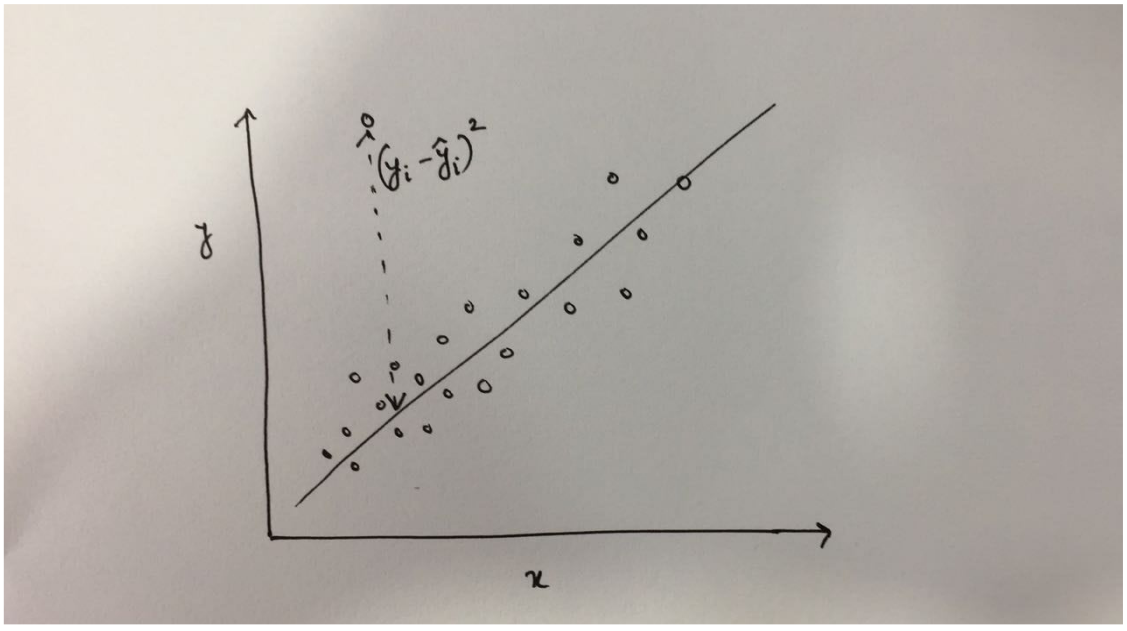
Question-1:
List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show *why* the assumption makes sense.

Answer:
Following are the main assumptions in linear regression:
1. Linearity: It is assumed that the dependent and the independent variables are linearly related, i.e. the relationship is linear in the *coefficients* (and not in the independent variables). For e.g. y = 3x^2 + 1 can be written as y = 3t +1, where t=x^2 is a new derived attribute. The relationship between y and t (and equivalently between y and x) is linear in nature. If the relationship is inherently nonlinear, then one cannot use linear regression. For e.g. if the actual relationship is y = ax^2 + bx + asin(bx), then one cannot reduce this relationship to a linear one.

2. Outliers: It is assumed that the dataset does not contain outliers - points that deviate from the general trend. For example, if the dataset looks like this:



Then the outlier point shown above will blow up the cost function since the value (yi-yi^2) will be very large.

3. Multicollinearity: It is assumed that the independent variables are not correlated with each other, i.e. they are linearly independent. In other words, one cannot predict any one of the independent variable using the other independent variables.

This assumption is useful because if some variables are correlated to others, the coefficients of those variables can vary a lot with changes in the dataset (and thus they are not reliable). For example, when predicting sales using TV and Newspaper advertising budget, if TV and Newspaper are correlated, one cannot infer how much each predictor is contributing to sales.

Some other key assumptions are as follows:

4. Autocorrelation:
One of the basic assumptions in linear regression model is that the random error components or disturbances are identically and independently distributed. The autocorrelation function is used to detect the non-randomness of the data.

Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times X_i and X_{i+k}. Thus, for a linear regression model to be applicable, the error terms should not have correlation with a lagged version of itself. In other words, one should not be able to predict the error e_i of the ith data point using the previous error terms e_{i-1}, e_{i-2} etc.

5. Heteroskedasticity: Heteroskedasticity refers to a situation where the error terms/residuals are not scattered uniformly with zero mean, but increase or decrease in a pattern with the independent predictors.

This is a problem because it indicates that there is some hidden pattern in the data which is not captured by the model, and it is rather leaking into the error terms. In other words, if the model has successfully captured the patterns in the data, the error terms should be normally distributed with zero mean and should not be predictable. Heteroskedasticity violated that condition.

Question-2:
Explain the gradient descent algorithm using the following pointers:

1. Illustrate at least two iterations of the algorithm using the univariate function $J(x) = x^2 + x + 1$. Assume a learning rate of 0.1 and an initial guess x=1 and demonstrate that the iterations converge towards the minima. Also, report the minima (which you can compute using the closed form solution).
2. Illustrate at least two iterations of the algorithm using the bivariate function of two independent variables $J(x, y) = x^2 + 2xy + y^2$. Assume a learning rate of 0.1 and an initial guess of (x, y) = (1, 1). Report the minima and show that the solution converges towards it.

Answer: Part-1

The closed form solution is x = -½ at which the cost is 0.75 (computed by differentiating the function with respect to x and equating the derivative to 0). The iterations starting from initial guess of x=1 are shown below.

$J(x) = x^2 + x + 1$

- starting point: $x_0 = 1$; $\eta = 0.1$
- $\partial J / \partial x = 2x + 1$
- Iteration scheme: $x_{i+1} = x_i - \eta(2x_i + 1)$

Iteration-1:

$$x_1 = x_0 - 0.1(2x_0 + 1)$$
$$= 1 - 0.30$$
$$= 0.70$$

Iteration-2 :

$$x_2 = x_1 - 0.1(2x_1 + 1)$$
$$= 0.70 - 0.24$$
$$= 0.46$$

It shows that the solution converges towards the minima of -½.

Answer: Part-2

The minimum value of the function is 0 at x=y=0. The function can be written as $J(x, y) = (x + y)^2$, and thus the minimum value it can take it is 0, which happens at all points where x = -y. Note that this problem does not have a unique solution since the cost of 0 wherever x+y = 0, which is an infinite set of points (x, y).

The iterations below show that starting from (x, y) = (1, 1), the solution gradually moves towards the minima of (x, y) = (0, 0).

$$J(x,y) = x^2 + 2xy + y^2$$

- Starting point: $x = y = 1$; $\eta = 0.1$
- $\partial J/\partial x = 2x + 2y$
- $\partial J/\partial y = 2x + 2y$
- $$\begin{bmatrix} x \\ y \end{bmatrix}^{new} = \begin{bmatrix} x \\ y \end{bmatrix}^{old} - 2\eta \begin{bmatrix} x+y \\ x+y \end{bmatrix}^{old}$$

Iteration-1 :

$$\begin{bmatrix} x \\ y \end{bmatrix}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 2(0.1) \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix}_1 = \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}$$

Iteration-2 :

$$\begin{bmatrix} x \\ y \end{bmatrix}_2 = \begin{bmatrix} x \\ y \end{bmatrix}_1 - 2(0.1) \begin{bmatrix} x \\ y \end{bmatrix}_1$$

$$\begin{bmatrix} x \\ y \end{bmatrix}_2 = \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix} - 2(0.1) \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix}_2 = \begin{bmatrix} 0.48 \\ 0.48 \end{bmatrix}$$