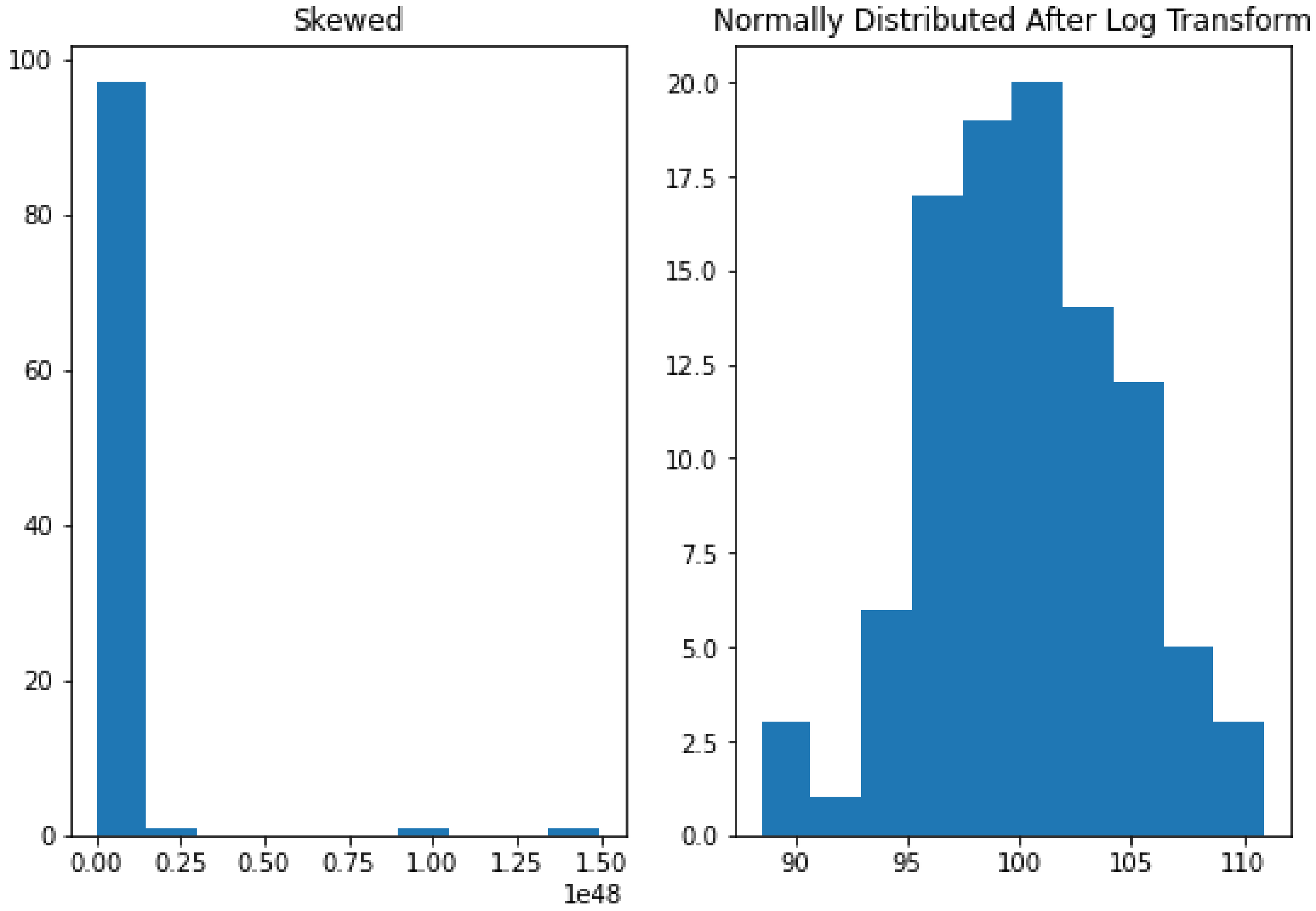
Logistic regression
Part 2: Probability and odds

By: Noureddin Sadawi, PhD
University of London

Data transformation

- Sometimes a variable in its original form does not satisfy some requirements (i.e. normality).
- A transformed form of this variable might actually do!

The idea is to transform a variable to another scale, perform some operations and then return to the original scale to interpret the results.



Skewed
100
80
60
40
20
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e48
Normally Distributed After Log Transform
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
90
95
100
105
110

Natural logarithm

- ln or $\log_e$ transformation.
- It is used in logistic regression.
- The two expressions are equivalent (i.e. can be used interchangeably).
- It must be distinguished from $\log_{10}$ which is a different transformation.

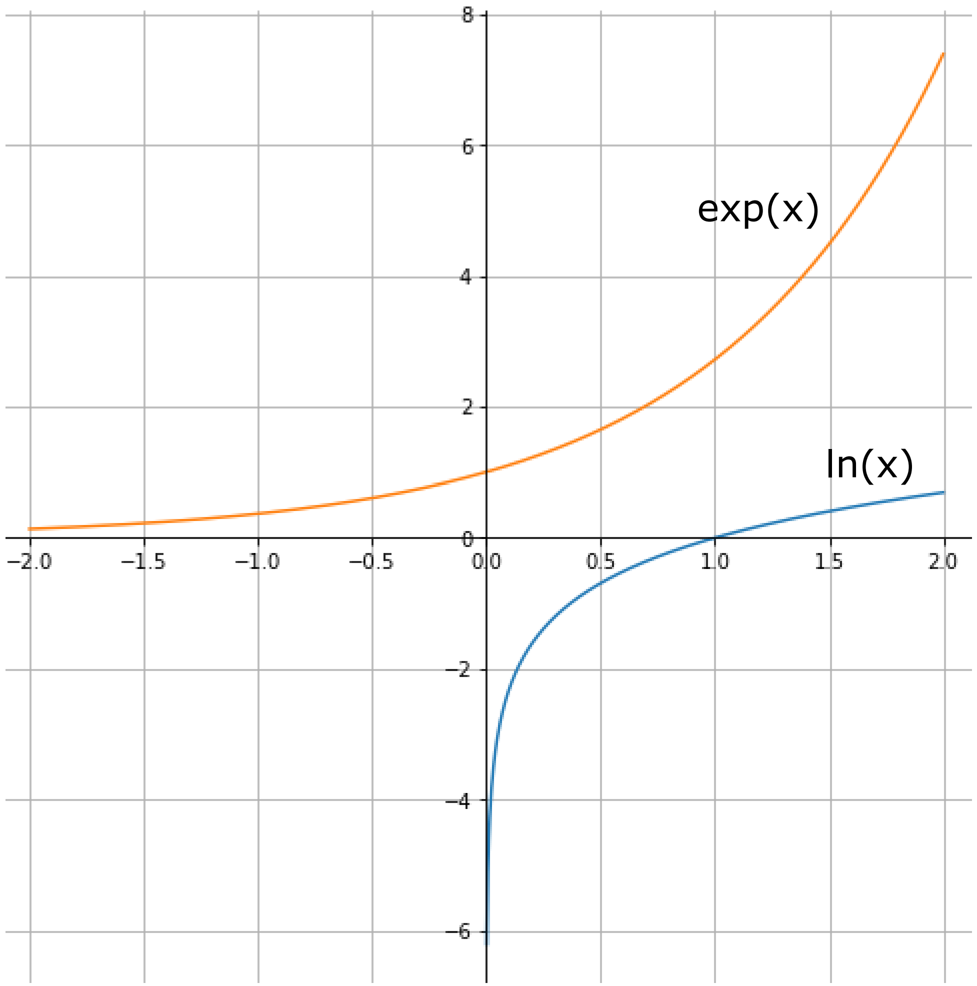
Natural logarithm

- When you take the natural log of a value, it changes to something new.
- If the original value needs to be retrieved then the we apply exponential function:

$$\boxed{x = exp(ln(x))}$$

Natural logarithm and exp

- The natural logarithm for a negative value is not defined.
- Values < 1 have -ve ln values.
- Values > 1 have +ve ln values.
- Exponentiated value never below zero.
- -ve values exp values<1.
- +ve values exp values>1.



Odds

Odds: probability of an event happening (p) divided by the probability of the event not happening (1-p).

Ratio of probabilities:

- Odds=1 the event is as likely to occur as not to occur.
- Odds>1 the event is more likely to occur than not to occur.
- Odds<1 the event is more likely not to occur than to occur.

Odds example

An auctioneer has 60% probability of winning an auction and 40% of losing it.

What are the odds of the auctioneer winning the auction?

- 60/40=1.5
  50% more likely to win than to lose.
- or 40/60=0.67 (1 - 0.67=0.33)
  33% less likely to lose than to win.

Odds example 1/3

	No Covid-19	Covid-19	Total
Smoker	187 (62%) [187/304]	117 (38%) [117/304]	304
Non-smoker	192 (69%) [192/278]	86 (31%) [86/278]	278
Total	379	203	582

Assume the above study was conducted.

Is Covid-19 more prevalent in people who smoke?

From the results table:

- 38% of smokers have Covid-19.
- 31% of non-smokers have Covid-19.

Odds example 2/3

	No Covid-19	Covid-19	Total
Smoker	187 (62%) [187/304]	117 (38%) [117/304]	304
Non-smoker	192 (69%) [192/278]	86 (31%) [86/278]	278
Total	379	203	582

Based on this study:

- What are the odds of being Covid-19 positive given non-smoker?
- The probability of a non-smoker being Covid-19 positive is 0.45 the probability of them being Covid-19 negative [31/69].
- Non-smokers are 55% less likely to be **Covid-19 positive** than not.
- The probability of non-smokers being **Covid-19 negative** is (1/0.45)=2.22 times the probability that they are **Covid-19 positive.**

Odds example 3/3

	No Covid-19	Covid-19	Total
Smoker	187 (62%) [187/304]	117 (38%) [117/304]	304
Non Smoker	192 (69%) [192/278]	86 (31%) [86/278]	278
Total	379	203	582

Based on this study:

- What are the odds of being Covid-19 positive given smoker?
- The probability of a smoker being Covid-19 positive is 0.61 the probability of them being Covid-19 negative [38/62].
- Smokers are 39% less likely to be **Covid-19 positive** than not.
- The probability of smokers being **Covid-19 negative** is (1/0.61)=1.64 times the probability that they are **Covid-19 positive.**

Odds ratio

Odds ratio: odds of an event in one group divided by the odds of the event in another group.

Ratio of probabilities:

- Odds=1 the event is as likely to occur as not to occur.
- Odds>1 the event is more likely to occur than not to occur.
- Odds<1 the event is more likely not to occur than to occur.

Odds ratio example 1

	No Covid-19	Covid-19	Total
Smoker	187 (62%) [187/304]	117 (38%) [117/304]	304
Non-smoker	192 (69%) [192/278]	86 (31%) [86/278]	278
Total	379	203	582

- Odds ratio = odds of Covid-19 positive given non-smoker/odds of Covid-19 positive given smoker = (86/192) / (117/187) = 0.72.
- The odds of becoming Covid-19 positive for non-smoker are 0.71 the odds for smoker.
- The odds of becoming Covid-19 positive for smoker are 1 / odds ratio.
- 1 / 0.72 = about 1.39 times as large as the odds for non-smoker.
- An increase of about 39%.

Odds ratio example 2

Example case control study:

- Odds of survival in the treatment group: 6/5.
- Odds of survival in the control group: 4/7.
- Odds ratio: (4/7)/(6/5) = 0.48.
- The odds of surviving in the control group are **less than half** the odds of surviving in the treatment group.
- Or, an individual in the treatment group has odds about 2.1 as high [(6/5)/(4/7) = 2.1] of surviving than individuals from the control group.

Odds ratio

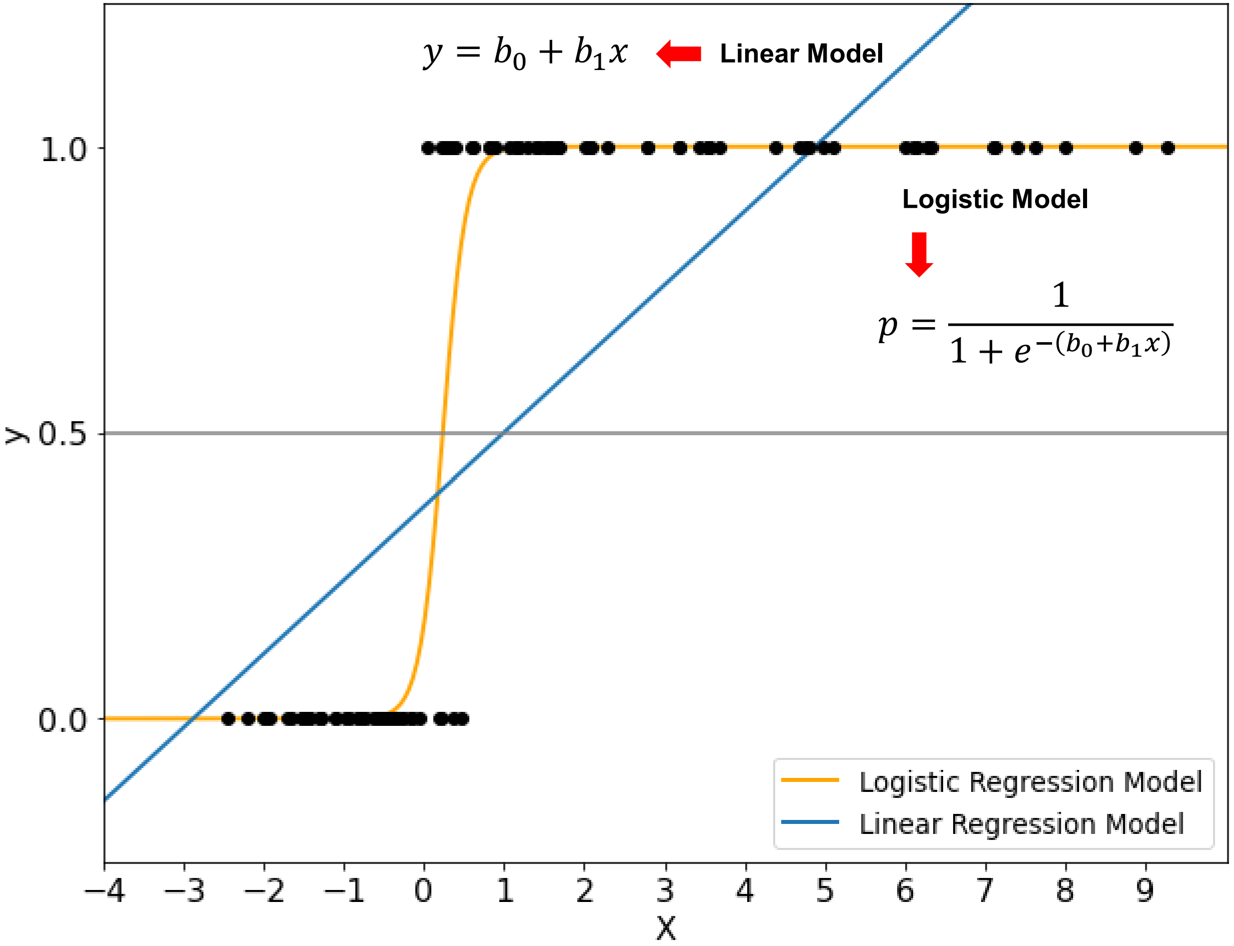
- Odds ratio < 1: odds of success in the first group are lower than in the second group.
- The closer the odds ratio to 0, the lower the odds of the first group to the second.
- Odds ratio = 1: the odds of both groups are the same.
- Odds ratio > 1: odds of the first group are higher than the second group.
- The higher the odds ratio, the higher the odds of the first group to the second.

Logistic regression

- Mathematical modelling approach that can be used to describe the relationship between several input variables (i.e. predictors) and a binary outcome.
- In more detail, logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy).
  - The prediction is based on the use of one or several predictors (numerical and categorical).

Logistic regression

- Probability values are always in the range 0 to 1 and linear regression predicts values outside it.
- As the outcome can only have one of two possible values for each data point, the residuals will not be normally distributed around the predicted line.
- Logistic regression produces a logistic curve, which is limited to values between 0 and 1.
- The curve is constructed using the natural logarithm of the 'odds' of the target variable.



Logistic regression

- Combination of natural log transformations and odds ratios.
- Remember: Outcome variable can have **only one of two** values – {0,1}.
- The main idea is to model the probability of being in one of the two categories.
- LOGIT transformation = Natural log of the odds.
- Logit makes probabilities into odds.
- **If *p* is the probability then *p/(1-p)* is the odds.**
- **The natural logarithm of the odds is the logit of the probability.**

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

Logistic regression

- The logistic regression equation can be written in terms of an odds ratio.
- The constant (b_0) moves the curve right and left and the slope (b_1) defines the steepness of the curve.
- The equation can be written in terms of log-odds (logit) which is a linear function of the predictors.
- The coefficient (b_1) is the amount the logit (log-odds) changes with a one unit change in *x.*
- Logistic regression can handle any number of numerical and/or categorical variables.

$$\frac{p}{1-p} = \exp(b_0 + b_1 x)$$

$$ln\frac{p}{1-p} = b_0 + b_1 x$$

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

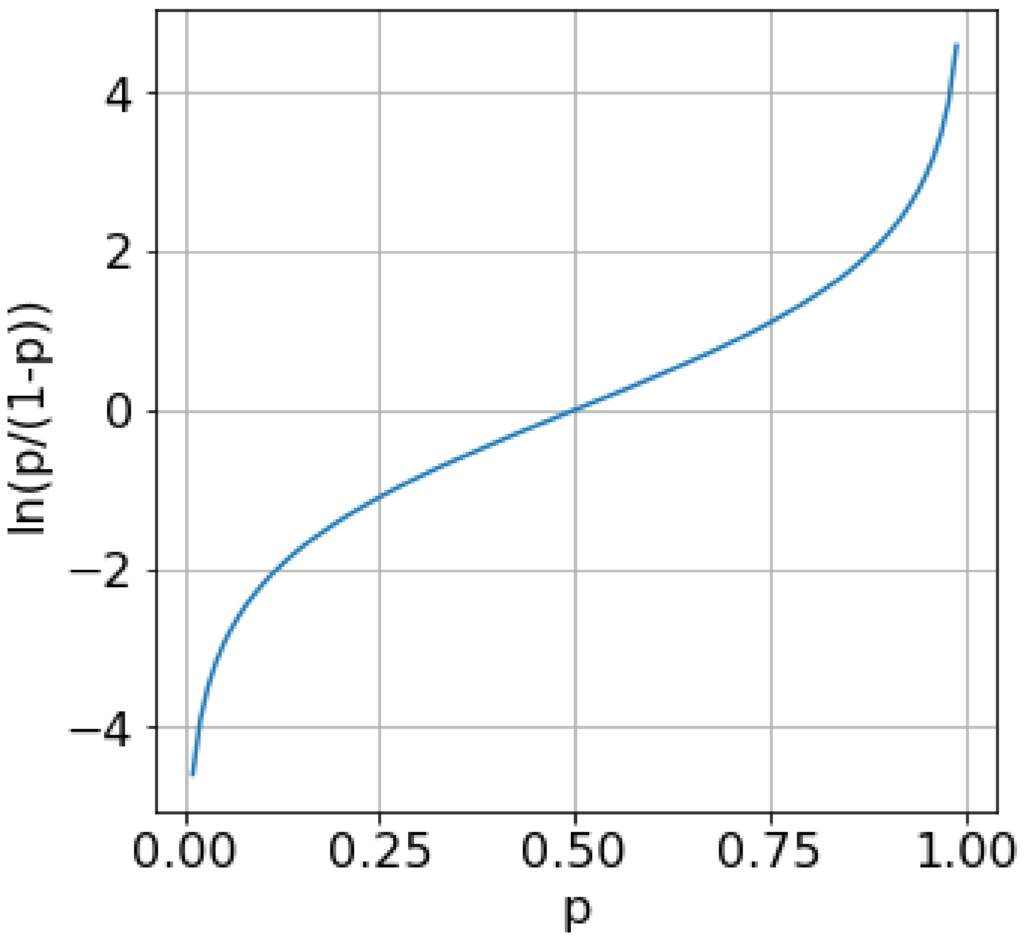
Logistic regression

- Y = f(X)
- logit (p) = f(X)
- ln (odds of event) = ln (p /(1 - p)) = f(X)
- Odds of event = exp(f(X))
- If you invert the logit transformation by exponentiating then you have the odds of the event of interest.
- p /(1 - p) = exp(f(X))
- p = exp(f(X)) / (1+ exp(f(X)))

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

The logit function

- Probability values are always in the range 0 to 1 and linear regression predicts values outside it.
- Notice the y axis is the natural log of the odds.
- The logit at probability 0.5 is 0.
- The logit at probability 0 is -∞.
- The logit at probability 0 is +∞.



Logistic regression

- Outcome variable: result of diabetes test
- logit(probability of positive diabetes test) =

  -0.268 + 0.018 * BMI.

- -0.268: logged odds of being diabetic for BMI=0 (no clinical interpretation).
- 0.018: change in the logged odds of diabetes **per unit** change in BMI.
- exp(0.018)=1.018: for each unit change in BMI the odds of becoming diabetic increase by approximately 1.8%.
- For a 5-unit increase in BMI, the odds of becoming diabetic increases by approx exp(0.018)=(1.018)^5=1.093 (i.e. 9.3%).

BMI	Outcome
28.1	0
43.1	1
31.0	1
30.5	1
30.1	1
…	…
43.3	1
36.5	1
28.4	0
32.9	0
26.2	0

Multivariate logistic regression

- Usually more than one input variable could be included in the model.
- Purpose: Determine which variables result in the best model within the scientific context of the problem.
- Example diabetes data:
  - number of pregnancies (Preg)
  - plasma glucose concentration (Gluc)
  - diastolic blood pressure (BP)
  - triceps skin fold thickness (ST)
  - BMI
  - diabetes pedigree function (DPF)
  - age in years.

Multivariate logistic regression

Preg	Gluc	BP	ST	BMI	DPF	Age	Outcome
1	89.0	66	23	28.1	0.167	21.0	0
0	137.0	40	35	43.1	2.288	33.0	1
3	78.0	50	32	31.0	0.248	26.0	1
2	197.0	70	45	30.5	0.158	53.0	1
1	189.0	60	23	30.1	0.398	59.0	1
…	…	…	…	…	…	…	…
0	181.0	88	44	43.3	0.222	26.0	1
1	128.0	88	39	36.5	1.057	37.0	1
2	88.0	58	26	28.4	0.766	22.0	0
10	101.0	76	48	32.9	0.171	63.0	0
5	121.0	72	23	26.2	0.245	30.0	0

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age

Multivariate logistic regression

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age.

- **Coefficients: effect on logit(p) for a unit change of one single predictor causes while keeping all the rest constant.**

For example:

- A unit change in BMI **increases** the logit of the probability of diabetes by 0.09 while keeping the other variables constant.
- A unit change in blood pressure **decreases** the logit of the probability of diabetes by 0.01 while keeping the other variables constant.

Multivariate logistic regression

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age.

- **exp(coefficient): same as before BUT in relation to the actual odds of having a positive diabetes test rather than the logit of this.**

For example:

- The odds of having a positive diabetes test go up by a factor of 1.094 by a unit change in BMI.
- The odds of the same outcome go down by a factor of 1.01 by a unit change in DBP.

Values of coefficient

- Linear regression uses least squares to estimate coefficients for the best fit line that relates input variables to the outcome.
- Logistic regression uses **maximum likelihood estimation** (MLE) to obtain the model coefficients.
- This function is initially estimated, then the process is repeated until LL (Log Likelihood) does not change significantly.

Goodness of fit

To what extent the fitted values under the model compare to the actual (i.e. observed) values.

- If the agreement between the observations and corresponding fitted values is good, the model may be acceptable.
- If not, the model is said to display ‘lack-of-fit’ and it needs to be revised.
- There are multiple diagnostic methods to measure the goodness of fit.

Variable importance

Example methods for measuring variable importance in logistic regression:

- If the input variables have the same scale, then coefficients can be used as a crude variable importance score.
- If the variables do not have the same scale, then a simple approach is to calculate variable importance as the magnitude of coefficient times the standard deviation of the corresponding variable in the data.
- The z score is also often used to determine variable importance
    - It is the regression coefficient divided by the standard error.

Model selection

Aim: find the simplest model that yields the best performance.

- Determine the smallest subset of input variables that produces the most accurate model.
- Multiple models can be created, the model with the lowest **Akaike information criterion** (**AIC**) is usually selected.
- Model building strategies:
    - forward selection
    - backward selection
    - stepwise selection.