IE605: Engineering Statistics

Linear Regression

Manjesh K. Hanawal

Simple Linear Regression

- Assume: Each sample has one feature/attribute ($x_i \in \mathcal{R}$)
- We will fit line of the from $y = \beta_1 x + \beta_0$

- x is called the independent/predictor variable
- y is called the dependent/response variable
- β_1 is the slope and β_0 is the intercept
- We will get different lines for different choice of (β_0, β_1)

- How to quantify how good is a line?
- Choose the best line!

Probabilistic Model for Linearly Related Data

- Instead of $y_i = \beta_1 x_i + \beta_0$ assume data is perturbed by noise
- $y_i = \beta_1 x_i + \beta_0 + \epsilon_i$, where ϵ_i is random perturbation (noise)
- perturbation denotes that data won't be fit the model perfectly
- We assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known

Quantify goodness of a line: Mean Squared Error

- Minimize the distance between the line and points
- distance of point (x_i, y_i) from line (β_0, β_1) (error)

$$y_i - (\beta_1 x_i + \beta_0)$$

- As staying above or below line are equally bad we can take

$$|y_i - (\beta_1 x_i + \beta_0)| \text{ absolute error}$$

$$(y_i - (\beta_1 x_i + \beta_0))^2 \text{ squared error}$$

- We take goodness of line (β_0, β_1) as sum of the squared errors

$$\frac{1}{m}\sum_{=1}^{n}(y_i - (\beta_1 x_i + \beta_0))^2$$

Mean Squared Error (MSE)

The best line: Least Squared Regression

$$\min_{(\beta_0,\beta_1)} \frac{1}{m}\sum_{=1}^{n}(y_i-(\beta_1 x_i+\beta_0))^2$$

Alternate derviation from MLE

- $y_i=\beta_1 x_i+\beta_0+\epsilon_i \implies y_i \sim \mathcal{N}(\beta_1 x_i+\beta_0,\sigma^2)$
- $(\epsilon_1,\epsilon_2,\ldots,\epsilon_n)$ are iid hence $(y_1,y_2,\ldots,y_n)$ are iid.
- Likelihood of $y=(y_1,y_2,\ldots,y_m)$ under the parameters $\beta=(\beta_0,\beta_1)$ is

$$L(y|\beta)=\prod_{i=1}^{m} f(y_i|\beta)=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-(y_i-\beta_1 x_i-\beta_0)^2/2\sigma^2\right\}$$

$$=\frac{1}{(2\pi\sigma^2)^{m/2}}\exp\left\{-\sum_{i=1}^{m}(y_i-\beta_1 x_i-\beta_0)^2/2\sigma^2\right\}$$

$$\arg\max_{\beta} L(y|\beta)=\arg\min_{\beta}\sum_{i=1}^{m}(y_i-\beta_1 x_i-\beta_0)^2$$

Least Squared Solution

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0,\beta_1)} \frac{1}{m} \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2$$

$$\hat{\beta}_1 = \frac{\frac{1}{m}\left(\sum_{i=1}^{m} x_i y_i\right) - \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right)}{\frac{1}{m}\left(\sum_{i=1}^{m} x_i^2\right) - \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^2}$$

$$\hat{\beta}_0 = \left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) - \hat{\beta}_1\left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)$$

Expressing the solutions in terms of statistics

Given a random sample $(X_1, X_2, \ldots, X_m)$

- Sample mean: $\bar{X} = \frac{1}{m}\left(\sum_{i=1}^{m} X_i\right)$
- Sample variance: $S_X^2 = \frac{1}{m-1}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2\right)$
- Sample standard deviations: $S_X = \sqrt{S_X^2}$.

For give data $S = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$

$$\bar{x} = \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right) \quad s_x = \frac{1}{m-1}\left(\sum_{i=1}^{m}(x_i - \bar{x})^2\right)$$

$$\bar{y} = \frac{1}{m}\left(\sum_{i=1}^{m} y_i\right) \quad s_y = \frac{1}{m-1}\left(\sum_{i=1}^{m}(y_i - \bar{y})^2\right)$$

$$r = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \text{ Correlation coefficient}$$

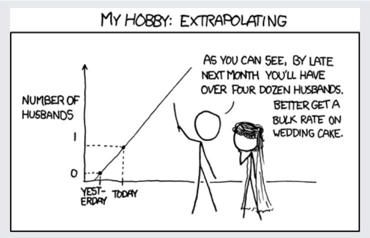
Prediction

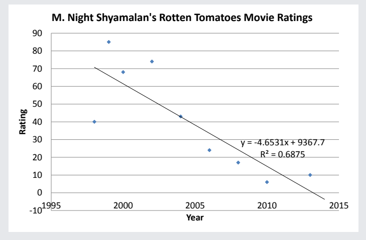
$$\hat{\beta}_1 = r\frac{s_y}{s_x} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Given any sample x, its predicted label is

$$y = \hat{\beta}_1 x + \hat{\beta}_0$$

For what all x we can get prediction?

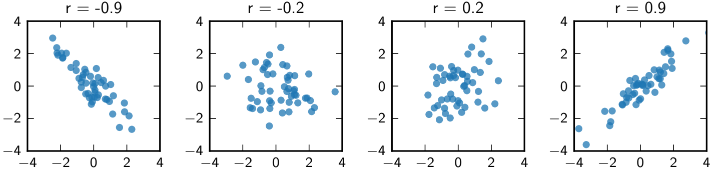




Correlation coefficient

$$r = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- $-1 \leq r \leq 1$. Measure how much y is related to x
- if r is positive y increases in x
- if r is negative y decreases in x



- r^2 is called coefficient of determination (explains how well data is fit).

Multiple Linear Regression

$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, x_i \in \mathcal{R}^d$, where $d > 1$.
Each sample point $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$.

- We can write linear relation: $y_i = \sum_{j=1}^{d} x_{ij}\beta_j + \beta_0$
- $y_i = \sum_{j=0}^{d} x_{ij}\beta_j$, where $x_{i0} = 1$ for all $i = 1, 2 \ldots, m$
- set $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_d)$ and $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{id})$
- Compactly $y_i = x_i\beta^T$ for all $i = 1, 2, \ldots, m$
- The probabilistic model is $y_i = x_i\beta^T + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots x_{1d} \\ 1 & x_{21} & x_{22} & \ldots x_{2d} \\ \vdots & & & \\ 1 & x_{m1} & x_{m2} & \ldots x_{md} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

$y = X\beta^T$ where X is data matrix

The probabilistic model is then

$$y = X\beta^T + \epsilon$$

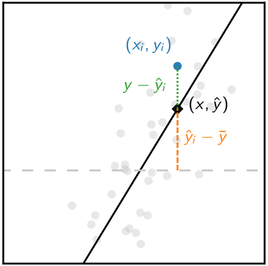
Solution of Multiple Linear Regression

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - x_i \beta^T)^2$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Model Evaluation:

Suppose every point y_i is very close to $\bar{y} \implies y_i$ does not dependent much on x_i and there is not much random error.



$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained by model}} + \underbrace{(y_i - \hat{y}_i)}_{\text{not explained by odel}}$$

Coefficient Determination

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{SSM}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{SSE}}$$

$$1 = \underbrace{\frac{SSM}{SST}}_{r^2} + \underbrace{\frac{SSE}{SST}}_{1-r^2}$$

- r^2 is called the coefficient of determination (square of coefficient of correlation!)
- Captures the fraction of variability explained by model
- It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph
- closer to 1, the better.