Nicolas Stoian CS780 AI

Dataset – 10% missing, given starting parameters

Starting points –
P(G) = 0.7
P(W|G) = 0.8
P(W|-G) = 0.4
P(H|G) = 0.7
P(H|-G) = 0.3

Final conditional probability tables –

Gender

M	0.64246
F	0.35754

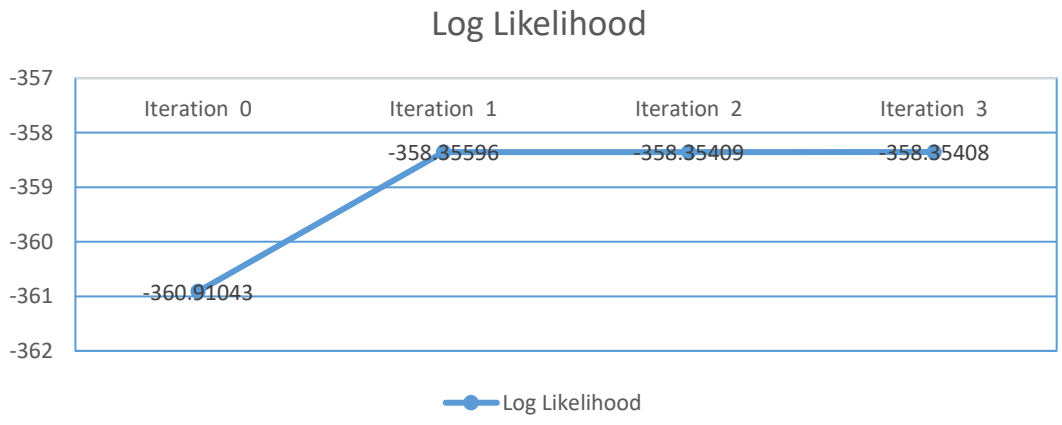
Weight | Gender

	W > 130	W < 130
M	0.79443	0.20557
F	0.34850	0.65150

Height | Gender

	H > 55	H < 55
M	0.66191	0.33809
F	0.26498	0.73502

Plot of likelihood vs iterations



Number of iterations = 3
Final log likelihood = -358.35408

Dataset – 10% missing, estimated starting parameters

Starting points –
P(G) = 0.64324
P(W|G) = 0.78991
P(W|-G) = 0.34848
P(H|G) = 0.66386
P(H|-G) = 0.27272

Final conditional probability tables –

Gender

M	0.64246
F	0.35754

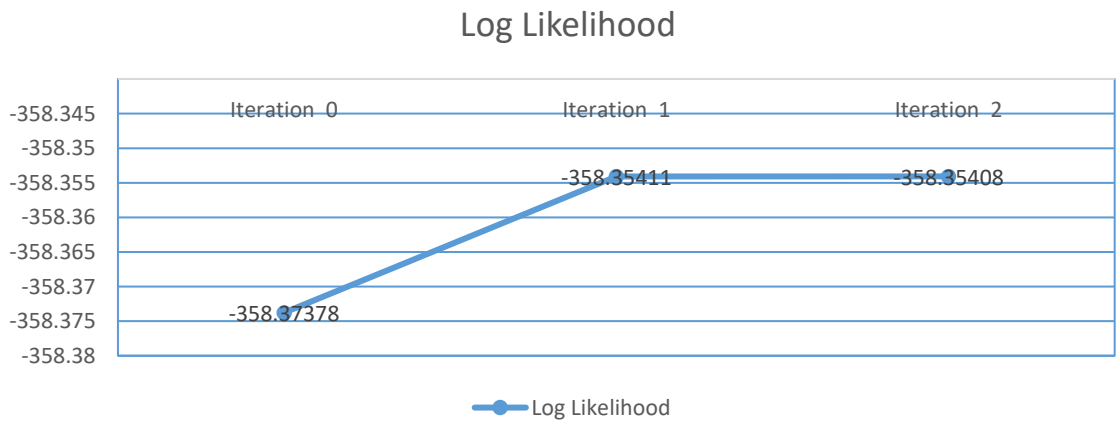
Weight | Gender

	W > 130	W < 130
M	0.79442	0.20558
F	0.34852	0.65148

Height | Gender

	H > 55	H < 55
M	0.66191	0.33809
F	0.26499	0.73501

Plot of likelihood vs iterations



Number of iterations = 2
Final log likelihood = -358.35408

Dataset – 30% missing, given starting parameters

Starting points –
P(G) = 0.7
P(W|G) = 0.8
P(W|-G) = 0.4
P(H|G) = 0.7
P(H|-G) = 0.3

Final conditional probability tables –

Gender

M	0.68187
F	0.31813

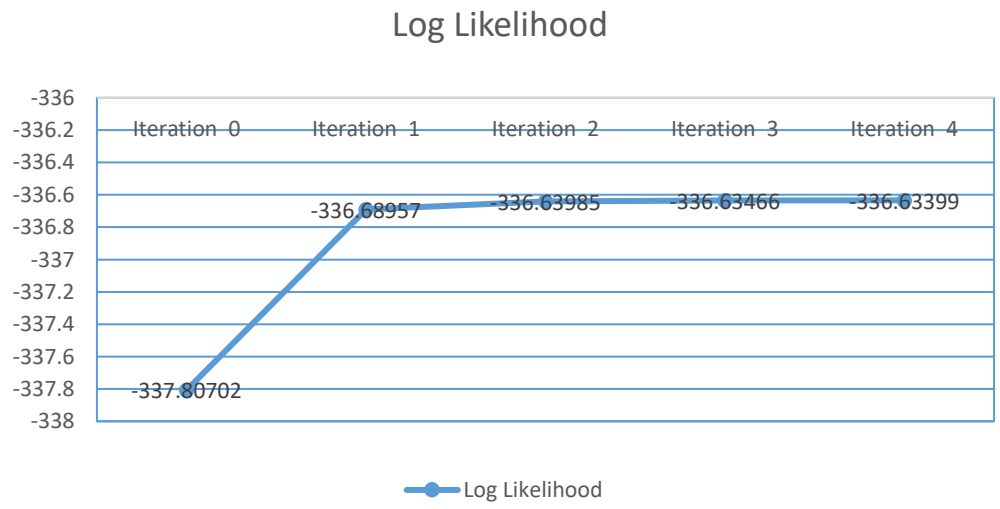
Weight | Gender

	W > 130	W < 130
M	0.78556	0.21444
F	0.37514	0.62486

Height | Gender

	H > 55	H < 55
M	0.65108	0.34892
F	0.34905	0.65095

Plot of likelihood vs iterations



Number of iterations = 4
Final log likelihood = -336.63399

Dataset – 30% missing, estimated starting parameters

Starting points –
P(G) = 0.68794
P(W|G) = 0.81443
P(W|-G) = 0.40909
P(H|G) = 0.62886
P(H|-G) = 0.31818

Final conditional probability tables –

Gender

M	0.68153
F	0.31847

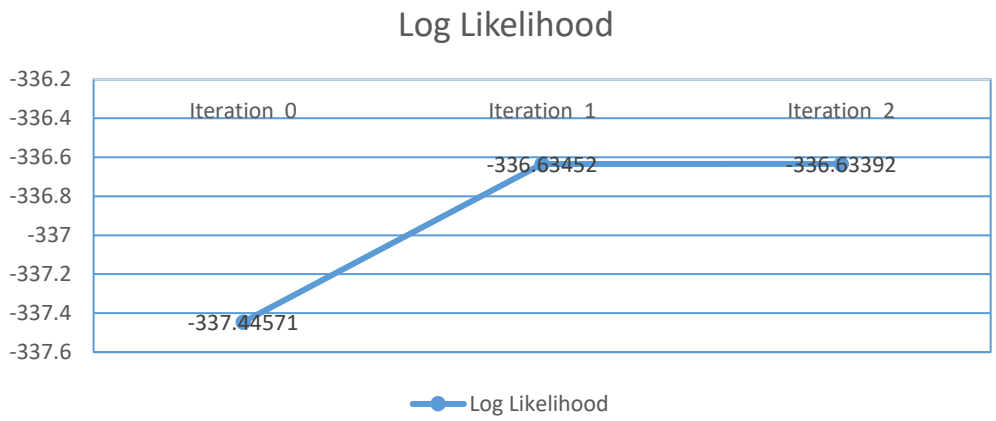
Weight | Gender

	W > 130	W < 130
M	0.78603	0.21397
F	0.37458	0.62542

Height | Gender

	H > 55	H < 55
M	0.65107	0.34893
F	0.34939	0.65061

Plot of likelihood vs iterations



Number of iterations = 2
Final log likelihood = -336.63392

Dataset – 50% missing, given starting parameters

Starting points –
P(G) = 0.7
P(W|G) = 0.8
P(W|-G) = 0.4
P(H|G) = 0.7
P(H|-G) = 0.3

Final conditional probability tables –

Gender

M	0.66299
F	0.33701

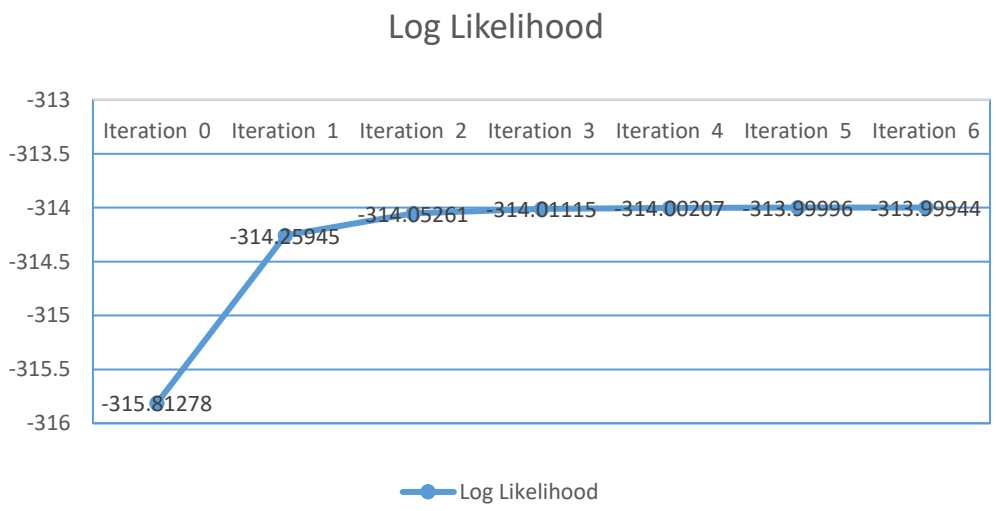
Weight | Gender

	W > 130	W < 130
M	0.74469	0.25531
F	0.47854	0.52146

Height | Gender

	H > 55	H < 55
M	0.68783	0.31217
F	0.33817	0.66183

Plot of likelihood vs iterations



Number of iterations = 6
Final log likelihood = -313.99944

Dataset – 50% missing, estimated starting parameters

Starting points –
P(G) = 0.655555
P(W|G) = 0.813559
P(W|-G) = 0.516129
P(H|G) = 0.644067
P(H|-G) = 0.258064

Final conditional probability tables –

Gender

M	0.66237
F	0.33763

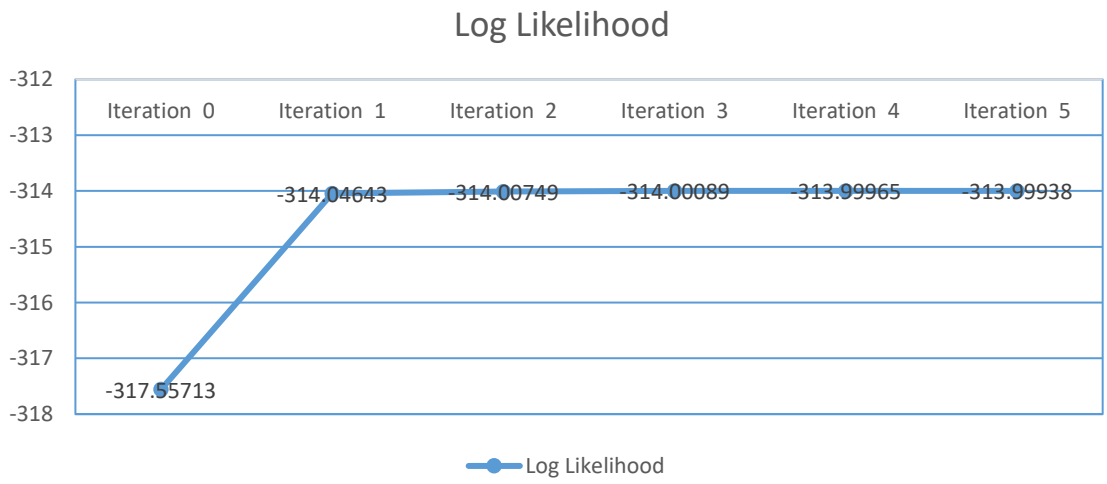
Weight | Gender

	W > 130	W < 130
M	0.74476	0.21397
F	0.47888	0.52112

Height | Gender

	H > 55	H < 55
M	0.68815	0.31185
F	0.33820	0.66180

Plot of likelihood vs iterations



Number of iterations = 5
Final log likelihood = -313.99938

Dataset – 70% missing, given starting parameters

Starting points –
P(G) = 0.7
P(W|G) = 0.8
P(W|-G) = 0.4
P(H|G) = 0.7
P(H|-G) = 0.3

Final conditional probability tables –

Gender

M	0.69084
F	0.30916

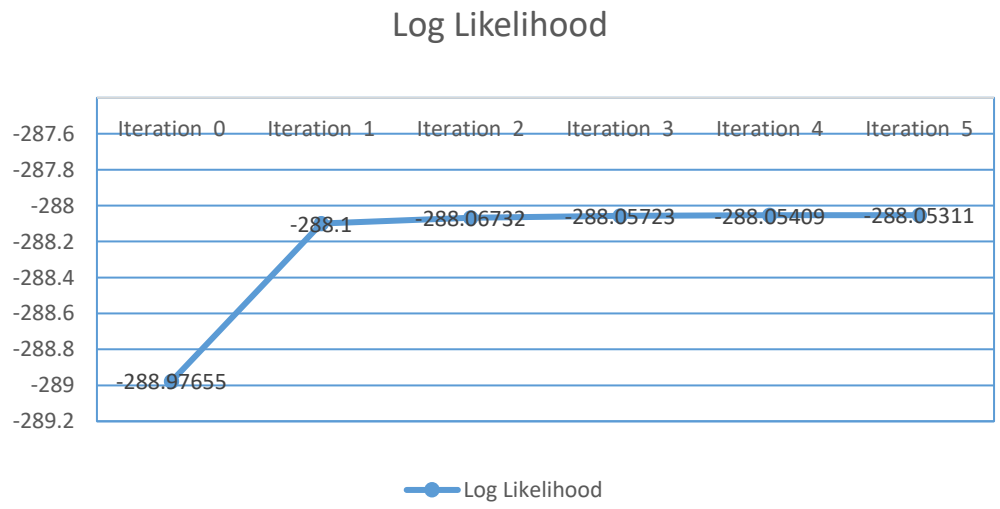
Weight | Gender

	W > 130	W < 130
M	0.82889	0.17111
F	0.42814	0.57186

Height | Gender

	H > 55	H < 55
M	0.65829	0.34171
F	0.30799	0.69201

Plot of likelihood vs iterations



Number of iterations = 5
Final log likelihood = -288.05311

Dataset – 70% missing, estimated starting parameters

Starting points –
P(G) = 0.68852
P(W|G) = 0.76190
P(W|-G) = 0.36842
P(H|G) = 0.71428
P(H|-G) = 0.42105

Final conditional probability tables –

Gender

M	0.69103
F	0.30897

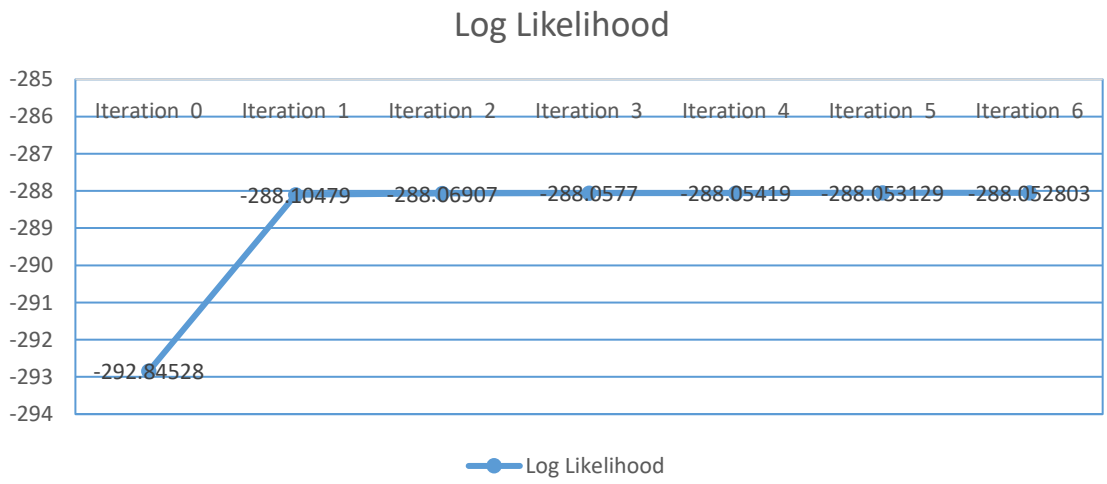
Weight | Gender

	W > 130	W < 130
M	0.82909	0.17091
F	0.42744	0.57256

Height | Gender

	H > 55	H < 55
M	0.65686	0.34314
F	0.31099	0.68901

Plot of likelihood vs iterations



Number of iterations = 6
Final log likelihood = -288.05280

Dataset – 100% missing, given starting parameters

Starting points –
P(G) = 0.7
P(W|G) = 0.8
P(W|-G) = 0.4
P(H|G) = 0.7
P(H|-G) = 0.3

Final conditional probability tables –

Gender

M	0.69703
F	0.30297

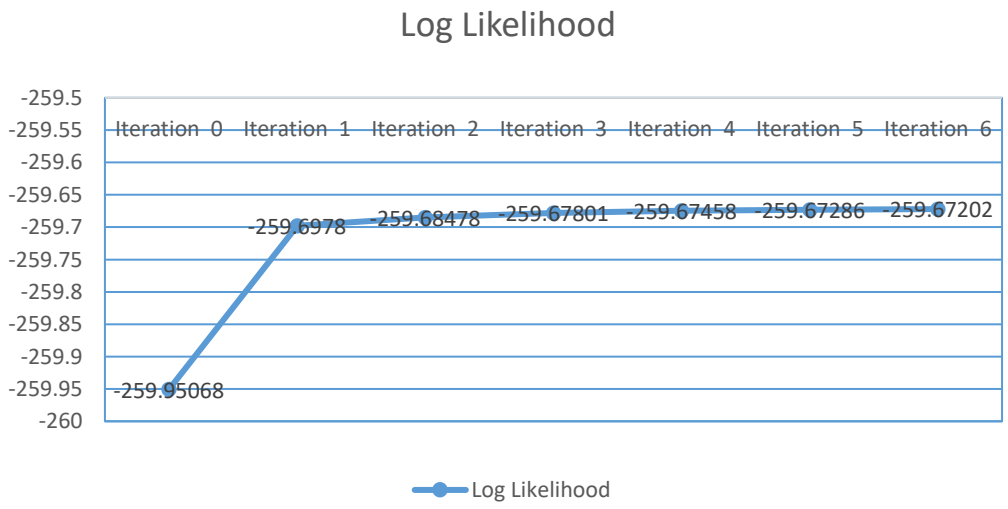
Weight | Gender

	W > 130	W < 130
M	0.79547	0.20453
F	0.36481	0.63519

Height | Gender

	H > 55	H < 55
M	0.72382	0.27618
F	0.29861	0.70139

Plot of likelihood vs iterations



Number of iterations = 6
Final log likelihood = -259.67202

Some other interesting cases –

Dataset – 70% missing, inverted given starting parameters

Starting points –
P(G) = 0.3
P(W|G) = 0.2
P(W|-G) = 0.6
P(H|G) = 0.3
P(H|-G) = 0.7

Final conditional probability tables –

Gender

M	0.68888
F	0.31112

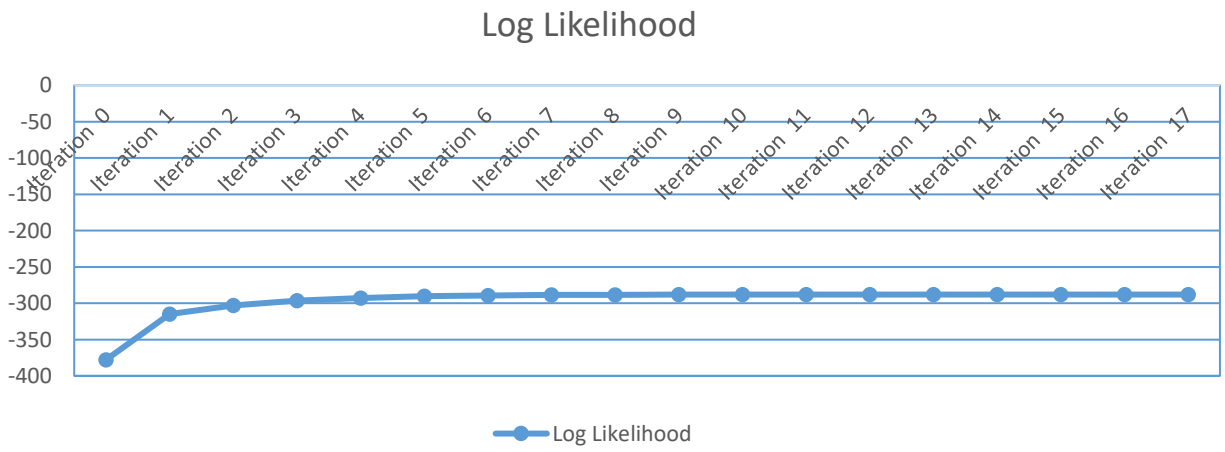
Weight | Gender

	W > 130	W < 130
M	0.82923	0.17077
F	0.42990	0.57010

Height | Gender

	H > 55	H < 55
M	0.65854	0.34146
F	0.30965	0.69035

Plot of likelihood vs iterations



Number of iterations = 17
Final log likelihood = -288.05336

Dataset – 10% missing, inverted given starting parameters

Starting points –
P(G) = 0.3
P(W|G) = 0.2
P(W|-G) = 0.6
P(H|G) = 0.3
P(H|-G) = 0.7

Final conditional probability tables –

Gender

M	0.64245
F	0.35755

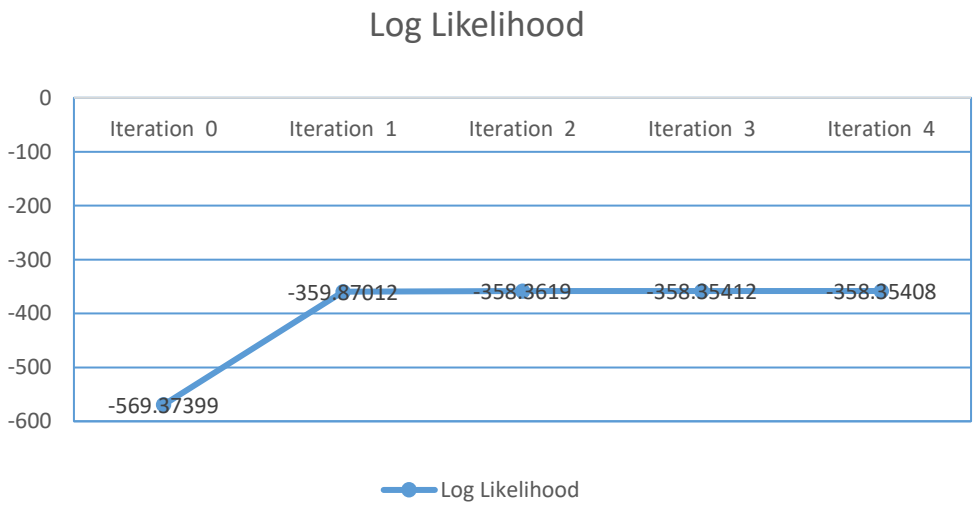
Weight | Gender

	W > 130	W < 130
M	0.79442	0.20558
F	0.34853	0.65147

Height | Gender

	H > 55	H < 55
M	0.66192	0.33808
F	0.26498	0.73502

Plot of likelihood vs iterations



Number of iterations = 4
Final log likelihood = -358.35408

Dataset – 100% missing, all starting parameters at half

Starting points –
P(G) = 0.5
P(W|G) = 0.5
P(W|-G) = 0.5
P(H|G) = 0.5
P(H|-G) = 0.5

Final conditional probability tables –

Gender

M	0.5
F	0.5

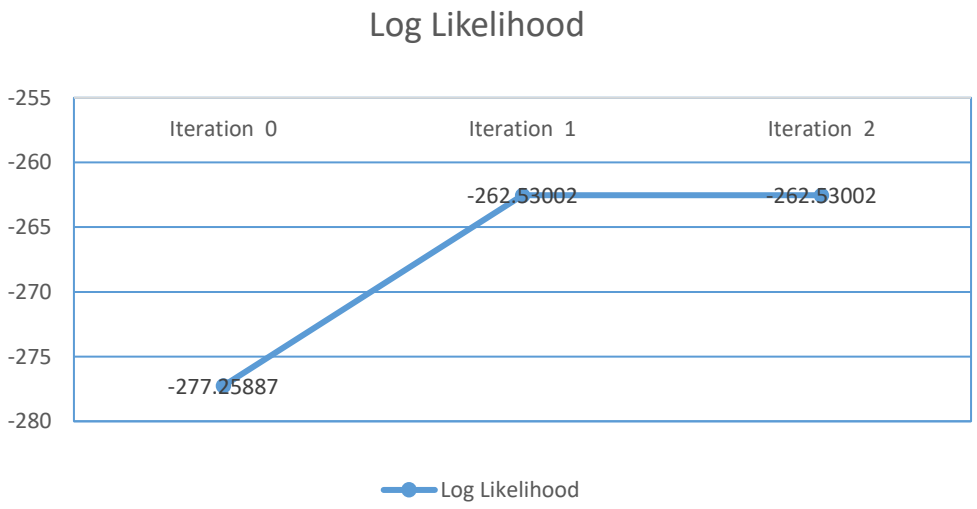
Weight | Gender

	W > 130	W < 130
M	0.665	0.335
F	0.665	0.335

Height | Gender

	H > 55	H < 55
M	0.595	0.405
F	0.595	0.405

Plot of likelihood vs iterations



Number of iterations = 2
Final log likelihood = -262.53002

Questions and answers

Do multiple starting points help in finding better solutions?

In all of the cases that contained some hard data the output ended up being close to each other for the given case regardless of the starting parameters. Even in the cases of inverted parameters the output would end up being close to the ones where the starting parameters were more accurate, it would just take more iterations.

Do some of the different solutions have the same likelihood scores?

Yes the solutions for each given set of data gives a likelihood score very close to each other. Different starting parameters have a wide effect of the initial likelihood, however ultimately the likelihood ends up being close to the same in the end.

How does the data missing rate affect your algorithm and the results?

The sets with less missing data required less iterations to come to a result in all of the cases. In the case with 100% missing data the results would always end up mirroring the starting parameters closely, so in this case having a good estimation is very important. In the cases with at least some data, regardless of the starting parameters, the end results would always end up close to whatever estimation the hard data contained.