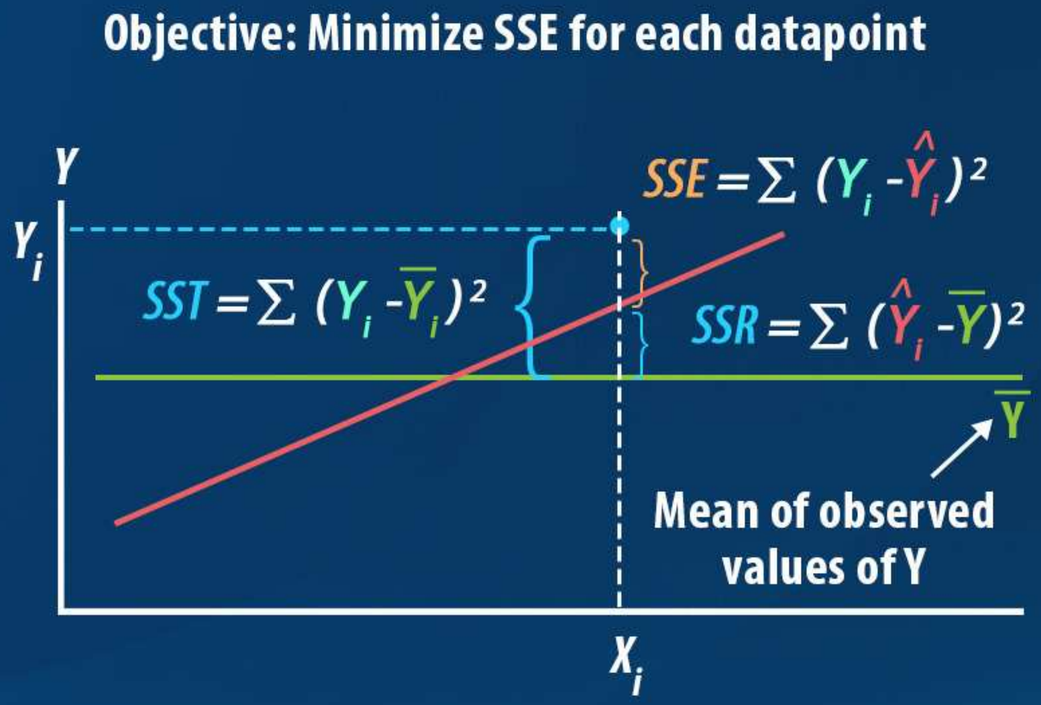
IMPORTANT TERMS

- SST = Total sum of squares
  - Measure of total variation in Y
- SSE = Sum of Squares Error
  - Measure of unexplained variation in Y
- SSR = Sum of Squares Regression
  - Measure of explained variation in Y



COEFFICIENT OF DETERMINATION

- Term represented by r^2 and calculated by SSR/SST is known as Coefficient of determination.
- Also, known as Goodness of fit.
- Ranges between 0 and 1.
- It indicates 'strength' of relationship between two variables (x and y).
- The higher is the value of r^2, more accurate is the regression model or equation found i.e. higher value of r^2 means model has captured variability in data well.
  - e.g. When $r^2 = 0.0233$ or 2.33%, it means, model has failed to capture whopping 97.7% of variability in the data and hence it is not a good model.

CORRELATION

- Linear relationship existence between two variables can be known by determining correlation coefficient (r).
- Also, known as Pearson Correlation Coefficient in Linear Regression. *(Should be always examined via scatter plot because non-linear data may show high value of r as well).*
- Value varies from -1 to 1:
  - Value of 1 means perfect positive correlation
  - Value of -1 means perfect negative correlation
  - Value of 0 means lack of correlation
- For linear regression, r is usually +/- of square root of r^2.

CORRELATIONS

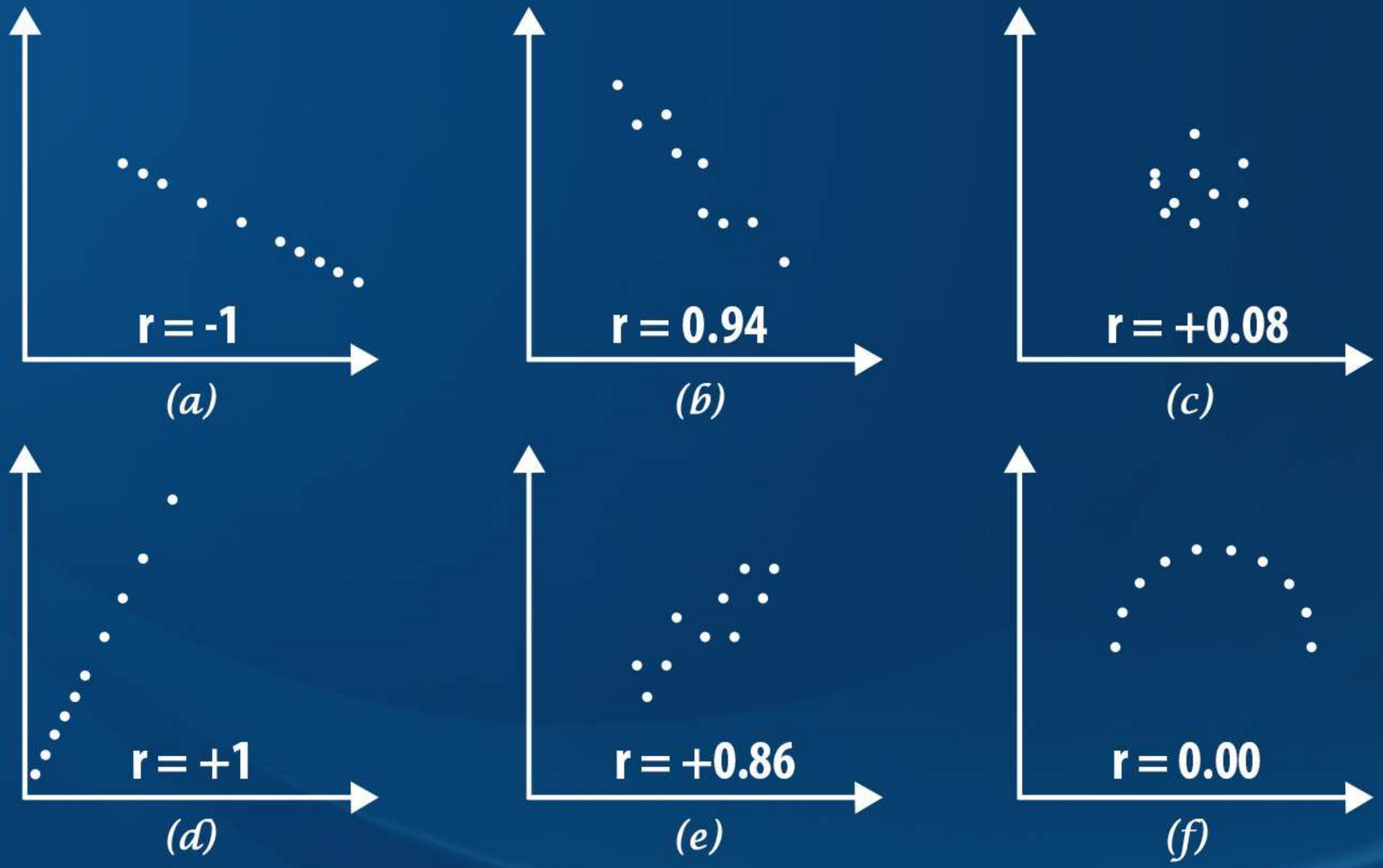


Figure:9

PREDICTION INTERVALS FOR $\hat{Y}$

Confidence Interval for the **mean** of Y, given X_i

$\hat{Y} = b_0 + b_1 X_i$

Prediction Interval for an **individual** Y, given X_i

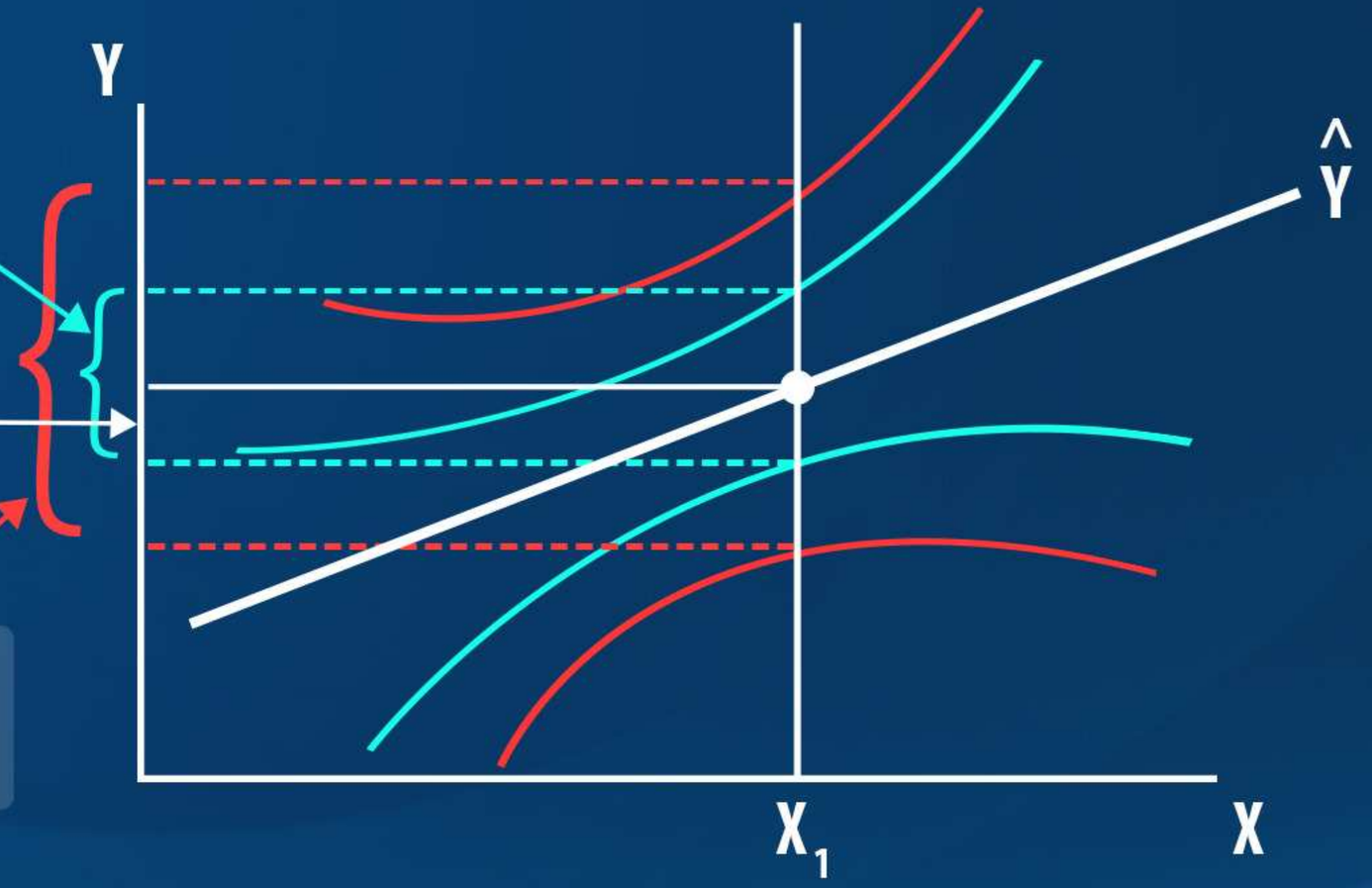


Figure:9

EXAMPLE – PREDICTION INTERVALS

House price (estimate) = 57.438 + 0.2136 * (Area in sq.ft.)

= 57.438 + 0.2136 * (2100)

= 506 +/- 102

= 362 to 608 (is the 95% PI for estimated value at 2100)

CAVEATS

- Whenever we use historical data to predict future values, we are assuming that the past is a reasonable predictor of the future. Thus, we should only use regression to predict the future if the general circumstances that held in the past, such as competition, industry dynamics, and economic environment, are expected to hold in the future.
- With regression, we can forecast dependent value (DV) for a given independent value (IV) provided IV value is within the range of IV levels we've seen historically.
- For example, even if we don't have data for house of size 280, we can still forecast a corresponding house price.

Note: Avoid having outliers in dataset. It tends to produce distorted equation for the model.