

MULTIPLE LINEAR REGRESSION

- Model $R^2 = 0.42$ i.e. 42% of variation in house price is explained by its size !
- Question - What factors exist that may contribute to explaining remaining variation in housing prices?
- Answer - House prices are also influenced by its location i.e. House price is also dependent on location of house (e.g. how far from city centre etc)
- This means, we have one more IV that can impact our DV
- When we have more than one IV for developing regression, we call it as Multiple Linear Regression
- It is concerned with developing equation of the form:

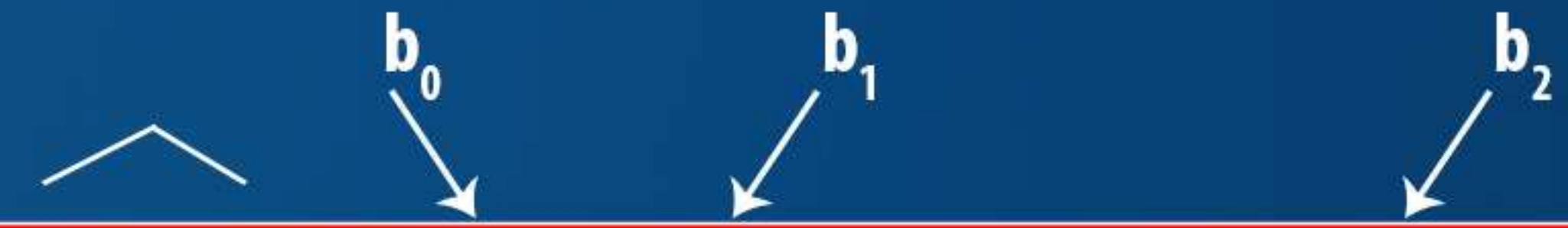
$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + \dots$$

MULTIPLE LINEAR REGRESSION

House Price (1000s) Y	Size (sft) X1	Distance (miles) X2
245	1400	5
312	1600	4
279	1700	6
308	1875	3
199	1100	2
219	1550	7
378	1910	6
364	1850	1
319	1425	8
...

MULTIPLE LINEAR REGRESSION EXAMPLE: FINDING BEST FIT LINE

- Regression line found for this data set is:


$$\text{house price} = 57.438 + 0.2832 * (\text{square feet}) - 0.1822 * (\text{distance})$$

- Notice how co-efficient b_1 has changed post introduction of b_2
 - This is known as net effect of size on housing price for houses at same distance (or controlling the distance)
 - The previous value of b_1 was telling us the gross effect
- Calculating p-Values for each co-efficient:
 - b_1 P-value = 0.03.
 - b_2 P-value = 0.01.
- Model adjusted- $R^2 = 0.62$ i.e. 62% of variation in house price is explained by its size and distance.
 - In MLR, we use adjusted- R^2 as true indicator of its performance instead of plain R^2 .

MULTIPLE LINEAR REGRESSION

► For each IV in MLR:

- Look at the **p-Value** to see if it is statistically significant.
- Check the **sign of coefficient** to see if makes sense as per given domain.
For example: $b_2 = -0.1822$
- Look at the **magnitude** of coefficient to understand amount of structural relationship between IV and DV.
- **Interpret** coefficient as net (i.e. effect of IV is net (on DV) considering other IVs are controlled) or gross.

Note: In order to assess if MLR is good model to study the relationship:

- Always, plot the residuals against each IV and look for presence of non-linear relationship between and IV and DV.

MULTICOLLINEARITY

- **Multicollinearity occurs when some of the independent variables are strongly interrelated.**
- **Multicollinearity is typically not a problem when we use regression for forecasting.**
- **When addition of new IV dramatically improves adjust- R^2 and at the same time, increases the p-Value of existing of IV, it mostly means that new IV and existing IV are likely correlated.**
- **When using regression to understand the net relationships between independent variables and the dependent variable, multicollinearity should be reduced or eliminated such increase the sample size by twice the size or remove the collinear variable.**