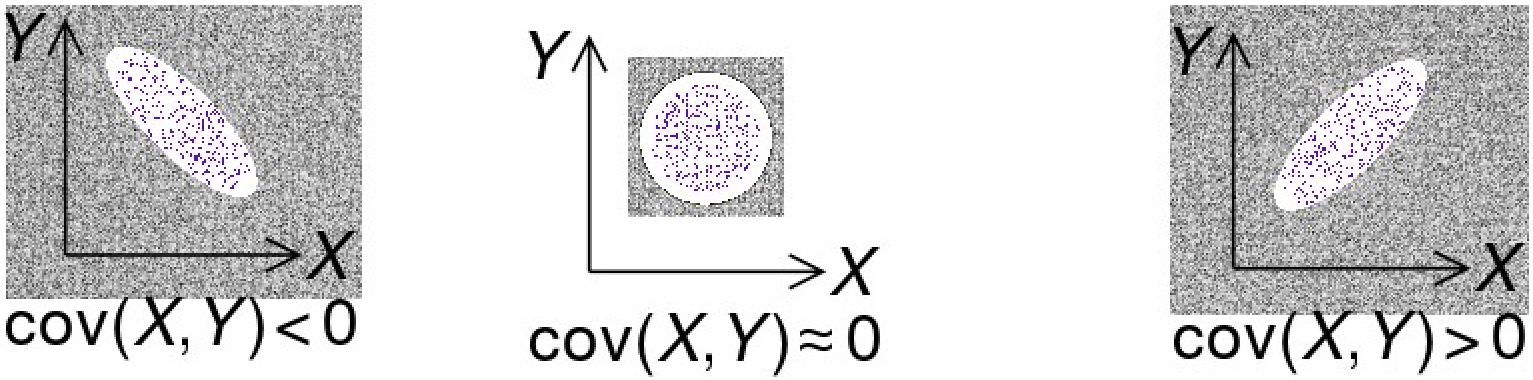
Covariance

- To measure the relationship between two mathematical variables or measured data values covariance and correlation are used
- Covariance and correlationare very similar
- Covariance is a measure of the joint variability of two random variables
- The sign of the covariance shows the tendency in the linear relationship between the variables.



Correlation

- The magnitude of the covariance is not easy to interpret because it is not normalized

$$\text{cov}_{XY} = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

- Correlation measures both the strength and direction of the linear relationship between two variables

$$\text{corr}_{XY} = \rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)]/(\sigma_X \sigma_Y)$$

- The value of covariance lies between -∞ and +∞.
- correlation valueswill be between -1 and +1
- covariance only measures how two variables change together

Correlation

- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations
- For a population
- Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho)
- $$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Where:
- **cov** is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Correlation

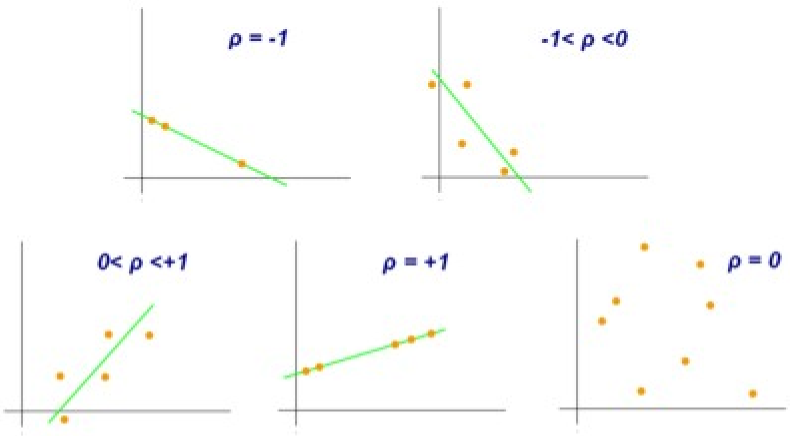
- For a sample
- Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient
- $$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

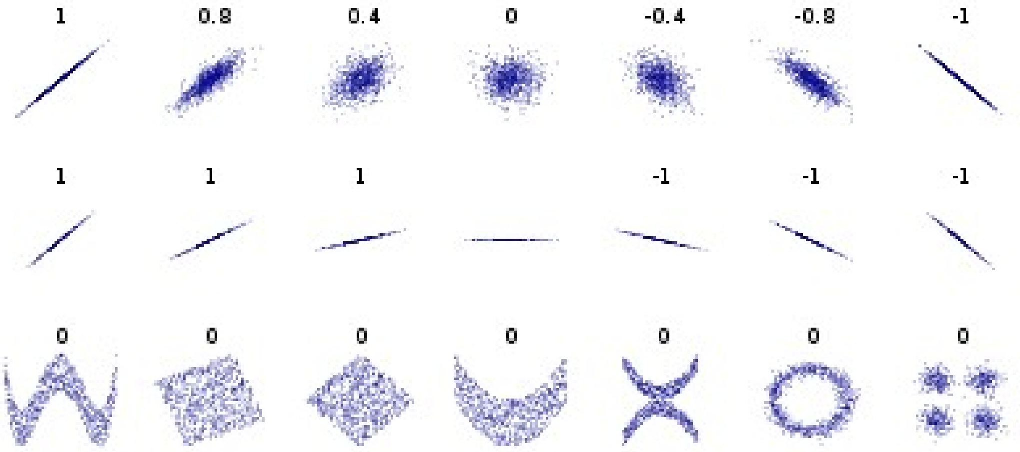
$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

- where:
- n is sample size
- x_i , y_i are the individual sample points indexed with i
- $\bar{x}$, $\bar{y}$ the sample means

Correlation

- Examples of scatter diagrams with different values of correlation coefficient (ρ)
- Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom)





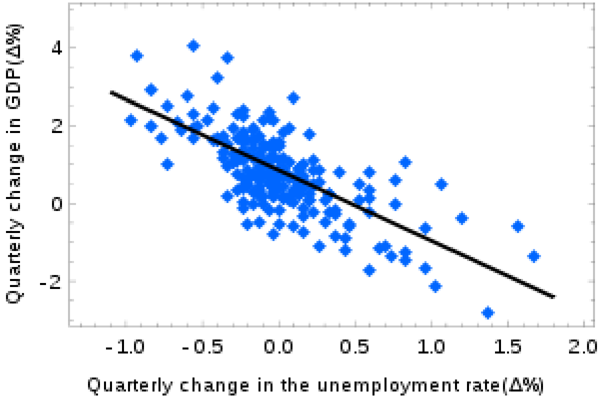
Correlation

- The absolute values of both the sample and population Pearson correlation coefficients are on or between 0 and 1.
- Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation).
- The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).
-

Correlation Coefficient Value (r)	Direction and Strength of Correlation
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

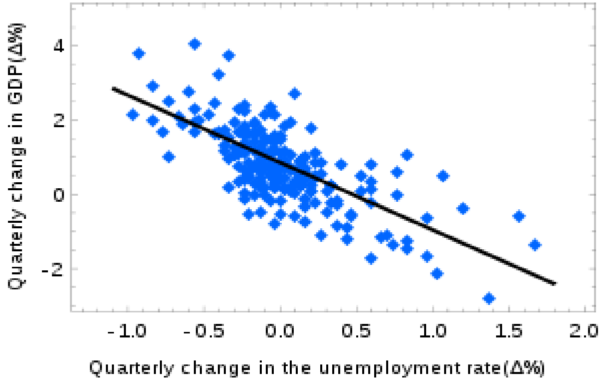
Simple Linear Regression

- simple linear regression is a linear regression model with a single explanatory variable.
- That is, it concerns two-dimensional sample points with one independent variable and one dependent variable and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable.
- The adjective simple refers to the fact that the outcome variable is related to a single predictor.



Simple Linear Regression

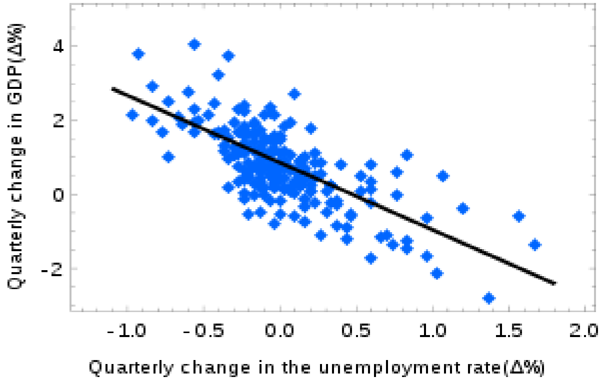
- Consider the model function
- $y = \alpha + \beta x$



- which describes a line with slope β and y-intercept α.
- In general such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables;
- we call the unobserved deviations from the above equation the errors

Simple Linear Regression

- Suppose we observe n data pairs and call them $\{(x_i, y_i), i = 1, \ldots, n\}$. We can describe the underlying relationship between y_i and x_i involving this error term ε_i by

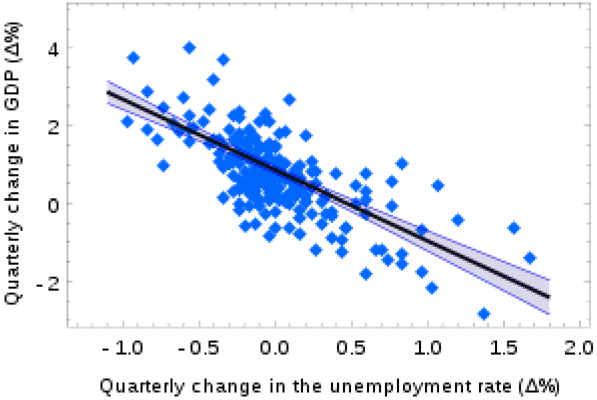


$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

- This relationship between the true (but unobserved) underlying parameters α and β and the data points is called a linear regression model.

Simple Linear Regression

- The goal is to find estimated values α^ and β^ for the parameters α and β which would provide the "best" fit in some sense for the data points.
-



- ordinary least squares (OLS) is method for estimating the unknown parameters in a linear regression model by making the sum of these squared deviations as small as possible.
- Finds a line that minimizes the sum of squared residuals ε_i^

$$\hat{\varepsilon}_i = y_i - \alpha - \beta x_i . \qquad SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Simple Linear Regression

- We look for $\widehat{\alpha}$ and $\widehat{\beta}$ that minimize the sum of squared errors (SSE):

$$\min_{\widehat{\alpha},\widehat{\beta}} \mathrm{SSE}\left(\widehat{\alpha},\widehat{\beta}\right) \equiv \min_{\widehat{\alpha},\widehat{\beta}} \sum_{i=1}^{n}\left(y_i - \widehat{\alpha} - \widehat{\beta}x_i\right)^2$$

To find a minimum take partial derivatives with respect to $\widehat{\alpha}$ and $\widehat{\beta}$

$$\frac{\partial}{\partial\widehat{\alpha}}\left(\mathrm{SSE}\left(\widehat{\alpha},\widehat{\beta}\right)\right) = -2\sum_{i=1}^{n}\left(y_i - \widehat{\alpha} - \widehat{\beta}x_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \widehat{\alpha} - \widehat{\beta}x_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\widehat{\alpha} + \widehat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\widehat{\alpha} + \widehat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i = \widehat{\alpha} + \frac{1}{n}\widehat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \bar{y} = \widehat{\alpha} + \widehat{\beta}\bar{x}$$

Simple Linear Regression

- Before taking partial derivative with respect to $\widehat{\beta}$, substitute the previous result for $\widehat{\alpha}$.

$$\min_{\widehat{\alpha},\widehat{\beta}} \sum_{i=1}^{n} \left[y_i - \left(\bar{y} - \widehat{\beta}\bar{x} \right) - \widehat{\beta} x_i \right]^2 = \min_{\widehat{\alpha},\widehat{\beta}} \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right) - \widehat{\beta} \left(x_i - \bar{x} \right) \right]^2$$

Now, take the derivative with respect to $\widehat{\beta}$:

$$\frac{\partial}{\partial \widehat{\beta}} \left(\mathrm{SSE}\left(\widehat{\alpha}, \widehat{\beta} \right) \right) = -2 \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right) - \widehat{\beta} \left(x_i - \bar{x} \right) \right] \left(x_i - \bar{x} \right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - \bar{y} \right) \left(x_i - \bar{x} \right) - \widehat{\beta} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = 0$$

$$\Rightarrow \widehat{\beta} = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y} \right) \left(x_i - \bar{x} \right)}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

And finally substitute $\widehat{\beta}$ to determine $\widehat{\alpha}$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

Simple Linear Regression

- Find the regression coffcients and pearson correlation coefficient

x	0	2	2	3
y	1	4	3	5

- $y=\alpha+\beta x$

$$\alpha=\overline{y}-\beta\overline{x}$$

$$\beta=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

X	y	x^2	y^2	xy
0	1	0	1	0
2	4	4	16	8
2	3	4	9	6
3	5	9	25	15
7	13	17	51	29

n= 4

$\Sigma x = 7$

$\therefore \bar{x} = \frac{7}{4} = 1.75$

$\Sigma y = 13$

$\therefore \bar{y} = \frac{13}{4} = 3.25$

Simple Linear Regression

- Find the regression coffcients and pearson correlation coefficient

x	0	2	2	3
y	1	4	3	5

$$a_1 = \frac{n \Sigma x y - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{4(29) - (7)(13)}{4(17) - (7)^2} = \frac{116 - 91}{68 - 49} = 1.316$$

$$Y = 0.947 + 1.316x$$

$$a_0 = \bar{y} - a_1 \bar{x} = 3.25 - (1.316)(1.75)$$

$$= 0.947$$

Coefficient of correlation (R) is

$$R = \frac{n \sum xy - (\sum x)(\sum y)}{[(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)]^{1/2}}$$

$$= \frac{4(29) - (7)(13)}{[4(17) - (7)^2][4(51) - 13^2]} = \frac{25}{[(19)(35)]^{1/2}}$$

$$= \frac{25}{25.8} = 0.969$$

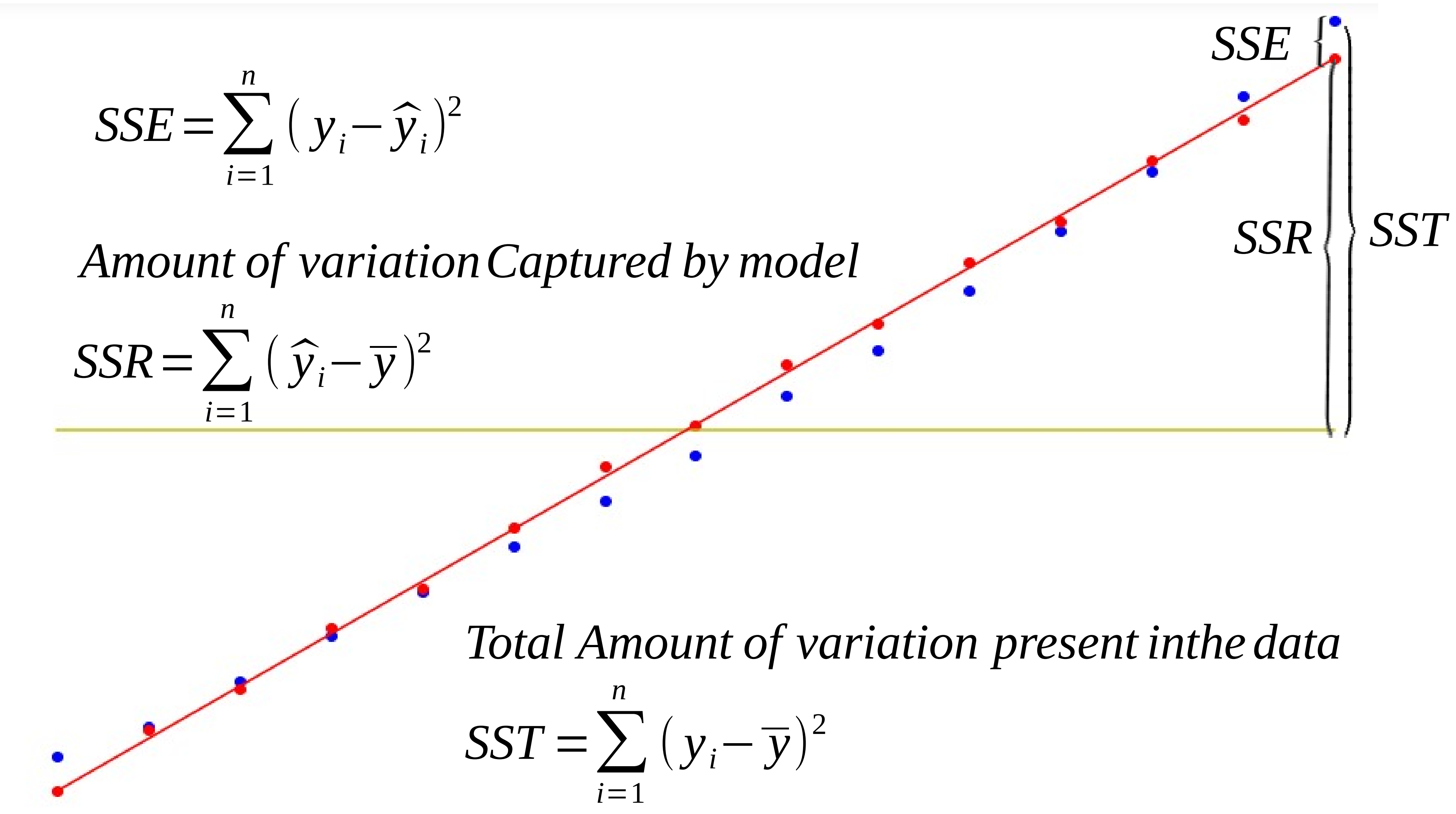
Simple Linear Regression

- No. X Height (m)   Y Mass (kg)
- 1   1.47   52.21
- 2   1.50   53.12
- 3   1.55   54.48
- 4   1.52   55.84
- 5   1.57   57.20
- mean of x   1.522
- mean of y   54.57
- correlation coefficient r 0.86441627
- A   -12.27197452
- B   43.91719745

$$SSE=\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2$$

Amount of variation Captured by model

$$SSR=\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2$$



Total Amount of variation present inthe data

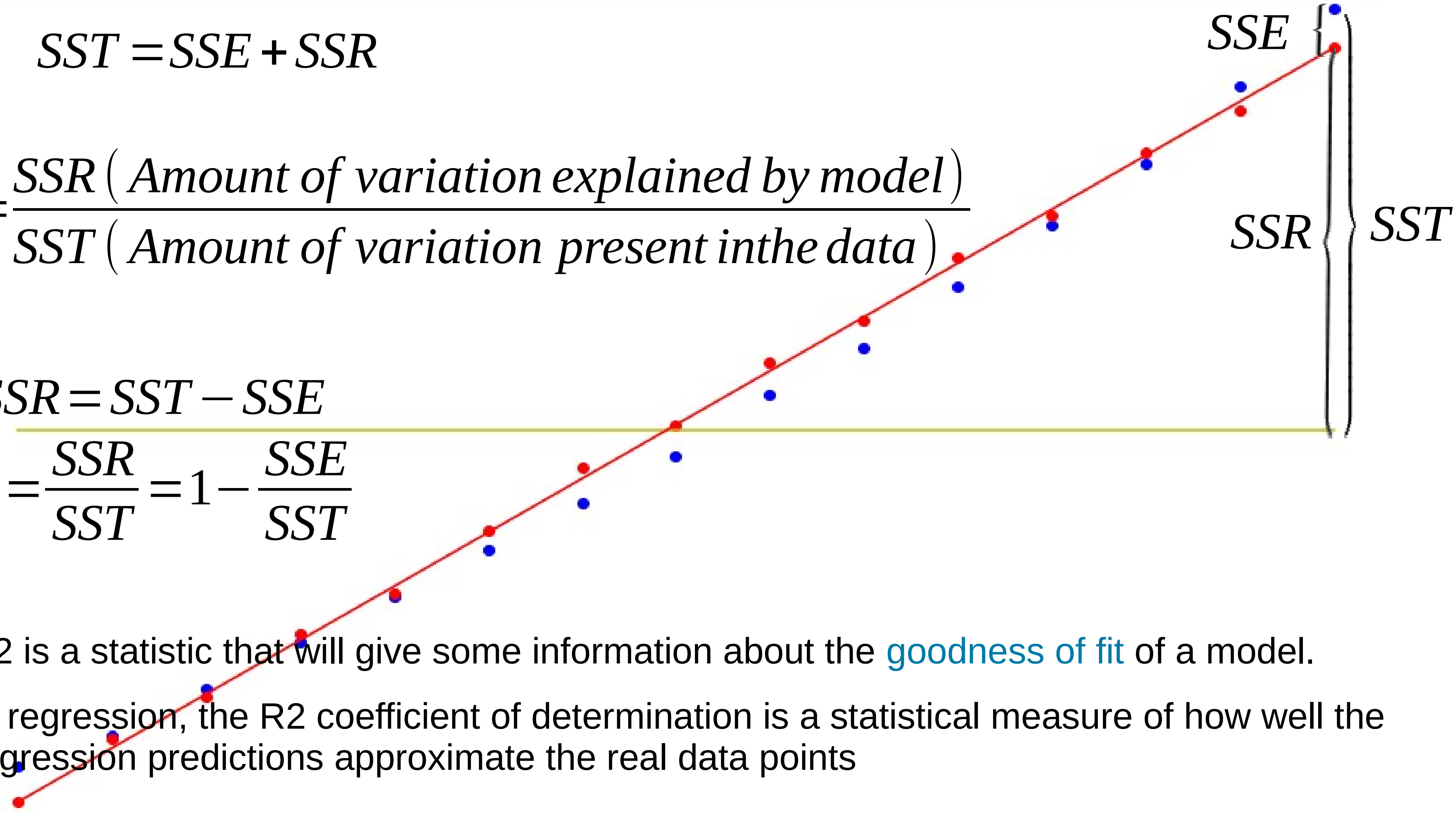
$$SST=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$$

$$SST = SSE + SSR$$

$$R^2 = \frac{SSR\left(\text{Amount of variation explained by model}\right)}{SST\left(\text{Amount of variation present in the data}\right)}$$

$$SSR = SST - SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



R2 is a statistic that will give some information about the goodness of fit of a model.

In regression, the R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points

Multiple Linear Regression

- Multiple linear regression (MLR) is a multivariate statistical technique for examining the linear correlations between two or more independent variables (IVs) and a single dependent variable (DV).
- Research questions suitable for MLR can be of the form "To what extent do X1, X2, and X3 (IVs) predict Y (DV)?"
- e.g., "To what extent does people's age and gender (IVs) predict their levels of blood cholesterol (DV)?"
- **When to use Multiple Linear Regression**
  - there should be one dependent and more than one independent variables
  - The relationship between dependent variable and independent variables is linear

Multiple Linear Regression

- **Linearity**
- Check scatterplots between the DV (Y) and each of the IVs (Xs) to determine linearity:
  - Are there any bivariate outliers? If so, consider removing the outliers.
  - Are there any non-linear relationships? If so, consider using a more appropriate type of regression.

Multiple Linear Regression

- **Homoscedasticity**
- Based on the scatterplots between the IVs and the DV:
  - Are the bivariate distributions reasonably evenly spread about the line of best fit?
  - Also can be checked via the the residuals plots.

- **Multicollinearity**
- IVs should not be overly correlated with one another.

Multiple Linear Regression

- Multiple linear regression is a generalization of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable.
- The basic model for multiple linear regression is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- for each observation i = 1, ... , n.

Multiple Linear Regression

- Example

Height(cm)	Gender	Weight(kg)
152	0	49
155	0	51
157	0	52
152	1	52
155	1	54
157	1	56

$$(49-(\beta_0+152*\beta_1+0*\beta_2))^2+(51-(\beta_0+155*\beta_1+0*\beta_2))^2+(52-(\beta_0+157*\beta_1+0*\beta_2))^2$$
$$+(52-(\beta_0+152*\beta_1+1*\beta_2))^2+(54-(\beta_0+155*\beta_1+1*\beta_2))^2+(56-(\beta_0+157*\beta_1+1*\beta_2))^2$$

Multiple Linear Regression

- the above equation can be minimized by taking partial derivatives with respect to β_0,β_1 and β_2 using the chain rule, and setting them equal to 0.
- we will get 3 equations with 3 unknowns. After solving them
- intercept (β_0) = -57.19
- coefficients([β_1, β_2]) = [0.7, 3.34]
- Final Regression equation for prediction is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$y = -57.19 + 0.7 X_1 + 3.34 X_2$$
$$y = 0.7 X_1 + 3.34 X_2 - 57.19$$

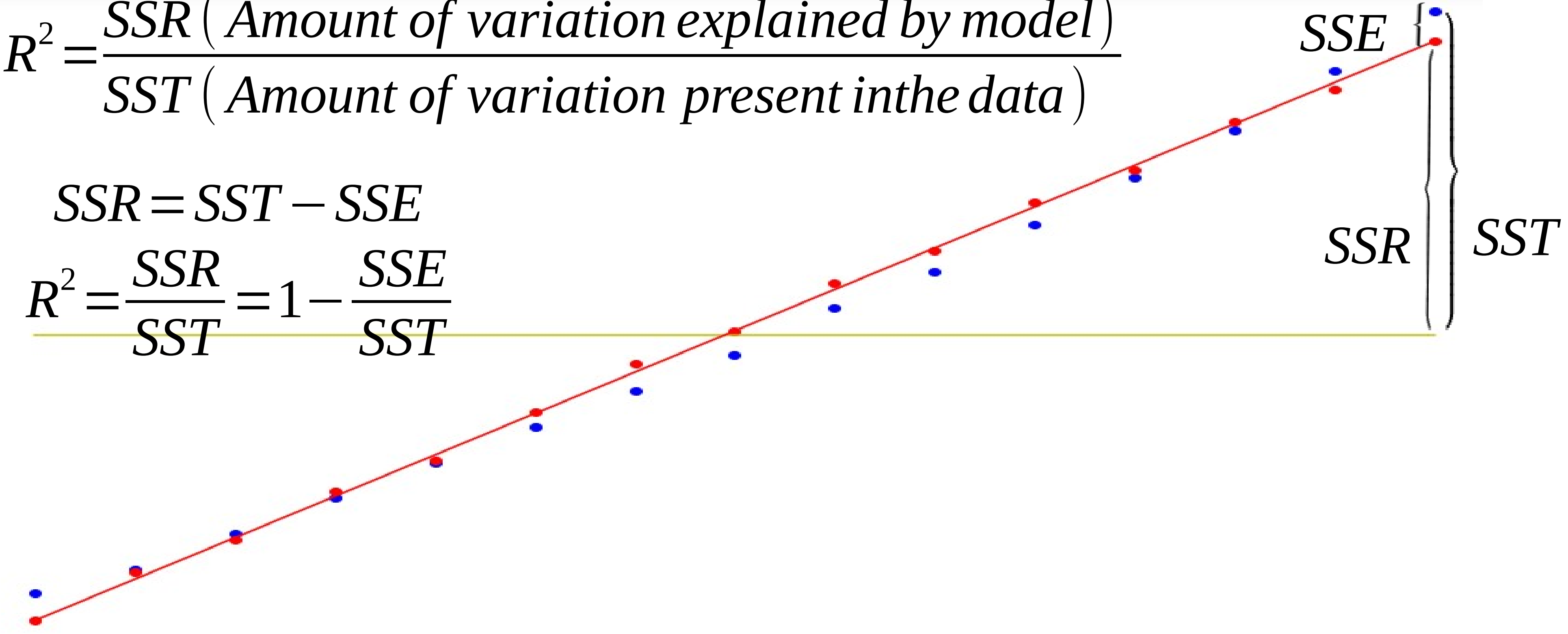
Validation of Multiple Linear Regression

1. R-Square and Adjusted R-Square
2. t test between response variable and individual explanatory variable at given significans level
3. F test to check the statistical significance of the overall model at given significans level
4. Conduct Residual Analysis for Normality , homoscedasticity
5. Check for presence of multi colinearity

$$R^2 = \frac{SSR\left(\text{Amount of variation explained by model}\right)}{SST\left(\text{Amount of variation present in the data}\right)}$$

$$SSR = SST - SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points

if R-square is 0.7, means 70% of the variation in dependent variable is explained by the independent variables

Adjusted R^2

- The problem with R^2 is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables.
- This is where "Adjusted R^2" comes to help. Adjusted R^2 penalizes you for adding variables which do not improve your existing model
- The adjusted R^2 is modified version of R^2 that has been adjusted for number of predictors in the model.
- The adjusted R^2 increases only if new term improves the model more than would be expected by chance
- It decreases when a preditor improves the model by less than expected by chance. Adjusted R^2 always lower than the R^2

Adjusted R^2

- The adjusted R^2 increases only when you add your model by relevant/significant variables
- Adjusted R^2 decrease if we add insignificant variables to our model while R^2 increase even if we add our model insignificant variables

$$Adjusted\, R^2 = 1 - \frac{SSE/(N-k-1)}{SST/(N-1)}$$