# Linear Regression

It is a supervised ML algorithm used to predict continuous values taking in consideration other factors.
for Eg: House prices, Demand Forecasting

→ It can also be used to find/calculate relationship between different parameters in biological systems.

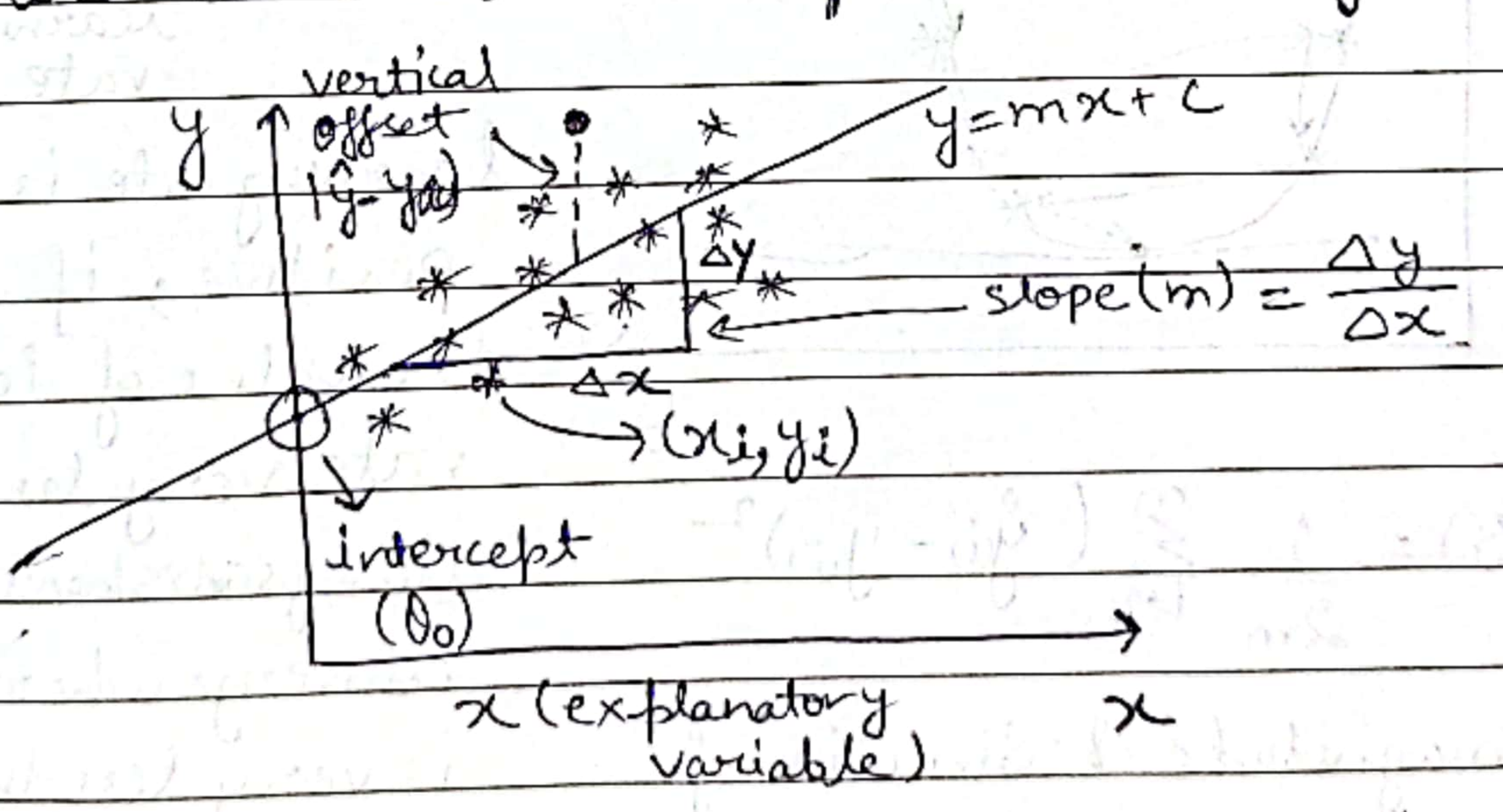
Linear regression is a linear model that assumes linear relationship between the input variables (X) and the output variable (Y).
The equation of linear regression is given as—

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n + \theta_0$$

- $y$: dependent variable
- $x_1, x_2, x_3, \ldots, x_n$: independent variables / features
- $\theta_1, \theta_2, \theta_3, \ldots, \theta_n$: coefficient
- $\theta_0$: constant / intercept on y-axis

* When there is only a single feature we call it as a **simple linear regression.**

** When there are multiple independent variables/features we call it as **multiple linear regression.**




→ The loss function for a linear regression model is called **OLS** (Ordinary Least Square) represented by—

$$\text{OLS} = (\hat{y} - y_{act})^2$$

→ The **cost function** used in a linear regression model is called

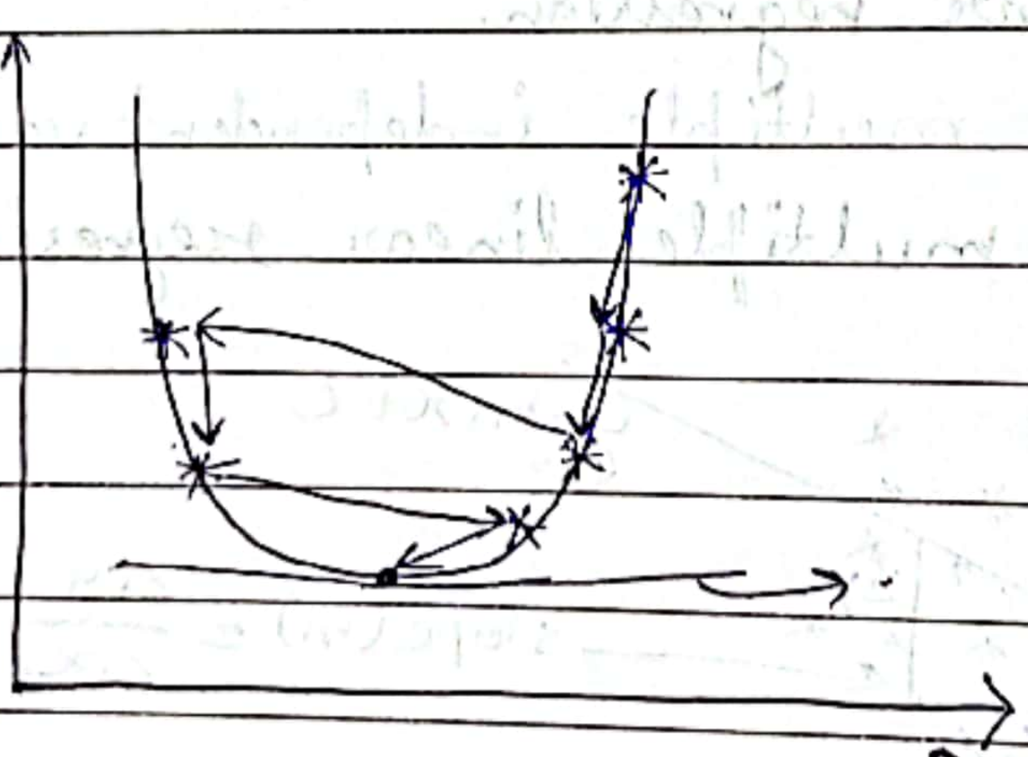
$$\text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y_{act})^2$$

* Loss function is used when we refer to the error for a single training example.

** Cost function is used to refer to an average of the loss function over an entire training data.

→ The idea of linear regression is to find a line (in case of 2D) or a hyperplane (in case of multiple dimensions) which best fits all the training points and minimizes the vertical offsets.
The idea is to reduce the cost function as much as possible to minimize errors.

**Gradient Descent :-** To minimize the cost function we use an optimization algorithm like gradient descent which repeatedly minimizes the cost function by moving in the direction of steepest descent.
In this the model weights are updated after each epoch.




$$\theta = \theta - \alpha \, \partial\theta$$

learning rate

learning rate is always positive, if we take the value of learning rate very large it will always overshoot and not converge whereas if it is very less then it will take a lot of iterations to converge to minima

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_{(i)} - y_{(i)})^2$$

* The magnitude & direction of weight update is computed by taking a step in the opposite direction of the cost gradient.

## Assumptions of Linear Regression:

Linear regression has five key assumptions:

(1) The relationship between the independent and dependent variables needs to be linear.
We can test linearity by plotting scatter plots.

(2) There should be no or very less **multicollinearity**.

* Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.

Disadvantage :- (a) The coefficient of the independent variable may not be reliable as the independance of predictors is itself violated due to correlation.

(b) Multicollinearity creates disturbance in the data and weakens the statistical power of the model.

→ Ways to detect and remove multicollinearity -

(1) Plotting a correlation matrix : It calculates Pearson's bivariate correlation among all independent variables. We can remove predictors with high correlation.

(2) Plotting scatter plot between predictors.

(3) **Tolerance**: Each feature is regressed against all other features. The main idea is we should not be able to explain one feature with respect to other feature. Its formula is given as —

$$\boxed{\text{Tolerance (T)} = 1 - R^2} \rightarrow \text{Goodness of fit}$$

Tolerance indicates how well a feature is explained by other features.

If $T < 0.1$, there might be multicollinearity

If $T < 0.01$, there certainly is collinearity.

(4) Variance Inflation Factor (VIF):

$$\boxed{VIF = \frac{1}{T} = \frac{1}{1-R^2}}$$ , if $VIF > 5$, then it is an indicator that multicollinearity is present in dataset.

If $VIF > 10$, then multicollinearity is certainly present among the variables.

(5) Feature transformation technique like PCA, t-SNE, UMAP.

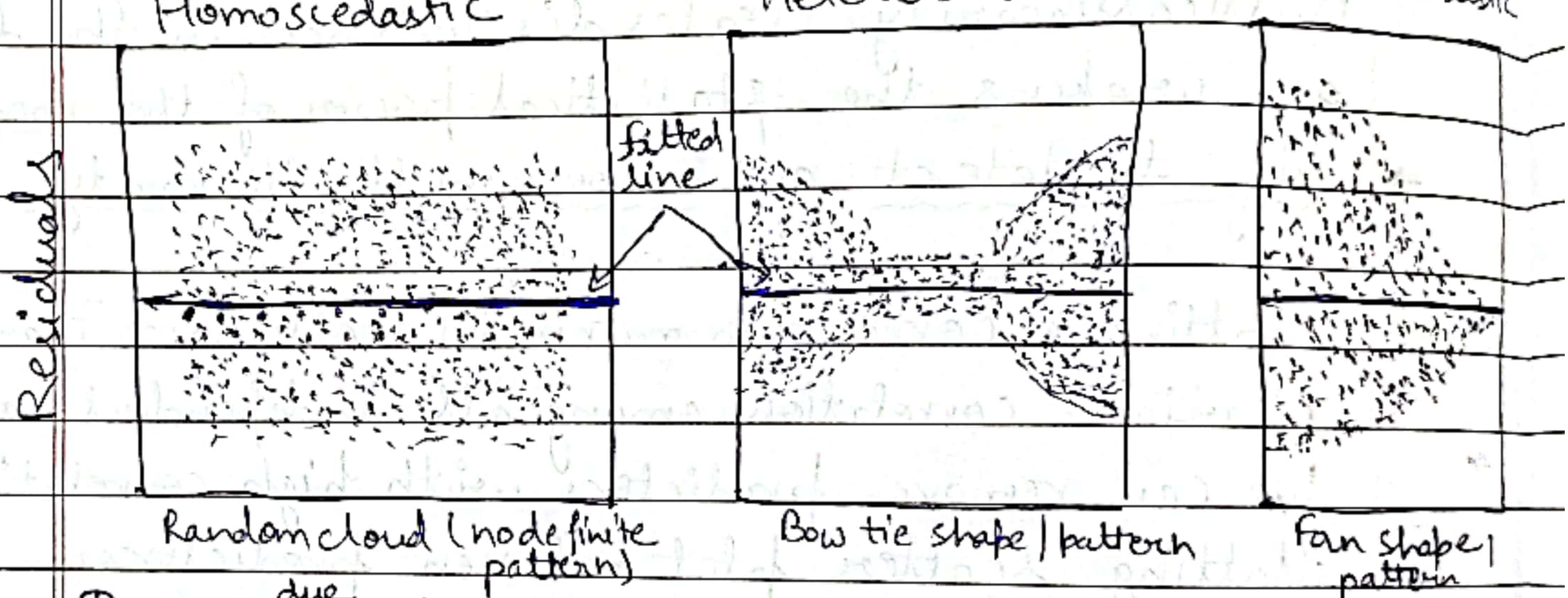
Date
Page 4

③ **Homoscedasticity:** Linear Regression assumes the residuals are homoscedastic (have same variance) and are randomly scattered along the regression line/hyperplane and does not follow any specific pattern.

* In contrast, heteroscedasticity is a systematic change in the spread of residuals.

→ Checks for homoscedasticity/heteroscedasticity:

* Residual plots are good way to check for this.
Residual plots display the residual values on y-axis and fitted values on x-axis.




Problems due to heteroscedasticity:—

ⓐ If there is pattern in the residuals, model has a problem and will be not able to explain the data patterns effectively. Hence, the **coefficients** values will be unreliable.

ⓑ Heteroscedasticity tends to reduce the **p-values** than they should be. thus reducing the statistical significance.

→ The **p-value** for each feature, tests the null-hypothesis that the coefficient is equal to zero (no effect has

→ A low p-value (<0.05) indicates that we can reject the null-hypothesis.

Possible solution that can be tried to fix it:

* Transformation of dependent variable may reduce the intensity of heteroscedasticity.
** We can add additional features that can better explain the variance in data.
*** Alternatively, a non-linear model may justify the data better.

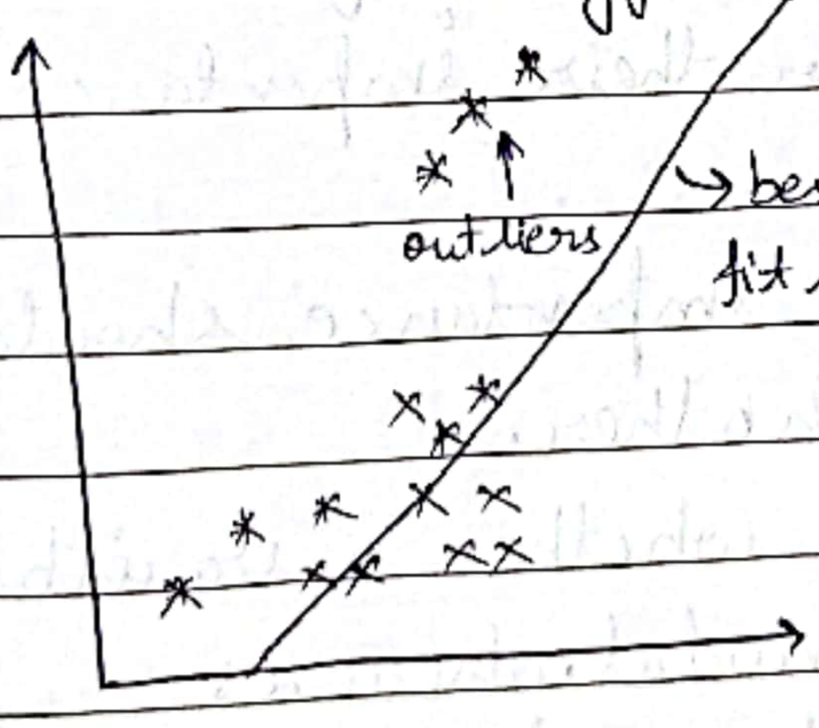
④ **Normality**: Linear Regression analysis requires the residuals to be normally distributed.

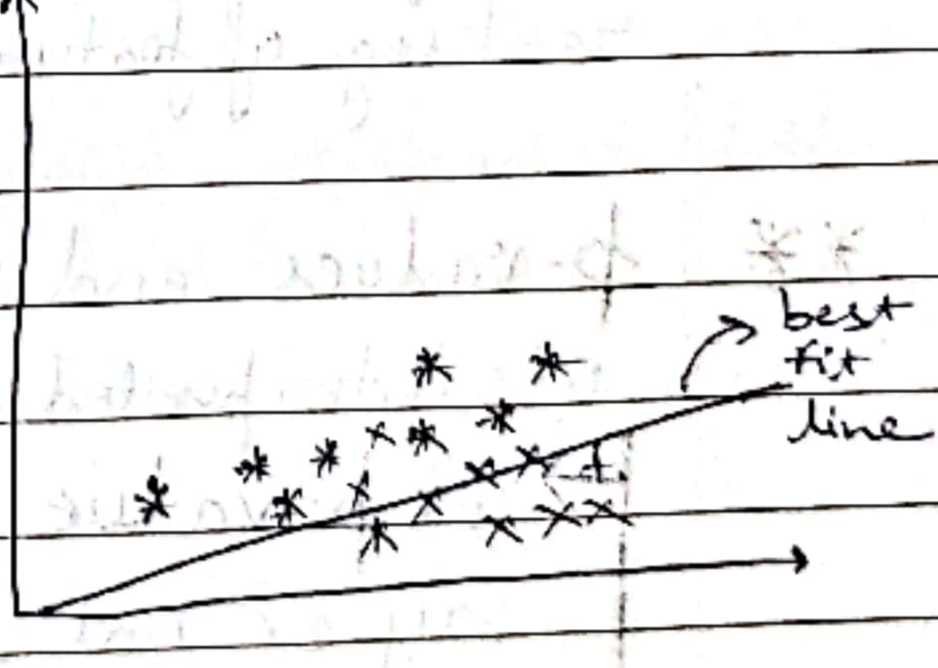
Ques. Why is multivariate normality needed?

→ Short Answer: It is important that the residuals don't contain extreme outliers since linear regression is sensitive to outlier effects.

→ Long Answer: Linear regression uses "OLS" for choosing the "best-fit". A candidate fit gets "penalized" on the basis of the sum of squarred residuals and outliers contributes to large squarred residuals. So, it is important that residuals follows normal distribution.

* How outliers affect the candidate fit —




To check for normality of residuals we can use Q-Q plots or Kolmogrov-Smirnov (KS) test.

(5) No or little autocorrelation: Autocorrelation occurs when the residuals are not independent from each other. The presence of correlation in error terms reduces model's accuracy. Auto correlation generally occurs in time-series models where next instant is dependent on previous instant.

To check for autocorrelation we can use Durbin-Watson test.

→ Possible ways to reduce autocorrelation.

* We can add more right features like lag variables.
* We can do smart feature engineering like adding week, months, cyclic features.

Variable Importance :- The feature/variable that impacts the response/dependent variable the most can be termed as the most important variable. Our goal should be to find out the ranking of features as per their importance.

** p-values and variable importance should not be misinterpreted for each other.
→ A p-value signifies whether a variable has its say or not in the model whereas
→ the variable importance helps us to know which variable has how much say.

Possible pitfalls to be careful of while finding variable importance:

(a) Don't compare regular regression coefficients to determine variable importance.
Suppose, we have a feature 'weight' in a model then it could be both in Kg and grams.

If we fit model for same dataset using grams in one model and kilograms in another, the coefficient for weight changes by a factor of 1000, even though the underlying fit of the model remains unchanged.

It should be noted that the **coefficient** value changes greatly but the **importance** of variable remains same.

(b) Don't link p-values to determine variable importance.

- Low p-value determines whether the variable should be included in the model or not (significance).
- We should not confuse **low p-value** with **high importance**. It basically tells that the feature should be **part of the model** and **not** how much its **contribution** will be.

Finding Variable Importance (the right way) —

(a) For a standardized linear regression model we can consider its coefficients for variable importance. In a standardized linear regression model, the independent variables have their standard deviation = 1 and all are on same scale, so they can be compared directly. We can take absolute of coefficient and then decide.

(b) We can consider the change in R-squared for the last variable added to the model.
Why? When an independent variable is the last one entered into the model, the associated change in R-Squared represents the improvement in the goodness-of-fit that is due solely to that last variable.

Suppose in a LR model, we have n features, we can add features one by one and check how much the value of R-Squared changes, this way we can know which feature improved how much R-squared thus having more importance. This

process of adding features one by one is also called **forward selection**. We can even do a **backward elimination** process where we remove features one by one and can track by what magnitude R-squared value is changing. The feature whose elemination causes the maximum decrease in the R-squared value can be termed as the most important variable.

## Goodness of Fit / $R^2$ as a measure of goodness of fit:

$R^2$ is a stastical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

$$R^2 = 1 - \frac{\sum_{I=1}^{n} (Y_I - Y_P)^2}{\sum_{i=1}^{n} (Y_I - \bar{Y}_I)^2} = 1 - \frac{SS_{res}}{SS_{tot}}$$

The value of $R^2$ ranges between 0 to 1 and the closer it is to 1, the better the model is. But, as the number of variables / features keep on increasing in a model, the $R^2$ value keeps on increasing which can be misleading as well. Thus, **adjusted $R^2$** comes into picuture which penalizes the less significant variables and

**Limitations of $R^2$:** (a) It increases every time on adding an independent variable to the model.

(b) $R^2$ value cannot determine whether the coefficient estimates and predictions are biased. To assess this, we can make use of the residual plot.

(c) A high $R^2$ value doesnot mean / indicate that the model has a good fit. It can be because of overfitting in training set.

**Adjusted R-squared →** The adjusted R-squared compares the explanatory power of regression models that contain different number of predictors.

→ The adjusted $R^2$ increases only when the new term improves the model more than it would be expected by chance.

→ It decreases when a predictor improves the model by less than expected by chance.

$$\overline{R^2} = 1 - \frac{SS_{res}/(n-p-1)}{SS_{tot}/(n-1)}$$

$$= 1 - \frac{SS_{res}}{SS_{tot}} * \frac{(n-1)}{(n-p-1)}$$

$$\boxed{\overline{R^2} = 1 - (1-R^2) * \frac{(n-1)}{(n-p-1)}}$$

n = sample size

p = total no. of independent variable not including the constant term.

## Variations of Linear Regression :

① **Polynomial Regression :-** It is a form of regression analysis in which the relationship between the independent variable, X, and the dependent variable, Y, is modelled as the $n^{th}$ degree polynomial in X.

Although polynomial regression fits a non-linear model to the data, as a statical estimation problem it is linear. For this reason, polynomial regression is a special case of multiple linear regression.

The equation for polynomial regression is given as—

$$\boxed{Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + \dots + \theta_n x_n + \varepsilon}$$

**Key Points about Polynomial Regression**

(a) It is able to model non-linearly separable data that linear regression can't. It is much more flexible than general and can model some fairly complex relationships.

(b) We have full control over the modelling of feature variable (which exponent to set for what variable). Domain knowledge can be beneficial in order to select exponents.

(c) Model can be prone to overfitting if exponents are poorly selected.

Why classical Linear Regression can fail?

(a) It can fail due to presence of multicollinearity. Multicollinearity is the phenomenon when the independent variable(s) can be expressed as the linear combination of other independent features.

(b) When the number of independent variables are larger than the number of samples/observations. When this happens, the OLS estimators are invalid mainly because there are infite solutions to the estimators.

## (2) Lasso Regression (L1 regularization):

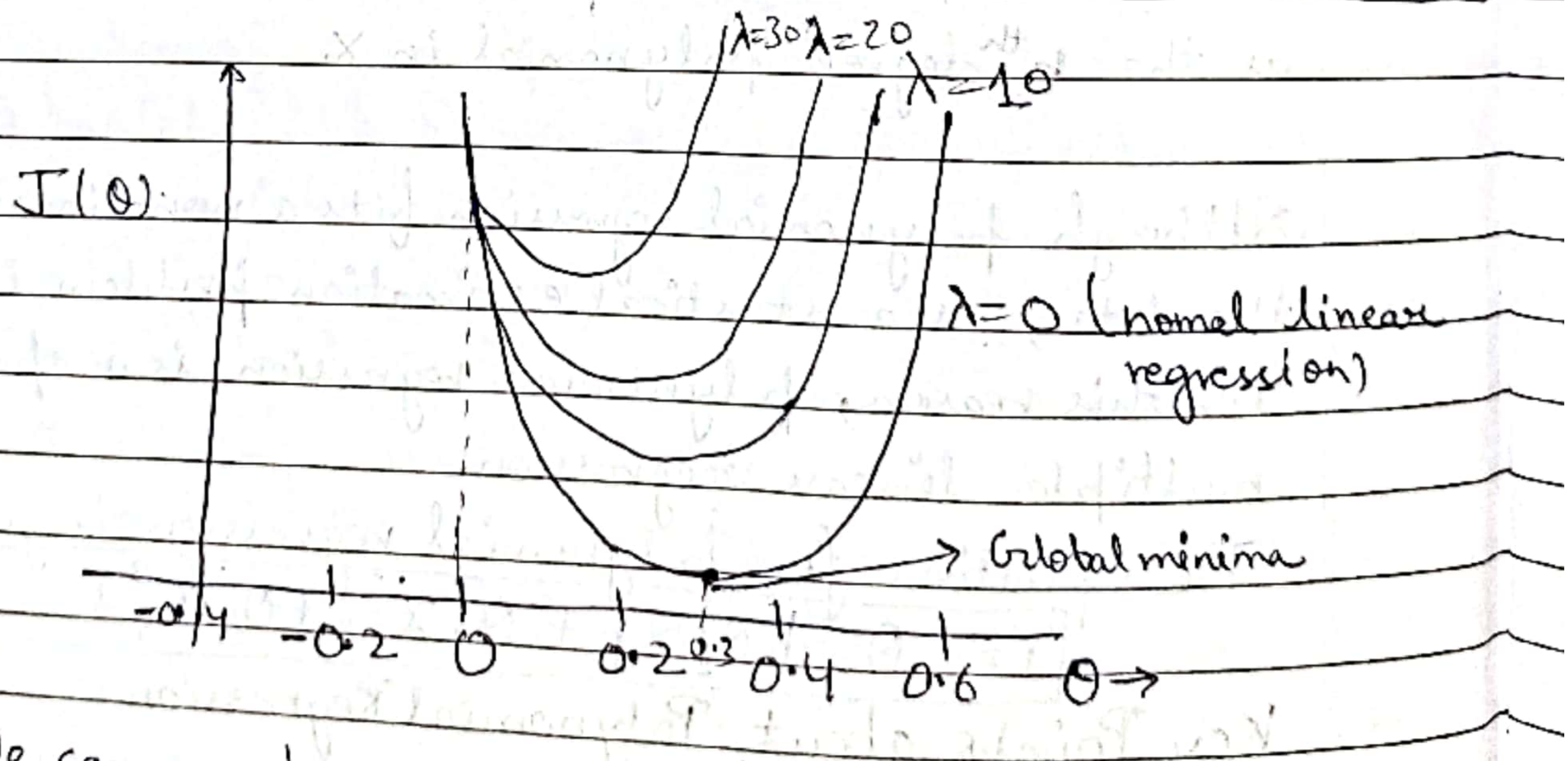
It penalizes the model by the absolute weight coefficients.

The cost function for Lass regression (L1 regularization) is given as —

hyperparameter

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_{(i)} - y_{(i)})^2 + \lambda \sum_{i=1}^{n} |\text{slope}|$$

↳ coefficient




We can use Lasso regression for feature selection as it forces the coefficients of some variables to become zero and thus we can remove that feature from the model to obtain a simpler model.

* As the value of $\lambda$ increases it causes the slope to decrease, thus coefficient of few variables become zero and they can be eleminated from the model.

When we have large number of features in a dataset we should use Lasso regression as it eleminates less useful features on its own from the dataset.

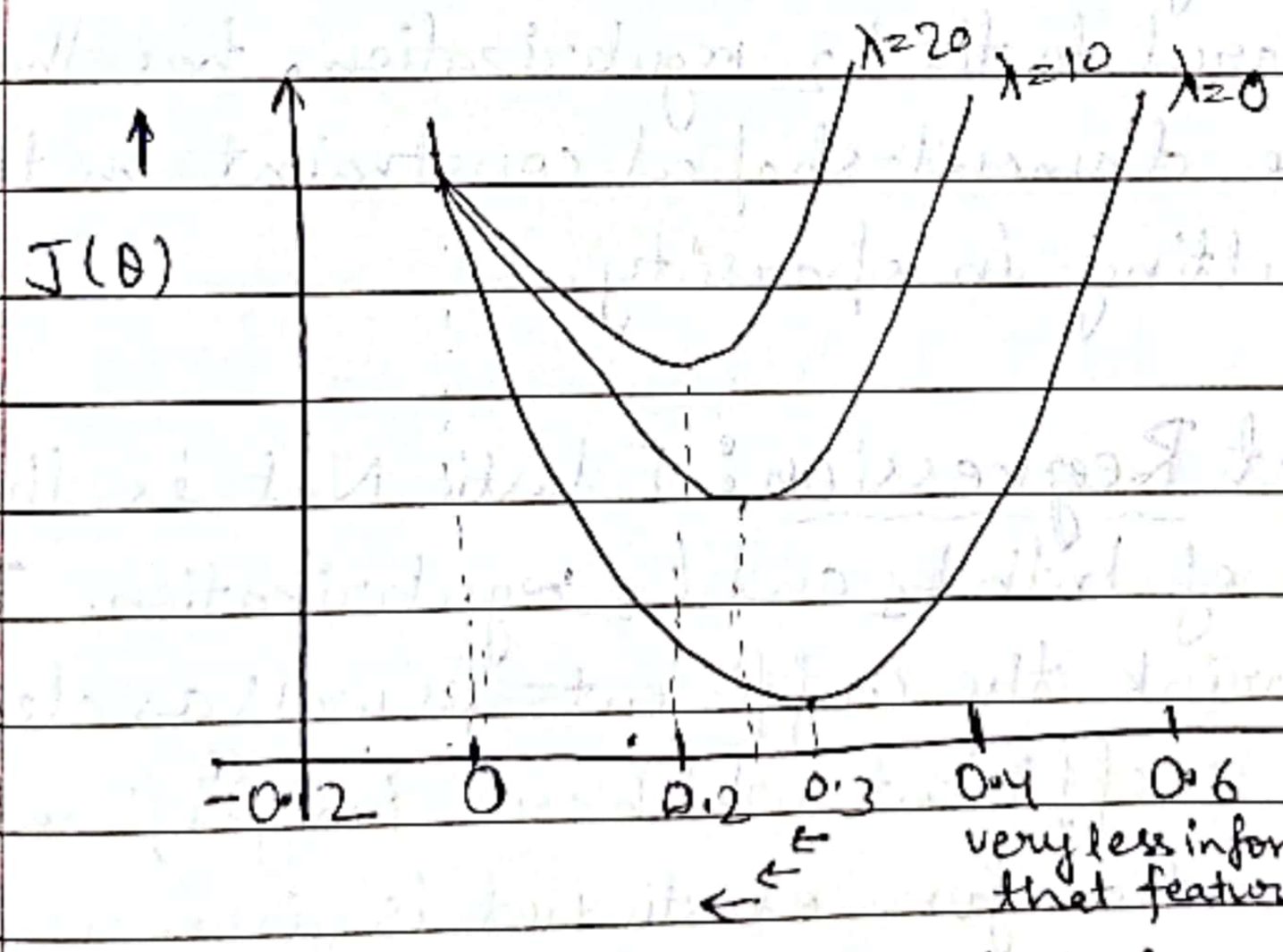
* Lasso Regression (L1-norm) does not have closed form solution.

## ③ Ridge Regression (L2 regularization):

It penalizes the model by adding a constraint that is a linear function of the squarred coefficients.

The cost function for Ridge regression is given as—

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}(i) - y(i))^2 + \lambda \sum_{i=1}^{n} |\text{slope}|^2$$




Note:-
In a model, addition of a variable should add information which is meaning ful and helps reduce the loss and the reduction in loss should be more significant than penalty incurred due to it.

If the less influential feature is adding very less information but penalty is more that feature/coefficient of feature is shrinked.

* Here, also as the value of $\lambda$ increases the slope decreases but it will never be zero at any point.

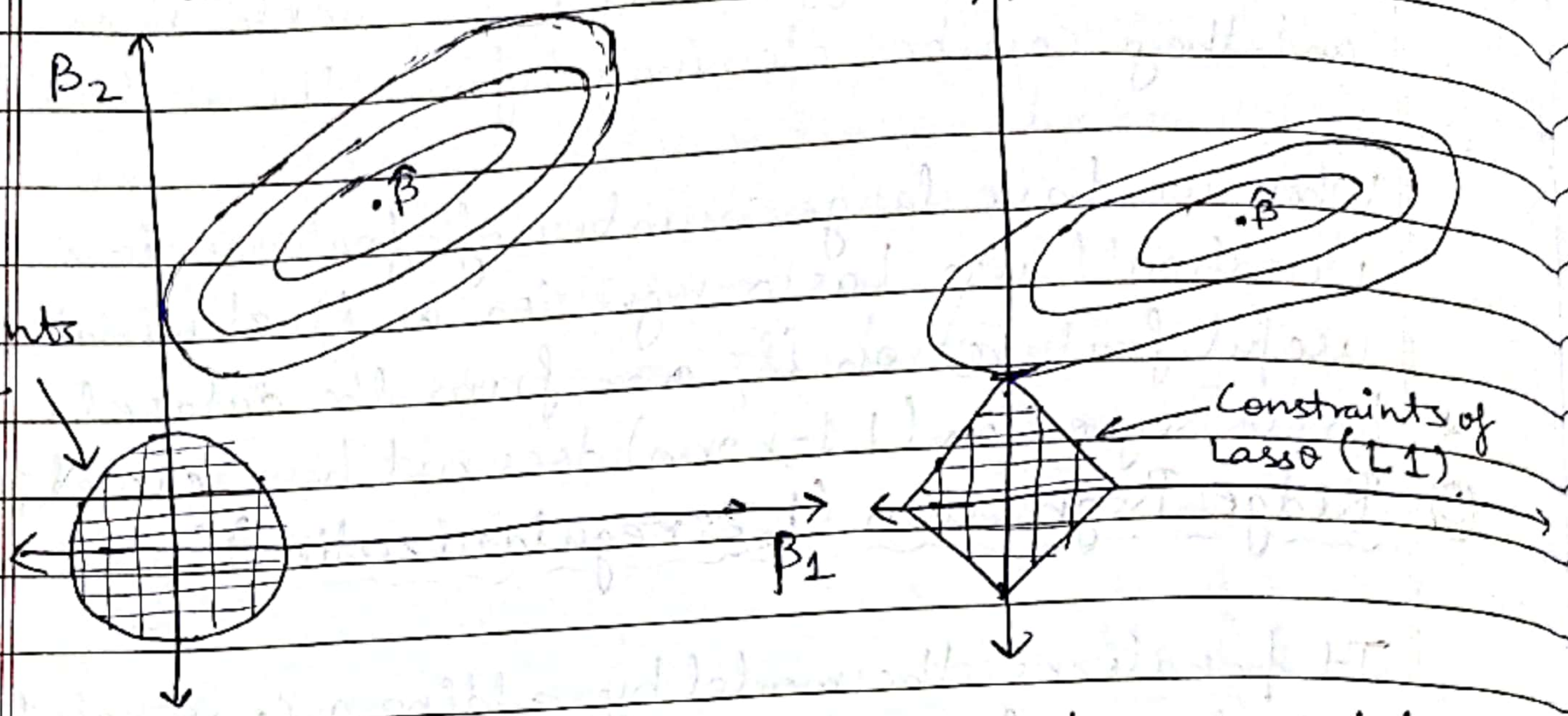
→ Ridge regression can be used to prevent overfitting because it shrinks the less influential features. The coefficients of less influential features shrinks but it does not become zero completely.

* It has a closed form solution and equation is given by—

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} + X^T Y$$

Ques. Why Lasso shrinks coefficient to 0 but Ridge does not?

Ans →




Shown are contours of errors and constraint functions. The shaded portions are the constraints $\beta_1^2 + \beta_2^2 <= t^2$ and $|\beta_1| + |\beta_2| <= t$ respectively, while the ellipsoids are the contours of the loss function (OLSE).

Here, $L_1$ compared to the $L_2$ regularization, has the tendency to touch the diamond-shaped constraint on the corner, resulting in sparsity.

③ Elastic Net Regression: Elastic Net is the combination of both $L_1$ and $L_2$ regularization. It can both shrink the coefficients as well as eleminate some of the coefficient which are insignificant.

The cost function for elastic net is given as —

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_{(i)} - y_{(i)})^2 + \lambda \sum_{i=1}^{m} (\text{slope})^2 + \lambda \sum_{i=1}^{m} |\text{slope}|$$