Regression

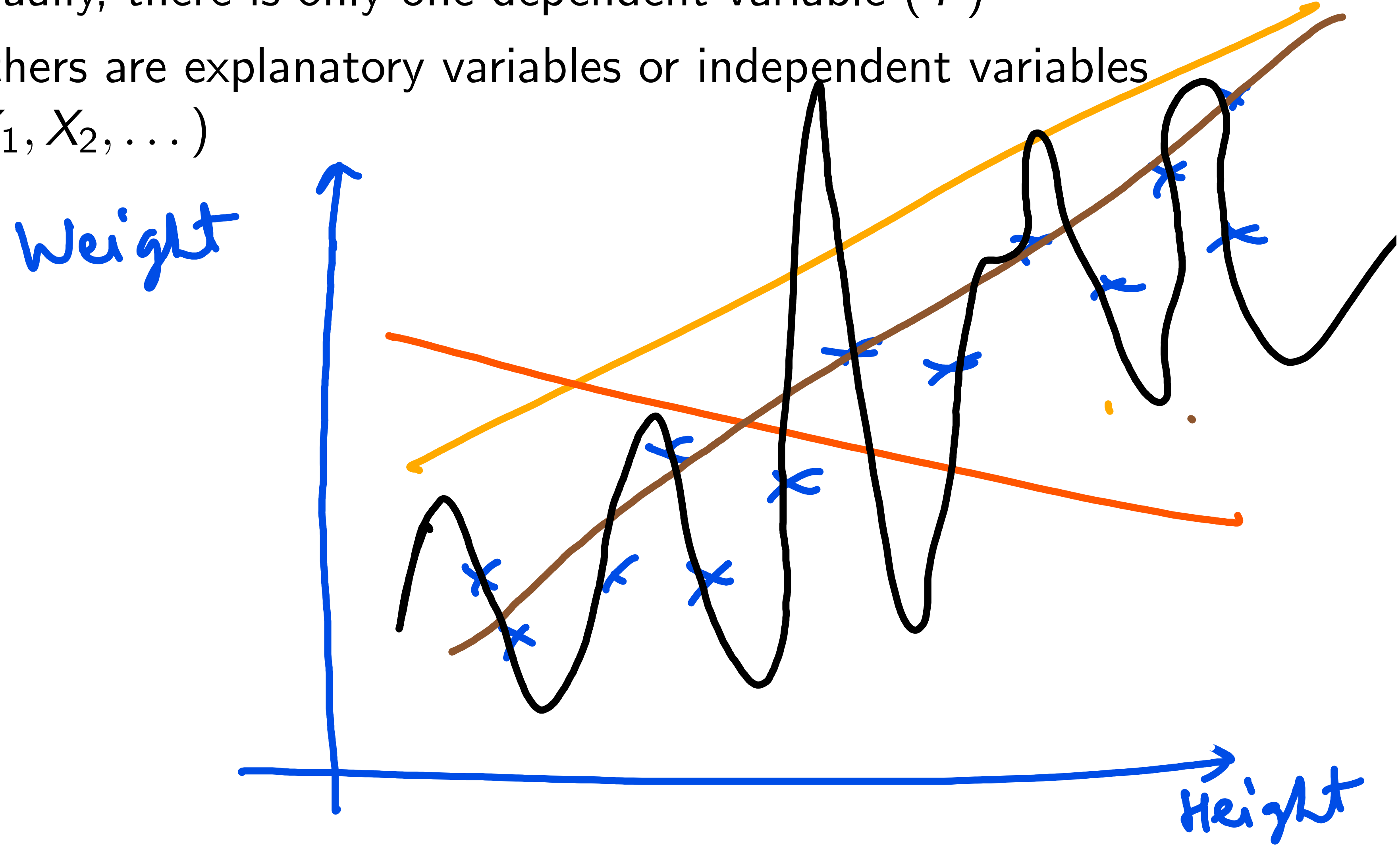
16 Feb 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Regression Analysis

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables ($X_1, X_2, \dots$)



Regression Analysis

$F = ma \qquad T = 2\pi\sqrt{\frac{l}{g}}$

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables ($X_1, X_2, \ldots$)
- Can we assume a functional relationship between Y and the independent variables? $Y = f(X_1, X_2, \ldots)$
- Usually, because of inherent nature of phenomena that we are trying to model, there is randomness and hence

Regression Analysis

- Looking for a relationship between a set of variables
- Usually, there is only one dependent variable (Y)
- Others are explanatory variables or independent variables ($X_1, X_2, \dots$)
- Can we assume a functional relationship between Y and the independent variables? $Y = f(X_1, X_2, \dots)$
- Usually, because of inherent nature of phenomena that we are trying to model, there is randomness and hence
- $Y = f(X_1, X_2, \dots) + \epsilon$
- ϵ is typically assumed to be a random variable with mean 0 and standard deviation σ
- Thus, $E(Y) = f(X_1, X_2, \dots)$

Simple Linear Regression

- If the assumed functional form is linear, we call it linear regression
- If the number of independent variables is one, we call it simple linear regression
- The linear form typically assumed is $Y = \alpha + \beta X + \epsilon$

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

Simple Linear Regression

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- β can be interpreted as average increase in Y for an unit increase in X
- α, in general, has no interpretation

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- α and β are population parameters, and hence are unknown
- Our task would be to estimate the values of α and β from the sample observations

Simple Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

- α and β are population parameters, and hence are unknown
- Our task would be to estimate the values of α and β from the sample observations
- When the number of independent variables is just 1, we can observe the scatter plot to observe if linear relationship can be assumed between the variables
- If the scatter plot doesn't indicate that a linear relationship can be assumed, we should possibly drop the idea of simple linear regression, and do something more to understand the relationship between the variables

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- We would estimate the values of α and β from sample observations
- Denote by $\hat{\alpha}$ and $\hat{\beta}$ the estimates of α and β respectively
- Note that α and β uniquely determine the line
- Thus, given the data, we would determine $\hat{\alpha}$ and $\hat{\beta}$, which would uniquely determine a line
- Which line to fit??

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- We would estimate the values of α and β from sample observations
- Denote by $\hat{\alpha}$ and $\hat{\beta}$ the estimates of α and β respectively
- Note that α and β uniquely determine the line
- Thus, given the data, we would determine $\hat{\alpha}$ and $\hat{\beta}$, which would uniquely determine a line
- The line which minimizes the sum of square of residuals

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- Given Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Minimize: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right)^2$
- Differentiate w.r.t $\hat{\alpha}$ and $\hat{\beta}$ and equate to zero
- We obtain 2 equations in 2 unknowns, which on solving give -
- $\hat{\beta} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$
- $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$

Simple Linear Regression - Estimation of Parameters

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- $\hat{\beta} = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$
- To fully specify the model, one more parameter needs to be estimated, which is ??

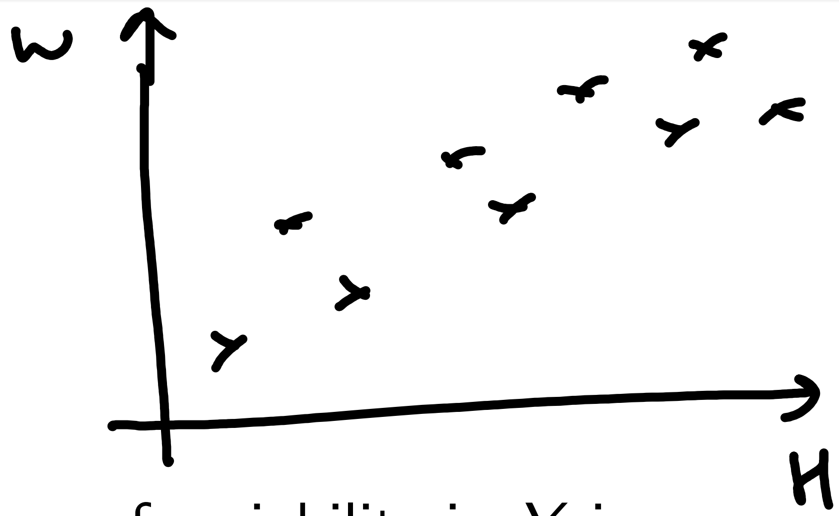
Simple Linear Regression - Estimation of Parameters

Simple Linear Regression Model

$Y = \alpha + \beta X + \epsilon$

- $\hat{\beta} = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$
- To fully specify the model, one more parameter needs to be estimated, which is σ
- σ is estimated using the standard deviation of residuals
- $\hat{\sigma} = s = \sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2}{n-2}}$

Goodness of fit - R^2



- Softwares will report a R^2 to you
- What does it mean??
- Gives an idea about what percentage of variability in Y is explained by the regression equation
- SST = SSR + SSE
- $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i + \hat{y}_i - \overline{y}\right)^2$$

Simple Linear Regression - Properties of Estimates of Parameters

Simple Linear Regression Model

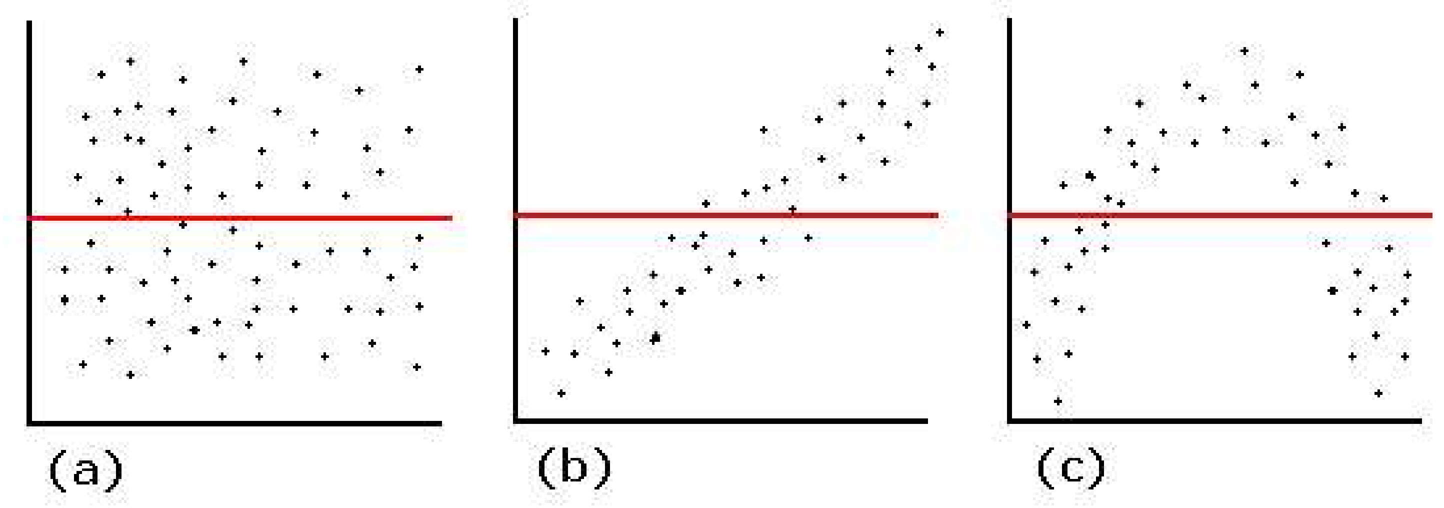
$Y = \alpha + \beta X + \epsilon$

- Sum of residuals is zero
- Residuals are uncorrelated with $x_i's$
- It can also be shown that $\hat{y}_i$ and e_i are uncorrelated
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, since $y_i = \hat{y}_i + e_i$

Assumptions of Regression

- ϵ is a random variable that is normally distributed with mean 0 and s.d. σ
- Variance of ϵ is same for all values of x

Examples of Residual Plots



Source - `http://analyticspro.org/2016/03/05/r-tutorial-residual-analysis-for-regression/`

Multiple Linear Regression

We now have more than 1 independent variables. (say k)

Multiple Linear Regression Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- Interpretation of $\beta' s$??
- How do you obtain α & $\beta' s$??
- Partial Differentiation to obtain $k+1$ equations in $k+1$ unknowns
- Example