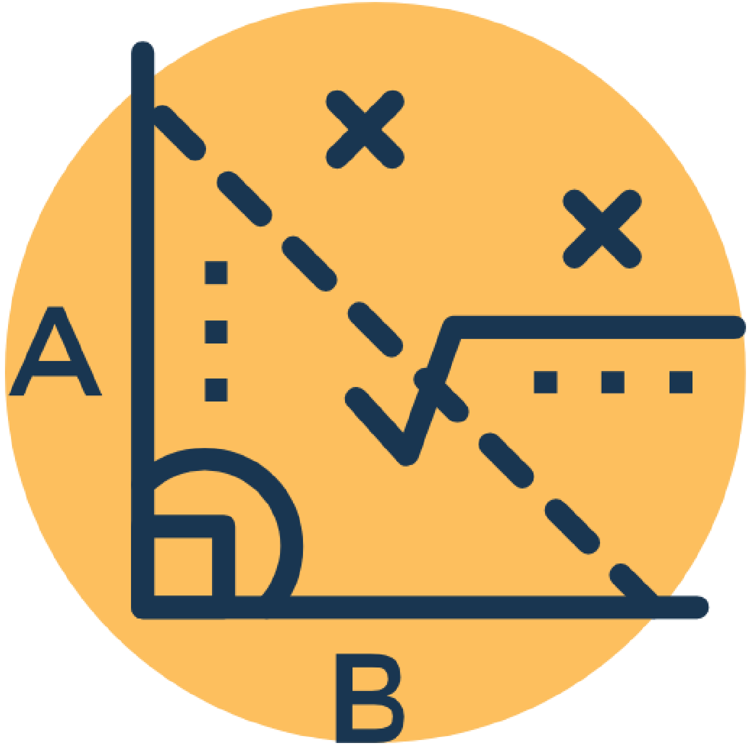
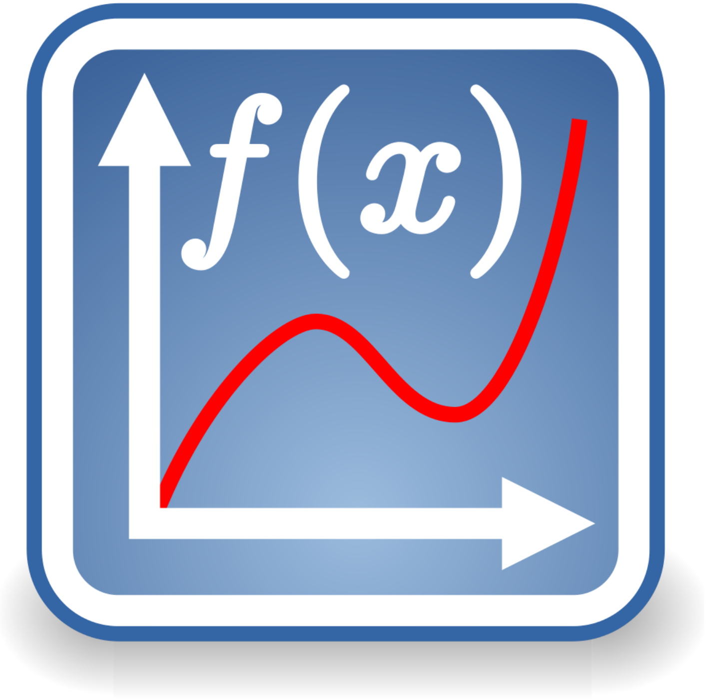




Logistic Regression

An introduction to the fundamentals

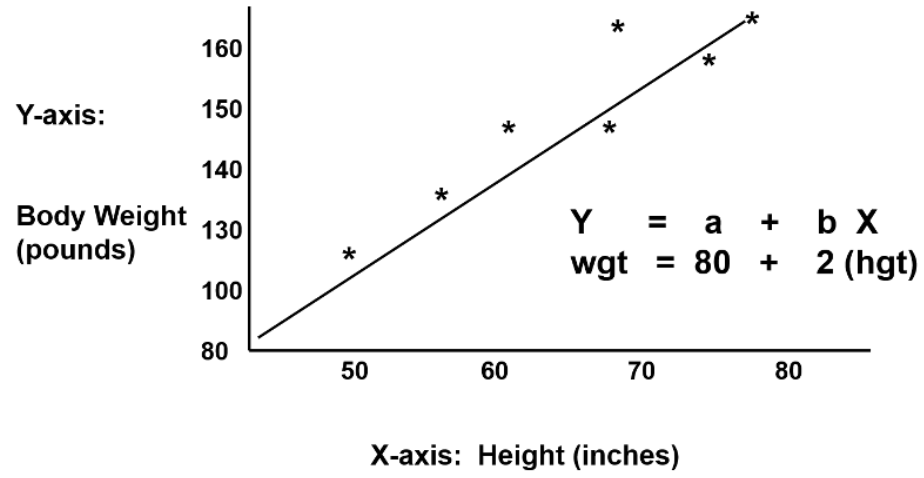
Presentation Outline

- Review of **Linear Regression.**
- **When to use** Logistic Regression?
- **Understanding the output** of a Logistic Regression.

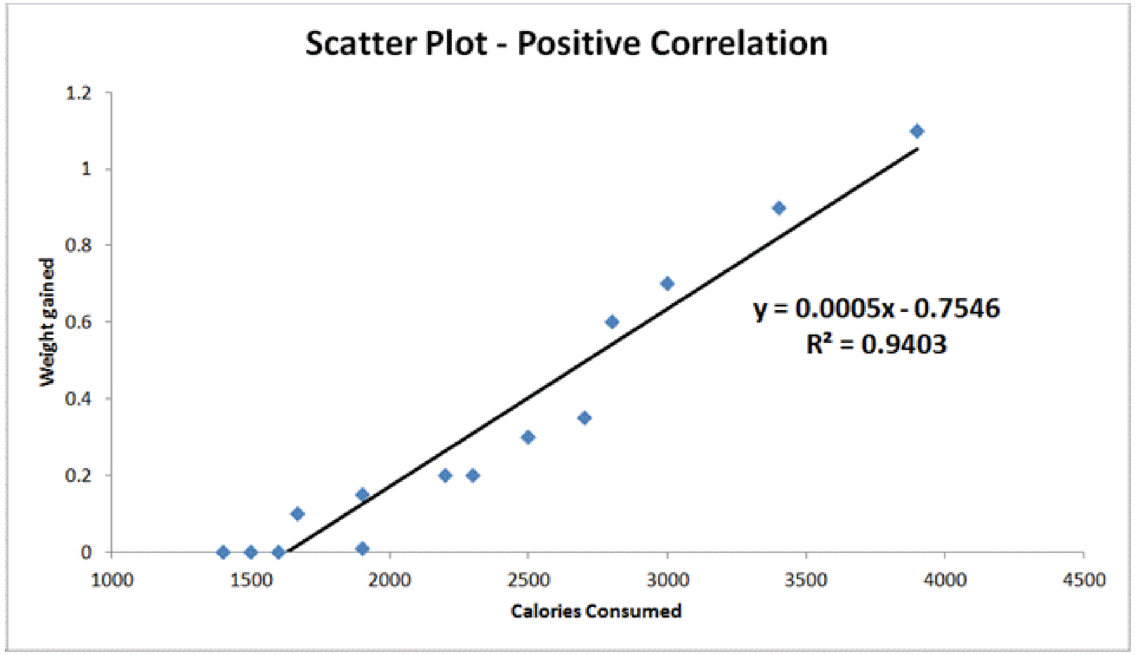
Review of Linear Regression

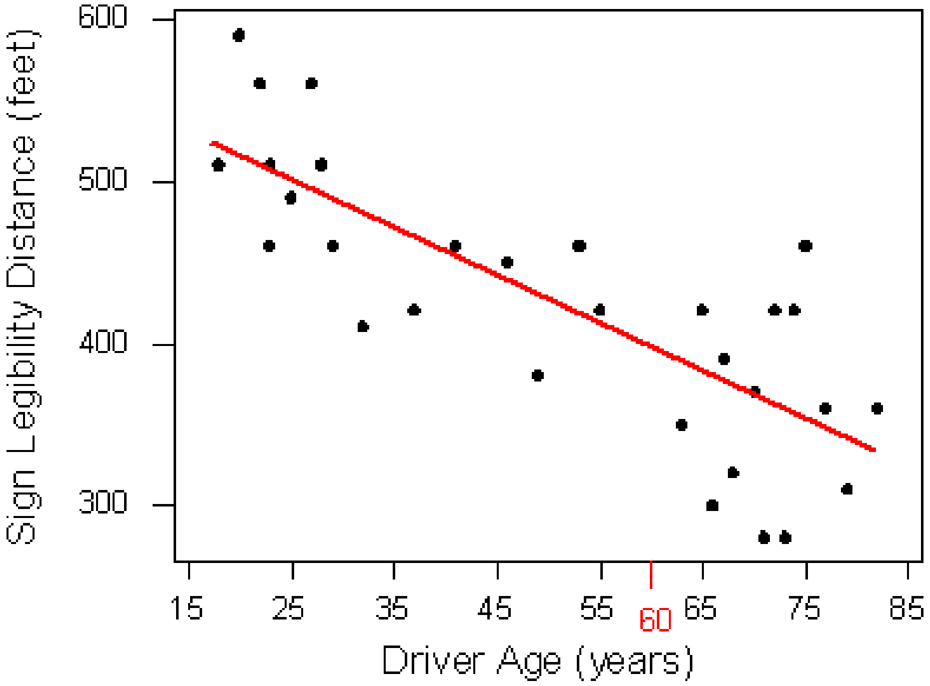
Linear Regression: a review

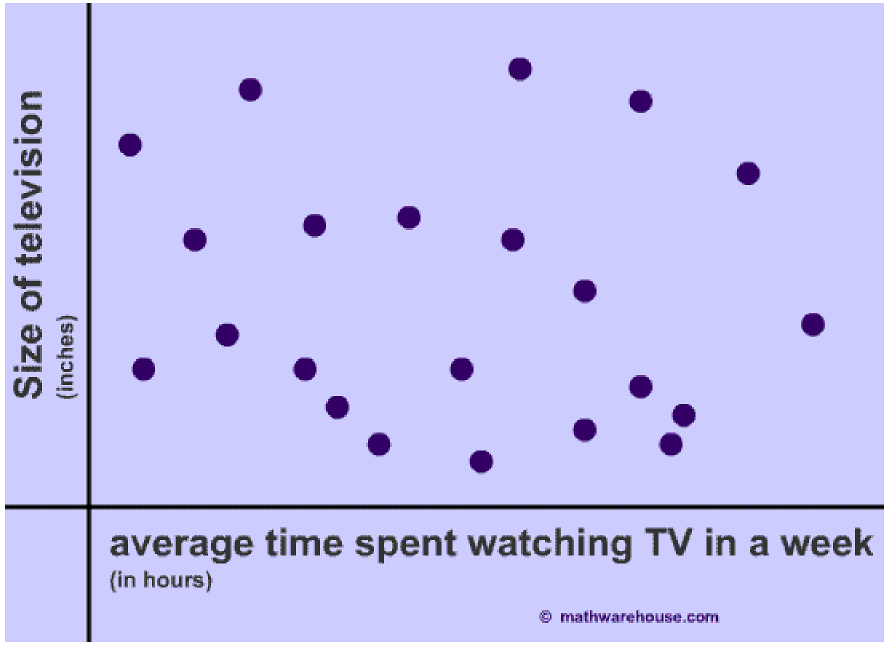
- A statistical technique used to understand **the relationship** between one set of variables upon another.
- This relationship is expressed in the form of a plotted line (hence “linear”), usually showing **positive or negative** relationships.
- At times, some relationships can be **neutral**, or simply just not exist.



Examples of Linear Relationships





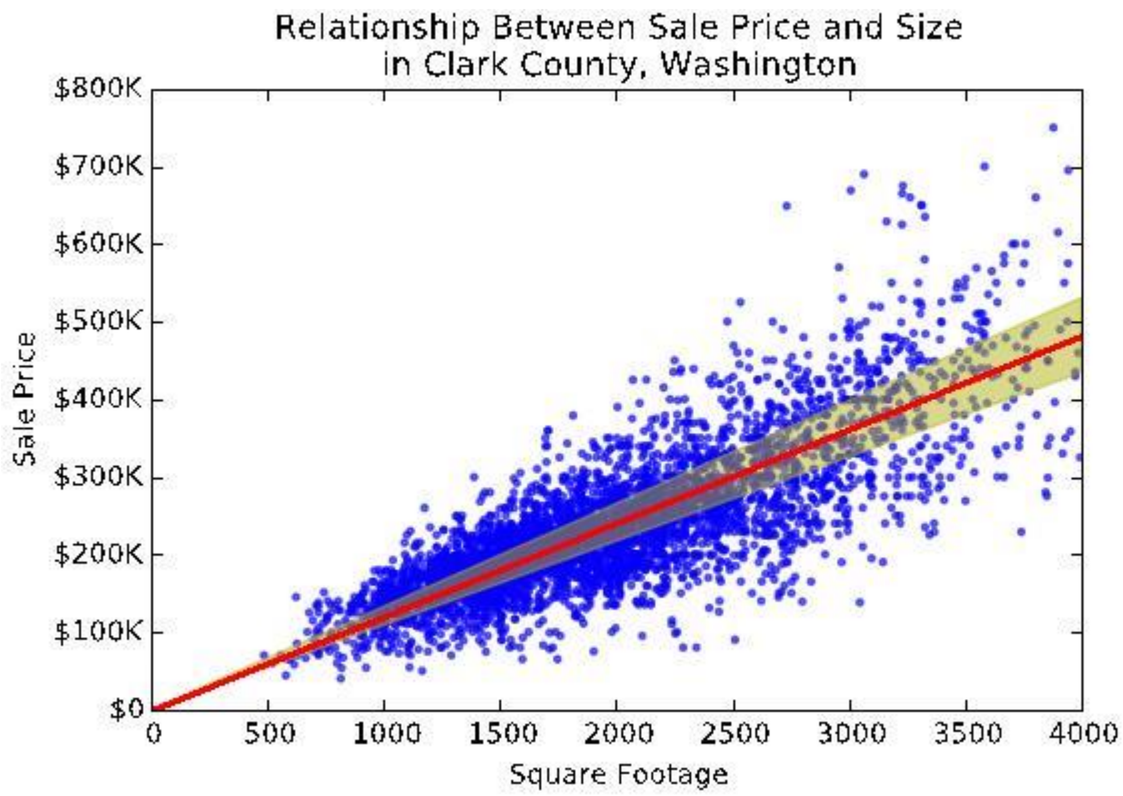


Why these Linear Relationships matter

- Understanding the relationships between variables can lead to the discovery of **valuable insights**.
  - e.g., **Predictive Pricing:** Square Footage vs. Price
  - e.g., **Sales Forecasting:** Temperature & Ice Cream Sales.

- These insights can **drive business decisions**.

- Statistics, through the technique of linear regression, can help us understand:
  - **Whether a relationship exists in the first place.**
  - **The extent to which the variables affect one another.**



Linear Regression: a review (cont'd)

- The variables that comprise a linear relationship are known as: **dependent** and **independent.**

  Y = **Dependent variable**.
  - The variable that we are trying to understand or predict.

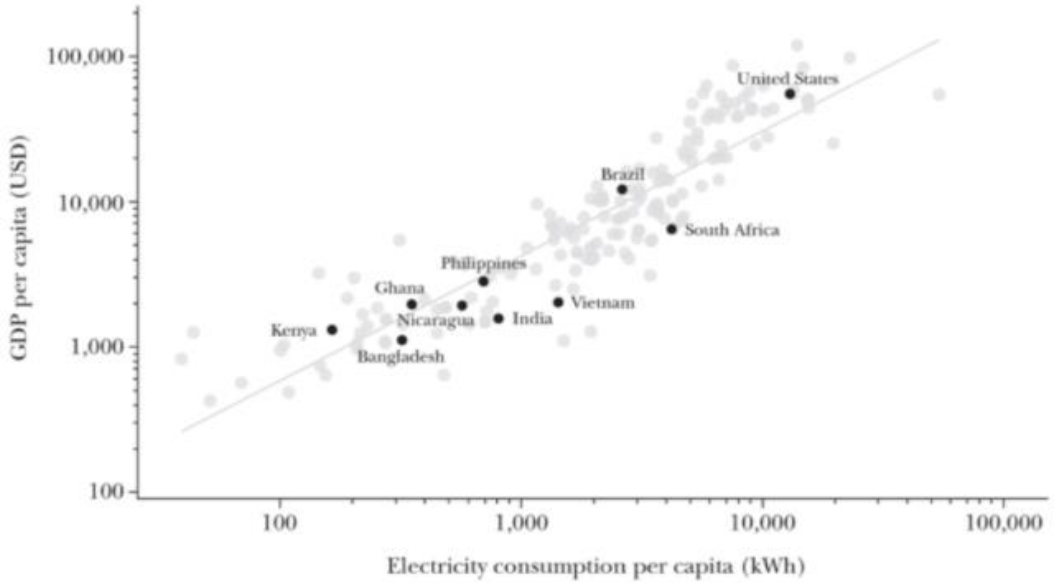
  X = **Independent variable**.
  - The variable that we suspect has an impact on Y (the dependent variable).

- The line that attempts to connect Y and X, is called the Regression Line, **or Line of Best Fit.**

- This Line of Best Fit is essentially **the best explanation of the relationship** between the Dependent and Independent variables.

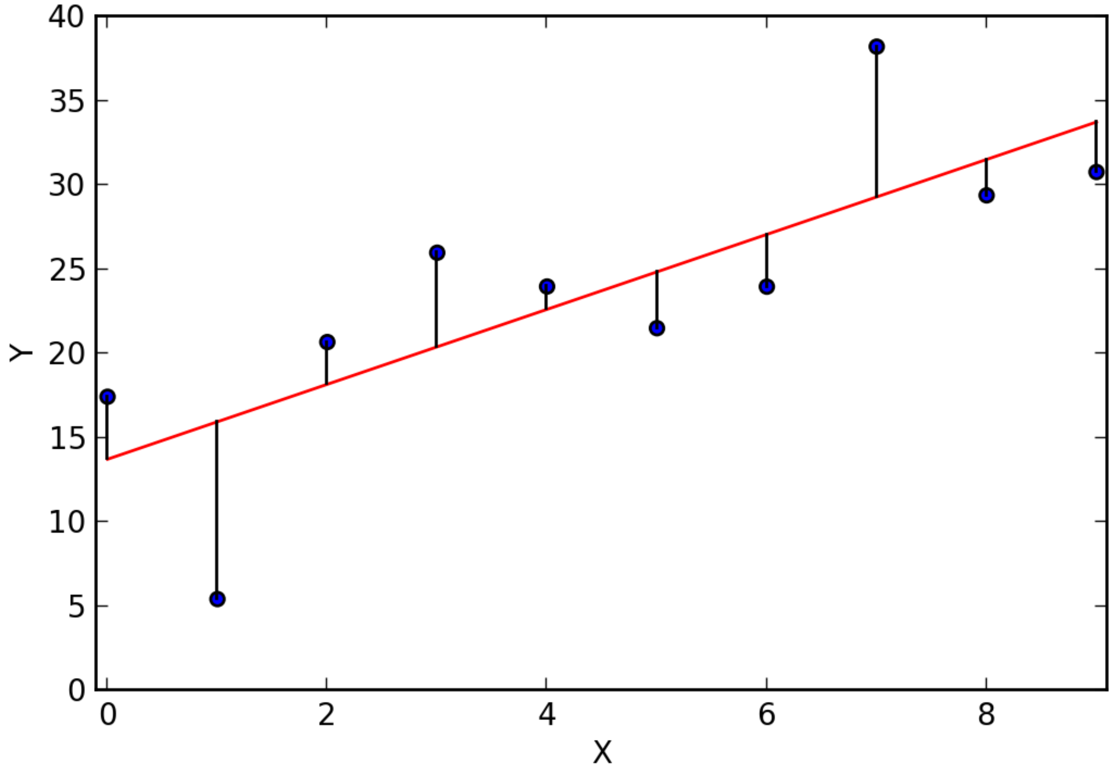
- The major task of Linear Regression is to find this Line of Best Fit!

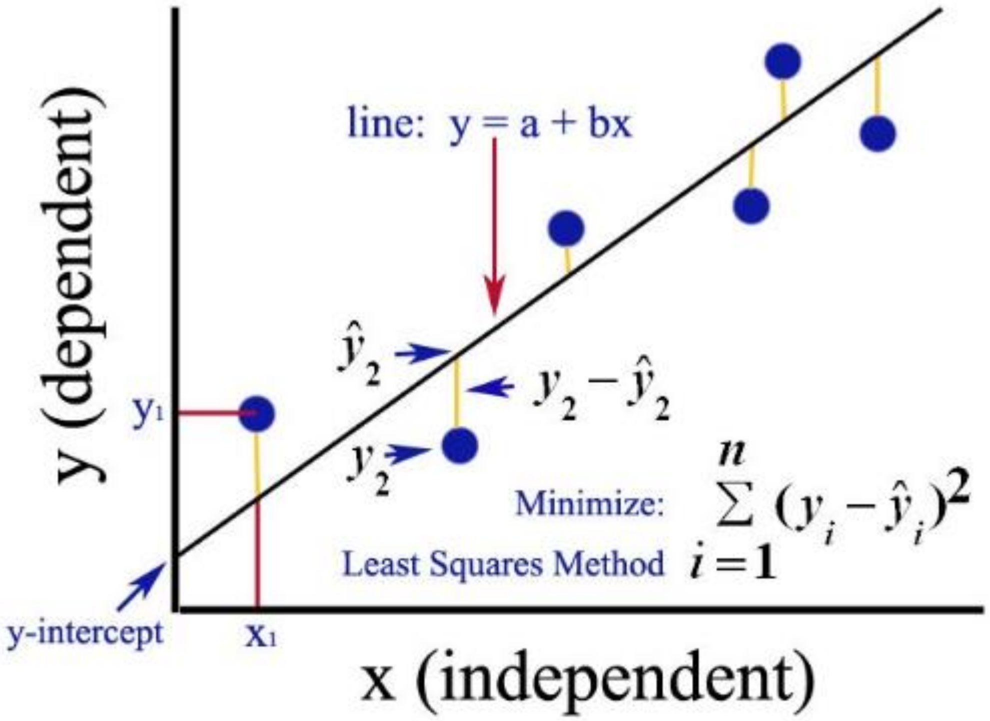
Figure 1
The Positive Correlation between Electricity Consumption and GDP per Capita



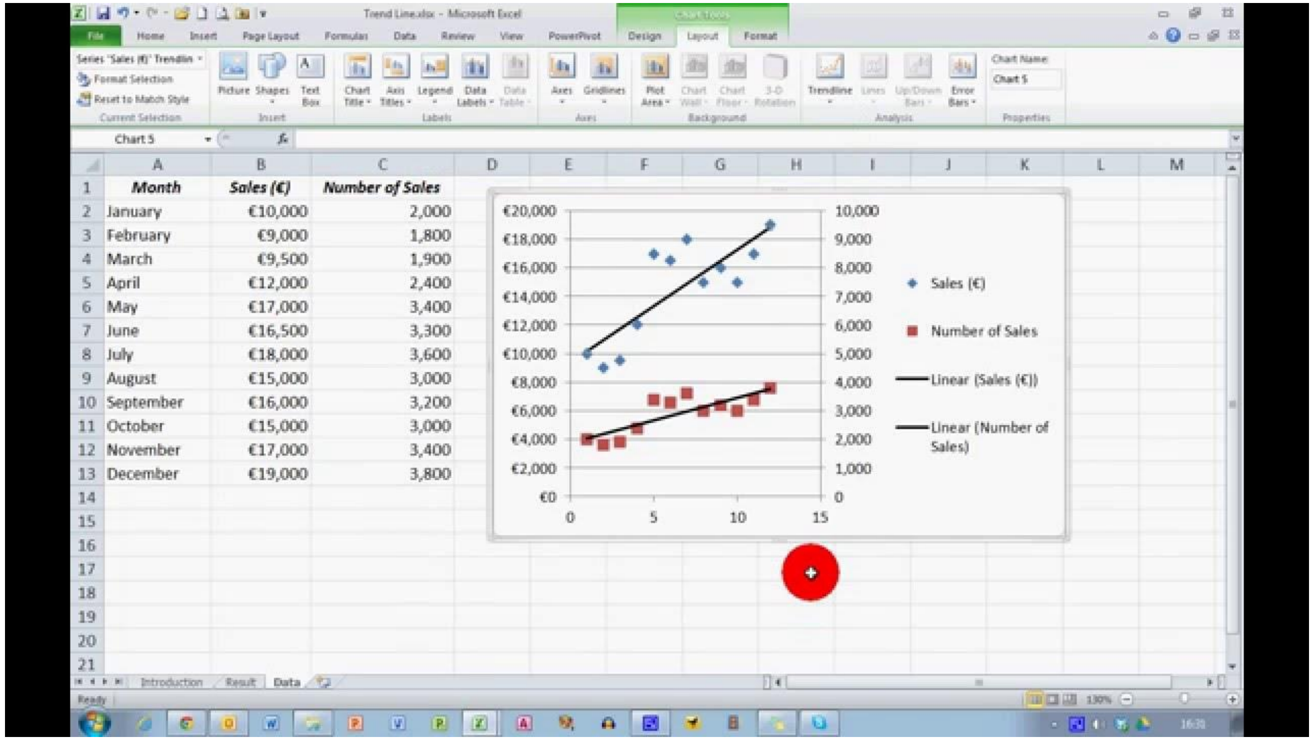
Source: 2014 data obtained from the World Bank DataBank.
Note: Both variables are presented on a logarithmic scale. GDP per capita data are in current US dollars.

Ordinary Least Squares (optional!)





Statistical Software : Microsoft Excel



19							
20	ANOVA						
21		*df*	*SS*	*MS*	*F*	*Significance F*	
22	Regression	2	9694299.568	4847149.784	50.269	0.001	
23	Residual	4	385700.432	96425.108			
24	Total	6	10080000.000				
25							
26		*Coefficients*	*Std Error*	*t Stat*	*P-values*	*Lower 95%*	*Upper 95%*
27	Intercept	8536.214	386.912	22.062	0.000	7461.975	9610.453
28	Price	-835.722	99.653	-8.386	0.001	-1112.404	-559.041
29	Advertising	0.592	0.104	5.676	0.005	0.303	0.882
30							
31							
32							
33							

Statistical Software : Python

In [23]:
```
results.summary()
```

Out[23]:

OLS Regression Results

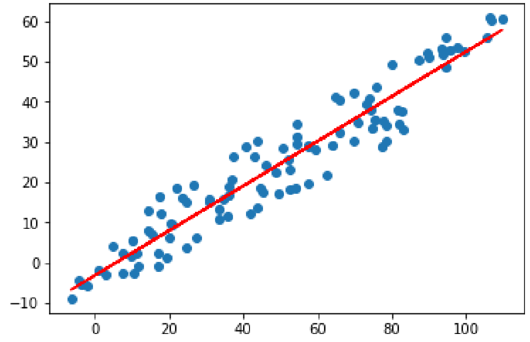
Dep. Variable:	medv	**R-squared:**	0.544
Model:	OLS	**Adj. R-squared:**	0.543
Method:	Least Squares	**F-statistic:**	601.6
Date:	Tue, 28 Jan 2020	**Prob (F-statistic):**	5.08e-88
Time:	22:35:45	**Log-Likelihood:**	-1641.5
No. Observations:	506	**AIC:**	3287.
Df Residuals:	504	**BIC:**	3295.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	34.5538	0.563	61.415	0.000	33.448	35.659
x1	-0.9500	0.039	-24.528	0.000	-1.026	-0.874

Omnibus:	137.043	**Durbin-Watson:**	0.892
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	291.373
Skew:	1.453	**Prob(JB):**	5.36e-64
Kurtosis:	5.319	**Cond. No.**	29.7

In [7]:
```
plt.scatter(X,y)
plt.plot(X,y_pred,'r')
```

Out[7]: [<matplotlib.lines.Line2D at 0x1869675b748>]



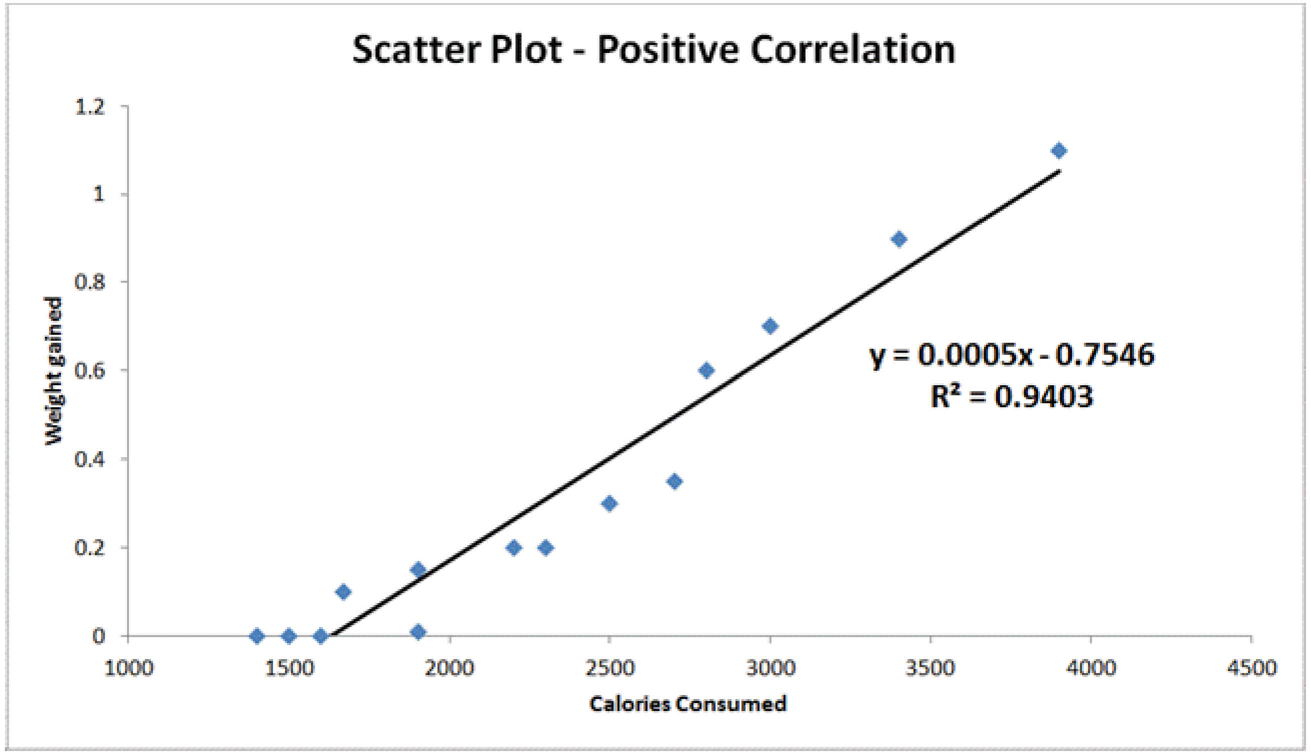
In [ ]:

Our two main goals:

Using Linear Regression (and eventually Logistic Regression) to understand:

1. **Whether a relationship exists in the first place.**
   - **R^2 (Coefficient of Determination)**

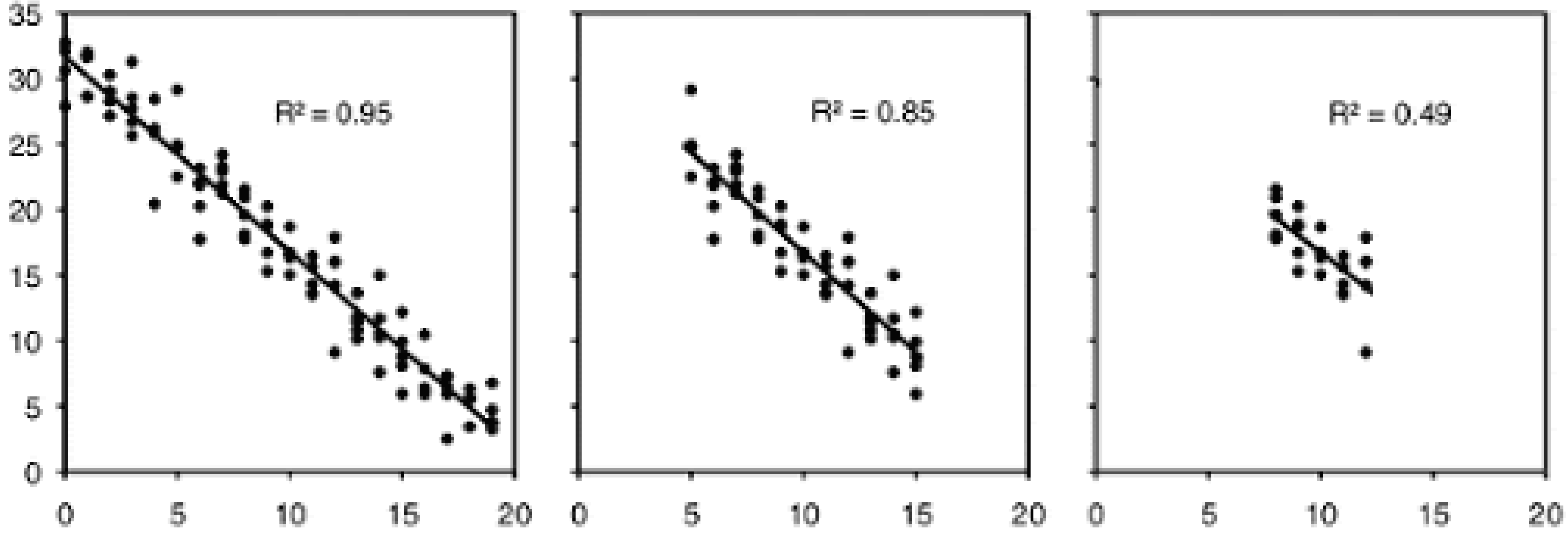
2. **The extent to which the X variables affect Y.**
   - **Regression Coefficients (Slopes).**



R Squared Formula

R Squared Formula = r^2

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$



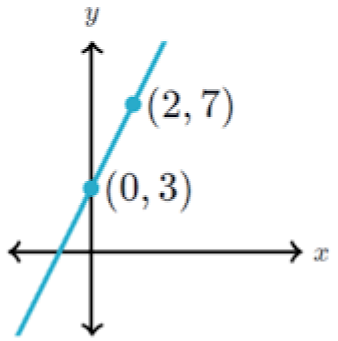
Linear Regression: a review (cont'd)

- The linear relationship between two sets of variables carries the following **mathematical form**:

$$y = mx + b$$

- It can be read as: **given a certain value for the variable "x", what is the corresponding value for "y"?**

$$y = 2x + 3$$



Sales = 2(degrees) + 3

$x = 0$

Sales = 2(0 degrees) + 3
= 2 + 3
= 5 ice creams sold

$x = 2$

Sales = 2(2 degrees) + 3
= 4 + 3
= 7 ice creams sold

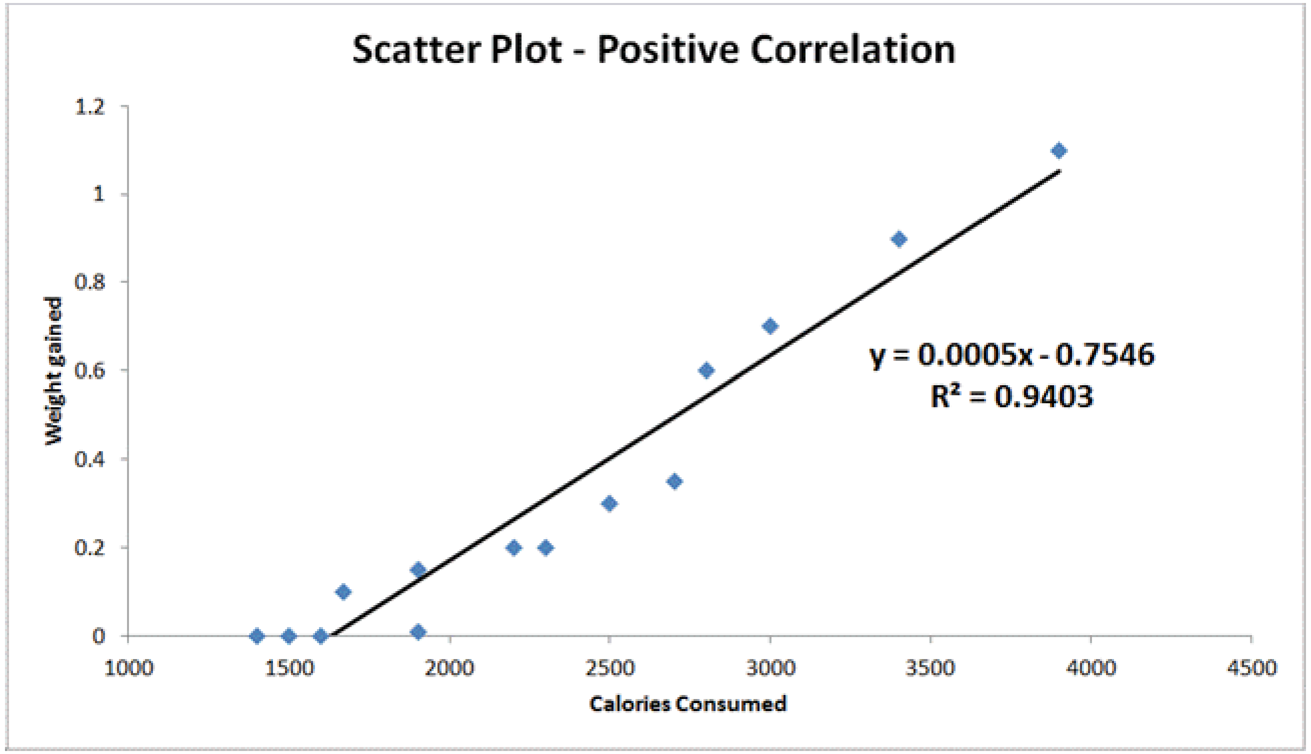
With each unit increase in degrees, we expect Sales to increase by a multiplicative factor of 2.

Recap of our two main goals:

Using Linear Regression (and eventually Logistic Regression) to understand:

1. **Whether a relationship exists in the first place.**
   - **R^2 (Coefficient of Determination)**

2. **The extent to which the X variables affect Y.**
   - **Regression Coefficients (Slopes).**
   - **P-values.**



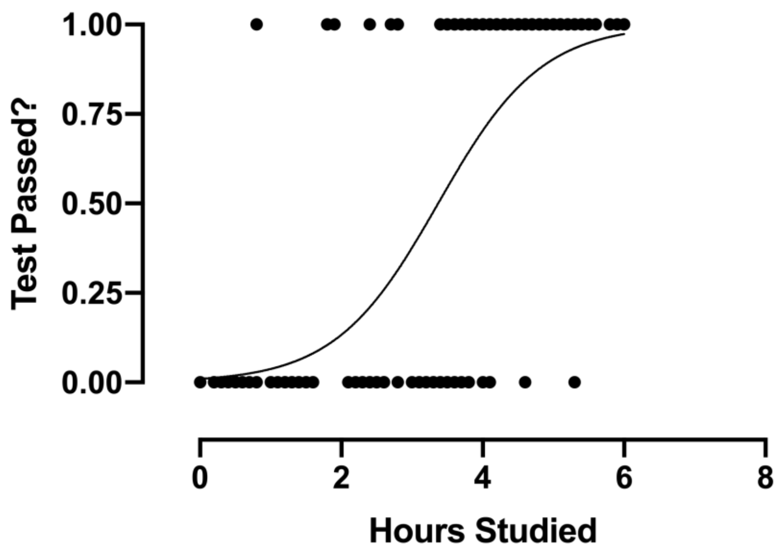
How does this now apply to <u>Logistic Regression</u>?

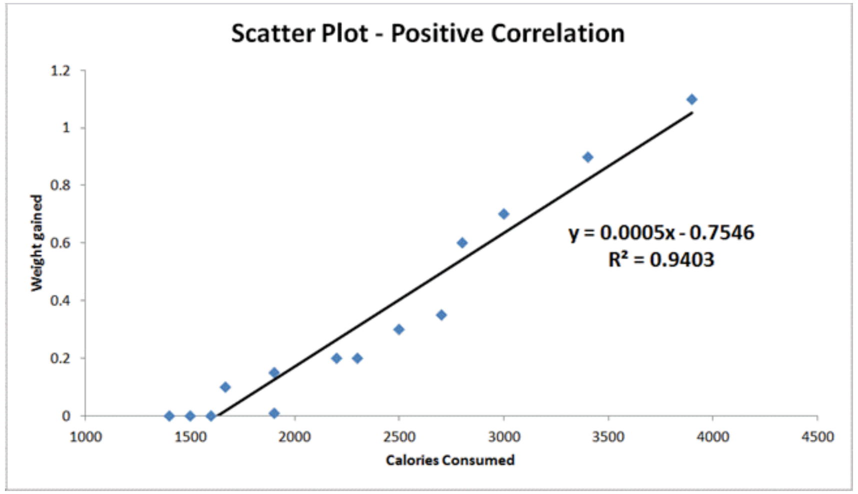
When to use Logistic Regression?

When to use Logistic Regression?

- Logistic Regression is used when our dependent variable, Y, is a **categorical variable** with two distinct categories (often yes/no or success/failure).
    - Did someone respond to our advertisement?
    - Did the patient respond to the vaccine?

- This is compared to Linear Regression, which has Y in the form of a **continuous variable**.
    - How does square footage relate to the sale price of a home?
    - How does calories consume relate to the amount of weight gained?

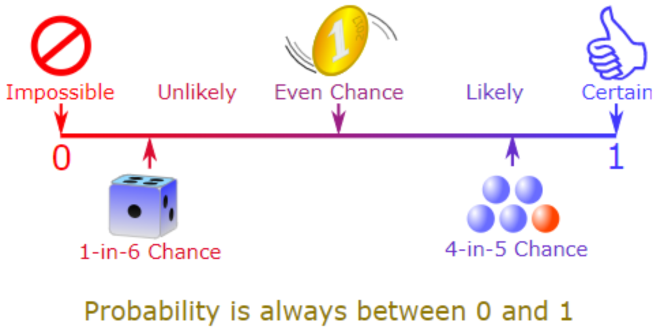


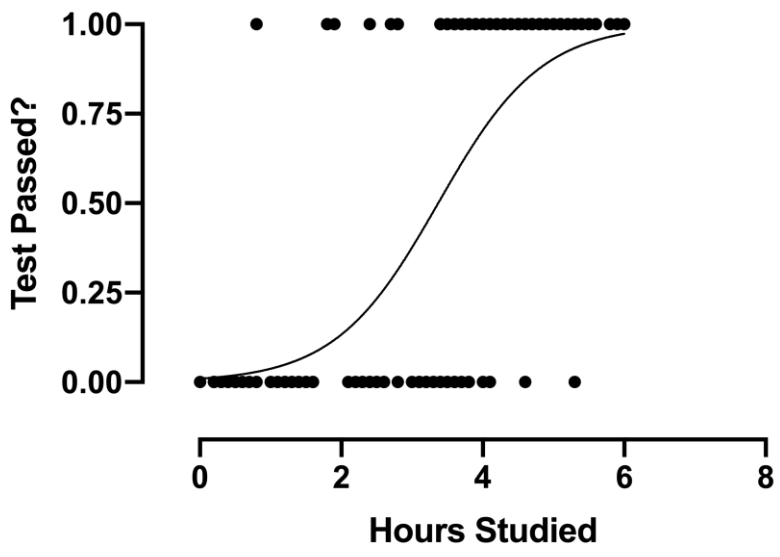


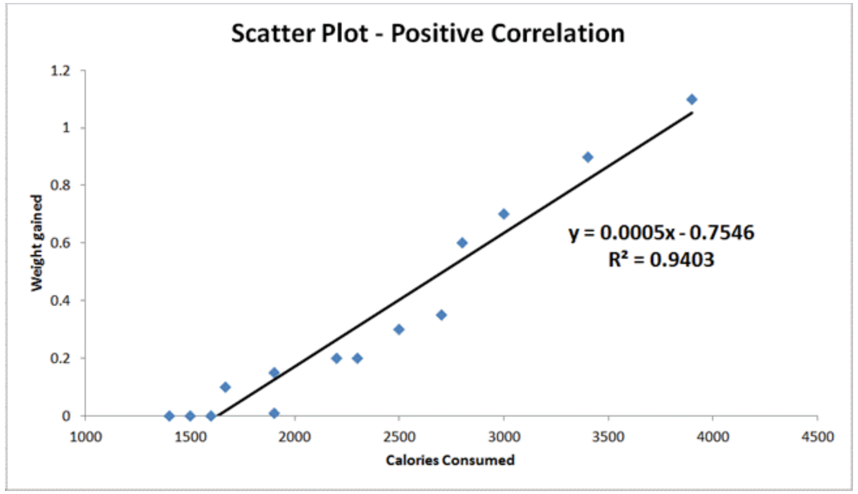
When to use Logistic Regression?

- In other words, we use Logistic Regression when we are interested in **understanding the probability/odds of success** (Y) given a set of variables (X).
  - Did someone respond to our advertisement?
    - Y = 50 Successes, 950 Failures. Odds of success = 50/950 = 0.05
    - X = previous customer status, amount spent before, age, gender, etc.
  - Did the patient respond to the vaccine?
    - Y = 5 Successes, 1995 Failures. Odds of success = 5/1995 = 0. 0025
    - X = age, gender, occupation, etc.

- Logistic Regression lends itself to questions of probability because it allows for calculations within one of probability's fundamental rules: it cannot be below 0% and cannot be above 100%.

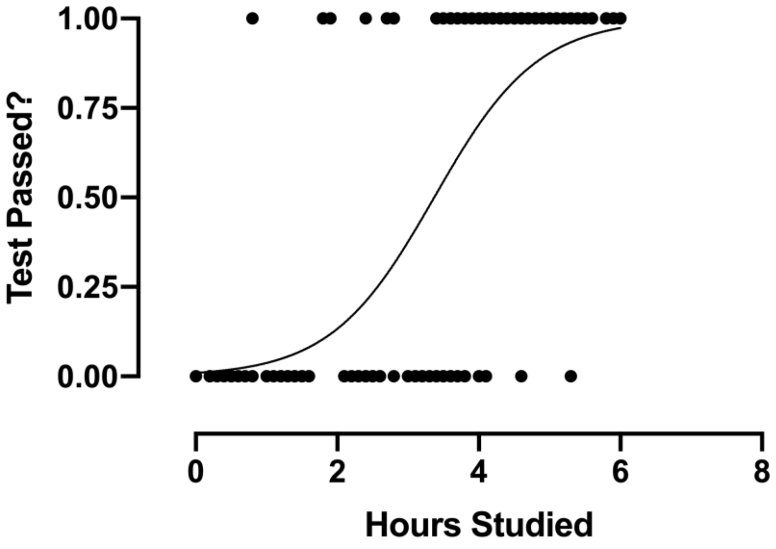


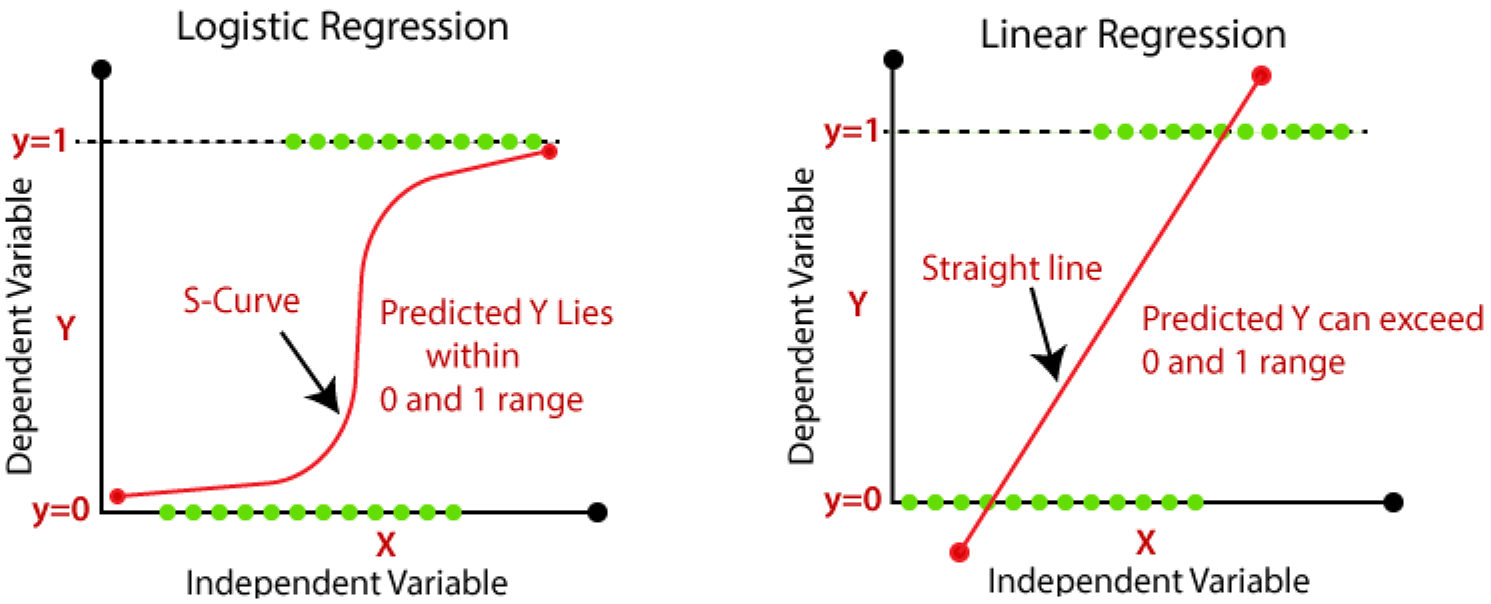




How does Logistic Regression suit itself to Probability?

- Essentially, straight lines are easier to work with than the curvy lines presented by probability data.
- Logistic Regression **helps us transform the curvy lines back into straight lines**.
- It accomplishes this by taking the **log of the odds of success** at each point along X.



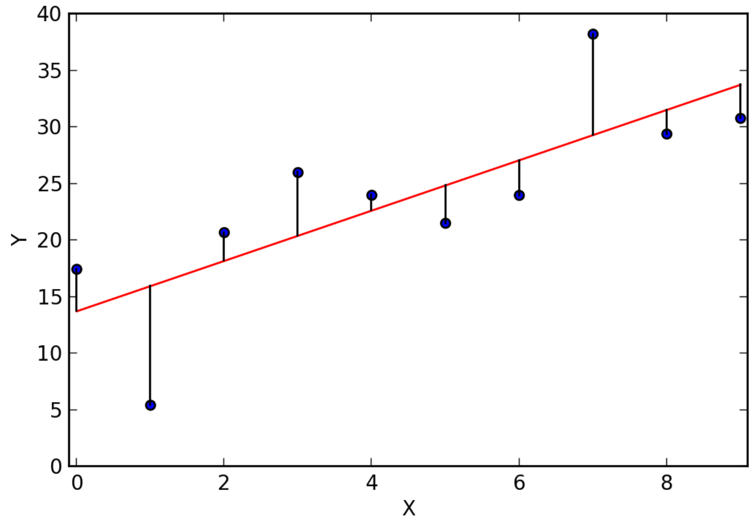


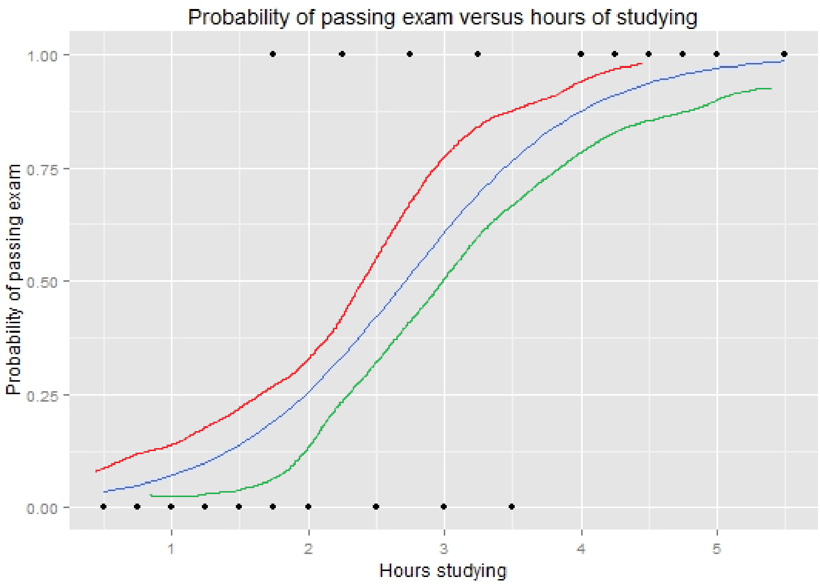
A Reminder of our goals

Using Linear Regression (and eventually Logistic Regression) to understand:

1. **Whether a relationship exists in the first place.**
   - **R^2 (Coefficient of Determination).**
   - **Pseudo R^2 (amongst others) for Logistic.**

2. **The extent to which the X variables affect Y.**
   - **Regression Coefficients (Slopes).**
   - **Exponent of Regression Coefficients for Logistic.**
   - **P-values + P-values**





Understanding the output of a Logistic Regression

Link to my Python Notebook hosted on my Github

$$\hat{y} = b + m \cdot x$$

$$\underset{cm}{\widehat{\text{height}}} = \underset{cm}{\underline{50.75}} + \underset{\frac{cm}{\cancel{mm}}}{0.9741} \cdot \underset{\cancel{mm}}{(\text{femur})}$$

Slope:

$m = 0.9741 \frac{cm}{mm}$ our model predicts that each additional mm of femur length is associated w/ an additional .9741 cm of height.

Y-Intercept:

A note on natural logs

Source: https://stats.stackexchange.com/questions/27682/what-is-the-reason-why-we-use-natural-logarithm-ln-rather-than-log-to-base-10

“We prefer natural logs (that is, logarithms base e) because, as described above, coefficients on the natural-log scale are directly interpretable as approximate proportional differences: with a coefficient of 0.06, a difference of 1 in x corresponds to an approximate 6% difference in y, and so forth.”

E= 2.71828

Formula >

$$e^{\ln x} = x$$

$\ln$ = natural logarithm

e = natural exponent

x = real number