Artificial Intelligence & Data Science (Sem VI)

ADC 601 : Data Analytics & Visualization

Module - 2 : Regression Models
(6 Hours)

Instructor : Mrs. Lifna C S

Topics to be covered

- Introduction to Regression
- Types of Regression
  - Simple Linear Regression Model
  - Multiple Linear Regression model,
  - Polynomial Regression Models;
  - Interaction models;
  - Weighted least squares;
  - ridge regression;
  - loess regression;
  - Bootstrapping
- Qualitative predictor variables.
- Model Evaluation Measures
- Model selection procedures
- Leverage;
  - influence measures;
  - diagnostics.
- Logistic Regression:
  - Logistic Response function and logit,
  - Logistic Regression and GLM,
  - Generalized Linear model,
  - Predicted values from Logistic Regression,
  - Interpreting the coefficients and odds ratios,
- Linear and Logistic Regression:
  - similarities and Differences,
  - Assessing the models.

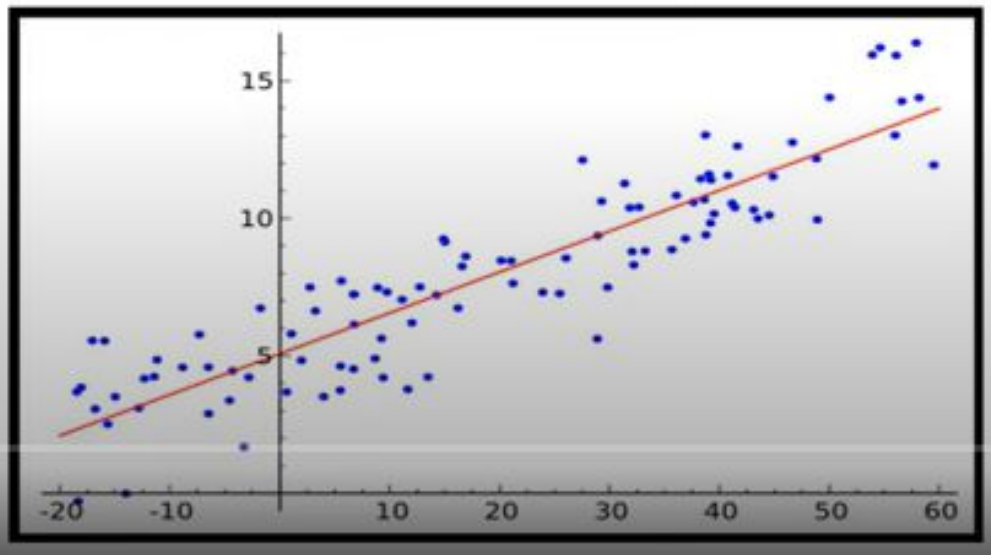
Regression

A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted y, is regarded as the **response**, **outcome**, or **dependent** variable.

Regression examples

- Relationship between uploading a **picture to your Facebook** and **number of friend requests**
- Relationship between **average food intake** and you **weight**
- Relationship of between the number of **hours of studies** and **marks scored**
- **Weight = a + b * (Food Intake)**



Regression - Types of relationships

1. Deterministic / functional

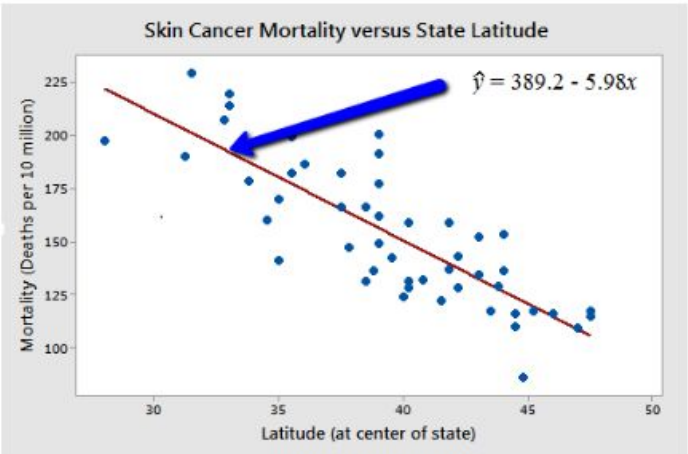
- the equation *exactly* describes the relationship between the two variables.
- Eg :

- Circumference = $\pi \times$ diameter
- **Hooke's Law:** $Y = \alpha + \beta X$, where Y = amount of stretch in a spring, and X = applied weight.
- **Ohm's Law:** $I = V/r$, where V = voltage applied, r = resistance, and I = current.
- **Boyle's Law:** For a constant temperature, $P = \alpha/V$, where P = pressure, α = constant for each gas, and V = volume of gas.

Regression - Types of relationships

2. Statistical

- Relationship between the variables is not perfect.
- **Eg :** The response variable *y* is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable *x* is the latitude (degrees North) at the center of each of 48 states in the United States (U.S. Skin Cancer data)
- Height and weight
- Alcohol consumed and blood alcohol content
- Vital lung capacity and pack-years of smoking
- Driving speed and gas mileage



Correlation and Regression

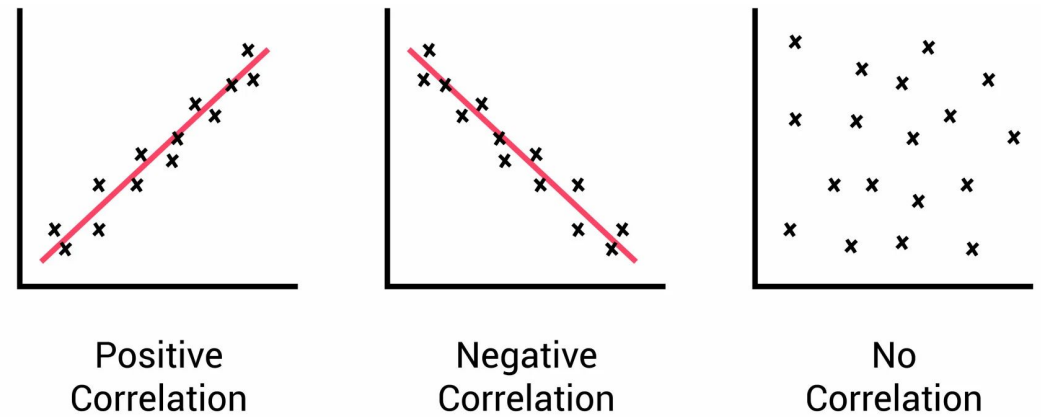
- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?
- **CORRELATION != CAUSATION**
- In order to infer causality: manipulate independent variable and observe effect on dependent variable
- Correlation tells you if there is an **association between x and y** but it **doesn't describe the relationship or allow you to predict one variable from the other.**

Correlation : Methods for studying Correlation

1. Scatter Diagrams
2. **Karl Pearson's Coefficient of Correlation (Covariance Method)**
3. Two way Frequency Table (Bivariate Correlation Method)
4. Rank Correlation Method
5. Concurrent Deviation Method

Correlation - Types of Correlation

- **Positive correlation:** the two variables change in the same direction.
- **Negative correlation:** the two variables change in opposite directions.
- **No correlation**: there is no association or relevant relationship between the two variables.



Scatter Diagrams

- plots the points on a X-Y plane
- Tells the nature of the relationship between X & Y
- Helps to obtain an approximate estimation

Karl Pearson's Coefficient of Correlation (Covariance Method)

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*. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i-1}^{n}(y_i - \bar{y})^2}}$$

Rearranging gives us this formula for r_{xy}:

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

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Karl Pearson's Coefficient of Correlation (Covariance Method)

1. Calculate Karl Pearson's Coefficient of Correlation between X and Y from the given table

X	39	65	62	90	82	75	25	98	36	78
Y	47	53	58	86	62	68	60	91	51	84

2. Calculate Karl Pearson's Coefficient of Correlation between Income and Expenditure of a worker from the given data

Month	Jan	Feb	Mar	April	May	June
Income	451	459	461	461	463	467
Expenditure	433	437	441	451	455	451

Karl Pearson's Coefficient of Correlation (Covariance Method)

1. Calculate Karl Pearson's Coefficient of Correlation between X and Y from the given table

X	39	65	62	90	82	75	25	98	36	78
Y	47	53	58	86	62	68	60	91	51	84

$$r_{xy} = 0.7804$$

Karl Pearson's Coefficient of Correlation (Covariance Method)

2. Calculate Karl Pearson's Coefficient of Correlation between Income and Expenditure of a worker from the given data

Month	Jan	Feb	Mar	April	May	June
Income	451	459	461	461	463	467
Expenditure	433	437	441	451	455	451

$$r_{xy} = 0.8066$$

Properties of Karl Pearson's Coefficient of Correlation

1. The correlation coefficient between X and Y is same as the correlation coefficient between Y and X (*i.e*, $r_{xy} = r_{yx}$).
2. The correlation coefficient is free from the units of measurements of X and Y
3. The correlation coefficient is unaffected by change of scale and origin.

Interpretation of Pearson's Correlation coefficient

The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$

- A positive value of '*r*' indicates positive correlation.
- A negative value of '*r*' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.
- If $r \geq 0.7$ then the correlation will be of higher degree. In interpretation we use the adjective 'highly'
- If X and Y are independent, then $r_{xy} = 0$. However the converse need not be true

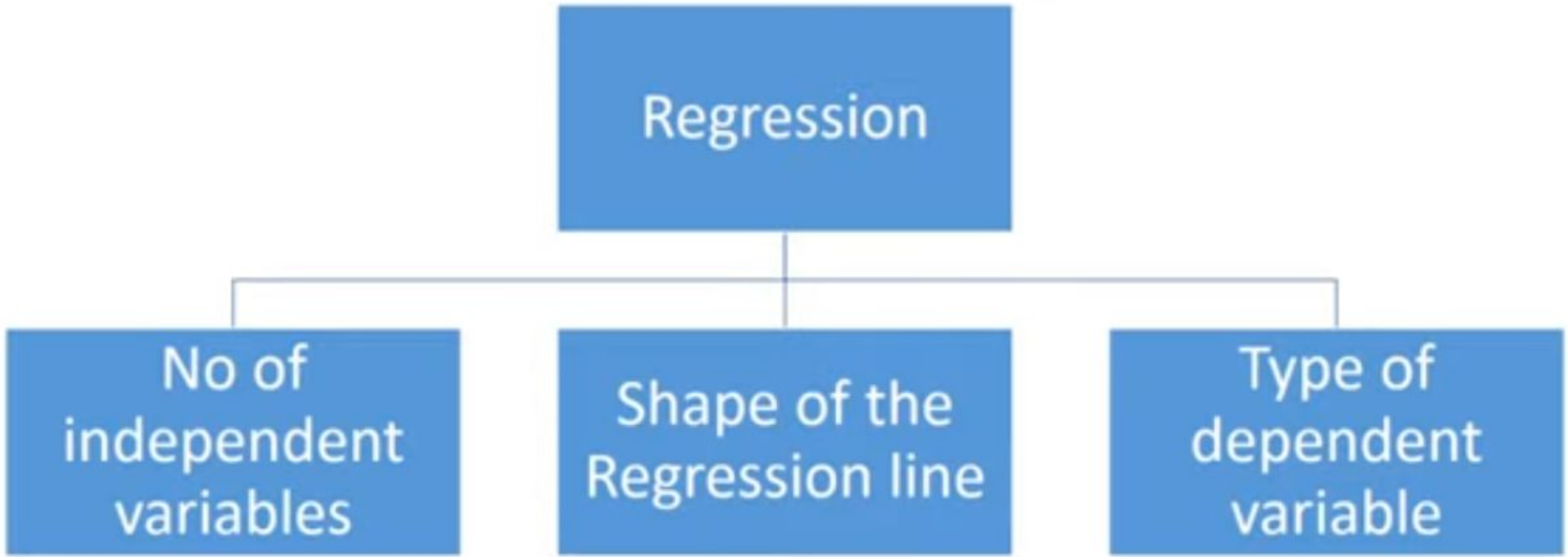
Limitations of Karl Pearson's Coefficient of Correlation

1. **Outliers (extreme observations) strongly influence** the correlation coefficient. If we see outliers in our data, we should be careful about the conclusions we draw from the value of *r*. The outliers may be dropped before the calculation for meaningful conclusion.

2. Correlation **does not imply causal relationship**. That a change in one variable causes a change in another.

Types of Regression Models

$$Y = a + b * X$$



Courtesy : Types of Regression

Types of Regression Models - based on # independent variables

1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Ridge Regression (L2 regularization) & Lasso Regression (L1 regularization)
5. Logistic Regression
6. Poisson Regression
7. Time Series Regression

Types of Regression Models - based on # independent variables

1. **Simple Linear Regression:**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Involves one dependent variable and one independent variable.
- models the relationship between the two variables as a straight line.
- Eg : predicting sales based on advertising expenditure.

2. **Multiple Linear Regression:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

- Involves one dependent variable and more than one independent variable.
- Models complex relationships by considering multiple factors simultaneously.
- Eg : predicting house prices based on features like square footage, number of bedrooms, and location.

Types of Regression Models - based on # independent variables

3. **Polynomial Regression:**

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n + \varepsilon$$

- Extends linear regression by introducing **polynomial terms of the independent variable**.
- captures non-linear relationships.
- represented by a polynomial curve rather than a straight line.

4. **Ridge Regression and Lasso Regression:**

- Similar to multiple linear regression but **includes regularization terms.**
- Used to **handle multicollinearity and prevent overfitting** in multiple linear regression models.
- **Ridge regression** ⇒ adds a penalty term based on the square of the coefficients.
- **Lasso regression** ⇒ adds a penalty term based on the absolute values of the coefficients.
- Commonly **used in situations where there are highly correlated independent variables.**

Types of Regression Models - based on # independent variables

Ridge Regression (L2 regularization):

$$\text{minimize}\left[\sum_{i=1}^{N}(y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_nx_{in}))^2 + \lambda\sum_{j=1}^{n} b_j^2\right]$$

- λ is the regularization parameter.

Lasso Regression (L1 regularization):

$$\text{minimize}\left[\sum_{i=1}^{N}(y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_nx_{in}))^2 + \lambda\sum_{j=1}^{n} |b_j|\right]$$

- λ is the regularization parameter.

Types of Regression Models - based on # independent variables

5. **Logistic Regression** $$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}}$$

- used for binary classification problems.
- models the probability of an event occurring.
- Eg : Predicting the probability of a customer making a purchase (yes/no), fraud detection, etc.

6. **Poisson Regression:** $$\lambda = e^{(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}$$

- Used when the dependent variable represents counts or frequencies
- Modeling the number of events within a fixed interval,
- Eg: the number of customer arrivals at a service center.

7. **Time Series Regression:**

- Depends on the specific time series model used
- Applied to data collected over time to model and predict future values based on past observations.
- Eg: Forecasting stock prices, sales, or any variable that exhibits temporal patterns.

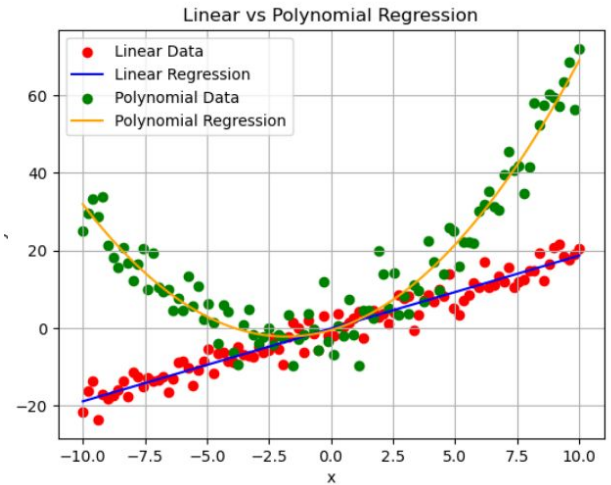
Types of Regression Models - based on shape of Regression Line

1. Linear Regression
2. Polynomial Regression
3. Exponential Regression
4. Logarithmic Regression
5. Power Regression (Power Law)
6. Sigmoidal (Logistic) Regression
7. Piecewise Regression
8. Quantile Regression

Types of Regression Models - based on shape of Regression Line

1. **Linear Regression:**

$$y = b_0 + b_1 x + \varepsilon$$



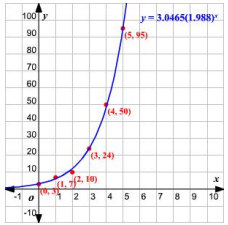
- Regression line : **straight line.**
- Used when there is a linear relationship between the independent and dependent variables.

2. **Polynomial Regression:**

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n + \varepsilon$$

- Regression line : **Curve**
- (degree of the polynomial determines its shape)
- when the relationship between variables is nonlinear and can be better represented by a polynomial.

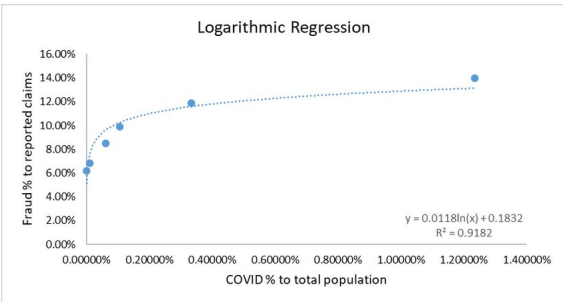
Types of Regression Models - based on shape of Regression Line



3. Exponential Regression:

$$y = ab^x + \varepsilon$$

- Regression line : **Exponential curve.**
- when the dependent variable changes at a constant percentage rate w.r.t independent variable.



4. Logarithmic Regression:

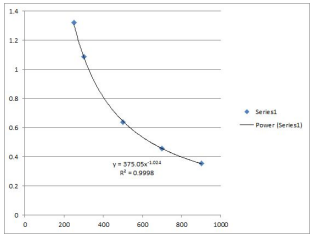
$$y = a + b\ln(x) + \varepsilon$$

a. Regression line : **logarithmic curve.**

b. when the rate of change of the dependent variable is proportional to its current value.

Types of Regression Models - based on shape of Regression Line

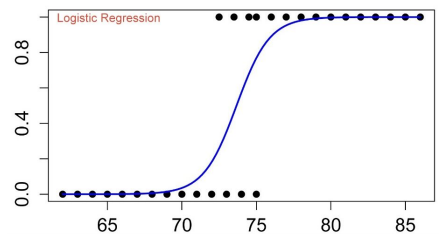
5. **Power Regression (Power Law):** $y = a \cdot x^b + \varepsilon$



- Regression line : **represents a power-law relationship.**
- modeling relationships where one variable's growth is proportional to a power of the other.

6. **Sigmoidal (Logistic) Regression:** $y = \frac{1}{1+e^{-(b_0+b_1 x)}} + \varepsilon$

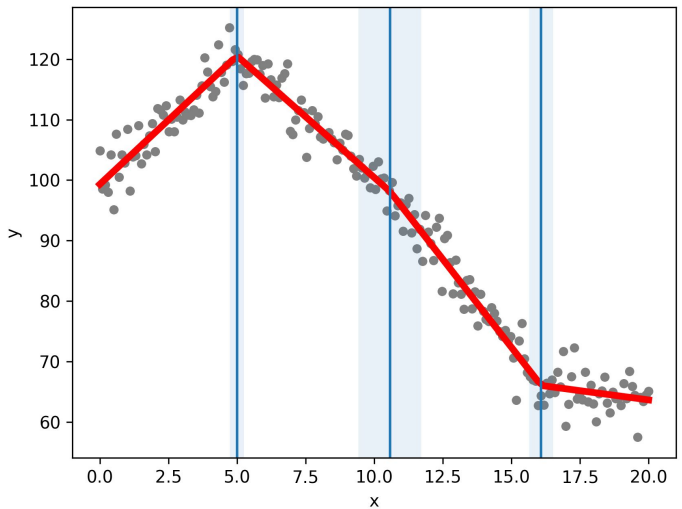
- Regression line : **S-shaped curve.**
- used for binary classification problems
- when the dependent variable has a logistic growth pattern.

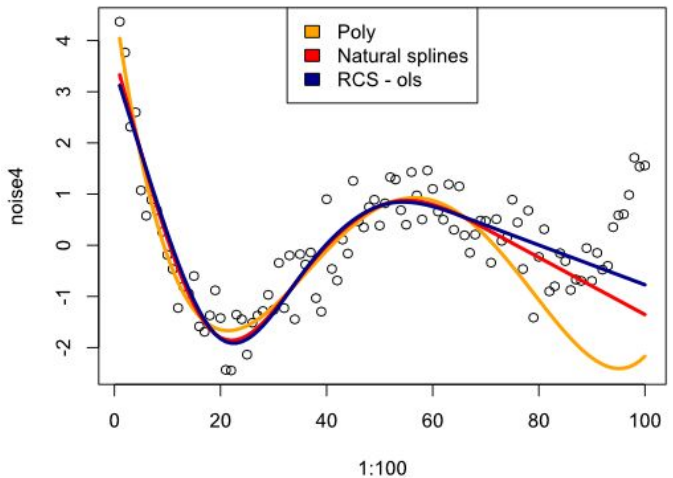


Types of Regression Models - based on shape of Regression Line

7. Piecewise Regression:

- Regression line : **multiple linear segments / piecewise continuous curves (Spline Regression)**
- Involves fitting multiple linear or nonlinear regression models to different segments of the data.
- Useful when the relationship between variables changes at certain points or intervals.

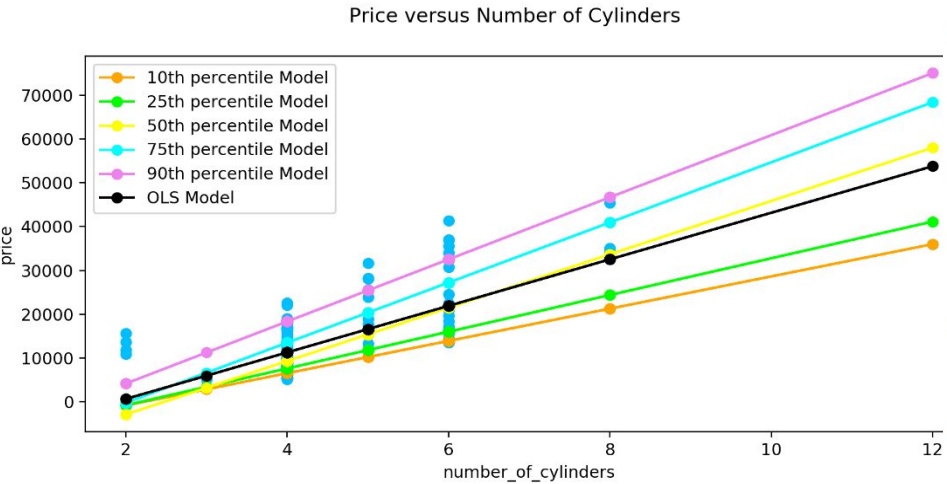


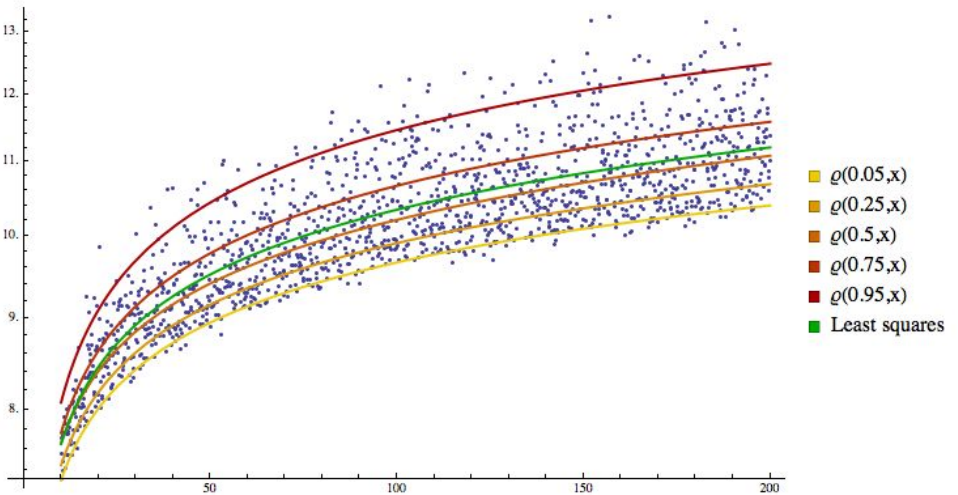


Types of Regression Models - based on shape of Regression Line

8. Quantile Regression:

- Regression Line: **different for the quantiles.**
- when the variability in the residuals is not constant across all values of the independent variable.
- estimates different quantiles of the dependent variable.





Types of Regression Models - based on type of dependent variable

1. Linear Regression
2. Logistic Regression
3. Multinomial Logistic Regression
4. Ordinal Regression
5. Poisson Regression
6. Negative Binomial Regression
7. Survival Analysis (Cox Proportional-Hazards Model)
8. Robust Regression
9. Quantile Regression
10. Ridge Regression and Lasso Regression

Types of Regression Models - based on type of dependent variable

1. **Linear Regression:**
   - Type of Dependent Variable: **Continuous**
   - Predicting a continuous outcome variable,
   - Eg: predicting sales, temperature, or height.
2. **Logistic Regression:**
   - Type of Dependent Variable: **Binary (0 or 1)**
   - Used for binary classification problems,
   - Eg: predicting whether a customer will buy a product (1) or not (0).
3. **Multinomial Logistic Regression:**
   - Type of Dependent Variable: **Categorical with more than two categories**
   - Suitable when the dependent variable has more than two unordered categories,
   - Eg: predicting the type of fruit (apple, orange, banana).

Types of Regression Models - based on type of dependent variable

4. **Ordinal Regression / Ordered Logistic Regression:**
   - Type of Dependent Variable: **Ordered categorical**
   - Used when the dependent variable has ordered categories, but the intervals between them are not assumed to be equal.
   - Eg: predicting the satisfaction level (low, medium, high).
5. **Poisson Regression:**
   - Type of Dependent Variable: **Count data (non-negative integers)**
   - Appropriate for modeling count data,
   - Eg: predict the number of customer arrivals, phone calls, or accidents in a day.
6. **Negative Binomial Regression:**
   - Type of Dependent Variable: **Count data with overdispersion**
   - Used when the count data exhibit more variability than expected in a Poisson regression, often due to unobserved heterogeneity.

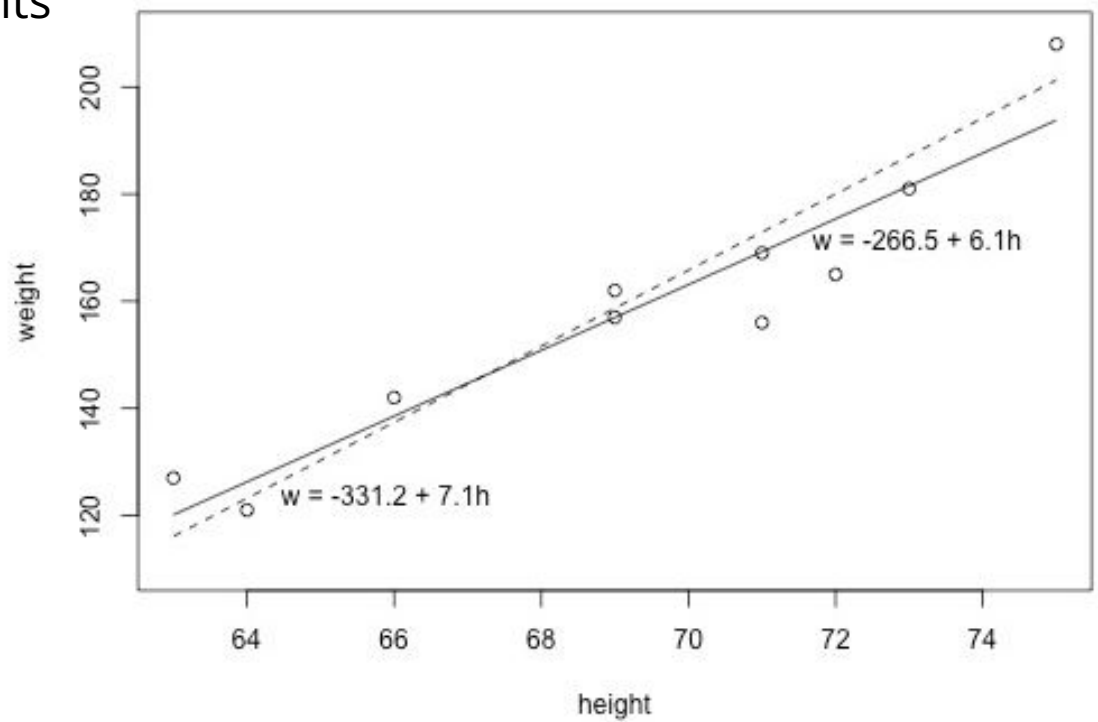
Types of Regression Models - based on type of dependent variable

7. **Survival Analysis (Cox Proportional-Hazards Model):**
   - Type of Dependent Variable: **Time until an event occurs**
   - Applied in medical research, economics, and other fields
   - to model the time until an event (e.g., death, failure, or relapse) occurs.
8. **Robust Regression:**
   - Type of Dependent Variable: **Continuous, resistant to outliers**
   - when there are outliers in the data, and traditional linear regression may be sensitive to them.
9. **Quantile Regression:**
   - Type of Dependent Variable: **Conditional quantiles**
   - Examining the effect of predictors on different quantiles of the dependent variable,
   - Provides insights into distributional changes.
10. **Ridge Regression and Lasso Regression:**
    - Type of Dependent Variable: **Continuous, with potential multicollinearity**
    - Used when dealing with multicollinearity in LR, and to prevent overfitting.

Simple Linear Regression - What is the “Best Fitting Line ?”

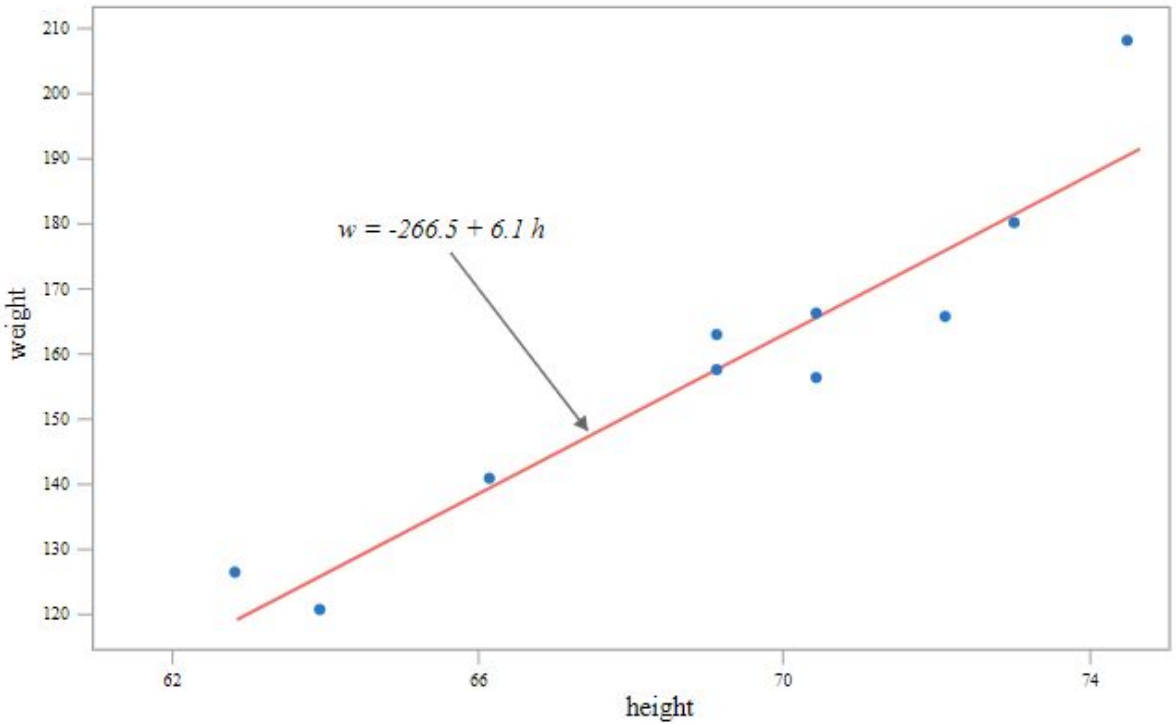
Given a set of heights (*x*) and weights (*y*) of 10 students, (Student Height and Weight Dataset).

Which line summarizes the trend accurately?



Simple Linear Regression - What is the “Best Fitting Line ?”

i	x_i	y_i	$\hat{y}_i$
1	63	127	120.1
2	64	121	126.3
3	66	142	138.5
4	69	157	157.0
5	69	162	157.0
6	71	156	169.2
7	71	169	169.2
8	72	165	175.4
9	73	181	181.5
10	75	208	193.8



Simple Linear Regression - What is the “Best Fitting Line ?”

the equation for the best-fitting line is:

$$\hat{y}_i = b_0 + b_1 x_i$$

- y_i denotes the observed response for experimental unit *i*
- x_i denotes the predictor value for experimental unit *i*
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit *i*

In general, when we use $\hat{y}_i = b_0 + b_1 x_i$ to predict the actual response y_i , we make a prediction error (or residual error) of size:

$$e_i = y_i - \hat{y}_i$$

- A line that fits the data "best" will be one for which the *n* prediction errors.
- The line which meets **"least squares criterion,"**

- We just need to find the values b_0 and b_1 which make the sum of the squared prediction errors the smallest they can be.
- That is, we need to find the values b_0 and b_1 that minimize:

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

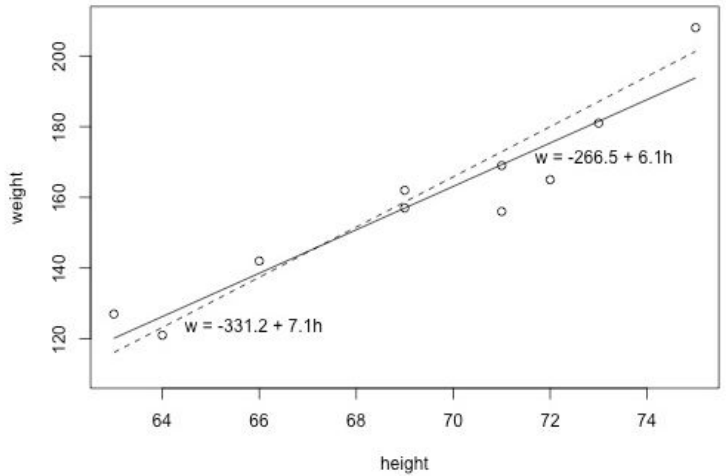
Simple Linear Regression - What is the “Best Fitting Line ?”

$\hat{w} = -331.2 + 7.1h$ (the dashed line)

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					766.5

$\hat{w} = -266.53 + 6.1376h$ (the solid line)

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					597.4



Simple Linear Regression - What is the “Best Fitting Line ?”

the equation for the best-fitting line is:

$$\hat{y}_i = b_0 + b_1 x_i$$

- y_i denotes the observed response for experimental unit *i*
- x_i denotes the predictor value for experimental unit *i*
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit *i*

- The formulas are determined using methods of calculus.

$$b_0 = \bar{y} - b_1 \bar{x}$$

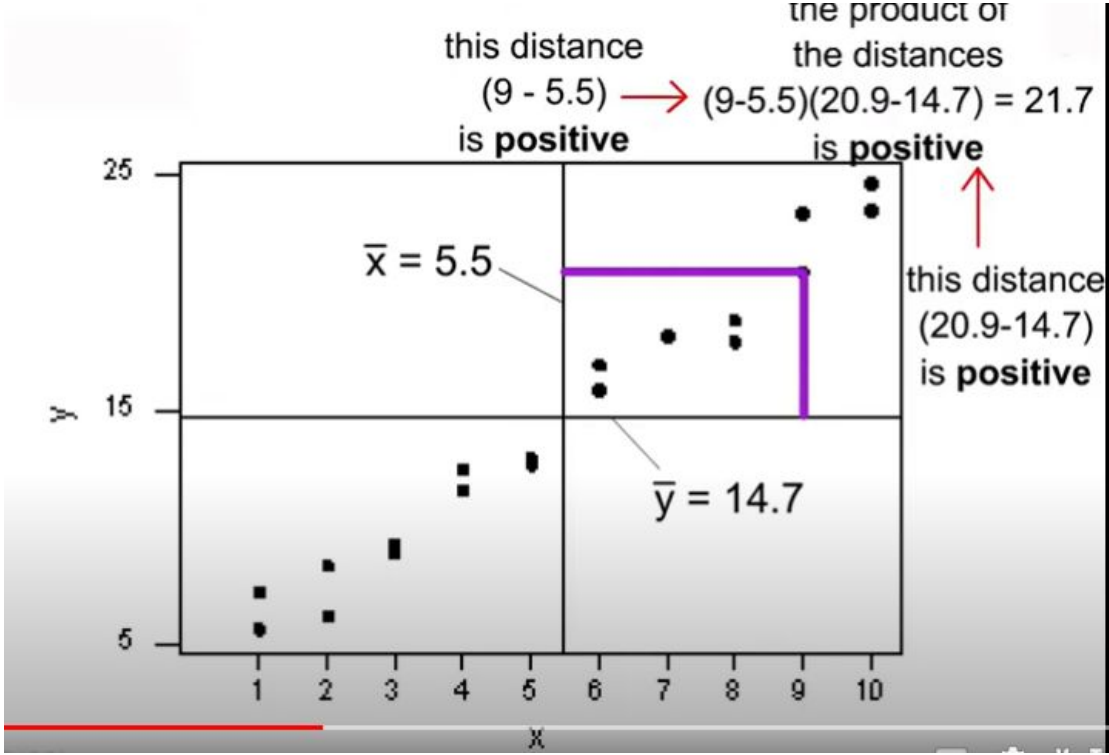
- Here, the Best Fitting Line is also known as
  - **"least squares regression line,"**
  - **"least squares line."**
  - **"estimated regression equation."**

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Simple Linear Regression - What is the “Best Fitting Line ?”

When is the slope $b_1 > 0$?

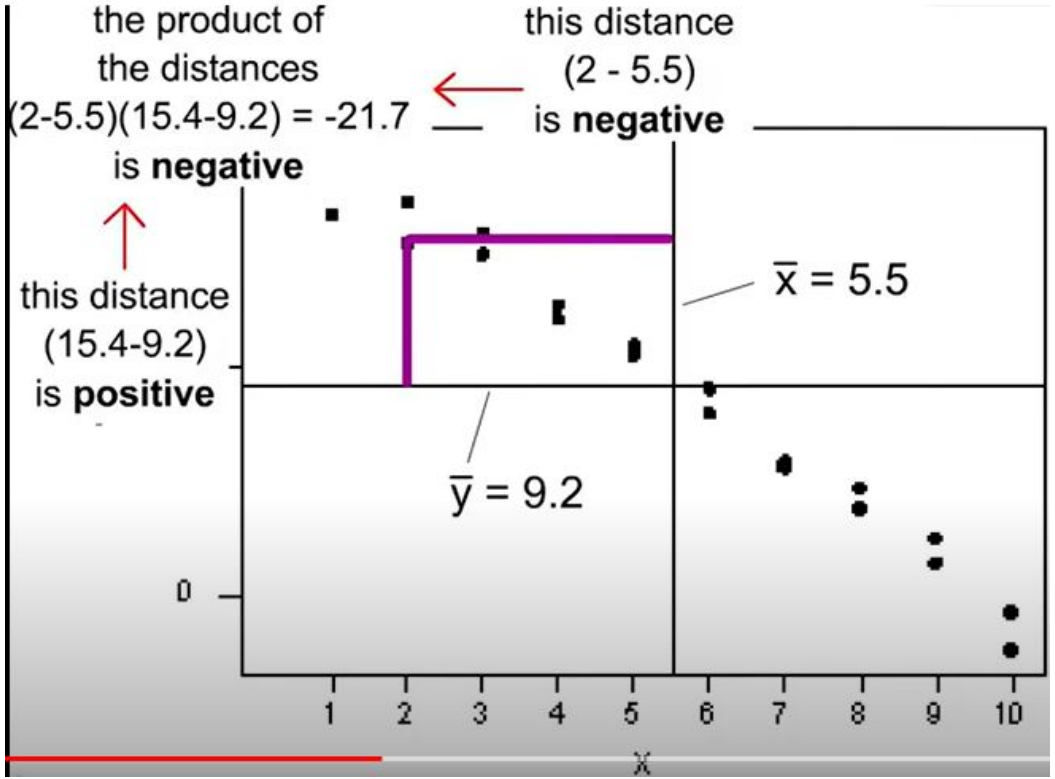
Trend is positive



Simple Linear Regression - What is the “Best Fitting Line ?”

When is the slope $b_1 < 0$?

Trend is negative



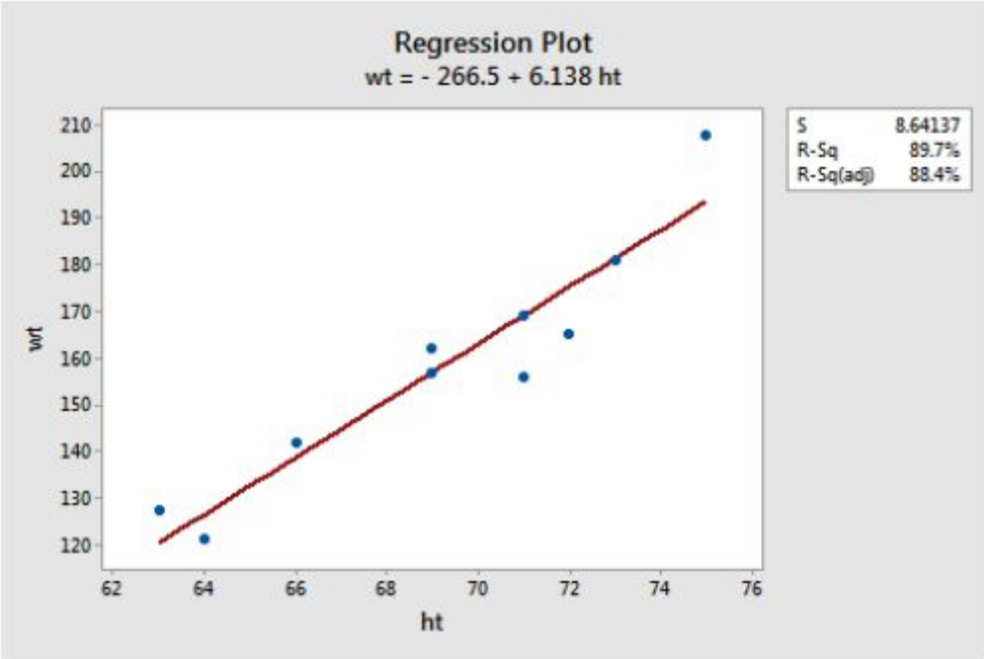
Simple Linear Regression - What is the “Best Fitting Line ?”

what does b_0 tell us?

- a person who is 0 inches tall is predicted to weigh -267 pounds!
- This happened because of **"extrapolation"**
- Scope of the model does not include $x = 0$.
- In general, if the "scope of the model" includes $x = 0$, then b0 is the predicted mean response when $x = 0$.

what does b_1 tell us?

- Mean weight to increase by 6.14 pounds for every additional one-inch increase in height.
- the mean response to increase or decrease by b1 units for every one-unit increase in x.



Simple Linear Regression

The equation of the least squares regression line : $\hat{y} = a + bx$

- $\hat{y}$ is the predicted value of y,
- $a = \bar{y} - b\bar{x}$,
- $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sum(x^2) - \frac{(\sum x)^2}{n}}$
- $\bar{x} = \frac{\sum x}{n}$,
- $\bar{y} = \frac{\sum y}{n}$,

Simple Linear Regression : Class Assignment - 1

Consider the example below where the mass, y (grams), of a chemical is related to the time, x (seconds), for which the chemical reaction has been taking place according to the table. Find the equation of the regression line.

Time	Mass
x (seconds)	y (grams)
5	40
7	120
12	180
16	210
20	240

Simple Linear Regression : Class Assignment - 1 Solution

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	40	$5 - 12 = -7$	$40 - 158 = -118$	$-7 \times -118 = 826$	$-7^2 = 49$
7	120	$7 - 12 = -5$	$120 - 158 = -38$	$-5 \times -38 = 190$	$-5^2 = 25$
12	180	$12 - 12 = 0$	$180 - 158 = 22$	$0 \times 22 = 0$	$0^2 = 0$
16	210	$16 - 12 = 4$	$210 - 158 = 52$	$4 \times 52 = 208$	$4^2 = 16$
20	240	$20 - 12 = 8$	$240 - 158 = 82$	$8 \times 82 = 656$	$8^2 = 64$
$\sum x = 60$	$\sum y = 790$			$\sum (x_i - \bar{x})(y_i - \bar{y}) = 1880$	$\sum (x_i - \bar{x})^2 = 154$

$$
\begin{aligned}
b &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{1880}{154} = 12.20779... \\
&= 12.208 \text{ (3.d.p.)}
\end{aligned}
$$

$$
\begin{aligned}
a &= \bar{y} - b\bar{x} \\
&= 158 - 12.208 \times 12 \\
&= 11.506... \\
&= 11.506 \text{ (3.d.p.)}.
\end{aligned}
$$

$$\hat{y} = a + bx = 11.506 + 12.208x.$$

Simple Linear Regression : Class Assignment - 2

To see how students' reaction skills have improved over a year, eight students took a reactions test at the start of the year and at the end of the year. These are their scores:

Student	Liam	Felicity	Adian	Mel	Leroy	Vic	Lawrie	Louise
First Test, x	56	75	61	61	67	72	62	61
Second Test, y	21	39	34	21	32	24	29	24

Find the equation of the regression line given that:

$$\sum x = 515,\ \sum y = 224,\ \sum x^2 = 33441,\ \sum y^2 = 6576\ and\ \sum xy = 14590.$$

Simple Linear Regression : Class Assignment - 2 Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sum(x^2) - \frac{(\sum x)^2}{n}}.$$

$$\begin{aligned} b &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sum(x^2) - \frac{(\sum x)^2}{n}} \\ &= \frac{14590 - \frac{515 \times 224}{8}}{33441 - \frac{515^2}{8}} \\ &= 0.590534... \\ &= 0.590 \text{ (3.d.p.)} \end{aligned}$$

$$\begin{aligned} \bar{x} &= \frac{\sum x}{n} = \frac{515}{8} = 64.375, \\ \bar{y} &= \frac{\sum y}{n} = \frac{224}{8} = 28 \\ a &= \bar{y} - b\bar{x} \\ &= 28 - (0.590 \times 64.375) \\ &= -10.015631... \\ &= -10.016 \text{ (3.d.p.)} \end{aligned}$$

$$\hat{y} = -10.106 + 0.590x.$$

Simple Linear Regression : Class Assignment - 3

The finance manager of ABC Motors wants to correlate variation in sales and variation in the price of electric bikes. For this purpose, he analyzes data pertaining to the last five years.

We assume there is no error. The price and sales volume for the previous five years are as follows:

Year	Price (in $)	Sales Volume
2017	2100	15000
2018	2050	16500
2019	2000	21000
2020	2200	19000
2021	2050	20000

Based on the given data, determine the regression line of Y on X,

Simple Linear Regression : Class Assignment - 3 Solution

Given:

- Y = Sales Volume
- X = Profit
- N = 5
- ε = 0

Year	Price (in $) (X)	Sales Volume (Y)	X^2	XY
2017	2100	15000	4410000	31500000
2018	2050	16500	4202500	33825000
2019	2000	21000	4000000	42000000
2020	2200	19000	4840000	41800000
2021	2050	20000	4202500	41000000
–	**10400**	**91500**	**21655000**	**190125000**

$$Y = a + bX + \varepsilon$$

Let us first find out the value of b and a:

$$b = (N\sum XY - (\sum X)(\sum Y) / (N\sum X^2 - (\sum X)^2)$$

- $b = ((5 \times 190125000) - (10400 \times 91500)) / ((5 \times 21655000) - 10400^2)$
- $b = (950625000 - 951600000) / (08275000 - 108160000)$
- $b = \mathbf{-8.478}$

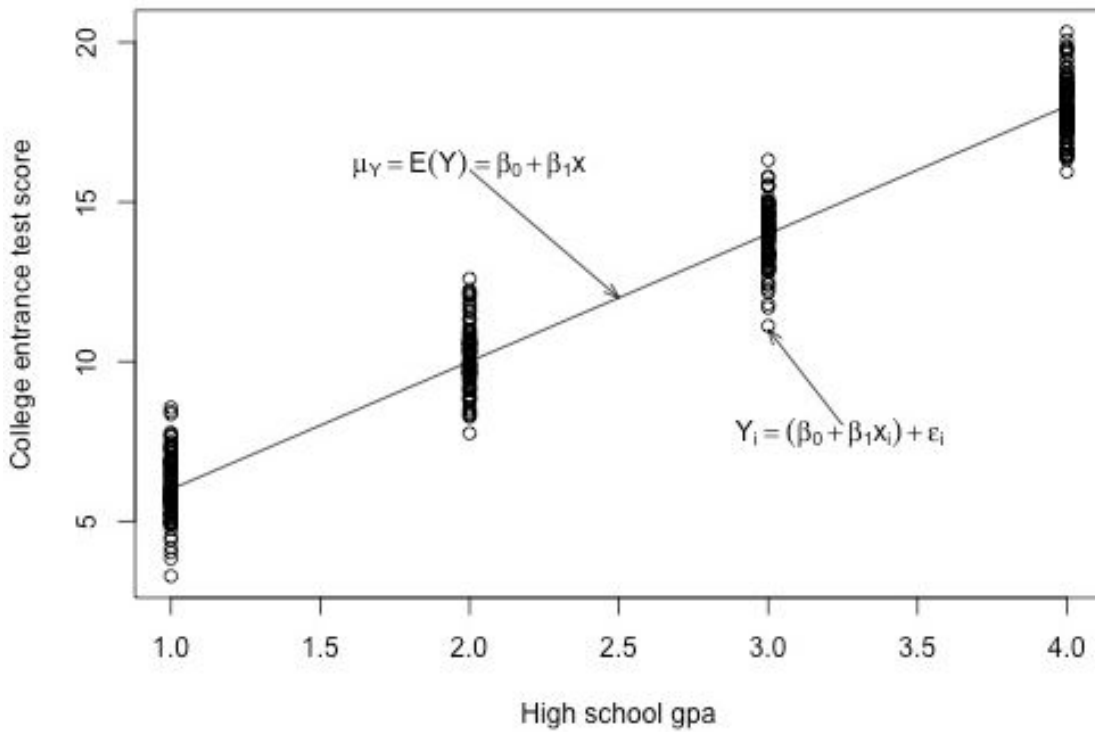
$$a = (\sum Y - b \sum X) / N$$

- $a = 91500 - (- 8.478 \times 10400) / 5$
- $a = \mathbf{35935}$
- $Y = 35935 + (- 8.478\ X) + 0$
- $\mathbf{Y = 35935 - 8.478X}$

Simple Linear Regression - What do b0 & b1 estimate?

- Plot illustrates the relationship between the predictor **"high school grade point average (GPA)"** and the response **"college entrance test score."**
- Only four groups of students are considered — those with a GPA of 1, those with a GPA of 2, ..., and those with a GPA of 4.
- Data on the entire subpopulation of students with a GPA of 1, 2, 3 & 4 are plotted.
- Here, $\mu_Y = \mathrm{E}(Y) = \beta_0 + \beta_1 x.$

  "population regression line" — summarizes the trend *in the population* between the predictor *x* and the mean of the responses μ_Y.

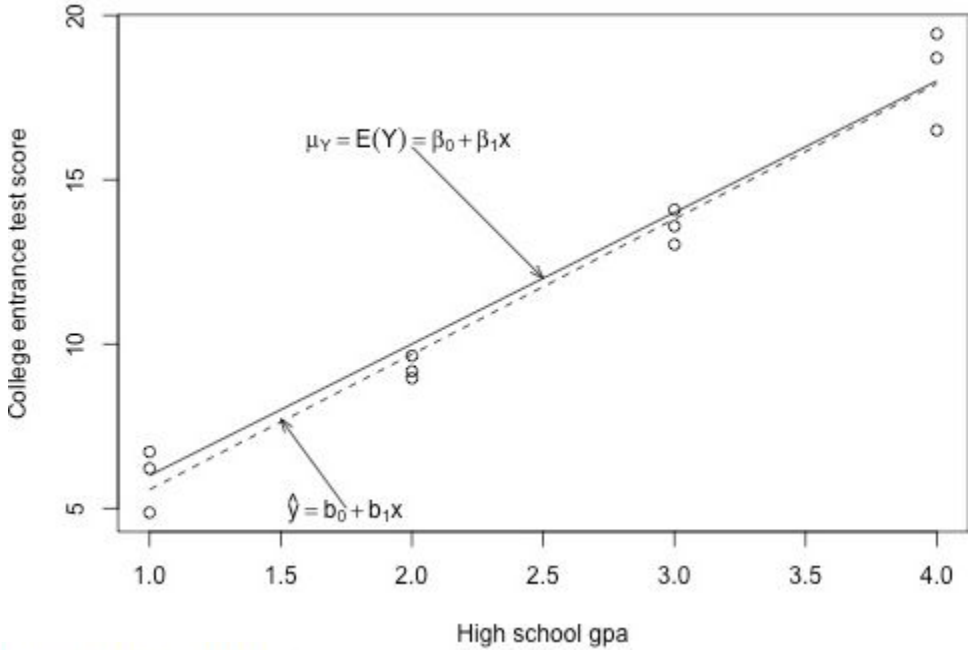


Simple Linear Regression - What do b0 & b1 estimate?

- Take a sample of three students from each of the subpopulations — that is, three students with a GPA of 1, three students with a GPA of 2, ..., and three students with a GPA of 4 — for a total of 12 students.
- Here,

least squares regression line $\hat{y} = b_0 + b_1 x$

population regression line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$.



- Least squares regression line **doesn't match** the population regression line perfectly, but it is a **pretty good estimate**

- That is, the sample intercept b_0 estimates the population intercept β_0 and the sample slope b_1 estimates the population slope β_1.

Simple Linear Regression - 4 Conditions

- **L inear Function:** The mean of the response, $\mathrm{E}(Y_i)$, at each value of the predictor, x_i, is a Linear function of the x_i.
- **I ndependent:** The errors, ϵ_i, are Independent.
- **N ormally Distributed:** The errors, ϵ_i, at each value of the predictor, x_i, are Normally distributed.
- **E qual variances (denoted σ^2):** The errors, ϵ_i, at each value of the predictor, x_i, have Equal variances (denoted σ^2).

An equivalent way to think of the first (linearity) condition is that the mean of the error, $\mathrm{E}(\epsilon_i)$, at each value of the predictor, x_i, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2.

Simple Linear Regression - 4 Conditions (contd…)

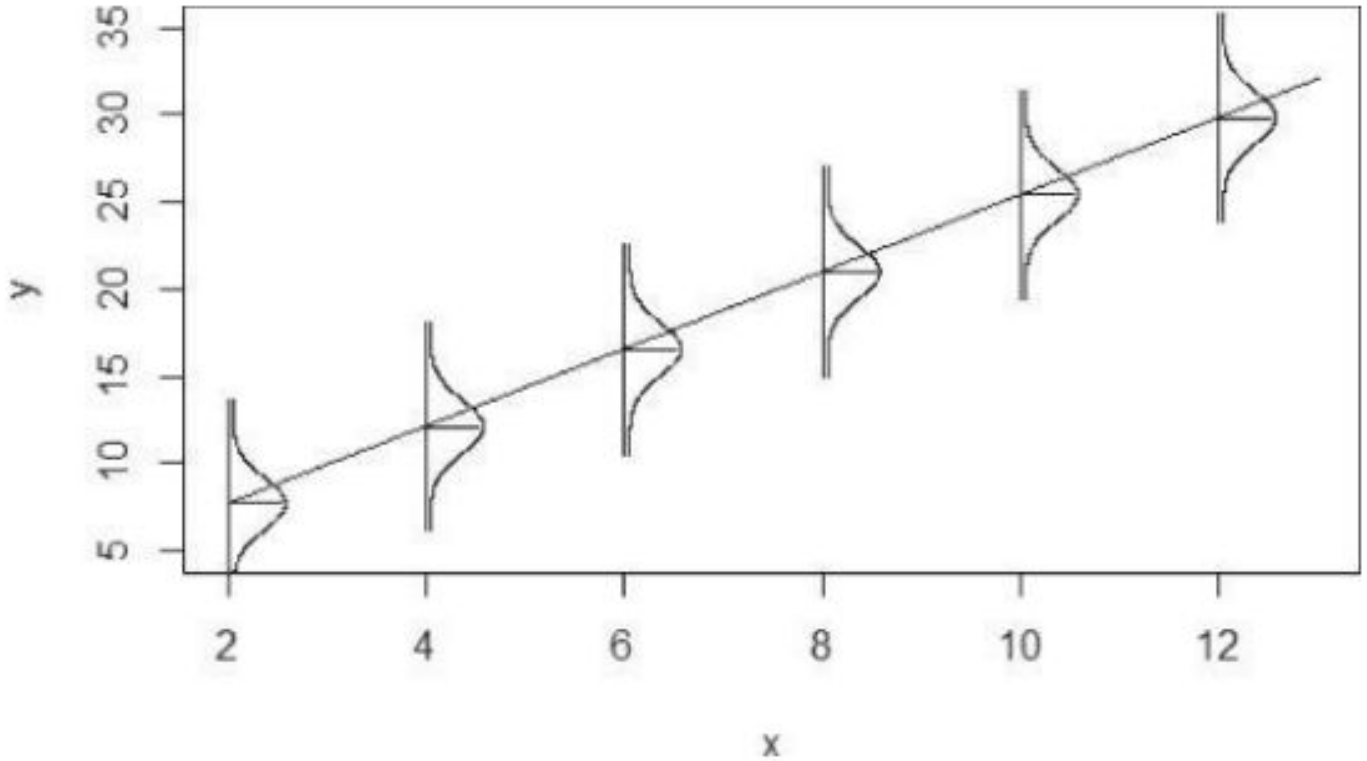
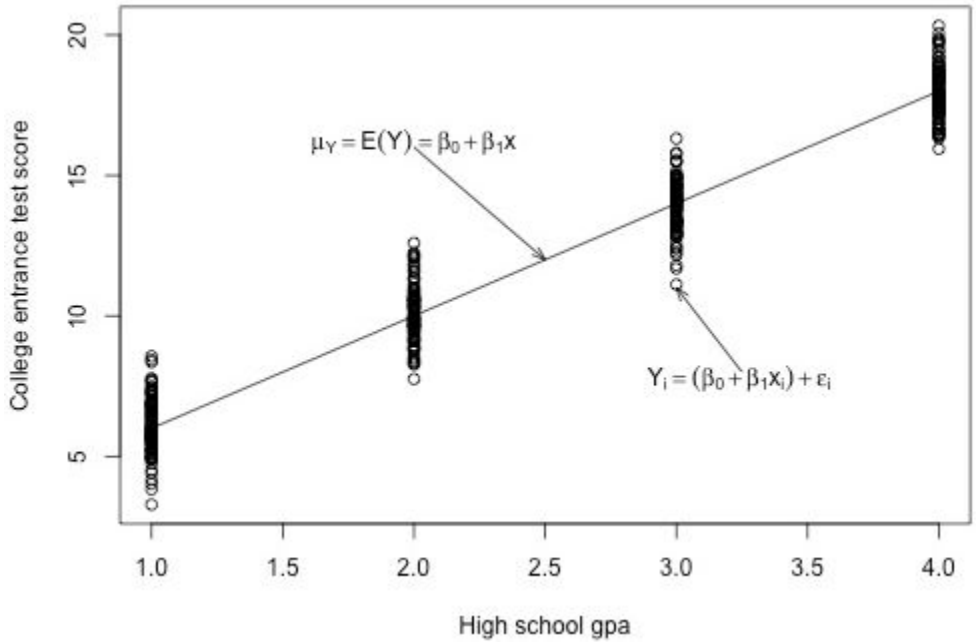


Figure 6.3 Normal distribution about y for a given value of x

Simple Linear Regression - What is The Common Error Variance?

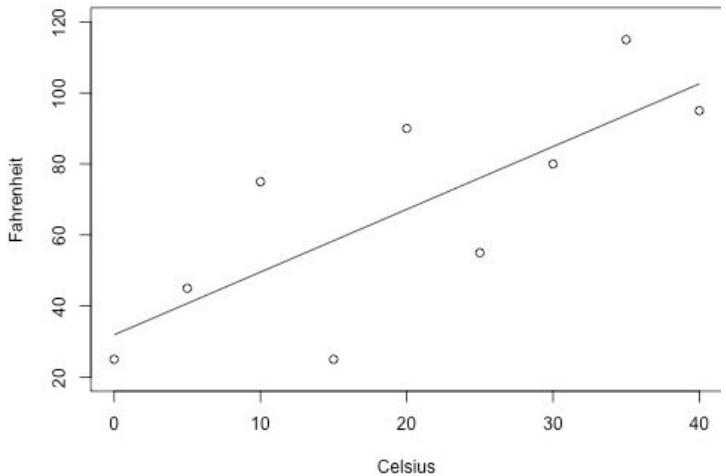
- The plot of college entrance test scores for each subpopulation have equal variance.
- Common variance is denoted as σ^2.

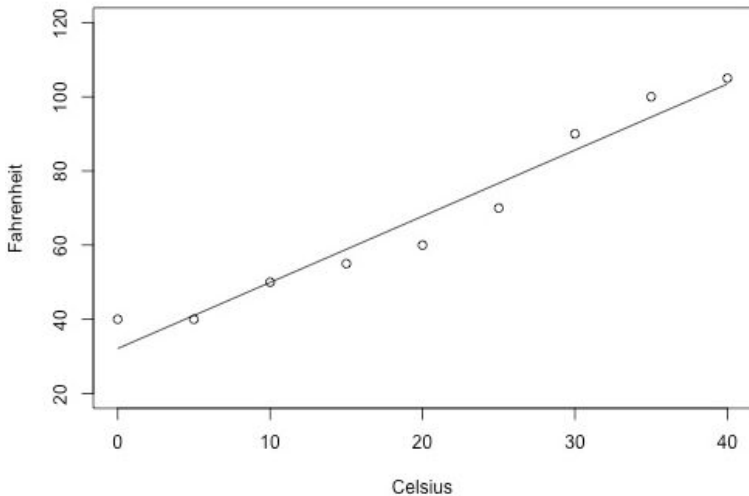


That is, σ^2 quantifies how much the responses (y) vary around the (unknown) mean population regression line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$.

Simple Linear Regression - Why should we care about σ^2 ?

- Suppose you have two brands (A and B) of thermometers
  - Measure the temperature in Celsius and Fahrenheit using each brand for 10 days.
  - **Will this thermometer brand (A) yield more precise future predictions ...? Or brand (B) ?**





- Here, brand B thermometer yield more precise future predictions than the brand A thermometer.

σ^2 quantifies this variance in the responses.

Simple Linear Regression - How to estimate σ^2 ?

- **Sample Variance :**
  - estimates σ^2, the variance of one population.
  - It is really close to being like an average.
  - Here, the population mean μ, is unknown, estimate with $\bar{y}$.
  - So, divide by *n*-1, and not *n*

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

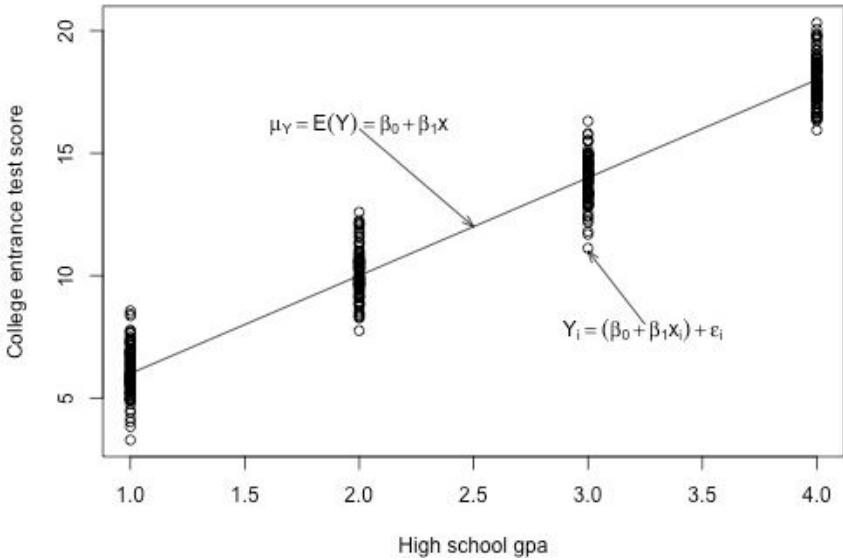
Simple Linear Regression - How to estimate σ^2 ?

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

- **Mean Square Error :**
  - estimates σ^2, the common variance of the many subpopulations.
  - Subpopulations has distinct *x* values.
  - Each subpopulation has its own mean μ_Y,

    $$\mu_Y = E(Y) = \beta_0 + \beta_1 x.$$
  - Subpopulation mean is estimated using the estimated regression equation $\hat{y}_i = b_0 + b_1 x_i$
  - As we use $\hat{y}_i$ to estimate μ_Y we lose two degrees of freedom. Hence, n-2 in the MSE equation.
  - $S = \sqrt{MSE}$ estimates σ ,

    Regression / Residual standard error



Simple Linear Regression - Coefficient of Determination, R^2

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

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 =$$

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

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line, $\hat{y}_i$, is from the horizontal "no relationship line," the sample mean or $\bar{y}$.
- SSE is the "error sum of squares" and quantifies how much the data points, y_i, vary around the estimated regression line, $\hat{y}_i$.
- SSTO is the "total sum of squares" and quantifies how much the data points, y_i, vary around their mean, $\bar{y}$.

Simple Linear Regression - Coefficient of Determination, R^2

Characteristics of R^2

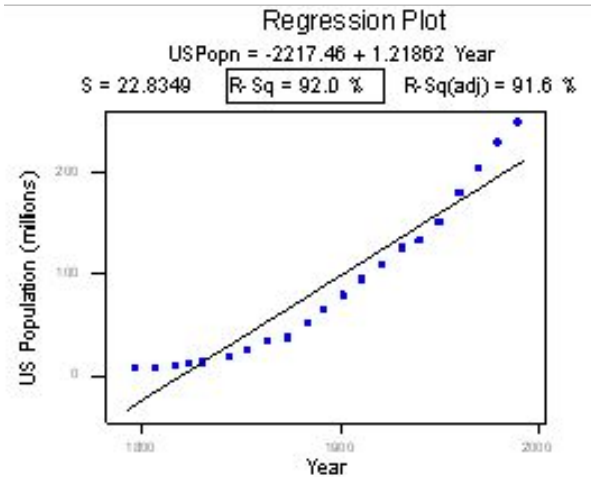
- Since R^2 is a proportion, it is always a number between 0 and 1.
- If R^2 = 1, all of the data points fall perfectly on the regression line.
- If R^2 = 0, the estimated regression line is perfectly horizontal.

Interpretation of R^2

- Association is not causation.
- A large *r*-squared value, it does not imply that *x causes* the changes in *y*.

Simple Linear Regression - R^2 Caution

1. The coefficient of determination R^2 and the correlation coefficient *r* quantify the strength of a *linear* relationship. **When R^2 = 0% and *r* = 0**, suggests that **there is no linear relation between *x* and *y***, and yet a perfect curved (or "curvilinear" relationship) exists.
2. A **large R^2 value** should not be interpreted as meaning that the **estimated regression line fits the data well.** Another function might better describe the trend in the data.



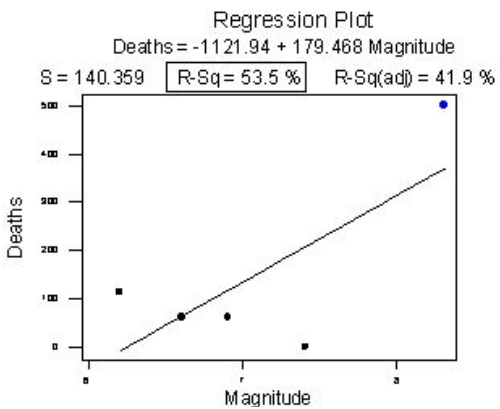
Simple Linear Regression - R^2 Caution

3. The coefficient of determination R^2 **and the correlation coefficient *r* can both be greatly affected by just one data point** (or a few data points).

Eg: Relationship between the number of deaths in an earthquake and its magnitude is examined.

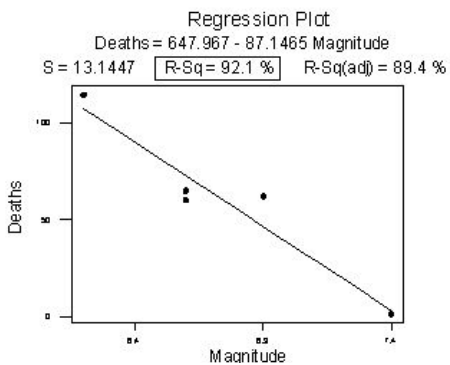
- Data on *n* = 6 earthquakes were recorded, and the fitted line plot on the left was obtained.
- Slope of the line $b_1 = 179.5$
- Correlation, r = 0.732

  Magnitude of earthquake increases ⇒ # deaths increases



Remove one unusual data point:

- slope of the line changes from +179.5 to -87.1
- *r* changes from a +0.732 to -0.960
- R^2 changes from 53.5% to 92.1%.



Simple Linear Regression - R^2 Caution

4. Correlation (or association) **does not imply** causation.

5. **Ecological correlations** — correlations that are based on rates or averages — tend to overstate the strength of an association.

6. A "statistically significant" R^2 value **does not imply** that the slope β_1 is meaningfully different from 0.

7. A **large** R^2 **value does not necessarily mean** that a useful prediction of the resp y_{new}, or estimation of the mean re μ_Y onse , can be made. It is still **possible to get prediction intervals or confidence intervals that are too wide to be useful.**

Simple Linear Regression - Pearson Correlation Coefficient

The correlation coefficient, *r*, is directly related to the coefficient of determination R^2

$$r = \pm\sqrt{R^2}$$

The sign of *r* depends on the sign of the estimated slope coefficient b_1:

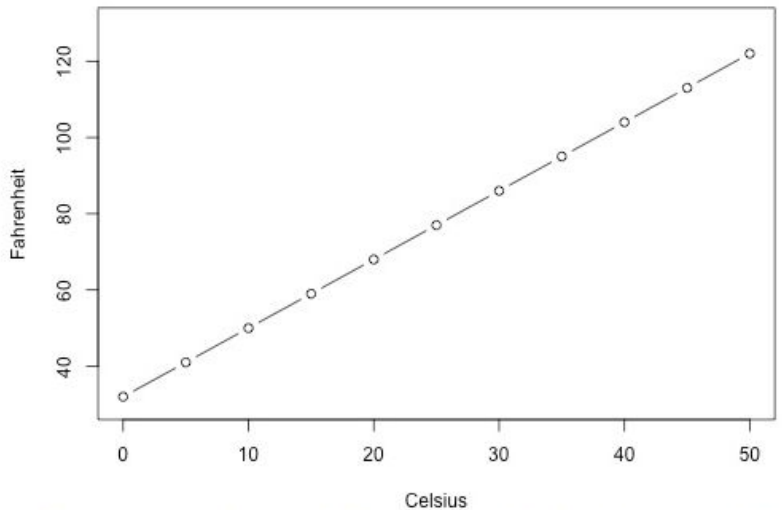
- If b_1 is negative, then *r* takes a negative sign.
- If b_1 is positive, then *r* takes a positive sign.

- If *r* = -1, then there is a perfect negative linear relationship between *x* and *y*.
- If *r* = 1, then there is a perfect positive linear relationship between *x* and *y*.
- If *r* = 0, then there is no linear relationship between *x* and *y*.

Simple Linear Regression - Pearson Correlation Coefficient

Example 1-1: Temperature in Celsius and Fahrenheit

How strong is the linear relationship between temperatures in Celsius and temperatures in Fahrenheit? Here's a plot of an estimated regression equation based on $n = 11$ data points:

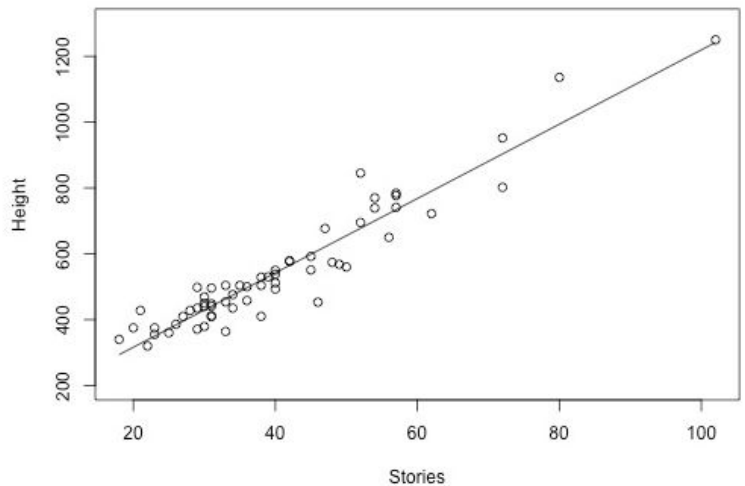


Pearson correlation of Celsius and Fahrenheit = 1.000

Simple Linear Regression - Pearson Correlation Coefficient

Example 1-2: Building Stories and Height

How strong is the linear relationship between the number of stories a building has and its height? One would think that as the number of stories increases, the height would increase, but not perfectly. Some statisticians compiled data on a set of $n = 60$ buildings reported in the 1994 World Almanac (Building Stories data). Minitab's fitted line plot and correlation output look like this:

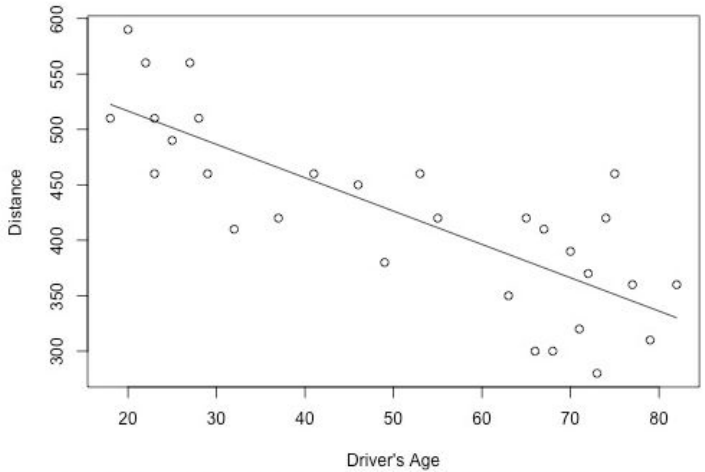


Pearson correlation of HEIGHT and STORIES = 0.951

Simple Linear Regression - Pearson Correlation Coefficient

Example 1-3: Drivers and Age

How strong is the linear relationship between the age of a driver and the distance the driver can see? If we had to guess, we might think that the relationship is negative — as age increases, the distance decreases. A research firm (Last Resource, Inc., Bellefonte, PA) collected data on a sample of $n = 30$ drivers (Driver Age and Distance data). Minitab's fitted line plot and correlation output on the data looks like this:

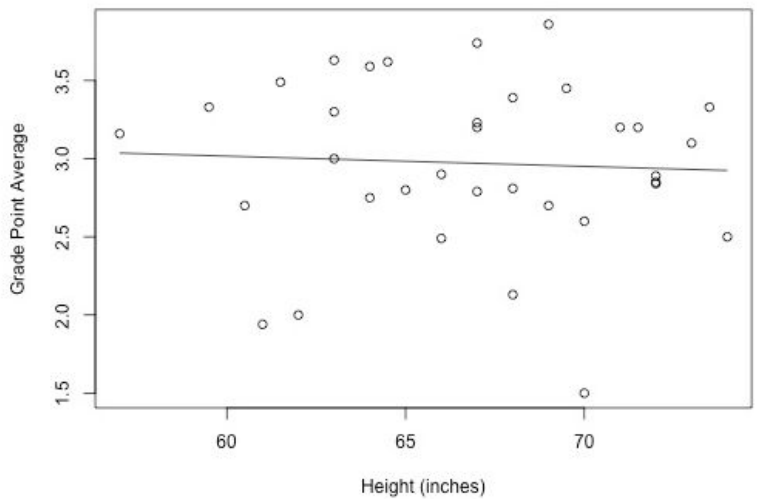


Pearson correlation of Distance and DrivAge = -0.801

Simple Linear Regression - Pearson Correlation Coefficient

Example 1-4: Height and GPA

How strong is the linear relationship between the height of a student and his or her grade point average? Data were collected on a random sample of $n = 35$ students in a statistics course at Penn State University (Height and GPA data) and the resulting fitted line plot and correlation output were obtained:



Pearson correlation of height and GPA = -0.053

Hypothesis Test for the Population Correlation Coefficient ρ

- Used to **learn of a linear association between two variables**, when it isn't obvious which variable should be regarded as the response.

- *t*-test for testing $H_0: \beta_1 = 0$
- ANOVA *F*-test for testing $H_0: \beta_1 = 0$

Steps for Hypothesis Testing for ρ

Step 1: Hypotheses

First, we specify the null and alternative hypotheses:

- Null hypothesis $H_0: \rho = 0$
- Alternative hypothesis $H_A: \rho \neq 0$ or $H_A: \rho < 0$ or $H_A: \rho > 0$

Step 2: Test Statistic

Second, we calculate the value of the test statistic using the following formula:

Test statistic: $t^* = \dfrac{r\sqrt{n-2}}{\sqrt{1-R^2}}$

Step 3: P-Value

Third, we use the resulting test statistic to calculate the *P*-value. As always, the *P*-value is the answer to the question "how likely is it that we'd get a test statistic *t** as extreme as we did if the null hypothesis were true?" The *P*-value is determined by referring to a *t*-distribution with *n*-2 degrees of freedom.

Step 4: Decision

Finally, we make a decision:

- If the *P*-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude that "there is sufficient evidence at theα level to conclude that there is a linear relationship in the population between the predictor x and response y."
- If the *P*-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor *x* and response *y*."

Hypothesis Testing for ρ

Example 1-5: Husband and Wife Data

Let's perform the hypothesis test on the husband's age and wife's age data in which the sample correlation based on n = 170 couples is r = 0.939. To test $H_0: \rho = 0$ against the alternative $H_A: \rho \neq 0$, we obtain the following test statistic:

$$\begin{aligned} t^* &= \frac{r\sqrt{n-2}}{\sqrt{1-R^2}} \\ &= \frac{0.939\sqrt{170-2}}{\sqrt{1-0.939^2}} \\ &= 35.39 \end{aligned}$$

To obtain the *P*-value, we need to compare the test statistic to a *t*-distribution with 168 degrees of freedom (since 170 - 2 = 168). In particular, we need to find the probability that we'd observe a test statistic more extreme than 35.39, and then, since we're conducting a two-sided test, multiply the probability by 2.

Hypothesis Testing for ρ

Student's t distribution with 168 DF

x	P(X<= x)
35.3900	1.0000

The output tells us that the probability of getting a test-statistic smaller than 35.39 is greater than 0.999. Therefore, the probability of getting a test-statistic greater than 35.39 is less than 0.001. As illustrated in

Since the *P*-value is small — smaller than 0.05, say — we can reject the null hypothesis. There is sufficient statistical evidence at the $\alpha = 0.05$ level to conclude that there is a significant linear relationship between a husband's age and his wife's age.

Hypothesis Testing for ρ

when it is okay to use the *t*-test for testing $H_0: \rho = 0$?

- When it is not obvious which variable is the response.
- When the (*x*, *y*) pairs are a random sample from a bivariate normal population.
  - For each *x*, the *y*'s are normal with equal variances.
  - For each *y*, the *x*'s are normal with equal variances.
  - Either, *y* can be considered a linear function of *x*.
  - Or, *x* can be considered a linear function of *y*.
- The (*x*, *y*) pairs are independent

Multiple Linear Regression

Multiple linear regression aims to find a linear relationship between variables in situations where there are several independent variables. The independent variables can either be continuous or qualitative, however the dependent variable must be measured on a continuous scale. A multiple regression model with k independent variables fits a regression "surface" in a $k + 1$ dimensional space.

The least squares regression line for multiple regression of n independent variables is

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

where:

- a is a constant,
- x_n is the nth independent variable,
- b_n is the coefficient of the nth independent variable.

Multiple Linear Regression - Interpretation

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

When all the independent variables $x_1, x_2, \ldots x_n$ are constant the predicted value of y is a .

The predicted value of y changes by b_1 for each one unit increase in x_1, when all other variables are constant. Similarly, the predicted value of y changes by b_2 for each one unit increase in x_2, when all other variables are constant. Therefore, the predicted value of y changes by b_n for each one unit increase in x_n, when all other variables are constant.

Multiple Linear Regression - Bivariate Model

A bivariate model is a model with two independent variables.

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

The values form a plane in a 3-dimensional space.

A multiple regression model with two predictor variables fits a regression plane in 3-dimensional space

The intercept a predicts where the plane will cross the y-axis. The value b_1 is gradient of the variable x_1, this predicts y with every change in unit of x_1 whilst x_2 is constant. The gradient of the variable x_2, b_2, predicts y with every change in unit of x_2 whilst x_1 is constant.

Multiple Linear Regression - Bivariate Model

Usage of Bivariate Model

- To show the correlation between **stress and number of working hours along with stress and days without exercise,** as both sets of variables have a positive correlation.

When Bivariate Model cannot be Used

- To show the correlation between depression and hours of sunlight along with depression and temperature, as hours of sunlight and temperature are not independent of each other.

Minimise the sum of square residuals

- Based on sum of the square residuals for simple regression.
- Used to calculate the equation of the regression line.
- It is very long and has complicated equations, which is why it is usually calculated by a computer.

Multiple Linear Regression - Class Assignment : 1

Q. Suppose we have following data:

Y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Find a multiple linear regression model for the dataset. Also find X_1^2, X_2^2, X_1y, X_2y and X_1X_2

Multiple Linear Regression - Class Assignment : 2

Q. Find multiple linear regression eqn of y on X_1 and X_2.

y	4	6	7	9	13	15
X_1	15	12	8	6	4	3
X_2	30	24	20	14	10	4

Polynomial Regression

- Form of Linear regression
- To handle the Non-linear relationship between dependent and independent variables
- Add some polynomial terms to linear regression to convert it into Polynomial regression.
- It is modeled as an **nth-degree polynomial function.**
    - When the polynomial is of degree 2, it is called a quadratic model;
    - when the degree of a polynomial is 3, it is called a cubic model, and so on.

Polynomial Regression

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Polynomial Regression

- The degree of the polynomial equation depends on **the complexity of the relationship between the independent variable and the dependent variable**
- The degree of order which to use is a **Hyperparameter,** choose it wisely.
  - High degree of polynomial tries to overfit the data,
  - Smaller values of degree, the model tries to underfit,
  - So, find the optimum value of a degree.
- Polynomial Regression models are usually fitted with **the method of least squares**.
  - **minimizes the variance of the coefficients under the Gauss-Markov Theorem.**

Polynomial Regression - Evaluation Metrics

- provide information about how well the model fits the data
- used to compare different models or to select the best model for a given problem.
- Some commonly used evaluation metrics for polynomial regression include:

1. **Mean Squared Error (MSE):**
   a. measures the average squared difference between the predicted and actual values.
   b. Calculated as the sum of the squared differences divided by the number of observations.
   c. The lower the MSE, the better the model performance.
2. **Root Mean Squared Error (RMSE):**
   a. square root of the MSE and provides a measure of the average deviation of the predictions from the actual values.
   b. The lower the RMSE, the better the model performance.
3. **R-squared (R2) Score:**
   - measures the proportion of the variance in the dependent variable that is explained by the independent variable(s) in the model.
   - It ranges from 0 to 1, with higher values indicating better model performance.
4. **Adjusted R-squared Score:**
   - similar to the R-squared score
   - takes into account the number of independent variables in the model.
   - It is adjusted for degrees of freedom and penalizes the model for including unnecessary independent variables.

Polynomial Regression - Quadratic Model

Let the quadratic polynomial regression model be

$$y=a_0+a_1*x+a_2*x^2$$

The values of **a_0, a_1, and a_2** are calculated using the following system of equations:

$$\sum y_i = na_0 + a_1(\sum x_i) + a_2(\sum x_i^2)$$
$$\sum y_i x_i = a_0(\sum x_i) + a_1(\sum x_i^2) + a_2(\sum x_i^3)$$
$$\sum y_i x_i^2 = a_0(\sum x_i^2) + a_1(\sum x_i^3) + a_2(\sum x_i^4)$$

Polynomial Regression - Solved Problem

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

Polynomial Regression - Solved Problem

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

	x	y	x^2	x^3	x^4	yx	yx^2
	3	2.5	9	27	81	7.5	22.5
	4	3.2	16	64	256	12.8	51.2
	5	3.8	25	125	625	19	95
	6	6.5	36	216	1296	39	234
	7	12	49	343	2401	80.5	563.5
Σ	25	27.5	135	775	4659	158.8	966.2

Polynomial Regression - Solved Problem

	x	y	x^2	x^3	x^4	yx	yx^2
	3	2.5	9	27	81	7.5	22.5
	4	3.2	16	64	256	12.8	51.2
	5	3.8	25	125	625	19	95
	6	6.5	36	216	1296	39	234
	7	12	49	343	2401	80.5	563.5
Σ	25	27.5	135	775	4659	158.8	966.2

$$\sum y_i = na_0 + a_1(\sum x_i) + a_2(\sum x_i^2)$$
$$\sum y_i x_i = a_0(\sum x_i) + a_1(\sum x_i^2) + a_2(\sum x_i^3)$$
$$\sum y_i x_i^2 = a_0(\sum x_i^2) + a_1(\sum x_i^3) + a_2(\sum x_i^4)$$

Using the given data we

$$27.5 = 5a_0 + 25a_1 + 135a_2$$
$$158.8 = 25a_0 + 135a_1 + 775a_2$$
$$966.2 = 135a_0 + 775a_1 + 4659a_2$$

Solving this system of equations we get

$$a_0 = 12.4285714$$
$$a_1 = -5.5128571$$
$$a_2 = 0.7642857$$

Polynomial Regression - Solved Problem

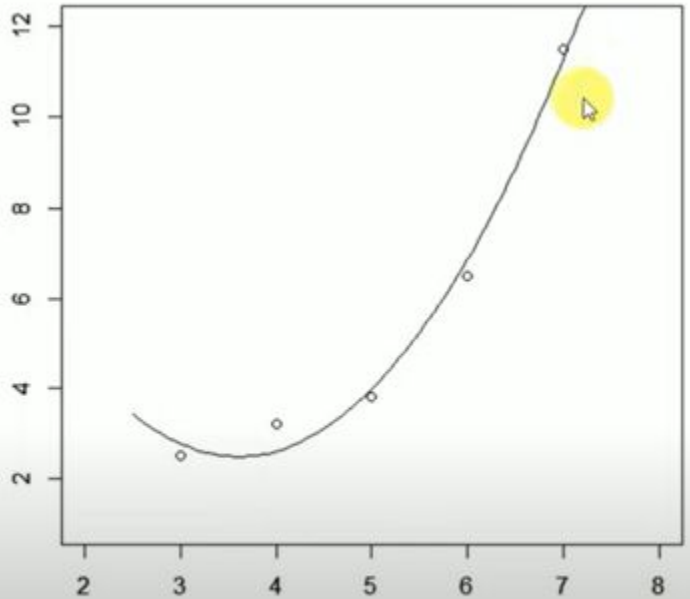
$a_0 = 12.4285714$

$a_1 = -5.5128571$

$a_2 = 0.7642857$

The required quadratic polynomial model is

$y = 12.4285714 - 5.5128571\,x + 0.7642857\,x^2$



Weighted Least Square (WLS) Regression

- an extension of the ordinary least squares (OLS) regression method
- allows for the incorporation of weights in the estimation process.
- useful when the **homoscedasticity** assumption of OLS is **violated**,
  - (**variance of the errors is not constant** across all levels of the independent variable(s).
- Basic idea
  - **assign different weights to different observations**
  - based on the assumption that the **variance of the errors is proportional to the inverse of the weights.**
    - Observations with **higher weights have lower variances,**
    - so, they are given **more influence in the estimation process.**

Weighted Least Square (WLS) Regression

- **Each term in the weighted least squares criterion** includes
  - an additional weight, that determines how much each observation in the data set influences the final parameter estimates
  - it can be used with functions that are either linear or nonlinear in the parameters.

- One of the common assumptions underlying most process modeling methods, including linear and nonlinear least squares regression,
  - Each data point provides equally precise information about the deterministic part of the total process variation.
- it is assumed that the **standard deviation of the error term is constant over all values of the predictor or explanatory variables**.
  - **This assumption, clearly does not hold,** even approximately, in every modeling application.

Weighted Least Square (WLS) Regression

- In a weighted fit,
  - **less weight** is given to the less precise measurements
  - **more weight** to more precise measurements when estimating the unknown parameters in the model.
- Using **weights that are inversely proportional** to the variance at each level of the explanatory variables yields the most precise parameter estimates possible.
- **∑ wi (yi -y'i)^2**
  - wi = Weighting factor for the ith calibration standard
    - (w=1 for unweighted least square regression)
  - yi = Observed instrument response for the i th calibration standard
  - y'i = Predicted (or calculated) response for the ith standard
  - ∑ = The sum of all individual values

Weighted Least Square (WLS) Regression

- Mathematics used in **unweighted least squares regression** has a tendency
  - to favor **numbers of larger value over numbers of smaller value.**
  - Thus the regression curves that are generated will tend to fit points that are at the upper calibration levels better than those points at the lower calibration levels.
- Examples of weighting factors which can place more emphasis on numbers of smaller value are:
  - wi = 1/yi or wi = 1/yi^2
    - wi = weighting factor for the ith calibration standard
      - (wi =1 for unweighted least squares regression).
    - yi = observed instrument response (area or height) for the ith calibration standard

Weighted Least Square (WLS) Regression

Different Types Of Weights

No Weights:	Default higher weighting of higher amounts or signal values
1/Amount:	Nearly cancels out the weighting of higher amounts
1/Amount^2 :	Causes over-proportional weighting of smaller amounts
1/Response:	Nearly cancels out the weighting of higher signal values

Ref: Chromeleon Manual

1/Response^2:	Causes over-proportional weighting of smaller signal values
1/RSD:	Weights signal values with small relative standard deviations more than those with large relative standard deviations
1/RSD ^{2} :	Weights signal values with small relative standard deviations clearly more than those with large relative standard deviation.

Ref: Chromeleon Manual

Weighted Least Square (WLS) Regression - Benefits

- Weighted least squares is an efficient method that makes **good use of small data sets**.
- It also **shares the ability to provide different types of easily interpretable statistical intervals for estimation, prediction, calibration and optimization.**
- The main advantage that weighted least squares enjoys over other methods is the **ability to handle regression situations in which the data points are of varying quality.**

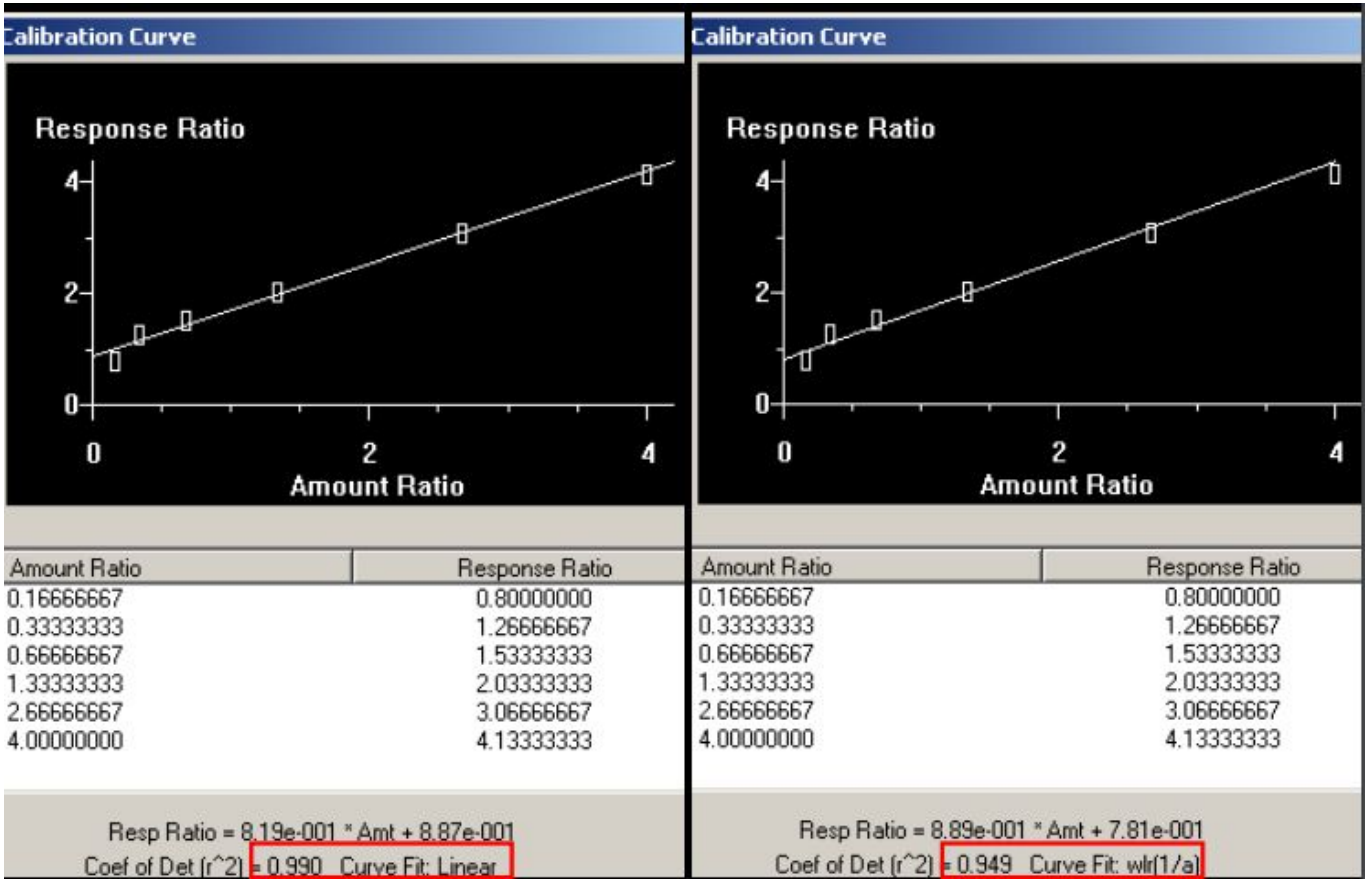
Weighted Least Square (WLS) Regression - Disadvantages

1. based on the **assumption that the weights are known exactly.**
   - The **exact weights are almost never known in real applications**, so estimated weights must be used instead.
   - The effect of using **estimated weights is difficult to assess,** but experience indicates that small variations in the weights due to estimation do not often affect a regression analysis or its interpretation.
2. When the weights are estimated from small numbers of replicated observations, the results of **an analysis can be very badly and unpredictably affected.**
3. **use weighted least squares when the weights can be estimated precisely relative to one another.**

Weighted Least Square (WLS) Regression - Disadvantages

- **Sensitive to the effects of outliers.**
  - If potential outliers are not investigated and dealt with appropriately, they will likely have a negative impact on the parameter estimation and other aspects of a weighted least squares analysis.
  - If a weighted least squares regression actually **increases the influence of an outlier,** the results of the analysis may be far inferior to an unweighted least squares analysis.

Weighted Least Square (WLS) Regression



Amount Ratio	Response Ratio
0.16666667	0.80000000
0.33333333	1.26666667
0.66666667	1.53333333
1.33333333	2.03333333
2.66666667	3.06666667
4.00000000	4.13333333

Resp Ratio = 8.19e-001 * Amt + 8.87e-001
Coef of Det (r^2) = 0.990 Curve Fit: Linear

Amount Ratio	Response Ratio
0.16666667	0.80000000
0.33333333	1.26666667
0.66666667	1.53333333
1.33333333	2.03333333
2.66666667	3.06666667
4.00000000	4.13333333

Resp Ratio = 8.89e-001 * Amt + 7.81e-001
Coef of Det (r^2) = 0.949 Curve Fit: wlr(1/a)

Ridge Regression Model

- also known as Tikhonov regularization or L2 regularization,
- is a **linear regression technique** that introduces a penalty term to the ordinary least squares (OLS) objective function.
- **Purpose :**
  - to address multicollinearity issues
  - prevent overfitting by adding a **regularization term that discourages the model from relying too heavily** on any particular predictor variable.

Ridge Regression Model

- Ridge Regression modifies the OLS objective function by adding a penalty term based on the squared magnitudes of the regression coefficients:

$$\text{Objective Function} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where:

- y_i is the observed value.
- $\hat{y}_i$ is the predicted value.
- λ is the regularization parameter (also known as the shrinkage parameter).
- β_j are the regression coefficients.

Ridge Regression Model

- **Shrinkage Parameter (λ)**:
  - controls the strength of the regularization.
  - A larger λ leads to greater shrinkage of the coefficients, effectively penalizing large coefficients more severely.
- **Bias-Variance Tradeoff:**
  - Bias to the model in exchange for reduced variance. This can be beneficial in situations where multicollinearity is present, as it helps stabilize the estimates of the regression coefficients.
- **Multicollinearity:**
  - Ridge Regression is particularly useful when there are high correlations among predictor variables, which can lead to unstable and unreliable estimates in OLS.
  - The regularization term helps to distribute the influence of correlated variables more evenly.

Ridge Regression Model

- **Standardization of Variables:**
  - It is common practice to standardize the predictor variables before applying Ridge Regression.
  - Standardization involves subtracting the mean and dividing by the standard deviation for each variable.
    - This **ensures that all variables are on a similar scale,**
    - regularization term has a **consistent impact across predictors.**
- **No Variable Selection:**
  - Ridge Regression does not perform variable selection in the same way as methods like LASSO (L1 regularization).
  - **It tends to shrink all coefficients toward zero**, but it rarely sets them exactly to zero.
- **Cross-Validation for λ Selection:**
  - The choice of the shrinkage parameter (λ) is critical.
  - Cross-validation techniques, such as k-fold cross-validation, are often employed to select an optimal λ that balances model complexity and performance.

Ridge Regression Model

- The Ridge Regression coefficient estimates ($\hat{\beta}_{\text{ridge}}$) are obtained by minimizing the modified objective function. The closed-form solution involves linear algebra, and it can be expressed as:

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

where:

- X is the matrix of predictor variables.
- y is the vector of observed values.
- I is the identity matrix.

Loess Regression Model

- Loess (**Locally Weighted Scatterplot Smoothing**) regression
- **non-parametric regression technique** used for estimating relationships between variables.
- **useful** when **dealing with complex and non-linear relationships in the data**.
- **fits a smooth curve to the data by locally fitting a polynomial regression model to subsets of the data.**
- **Key Characteristics:**

  1. **Local Regression:**
     - by fitting a polynomial model to a subset of the data points within a specified neighborhood (window) around each point.
     - This **allows the model to capture local trends in the data.**

  2. Weighted Regression:
     - The fitting process in Loess **involves assigning weights to data points based on their proximity to the point being predicted**.
     - Points closer to the target point have **higher weights**, while those farther away have lower weights.
     - This weighting emphasizes the **influence of nearby points** in the local regression

Loess Regression Model

3. **Polynomial Fitting:**
   - In each local subset of the data, a **polynomial regression model is fitted**.
   - The degree of the polynomial is typically low (e.g., quadratic or cubic) **to avoid overfitting.**
4. **Smoothing Parameter:**
   - The degree of smoothing in Loess is controlled by a parameter often denoted as **α or τ.**
   - This parameter determines the size of the local neighborhood and influences the degree of flexibility in the fitted curve.
   - A larger smoothing parameter results in a **smoother curve.**
5. **Residual Weighting:**
   - beneficial when dealing with **heteroscedasticity** (varying levels of variability across the data).
6. **Adaptive Bandwidth:**
   - The bandwidth or window size can vary across different regions of the dataset.
   - This adaptability helps to capture local features accurately, especially in **areas where the relationship between variables changes.**

Loess Regression Model

7. **Iterative Process:**

- The fitting process in Loess is typically iterative.
- After the initial fit, the **weights are adjusted based on the residuals**, and the process is repeated. This iteration helps in refining the model and improving the fit.

8. Robustness

- Loess regression is generally **robust to outliers** since the influence of each point is locally determined.
- Outliers in one region may have minimal impact on the fit in other regions.

Hypothesis Testing for ρ

Returning to the Income example, in addition to the variables age and education, the person's gender, female or male, is considered an input variable. The following code reads a comma-separated-value (CSV) file of 1,500 people's incomes, ages, years of education, and gender. The first 10 rows are displayed:

```
income_input = as.data.frame( read.csv("c:/data/income.csv") )
income_input[1:10,]
```

```
ID Income Age Education Gender
1 1 113 69  12  1
2 2  91 52  18  0
3 3 121 65  14  0
4 4  81 58  12  0
5 5  68 31  16  1
6 6  92 51  15  1
7 7  75 53  15  0
8 8  76 56  13  0
9 9  56 42  15  1
10 10  53 33  11  1
```

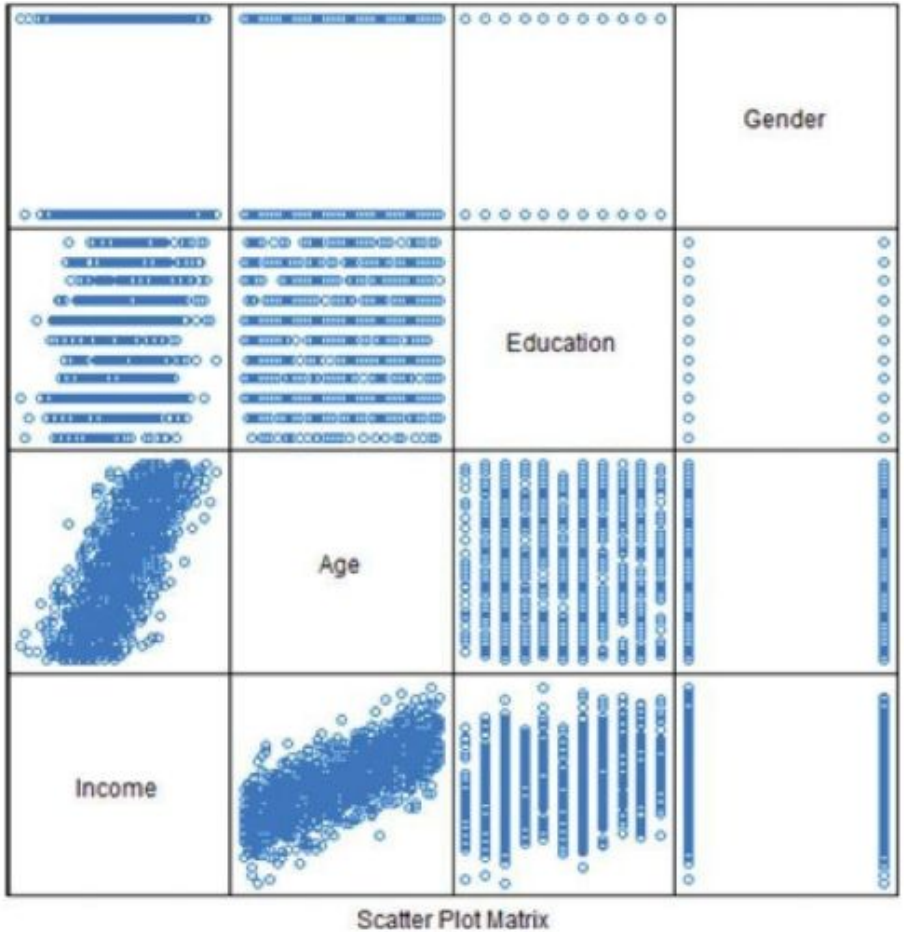
Hypothesis Testing for ρ

```
summary(income_input)

      ID            Income          Age          Education
 Min.   :   1.0   Min.   :  14.00   Min.   :18.00   Min.   :10.00
 1st Qu.: 375.8   1st Qu.:  62.00   1st Qu.:30.00   1st Qu.:12.00
 Median : 750.5   Median :  76.00   Median :44.00   Median :15.00
 Mean   : 750.5   Mean   :  75.99   Mean   :43.58   Mean   :14.68
 3rd Qu.:1125.2   3rd Qu.:  91.00   3rd Qu.:57.00   3rd Qu.:16.00
 Max.   :1500.0   Max.   : 134.00   Max.   :70.00   Max.   :20.00
     Gender
 Min.   :0.00
 1st Qu.:0.00
 Median :0.00
 Mean   :0.49
 3rd Qu.:1.00
 Max.   :1.00
```

Hypothesis Testing for ρ

```
library(lattice)
splom(~income_input[c(2:5)], groups=NULL, data=income_input,
  axis.line.tck = 0,
  axis.text.alpha = 0)
```



Scatter Plot Matrix

Hypothesis Testing for ρ

```
results <- lm(Income~Age + Education + Gender, income_input)
summary(results)
```

```
Call:
lm(formula = Income ~ Age + Education + Gender, data = income_input)

Residuals:
    Min      1Q  Median      3Q     Max
-37.340  -8.101   0.139   7.885  37.271

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.26299    1.95575   3.714 0.000212 ***
Age          0.99520    0.02057  48.373  < 2e-16 ***
Education    1.75788    0.11581  15.179  < 2e-16 ***
Gender      -0.93433    0.62388  -1.498 0.134443
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.07 on 1496 degrees of freedom
Multiple R-squared:  0.6364,    Adjusted R-squared:  0.6357
F-statistic:   873 on 3 and 1496 DF,  p-value: < 2.2e-16
```

Hypothesis Testing for ρ

The residuals are the observed values of the error term for each of the n observations and are defined for i = 1, 2, …n, as shown in Equation 6.6.

$$6.6 \quad e_i = y_i - (b_0 + b_1 x_{i,1} + b_2 x_{i,2} \ldots + b_{p-1} x_{i,p-1})$$

where bj denotes the estimate for parameter β_i for $j = 0, 1, 2, \ldots p - 1$

Note :
Residuals are assumed to be normally distributed with a mean near zero and a constant variance.
Intercept

- estimated income of $7,263 for a newborn female with no education.
- It is important to note that the available dataset does not include such a person.
- The minimum age and education in the dataset are 18 and 10 years, respectively

Hypothesis Testing for ρ

Coefficient for Age

- is approximately equal to one.
- For every one unit increase in a person's age, the person's income is expected to increase by $995.

Coefficient for Education:

- for every unit increase in a person's years of education, the person's income is expected to increase by about $1,758.

Coefficient for Gender

- When Gender is equal to zero, the Gender coefficient contributes nothing to the estimate of the expected income.
- When Gender is equal to one, the expected Income is decreased by about $934.

Hypothesis Testing for ρ

Note:

- **Coefficient** values are only estimates based on the observed incomes in the sample, there is some uncertainty or sampling error for the coefficient estimates.
- **Std. Error**
  - provides the sampling error associated with each coefficient
  - used to perform a hypothesis test, using the t-distribution, to determine if each coefficient is statistically different from zero.

Hypothesis Testing for ρ

- If a **coefficient is not statistically different from zero**, the coefficient and the associated variable in the model should be excluded from the model.
- In this example, the associated hypothesis tests' p-values, Pr(>|t|), are very small for the Intercept, Age, and Education parameters

for a given j = 0, 1, 2, …, p – 1, the null and alternate hypotheses

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_A : \beta_j \neq 0$$

- For small p-values, as is the case for the Intercept, Age, and Education parameters, the null hypothesis would be rejected.
- For the Gender parameter, the corresponding p-value is fairly large at 0.13.
- In other words, at a 90% confidence level, the null hypothesis would not be rejected.
- So, dropping the variable Gender from the linear regression model should be considered

Hypothesis Testing for ρ

```
results2 <- lm(Income ~ Age + Education, income_input)
summary(results2)

Call:
lm(formula = Income ~ Age + Education, data = income_input)

Residuals:
    Min      1Q  Median      3Q     Max
-36.889  -7.892   0.185   8.200  37.740

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

Hypothesis Testing for ρ

```
(Intercept)  6.75822    1.92728   3.507 0.000467 ***
Age          0.99603    0.02057  48.412  < 2e-16 ***
Education    1.75860    0.11586  15.179  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.08 on 1497 degrees of freedom
Multiple R-squared:  0.6359,  Adjusted R-squared:  0.6354
F-statistic:  1307 on 2 and 1497 DF,  p-value: < 2.2e-16
```

Hypothesis Testing for ρ

- **Residual standard error :** is the standard deviation of the observed residuals.
  - This value, along with the associated degrees of freedom, can be used to examine the variance of the assumed normally distributed error terms.
- **R-squared** (R 2)
  - measures the variation in the data that is explained by the regression model.
  - vary from 0 to 1,
  - closer to 1 : the model is better at explaining the data than values closer to 0.
  - exactly 1 : the model explains perfectly the observed data (all the residuals are equal to 0).
  - R squared can be increased by adding more variables to the model.
  - However, just adding more variables results in **overfitting.**
- **F-statistic :** provides a method for testing the entire regression model.
  - t-tests, individual tests were conducted to determine the statistical significance of each parameter.
  - F-statistic and corresponding p-value enable the analyst to test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \quad \text{versus} \quad H_A : \beta_j \neq 0$$
$$\text{for at least one } j = 1, 2, \ldots, p-1$$

Hypothesis Testing for ρ

- **F-statistic :** provides a method for testing the entire regression model.
  - t-tests, individual tests were conducted to determine the statistical significance of each parameter.
  - F-statistic and corresponding p-value enable the analyst to test the following hypotheses:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \quad \text{versus} \quad H_A: \beta_j \neq 0$$
$$\text{for at least one } j = 1, 2, \ldots, p-1$$

In this example, the p-value of 2.2e – 16 is small, which indicates that the null hypothesis should be rejected.

Interaction Models

- Examine the **combined effects of two or more independent variables on the dependent variable.**
- provide insights into how the relationship between one independent variable and the dependent variable depends on the level or presence of another variable.
- Interaction terms are created by multiplying the values of two or more variables.
- Common types of Interaction Models
  1. Two-Way Interaction:
  2. Three-Way Interaction:
  3. Moderation (Interaction with a Moderator):
  4. Mediation (Interaction with a Mediator):
  5. Cross-Product Terms:

Interaction Models - Types

1. **Two-Way Interaction:**
   - Involves the interaction between two independent variables.
   - includes the main effects of both variables as well as their interaction term.
   - $Y=\beta 0+\beta 1\ X1+\beta 2\ X2+\beta 3\ (X1\times X2)+\epsilon$
2. **Three-Way Interaction:**
   - Involves the interaction between three independent variables.
   - The model includes the main effects of all three variables as well as their pairwise and triple interaction terms.
   - $Y=\beta 0+\beta 1\ X1+\beta 2\ X2+\beta 3\ X3+\beta 4\ (X1\times X2)+\beta 5\ (X1\times X3)+\beta 6\ (X2\times X3)+\beta 7(X1\times X2\times X3)+\epsilon$
3. **Moderation (Interaction with a Moderator):**
   - Examines whether the effect of one independent variable on the dependent variable is moderated by the presence or level of another variable.
   - $Y=\beta 0+\beta 1\ X1+\beta 2\ X2+\beta 3\ (X1\times X2)+\epsilon$, Where, $X2$ moderates the relationship between $X1$ and Y.

Interaction Models - Types

4. **Mediation (Interaction with a Mediator):**
   - Investigates whether the effect of one independent variable on the dependent variable is mediated by the presence or level of another variable.
   - $Y=\beta0 +\beta1\ X1 +\beta2\ X2 +\beta3\ (X1 \times X2)+\epsilon$, where X2 mediates the relationship between X1 and Y.
5. **Cross-Product Terms:**
   - More general term for interaction terms
   - involve multiplication of various independent variables.
   - $Y=\beta0 +\beta1\ X1 +\beta2\ X2 +\beta3\ (X1 \times X2)+\beta4\ (X1 \times X3)+\epsilon$

Interaction models

- Helps to explore how the effects of one variable may change depending on the values of another variable.
- involves examining the coefficients of the interaction terms and understanding how they influence the overall relationship between variables.

Qualitative Predictor variables

- also known as **categorical variables,**
- pose some unique challenges compared to continuous variables.
- can be either nominal or ordinal,
- their inclusion in regression models requires special considerations.
- Some techniques used with qualitative predictor variables:

1. **Dummy Coding (Indicator Variables):**
   - common technique for representing categorical variables in regression models.
   - For a **categorical variable with *k* levels, *k*−1 dummy variables are created.**
   - Each dummy variable takes the value 0 or 1, ⇒ absence or presence of a particular level of the categorical variable.
   - Example: For a variable "Color" with three levels (Red, Green, Blue), you might create two dummy variables: D_1 for Green and D_2 for Blue, with Red as the reference level.

Qualitative Predictor variables

2. **Effect Coding:**
   - Similar to dummy coding,
   - reference category is assigned a value of -1, and the other categories are coded as 0 or 1.
   - useful when you are interested in the overall average effect of the categorical variable.
3. **Contrast Coding:**
   - Involves creating contrasts that represent specific comparisons of interest among the levels of the categorical variable.
   - Popular contrast codings include treatment (dummy) coding and Helmert coding.
4. **Interaction with Dummy Variables:**
   - Used when both continuous and categorical predictors,
   - to capture potential differential effects.

Model evaluation Measures

1. **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

   - Represents the average absolute differences between the observed and predicted values.

2. **Mean Squared Error (MSE):**

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

   - Measures the average squared differences between observed and predicted values.

3. **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{MSE}$$

   - Square root of the mean squared error, providing a more interpretable scale.

4. **R-squared (R2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   - Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.
   - Ranges from 0 to 1, where 1 indicates a perfect fit.

Model evaluation Measures

5. **Adjusted R-squared:**

$$R^2_{\text{adj}} = 1 - \left(\frac{(1-R^2)\cdot(n-1)}{n-k-1}\right)$$

   - Adjusts R-squared for the number of predictors in the model, providing a more realistic measure.
   - Penalizes the addition of irrelevant variables that do not improve the model significantly.

6. **Mean Absolute Percentage Error (MAPE):**

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|} \times 100$$

   - Represents the average percentage difference between observed and predicted values.
   - Useful for expressing errors as a percentage of the observed values.

7. **Mean Bias Deviation (MBD):**

$$MBD = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

   - Represents the average difference between predicted and actual values.

Model Selection Procedures

- choosing the most appropriate model from a set of candidate models.
- find a model that balances goodness of fit with simplicity to avoid overfitting.

1. **Stepwise Regression**
2. **Subset Selection:**
3. **Regularization Techniques**
4. **Information Criteria**
5. **Cross-Validation**
6. **Bootstrap Resampling**
7. **Model Comparison**
8. **Domain Knowledge**
9. **Model Diagnostics**

Model Selection Procedures

1. **Stepwise Regression:**
   - **Forward Selection:**
     - Starts with an empty model
     - Adds predictors one at a time,
     - selecting the one that most improves the model fit at each step.
   - **Backward Elimination:**
     - Starts with all predictors in the model
     - Removes the one that contributes least to the model fit at each step.

Model Selection Procedures

2. Subset Selection

- **Best Subset Selection:**
  - Fits all possible combinations of predictors
  - selects the model with the best fit based on a criterion such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).
  - balance the goodness of fit and the number of parameters in the model.
  - Model with the lowest AIC or BIC is often selected.
  - Formula : $AIC = -2 \cdot \ln(L) + 2k$ $BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$
    - L is the likelihood of the model.
    - k is the number of parameters in the model.
    - n is the number of observations in the dataset.
- **Recursive Feature Elimination (RFE):**
  - Iteratively removes the least important variable until the desired number of features is reached.

Model Selection Procedures

3. **Regularization Techniques:**
   - **Ridge Regression:**
     - Introduces a regularization term to the least squares equation,
     - preventing overfitting by penalizing large coefficients.
   - **Lasso Regression:**
     - Similar to ridge regression
     - but uses the absolute values of coefficients, ***promoting sparsity and variable selection.***
   - **Elastic Net Regression:**
     - A combination of ridge and lasso regularization, balancing their strengths.
4. **Information Criteria:**
   - **Akaike Information Criterion (AIC):**
     - Penalizes models for complexity,
     - favoring simpler models that explain the data well.
   - **Bayesian Information Criterion (BIC):**
     - Similar to AIC
     - But places a stronger penalty on model complexity.

Model Selection Procedures

5. **Cross-Validation:**
   - **k-Fold Cross-Validation:**
     - Divide the data into k folds,
     - train the model on k-1 folds,
     - Validate on the remaining fold.
     - Repeat this process k times, rotating the validation set.
   - **Leave-One-Out Cross-Validation (LOOCV):**
     - Special case of k-fold where k equals the number of observations.
     - Each observation serves as a validation set in turn.

Model Selection Procedures

6. **Bootstrap Resampling:**
   - **Bootstrap Aggregating (Bagging):**
     - Create multiple bootstrap samples from the dataset,
     - train models on each sample,
     - average their predictions to reduce variance.
   - **Bootstrap Confidence Intervals:**
     - Assess the stability and reliability of regression coefficients using bootstrap resampling.
7. **Model Comparison:**
   - Compare different regression models using statistical tests or information criteria to determine the most suitable model for your data.
8. **Domain Knowledge:**
   - Consider the theoretical aspects of the problem and domain knowledge to guide variable selection and model specification.
9. **Model Diagnostics:**
   - Use residual analysis, leverage plots, and other diagnostic tools to identify potential issues with the chosen model.

Leverage in Regression

- **influence that a single data point can have on the overall fit of a regression model.**
- It is a measure of **how much a particular observation can affect the estimated regression coefficients.**
- important in multiple linear regression, where there are **more than one independent variable.**
- measure of **how far an independent variable value is from the mean of the independent variables.**
- Leverage for an observation i in a dataset with n observations is given by **Hat Matrix**
  - Also known as **Projection Matrix**
  - defined in terms of the data matrix *X*:
    - p is the number of coefficients,
    - n is the number of observations (rows of X) in the regression model.
  - HatMatrix is an n-by-n matrix in the Diagnostics table.

$$H = X(X^TX)^{-1}X^T$$

and determines the fitted or predicted values since

$$\hat{y} = Hy = Xb.$$

The diagonal elements of H, h_{ii}, are called leverages and satisfy

$$0 \leq h_{ii} \leq 1$$

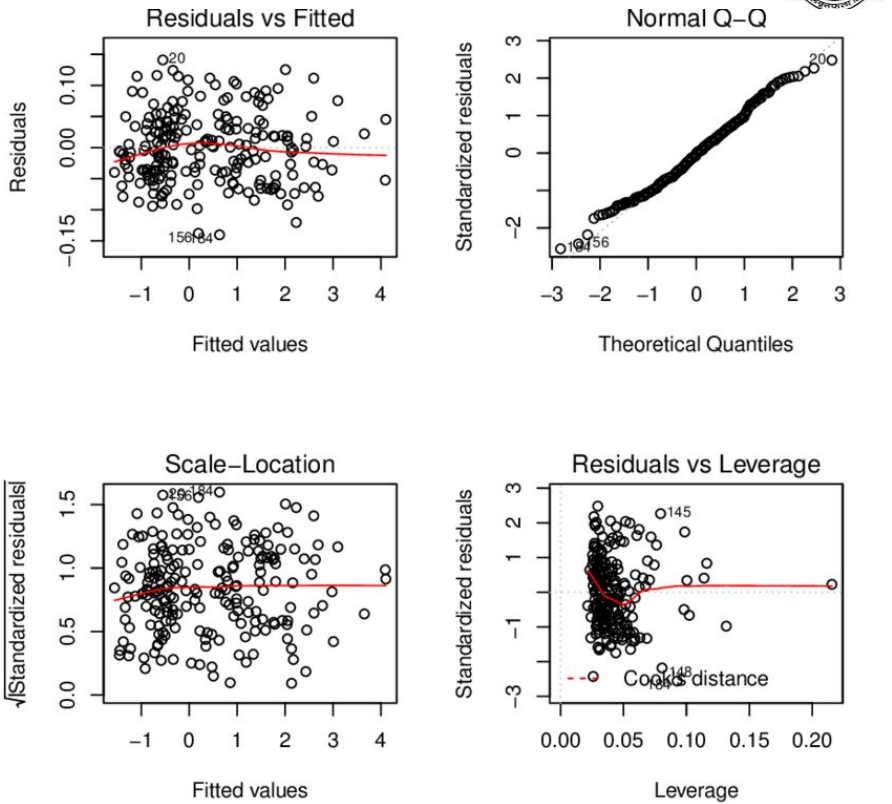
$$\sum_{i=1}^{n} h_{ii} = p,$$

Leverage in Regression

- **Influence and Outliers:**
  - Observations with high leverage can have a strong influence on the estimated coefficients.
  - If a data point has high leverage and an extreme value in the dependent variable,
    - it can significantly impact the regression model,
    - potentially **leading to outliers or influential points.**
- **High Leverage Points:**
  - Points with high leverage typically have extreme values in one or more independent variables.
  - These points have the potential to **disproportionately influence the regression model,**
    - especially if they deviate from the overall pattern of the data.

Leverage in Regression

- **Diagnostic plots,**
  - such as **leverage-residual plots,**
  - Helps to identify observations with high leverage.
  - Observations with both high leverage and large residuals may have a substantial impact on the model.



Residual Leverage Plot (Regression Diagnostic)

Regression analysis requires some assumptions to be followed by the dataset.

- Observations are independent of each other. It should be correlated to another observation.
- Data is normally distributed.
- The relationship b/w the independent variable and the mean of the dependent variable is **linear.**
- The data is in **homoscedasticity**, (variance of the residual is the same for each value of the dependent variable.)

To perform a good linear regression analysis, check whether these assumptions are violated:

- If the data contain non-linear trends then it will not be properly fitted by linear regression resulting in a high residual or error rate.
- To check for the normality in the dataset, draw a **Q-Q plot on the data.**
- The presence of correlation between observations is known as autocorrelation. **autocorrelation plot.**
- The presence of homoscedasticity (**Scale Location plot, the Residual vs Legacy plot.)**

Residual Leverage Plot (Regression Diagnostic)

1. **Residual vs fitted plot:**
   - This plot is used to check for linearity and homoscedasticity,
   - if there is a linear relationship then it should have a **horizontal line** with much deviation.
   - If the model meets the condition for homoscedasticity, the graph should be **equally spread around the y=0 line.**

- **Q-Q plot:**
  - This plot is used to **check for the normality** of the dataset,
  - if there is normality that exists in the dataset then, the scatter points will be distributed along the 45 degrees dashed line.

Residual Leverage Plot (Regression Diagnostic)

3. **Scale-Location plot:**
   - It is a plot of square rooted standardized value vs predicted value.
   - This plot is used for **checking the homoscedasticity of residuals.**
   - Equally spread residuals across the horizontal line indicate the **homoscedasticity of residuals**.
4. **Residual vs Leverage plot:**
   - plot between standardized residuals and leverage points of the points.

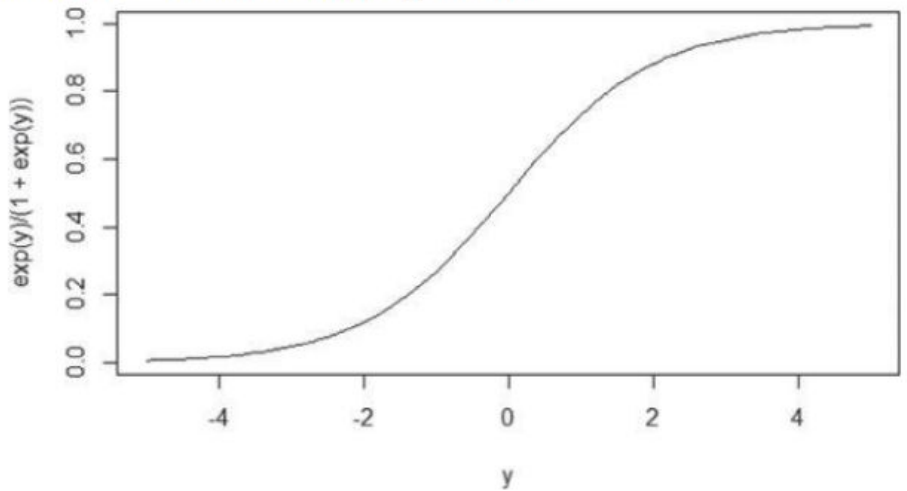
Logistic Regression

- Statistical method used for binary classification, predicting the probability that an instance belongs to a particular class.
- commonly used for classification problems rather than regression problems.
- **Sigmoid Function (Logistic Function):**
  - Is used to model the probability that a given input belongs to a particular class.
  - y is expressed as a linear function of the input variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1}$$

$$f(y) = \frac{e^y}{1+e^y} \quad \text{for } -\infty < y < \infty \qquad (6.7)$$

that as $y \to \infty, f(y) \to 1$, and as $y \to -\infty, f(y) \to 0$. So, as Figure 6.14 illustrates, the value of the ic function $f(y)$ varies from 0 to 1 as y increases.



e 6.14 The logistic function

Logistic Regression

- **Probability Prediction:**
  - The output of the sigmoid function represents the probability that the instance belongs to the positive class (class 1): $P(y=1)=\sigma(z)$
  - The probability of belonging to the negative class (class 0) is then $1-P(y=1)$.
  - based on the input variables , the probability of an event

$$p(x_1, x_2, \ldots, x_{p-1}) = f(y) = \frac{e^y}{1+e^y} \quad \text{for} -\infty < y < \infty$$

- **Basic difference with Linear Regression**
  - the values of y are not directly observed.
  - Only the value of in terms of success or failure (typically expressed as 1 or 0, respectively) is observed.
- **Log odds Ratio**
  - Using p, f(y) can be rewritten as : $$ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1}$$

Logistic Regression

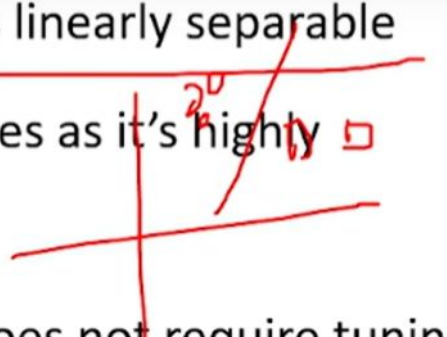
- **Maximum Likelihood Estimation (MLE)**
  - used to estimate the model parameters, that maximize the chances of observing the given dataset.
- **Decision Boundary:**
  - A threshold (usually 0.5) is set, and if the predicted probability is above this threshold, the instance is classified as belonging to the positive class; otherwise, it is classified as belonging to the negative class.
- **Training:**
  - The model is trained by adjusting the weights (b0, b1,…,bn) using an optimization algorithm such as **gradient descent.**
  - The objective is to minimize a cost function that penalizes the model for making incorrect predictions.

Logistic Regression Applications

- Is the mail spam or not spam? The answer is yes or no. Thus, categorical dependent variable is a binary response of yes or no.
- If the student should be admitted or not is based on entrance examination marks. Here, categorical variable response is admitted or not.
- Ecommerce companies can identify buyers if they are likely to purchase a certain product
- Companies can predict whether they will gain or lose money in the next quarter, year, or month based on their current performance

Logistic Regression Adavantages

- Logistic regression performs better when the data is linearly separable
- It does not require too many computational resources as it's highly interpretable
- There is no problem scaling the input features—It does not require tuning
- It is easy to implement and train a model using logistic regression
- It gives a measure of how relevant a predictor (coefficient size) is, and its direction of association (positive or negative)



Logistic Regression - Problem

- The dataset of pass or fail in an exam of 5 students is given in the table.
- Use logistic regression as classifier to answer the following questions.

1. Calculate the probability of pass for the student who studied 33 hours.
2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

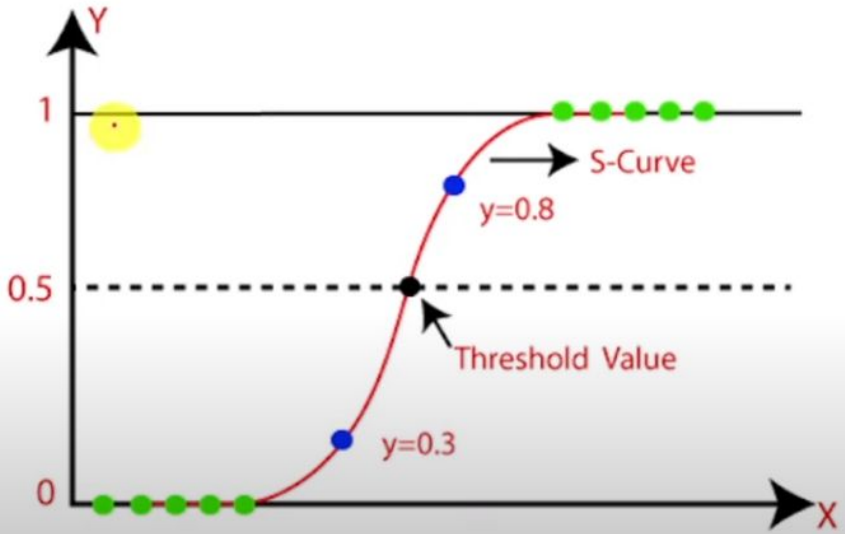
Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(odds) = -64 + 2 * hours$$

Logistic Regression - Problem

- We use Sigmoid Function in logistic regression

- $s(x) = \frac{1}{1+e^{-x}}$



Logistic Regression - Problem

1. Calculate the probability of pass for the student who studied 33 hours.

- $p = \frac{1}{1+e^{-z}}$ $\qquad s(x) = \frac{1}{1+e^{-x}}$
- $z = -64 + 2 * 33 = -64 + 66 = 2$
- $p = \frac{1}{1+e^{-2}} = 0.88$
- That is, if student studies 33 hours, then there is **88% chance** that the student will pass the exam

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

$$\log(odds) = z = -64 + 2 * hours$$

Logistic Regression - Problem

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- $p = \frac{1}{1+e^{-z}} = 0.95$
- $0.95 * (1 + e^{-z}) = 1$
- $0.95 * e^{-z} = 1 - 0.95$
- $e^{-z} = \frac{0.05}{0.95} = 0.0526$
- $\ln(e^{-z}) = \ln(0.0526)$

$\ln(e^x) = x$

$-z = ln(0.0526) = -2.94$

$z = 2.94$

Exit full screen (f)

Logistic Regression - Problem

- $z = 2.94$
- $\log(odds) = z = -64 + 2 * hours$
- $2.94 = -64 + 2 * hours$
- $2 * hours = 2.94 + 64$
- $2 * hours = 66.94$
- $hours = \frac{66.94}{2}$
- $\boldsymbol{hours = 33.47\ Hours}$

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- The student should study **at least 33.47 hours**, so that he will pass the exam with more than 95% probability

Generalized Linear Model

- statistical modeling framework that extends the classical linear regression model
- to handle a broader range of data types and distributions.
- assumes normally distributed errors and continuous response variables.
- accommodating various types of response variables and distributional assumptions.
- Key components of a Generalized Linear Model include:
  - Random Component:
    - This part of the model specifies the distributional family of the response variable, which can be chosen based on the nature of the data. Examples of distribution families include Gaussian (for continuous data), Binomial (for binary data), Poisson (for count data), and Gamma (for positively skewed continuous data).
  - Systematic Component:
    - describes how the linear predictor is related to the predictors.
    - It includes a linear combination of the predictor variables, each multiplied by a regression coefficient.
  - Link Function:
    - connects the mean of the distribution (specified by the random component) to the linear predictor in the systematic component.
    - The choice of link function depends on the distributional family and the characteristics of the data. Common link functions include the identity, logit, log, and inverse.

Generalized Linear Model

The general form of a GLM can be expressed as:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Here, $g(\mu)$ is the link function, μ is the mean of the distribution, and $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients.

Logistic Regression Vs Generalized Linear Model

Feature	Logistic Regression	Generalized Linear Models (GLM)
Objective	Binary classification	Generalized framework for regression models
Type of Model	Specific case of GLM for binary outcomes	General framework that includes LR as a special case
Dependent Variable	Binary (0 or 1)	Can handle various types (e.g., continuous, count)
Link Function	Logit function (sigmoid)	Can use different link functions based on the distribution (e.g., log, identity)
Distribution	Binomial	Various distributions (e.g., Gaussian, Poisson)
Assumption	Assumes a binomial distribution of the response variable	More flexible in terms of distribution assumptions
Use Cases	Binary classification problems	Regression problems with different types of response variables
Interpretability	Coefficients represent log-odds	Interpretation depends on the chosen link function
Examples	Predicting whether an email is spam or not	Predicting house prices, count of events, etc.

Linear Regression Vs Logistic Regression

Feature	Linear Regression	Logistic Regression
Type of Output	Continuous (Real Numbers)	Binary or Multinomial (Categorical)
Nature of Dependent Variable	Continuous	Categorical (Binary or Multinomial)
Equation Form	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$	$p(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots)}}$
Objective Function	Minimize the sum of squared residuals	Maximize the likelihood function (cross-entropy)
Error Term	Residuals $(Y - \hat{Y})$	Log-odds (logit function)

Linear Regression Vs Logistic Regression

Feature	Linear Regression	Logistic Regression
Assumption of Linearity	Assumes a linear relationship between variables	Assumes a linear relationship between log-odds and predictors
Assumption of Homoscedasticity	Assumes constant variance of errors	Does not assume constant variance of errors
Assumption of Independence	Assumes independence of errors	Assumes independence of errors
Predicted Values	Unrestricted range (-∞ to +∞)	Constrained between 0 and 1 (log-odds transformed)

Linear Regression Vs Logistic Regression

Feature	Linear Regression	Logistic Regression
Application	Prediction of continuous outcomes	Prediction of binary or categorical outcomes
Model Evaluation	Mean Squared Error (MSE), R-squared	Accuracy, Precision, Recall, F1 Score, ROC-AUC, etc.
Gradient Descent Usage	Used for optimization in some cases	Commonly used due to the non-linear log-odds transformation
Example Use Cases	Predicting house prices, temperature, etc.	Binary classification (spam or not spam), disease prediction, etc.