This document highlights the important areas in Linear Regression and Logistic Regression.

Linear Regression:

Linear regression is a statistical model used for finding linear relationship between target and one or more predictors. There are two types of linear regression - Simple and Multiple.

When in your linear regression model you have many independent variables, this will be called multiple linear regression. And when you have only one independent variable in your linear regression model, the model will be called simple linear regression. Simple linear regression is a special case of multiple linear regression.

The following is the linear regression equation

$y_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3+\ldots\ldots} + \beta_k x_k + \varepsilon_i$ for $i = 1,2, \ldots, n$
n is number of observations

In vector and matrix notation
$y = X\beta + \varepsilon$

where,
y is $n \times 1$ vector of dependent variable
X is $n \times k$ design matrix of k independent variables
β is $k \times 1$ vector of unknown slope parameters
ε is $n \times 1$ error vector,
errors are independent normal with zero mean and constant variance
OLS estimates of parameters
$\hat{\beta} = (X'X)^{-1}X'y$

Some Useful Statistics in Linear Regression

R-square and Adjusted R-square. Both are used to measure goodness of model fit. These values usually lie between 0 and 1. Value 1 indicates independent variables perfectly explains the dependent variable and value 0 indicates the opposite that is independent variables are not able to explain dependent variables at all. Higher values indicate better fit. The difference between R-square and adjusted R-square is that, adjusted R-square is penalized by number of independent variables. By increasing the number of independent variables, we can achieve better fit for training data. But the model performs poorly later on test data. This is called overfitting of the model. To avoid such overfitting, R-square value is adjusted or penalized by number of independent variables in the model, and it is considered as a better measure of goodness of fit. This is called adjusted R-square.

p-values of slope parameters. Low p-value indicates the parameter is close to zero. If the parameter is zero or very close to zero this means the associated predictor does not have any impact on target variable. So, the presence of such predictor in the model is meaningless.

Worked-out Example:

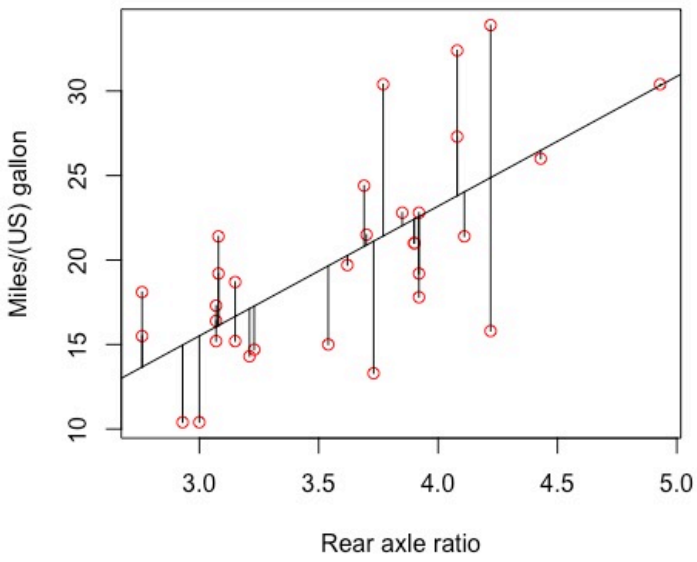
Launching a prototype car is very expensive. While testing a prototype car, if the car manufacturing company finds that the mileage is very low then it will be a waste of money to build this car. That is why the car company wants to estimate mileage (mpg) of the prototype car in advance. For that, the company collects mileages of existing cars and few important features of the car such as number of cylinders (cil), displacement (disp), rear axle ratio (rar), bhp etc. The multiple regression model is as follows.

$$mpg = a + b_1*cil + b_2*disp + b_3*rar + b_4*bhp + error$$

Let's consider, a simple version of the above model where we have only one independent variable which is rear axle ratio (rar).

$$mpg = a + b*rar + error$$

Where a is the intercept parameter and b is slope parameter associated with 'rar' independent variable. This is an example of simple linear regression analysis. Because this model has only one independent variable.



Using the available data, the company estimates the parameters and gets expected value of mpg, which is expressed as follows.

$$E(mpg) = -7.5 + 6.5*rar$$

Now from this model, the company can find expected mileage of the prototype car, by inputting the value of independent variable 'rar' in the above equation. If 4 is the value of 'rar' the expected mileage is – 7.5 + 6.5 * 4 = – 7.5 + 26 = 18.5. If the company changes 1 unit in 'rar' then expected mileage will be changed by 6.5 mpg.

Logistic Regression

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is binary (dichotomous). For example, to predict whether an email is

a spam (1) or not (0) or whether the new movie will be a hit (1) or not (0). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data to explain the relationship between one dependent binary variable and independent variables.

We should not use linear regression when dependent variable is binary. Under linear regression, estimated output can be anything even less than zero or greater than 1. So, it will be very difficult to interpret the result. In linear regression, we mentioned that the error is normal distribution with zero mean and constant variance. This assumption is also violated if dependent variable takes only 0 and 1. So, instead of applying multiple linear regression on dependent variable, the same is applied to log of odds.

Odds of an event happening is defined as the likelihood that an event will occur, expressed as a proportion of the likelihood that the event will not occur. Therefore, if p is the probability of an email to be spam and q is the probability

of the email is not a spam, then odds = p/q, which is eventually p/(1-p), since p lies between 0 and 1, log of odds is unbounded and varies from – infinity to infinity. Odds > 1 implies increased occurrence of the event (i.e. spam). Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$\log\left(\frac{Prob(y=1)}{1-Prob(y=1)}\right) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$$

Or,

$$Prob(y=1) = \frac{1}{1+e^{-(\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_p x_p)}}$$

Goodness of Fit/ Model Validation

R-square: We can compute the usual R-square or adjusted R-square statistic, although it is rarely used in logistic

regression. It is almost always low, since observed values need to be either 0 or 1, but predicted values are always in between these extremes. R-square is not a suitable measure for goodness of fit when dependent variable is binary or categorical.

Overall accuracy: One can start by fitting a model and calculating all fitted values. Then, one can choose a cut-off value on the probability scale, say 50%, and classify all predicted values above that as predicting an event, and all below that cut-off value as not predicting the event. Now, we construct a two-by-two table of data as follows:

	Actual	
Predict	y = 1	y =0
y = 1	a	b
y = 0	c	d

Values 'a' and 'd' are the number of observations that classified correctly. So, misclassification rate is: $1 - \frac{a+d}{a+b+c+d}$, lower values imply better fit.

This metric is used for model validation for all classification models in statistics or machine learning.

Misclassification rate or Overall accuracy (1 - misclassification rate) is a better measure of goodness of fit (or model validation) than R-Square.

Example:

Consider the following simple dataset where we have only one explanatory variable 'weight' that says about weights (kg) of cat and dog and target variable that indicates cat (0) and dog (1). Let us apply logistic regression model to predict the target variable based on this dataset.

Obs	weight	target
1	6.18	1
2	3.76	0
3	7.86	1
4	2.04	0
5	8.97	1
6	5.96	1
7	9.46	1
8	4.25	0
9	4.2	0
10	3.59	0

The corresponding logistic regression model is

$$\log\left(\frac{Prob(y = \text{'dog'})}{1 - Prob(y = \text{'dog'})}\right) = \text{a} + \text{b*weight}$$

The estimated values of parameters from the above dataset are,
a = -9.0
b = 1.4

The following table shows how to compute probabilities and prediction.

weight	target	a+b*weight	probability 1/(e(-(a+b*weight)+1)	prediction if(probability<0.5,0,1)	iscorrect
6.18	1	-0.348	0.4138675	0	0
3.76	0	-3.736	0.02329377	0	1
7.86	1	2.004	0.881216413	1	1
2.04	0	-6.144	0.002141724	0	1
8.97	1	3.558	0.972293751	1	1
5.96	1	-0.656	0.341638728	0	0

9.46	1	4.244	0.985852935	1	1
4.25	0	-3.05	0.045217473	0	1
4.2	0	-3.12	0.042289772	0	1
3.59	0	-3.974	0.018451242	0	1

From the above table it is clear that 8 out of 10 predictions are correct. So, the model has 80% accuracy.