Linear Regression

Agenda

After this session, you will know:

- Business Context
- Covariance & Correlation
- Transformations
- Math behind Linear Regression
- SSE, SST & R-squared

Business Context (When to use Linear Regression)

Need to predict medical expense

For a health insurance company to be profitable, it needs to gather more premium than it spends on medical care for its customers. So its important for insurance companies to develop predictive models that forecast medical expenses for the insured people.

Our objective of using Linear Regression is to use historical patient data to estimate the average medical care expenses people.

These estimates can be used to create premium tables for certain segment of people that can set the price of the premiums higher or lower depending on the expected treatment costs.

A few basics

- Covariance & Correlation
- Distribution – Gaussian Distribution – Skewness
- Dependent variable: the variable we wish to explain usually denoted by Y
- Independent variable(s): the variable(s) used to explain the dependent variable. Denoted by X

Covariance & Correlation

Covariance is a measure of how changes in one variable are associated with changes in a second variable.

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

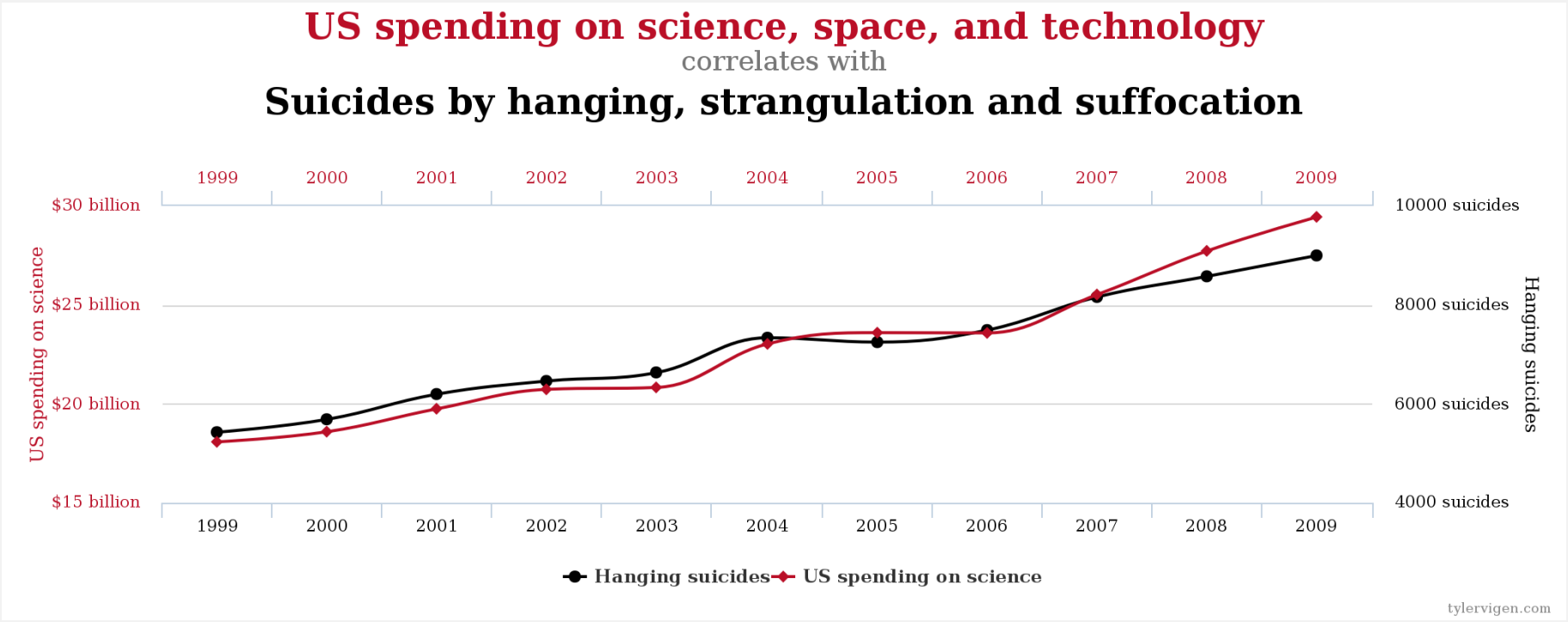
Correlation is a scaled version of covariance that takes on values in [−1,1] with a correlation of ±1 indicating perfect linear association and 0 indicating no linear relationship.

$$\rho=\frac{\mathrm{cov}(X,Y)}{\sigma_X\cdot\sigma_y}$$

Understanding Correlation

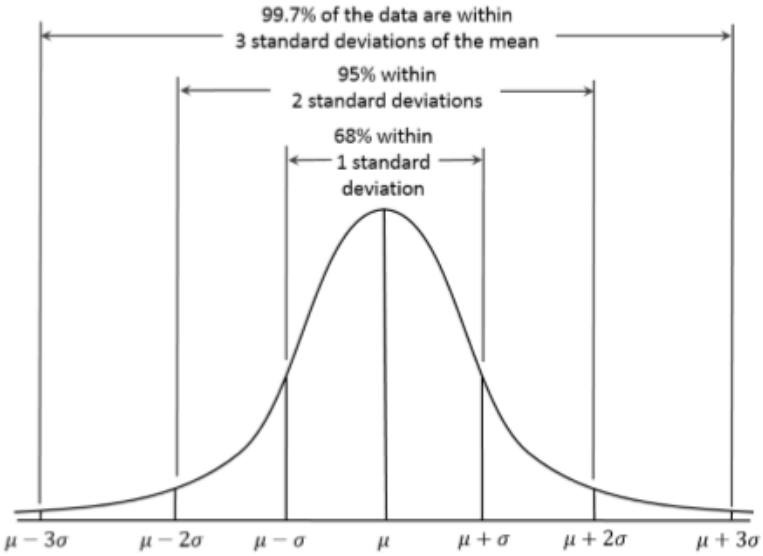
Correlation is not causation

Funny Correlation

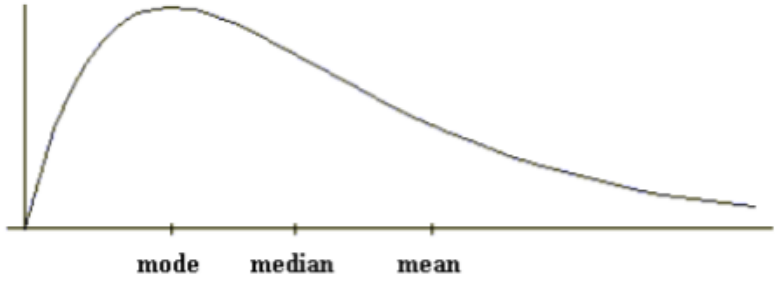


Correlation is not causation

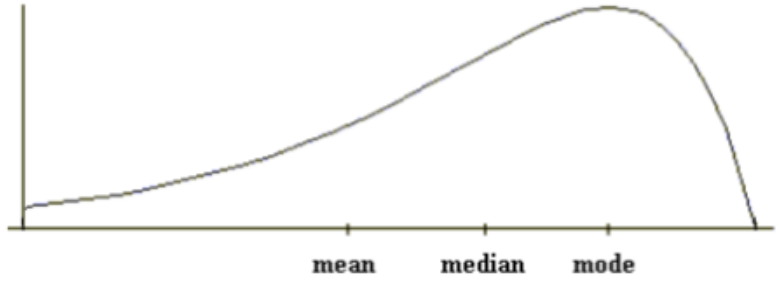
Gaussian/Normal Distribution



$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Right Skewed

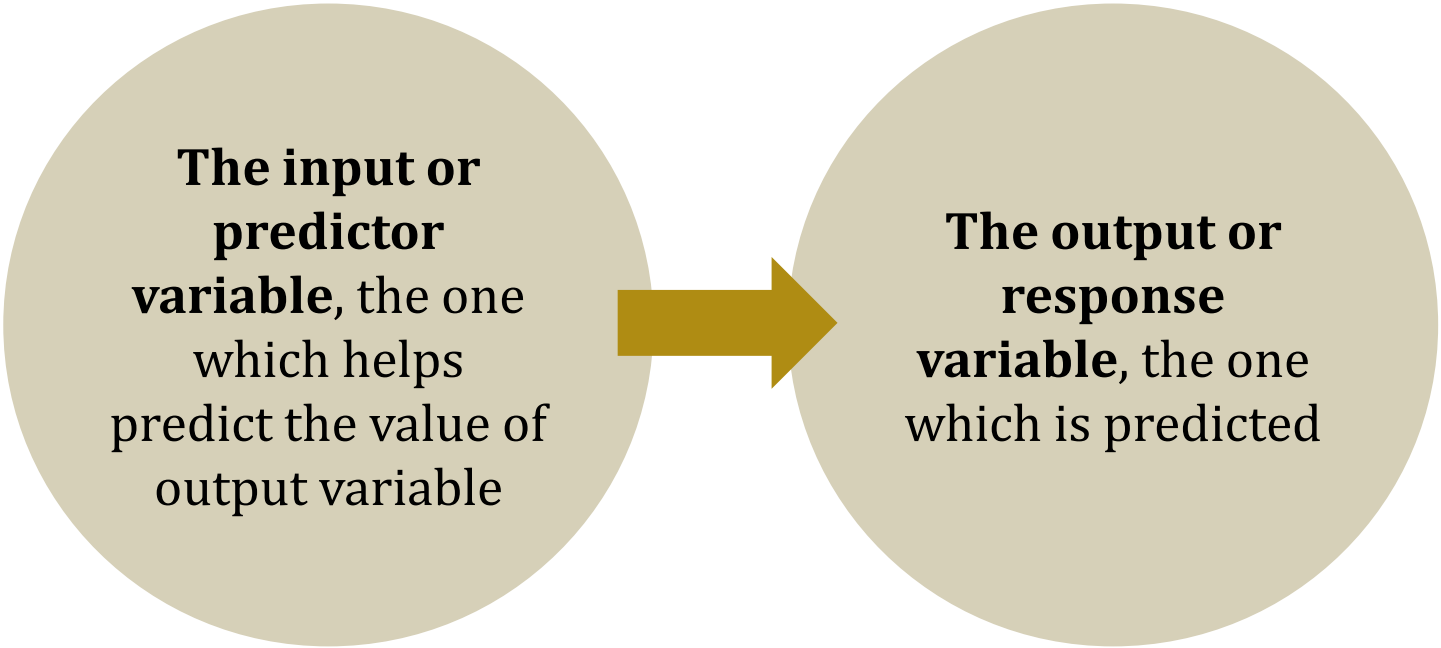


Left Skewed

Math behind Linear Regression

Math behind Linear Regression

There are 2 kinds of variables in a model:



Math behind Linear Regression (Contd.)

In case of Linear Regression, we assume that Y (Expense) is a linear function of X (BMI) and to estimate Y, we write:

$$Y = \alpha + \beta X + \varepsilon \qquad \text{i.e.} \qquad \text{Expense} = \alpha + \beta\,(\text{BMI}) + \varepsilon$$

Y: The dependent variable

X: Independent variable

β: Model Parameters

ε: Random Error, how the observation deviate from population mean

BMI	Medical Expense
19	10295
27	20627
27	18859
26	22136
30	21171
29	13422
27	12050
22	21820
24	18950
20	10467
28	17985
26	22737

The Error Term

Fixed:$\alpha + \beta X$

-- mean of Yi, (E[Yi])

Random: ε_i

-- Variability of Yi

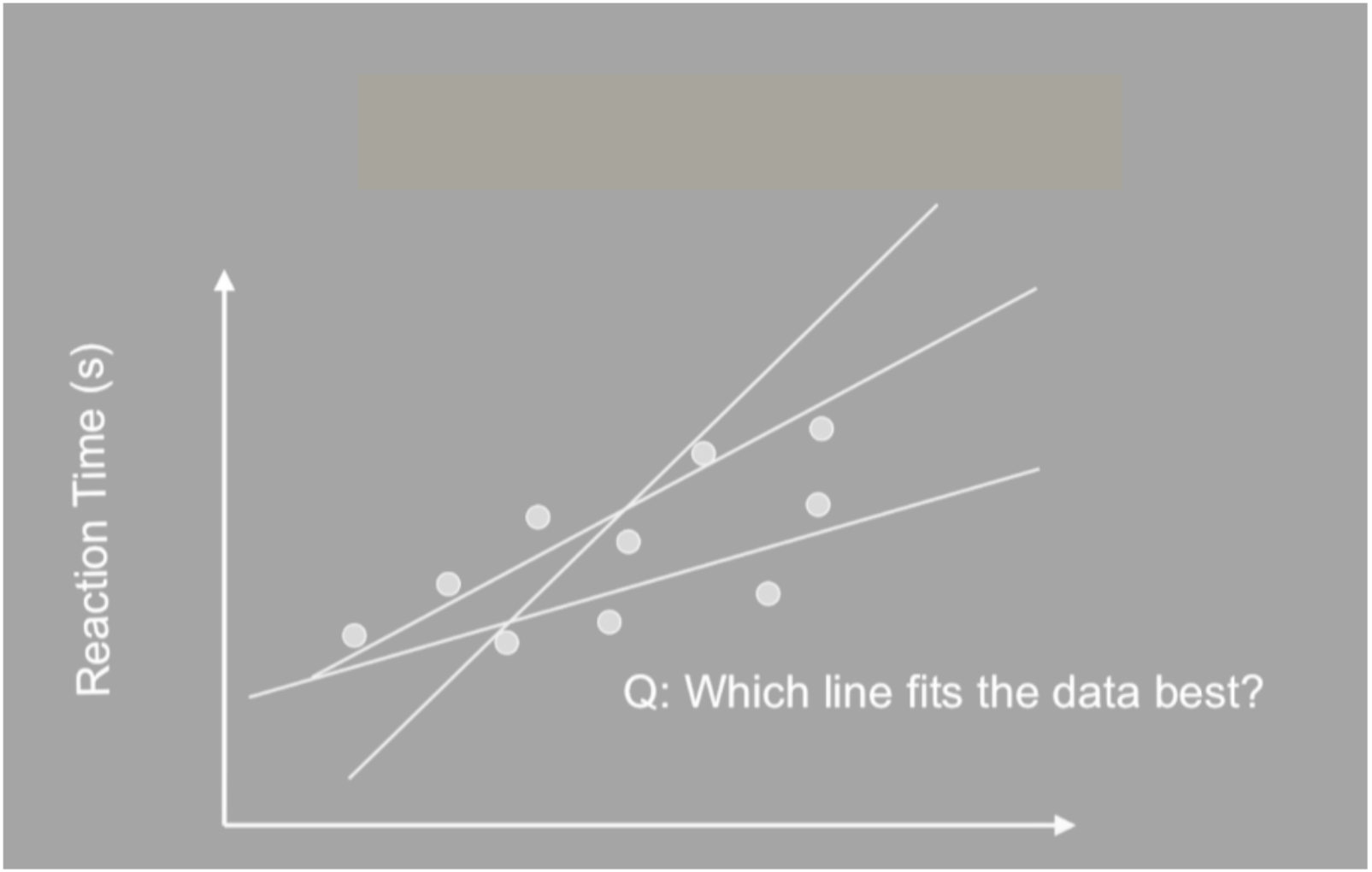
-- $E(\varepsilon_i) = 0$

--$Var(\varepsilon_i) = \sigma^2$

--$Cov(\varepsilon_i, \varepsilon_j) = 0$

--It follows that variance of Y is σ^2

Fitting the Model



Least Squares Linear Equation

The Least Square Regression line is given by:

$$f(x)=b + mx$$

Which will minimize the sum of squared error, which are the errors in using the regression function $f(x)$ to estimate the true y values

$$(y_1-f(x_1))^2 + (y_2-f(x_2))^2 + (y_3-f(x_3))^2 + \ldots\ldots + (y_n-f(x_n))^2$$

where $e_i = y_1-f(x_1)$ is the error approximating y_i

Derivation

Sum of squared Errors is given by:

$$SS_{(residuals)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

If $\hat{y} = b_0 + b_1 x$ then error in estimate for x_i is $e_i = y_i - \hat{y}_i$

Minimize Sum of Squared Errors (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Derivation (Contd.)

To minimize the error, 1st order derivative should be equal to zero:

$$\delta e/\delta b_0 = 0$$

$$\delta e/\delta b_1 = 0$$

So, we get 2 equations and 2 unknowns –b_0 and b_1

$$\delta e/\delta b_0 = \sum_{i=1}^{n} 2\,(y_1 - b_0 - b_1 x_1)\,(-1) = 0 \ldots\ldots\ldots\ldots (1)$$

$$\delta e/\delta b_1 = \sum_{i=1}^{n} 2\,(y_1 - b_0 - b_1 x_1)\,(-x_1) = 0 \ldots\ldots\ldots\ldots (2)$$

Expanding these equations, we calculate the b_0 & b_1

It is more convenient to deal with Multiple Regression Models if they are expressed in matrix notation.

Matrix Equation for Simple Linear Regression

Using (x1,y1), (x2,y2), (x3,y3)……(xn,yn) we would have the following equations:

$$y_1 = (\beta_0 + \beta_1 x_{11}) + \varepsilon_1$$
$$y_2 = (\beta_0 + \beta_1 x_{12}) + \varepsilon_2$$
$$y_3 = (\beta_0 + \beta_1 x_{13}) + \varepsilon_3$$
$$\vdots$$
$$y_n = (\beta_0 + \beta_1 x_{1n}) + \varepsilon_n$$

In Matrix form, it would look as follows:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad x = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \quad M = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

n x 1

This gives us the Matrix equation: Y=XM+E

Using Linear regression technique, we need to solve for M which will give us the β and m

Matrix Equation for Multiple Linear Regression

Using more than 1 predictor variable, we would have the following equations:

$$
\begin{aligned}
y_1 &= (\beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31}) + \varepsilon_1 \\
y_2 &= (\beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32}) + \varepsilon_2 \\
y_3 &= (\beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \beta_3 x_{33}) + \varepsilon_3 \\
&\vdots \\
y_n &= (\beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_3 x_{3n}) + \varepsilon_n
\end{aligned}
$$

In Matrix form, it would look as follows:

$$
\underset{n \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{n \times (3+1)}{\begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix}} \underset{(3+1) \times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}} + \underset{n \times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}}
$$

Solution to Least Square Equation

Solution to the least square equation $Y = \beta X + \varepsilon$ is:

Least square estimator of $\beta = \left(X^T X\right)^{-1} X^T Y$

provided that the inverse matrix $\left(X^T X\right)^{-1}$ exists.

The $\left(X^T X\right)^{-1}$ matrix will always exist if the regressors are linearly independent

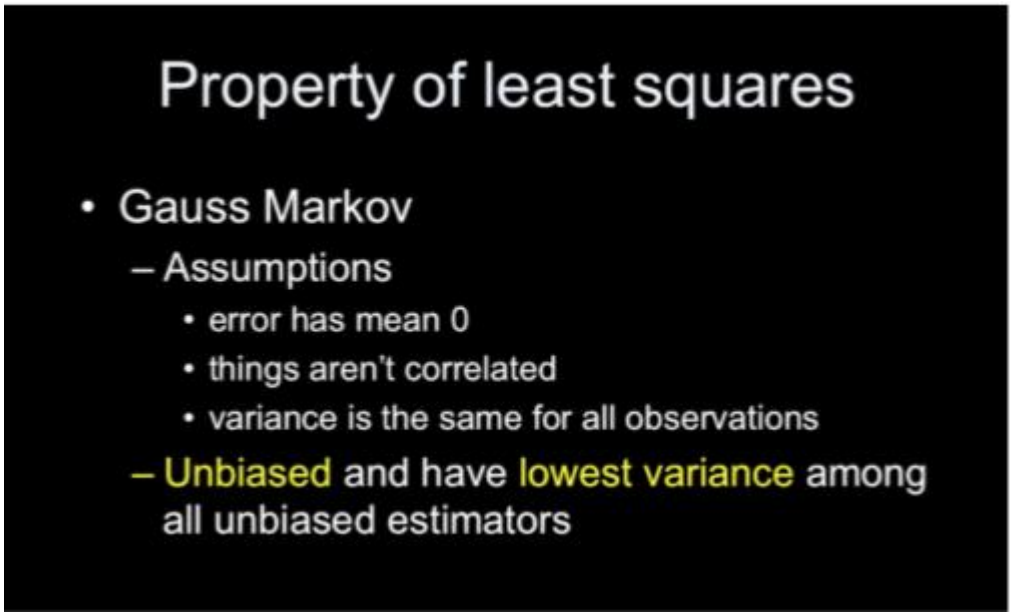
i.e. if no columns of the X matrix is a linear combination of the other columns.

The sum of squared errors is: $SSE = \varepsilon^T \varepsilon$ (scalar value)

Properties of Least Squares – Gauss Markov Theorem

Gauss Markov Theorem says that the OLS estimator is BLUE

Best (minimum variance)
Linear (linear function of the data)
Unbiased
Estimator (estimator of the coefficients of β



Variance is same for all observations i.e. homoscedasticity

Estimators are unbiased & lowest variance. Not low biased, low variance.

Sum of Squared Total Error

What is Total Error?

Calculate how much error would have if don't even try to do regression, and instead just guess the mean of all the values

To get the total sum squared error:

- Start with the mean value
- For every data point subtract that mean value from the data point value
- Square that difference
- Add up all the differences

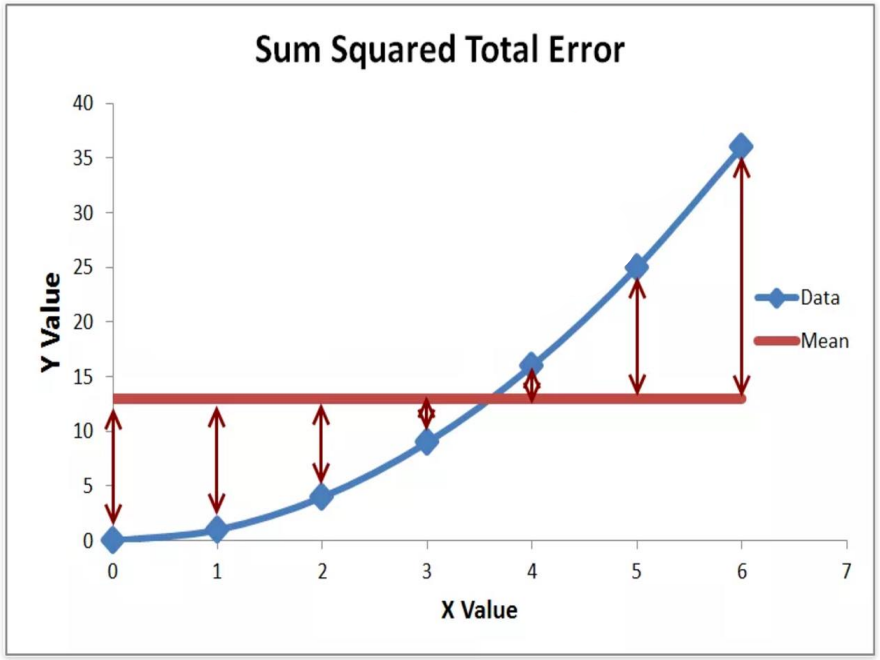
Sum of Squared Total Error

The distance between the blue dots and the mean value (brown line) is the error

Square that value for squared error

Sum the squared error to get Sum of Squared Total Error

We will denote it with SST

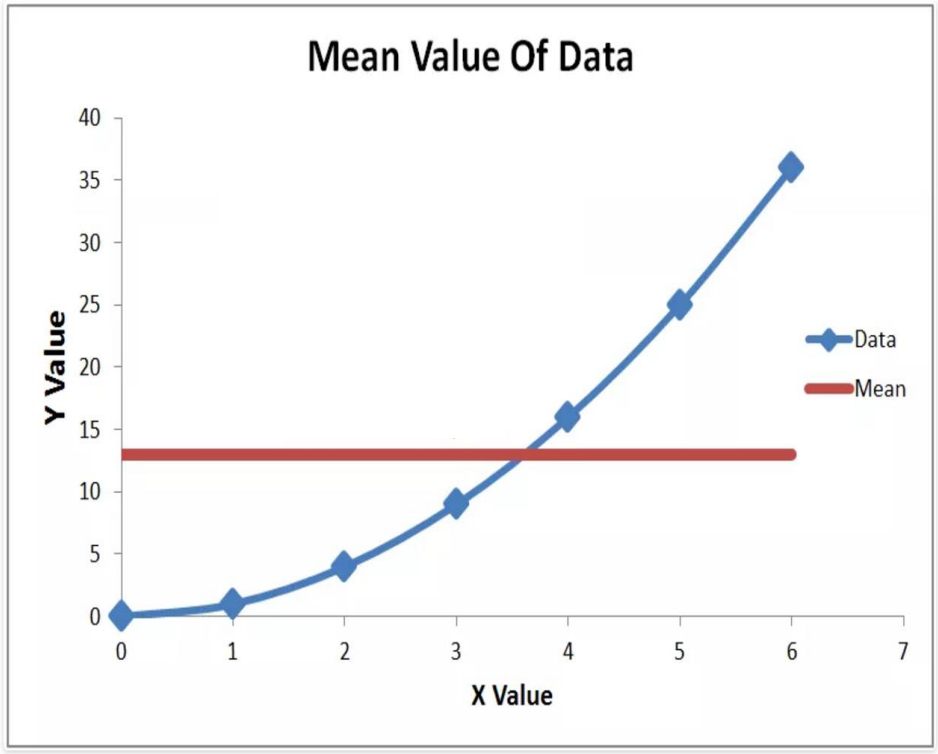


What is R-squared

What is R-squared?

What is R squared?

It is how much better your regression line is than a simple horizontal line through the mean of the data.

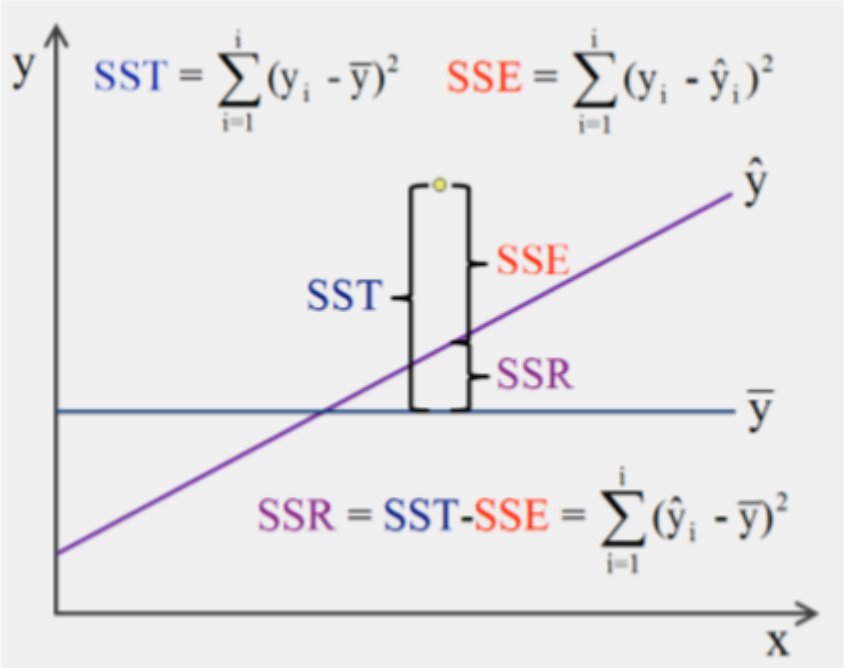


Blue line denotes the data for which we are trying to generate a regression to and the brown line is the average of data

The brown line is the value that gives the lowest summed squared error to the blue data points

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Checking its Efficacy



$$SST = \sum(yact - yavg)^2$$

$$SSR = \sum(ypred - yavg)^2$$

$$SSE = \sum(yact - ypred)^2$$

$$R^2 = 1 - (SSE/SST)$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

- Since SST > SSR, R^2 will be between 0 and 1. Closer to 1, better the model.
- A good R^2 does not mean that the model is a very efficient one.
- May other factors need to be analyzed

Another version of R^2 is: $R^2 = [corr(y, \hat{y})]$

What is a good R-Squared value?

- Lower the error in your regression analysis relative to total error, higher the R squared value

- The best R squared value we get is 1

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

To get R squared as 1, we need to have SSE i.e. $SS_{regression}$ be Zero

$$R^2 = 1 - \frac{0}{SS_{Total}} \rightarrow 1.0$$

Assumptions of Linear Regression

Assumptions of Linear Regression - 1

Assumption #1	One dependent variable that is measured at continuous level
Two or more independent variables that are measured either at the continuous or nominal level	Assumption #2
There needs to be a linear relationship between the dependent variable and each of your independent variables ***(Scatter Plot)***	Assumption #3

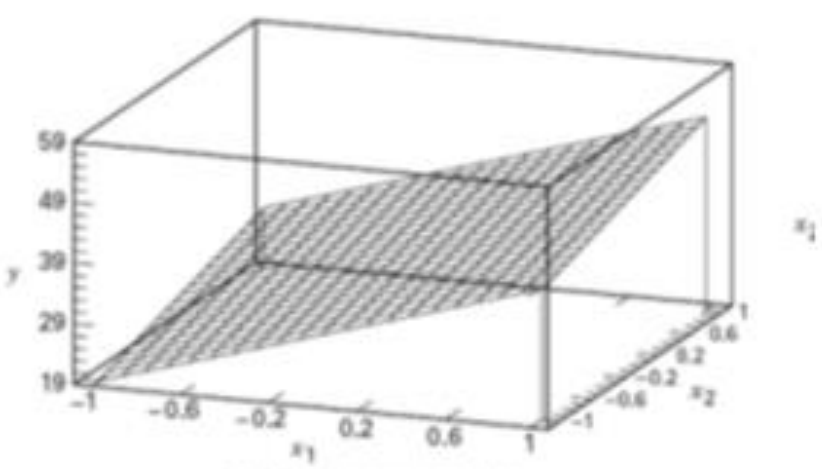
Assumptions of Linear Regression - 2

Assumption #4	Data should not show heteroscedasticity
Predictors must not show multicollinearity, i.e. the matrix would be a full rank matrix ***(VIF)***	Assumption #5
Assumption #6	There should be no significant outliers
Check that the residuals (errors) are approximately normally distributed	Assumption #7

Interaction

Interaction effects occur when the effect of one variable depends on the value of another variable.

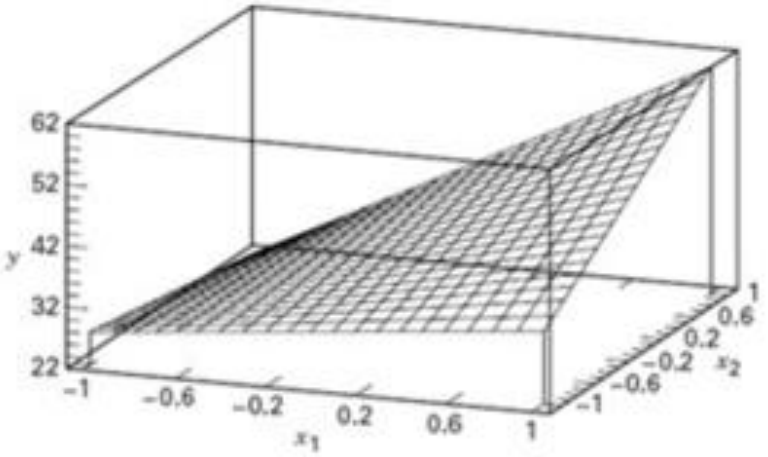
Response Surface for a model without interaction, say, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



(a) The response surface

Response Surface for a model without interaction, say,

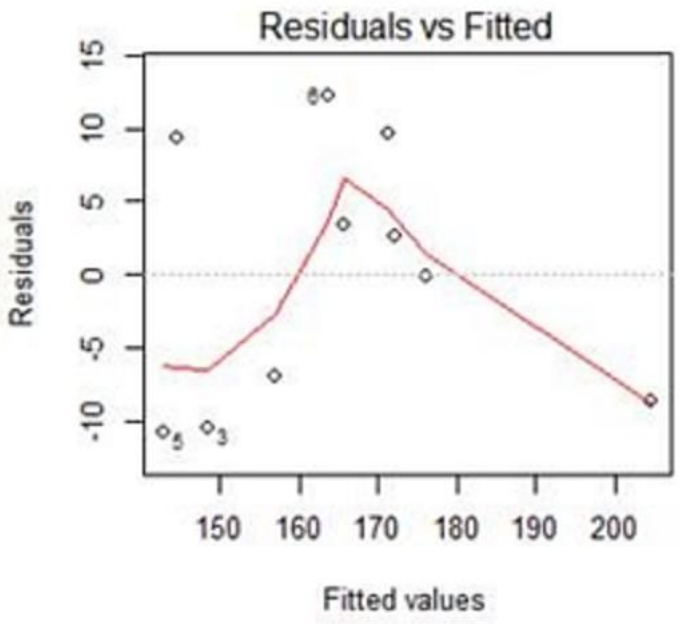
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

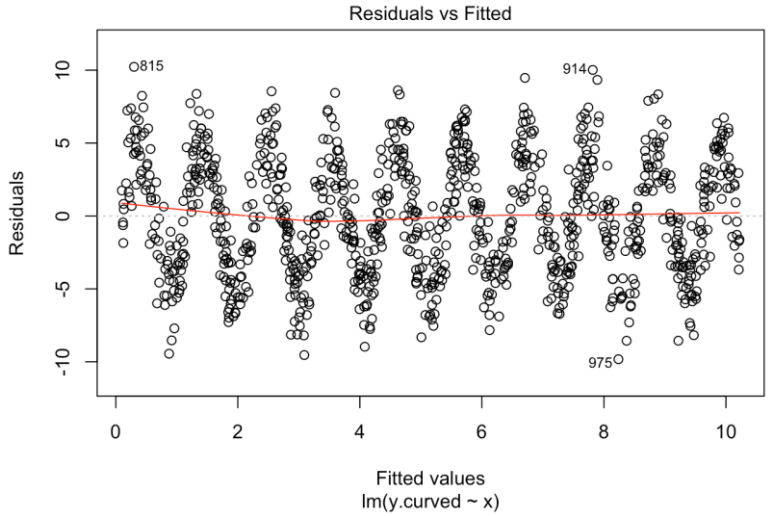


(a) The response surface

Diagnostic Plots

Diagnostic Plot – Residual Vs. Fitted





- This is a simple scatter plot between residuals & predicted values
- It should look more or less random and should not exhibit much distinctive pattern, no non-linear trends or changes in variability

- We see a clear trend in the residuals
- Have a periodic trend
- Unfortunately the scatterplot smoother (red line) isn't doing a good job here
- Don't always trust the red curve

Steps to Build The Model