Discriminant Analysis using Logistic Regression

by
Milo Schield
Member: International Statistical Institute
US Rep: International Statistical Literacy Project
Director, W. M. Keck Statistical Literacy Project

Slides, output and data at: www.StatLit.org/
pdf/2016-Schield-Logistic-OLS1D-Excel2013-Slides.pdf
pdf/2016-Schield-Logistic-OLS1D-Excel2013-Demo.pdf
Excel/2016-Schield-Logistic-OLS1D-Excel2013-Data.xlsx

Discriminant Analysis: Outcome must be Categorical

Definition: A statistical technique used to classify objects into groups (to predict membership in groups).

Two-Group (Binary) Examples:
Admission to grad, law or medical school
Passing a test (CPA, CMA, etc.)
Toxicity of a substance on insects (causes death in some)

Making a loan; Bankruptcy
Winning an election; Being unemployed
Use of contraceptives; Driving drunk
Pregnancy or divorce; Heart attack or Alzheimer's

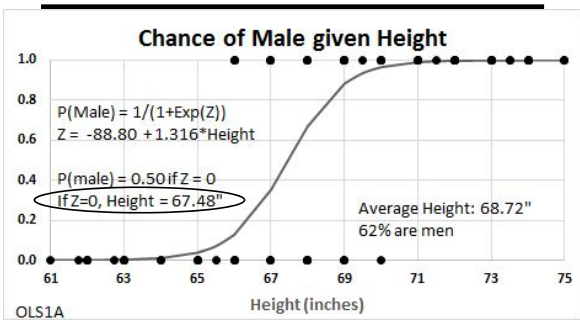
Discriminant Analysis Uses Regression

Modelling a binary outcome (loan vs. no-loan) requires logistic regression.

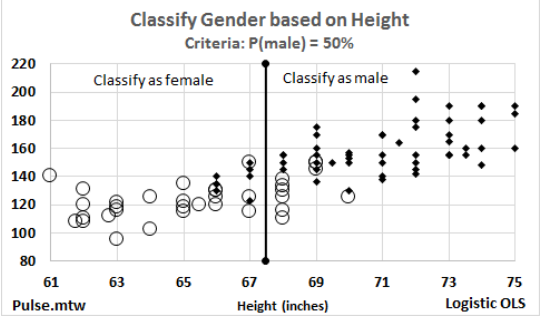
This presentation classifies college students by gender based on their height and weight.

Three logistic models are referenced:
* www.statlit.org/pdf/2015-Schield-Logistic-OLS1A-slides.pdf
* www.statlit.org/pdf/2015-Schield-Logistic-OLS1B-slides.pdf
* www.statlit.org/pdf/2015-Schield-Logistic-OLS1C-slides.pdf

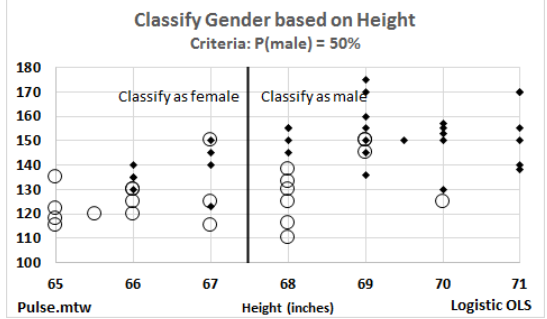
1a: Model gender on Height



1b: Predict Sex given Height Diamond=Male; Circle=Female



1c: Predict Sex given Height: Error Analysis Close-up



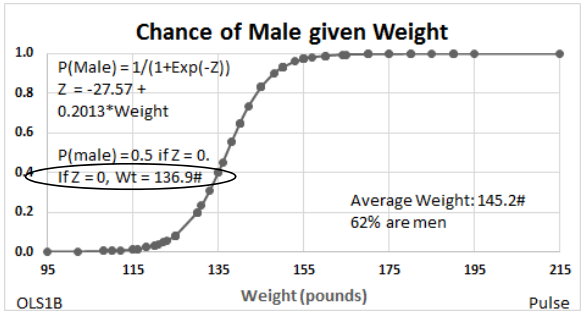
1d. Predict Sex given Height: Error Analysis Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	25	**10**	35
Male	**8**	49	57
All	33	59	92

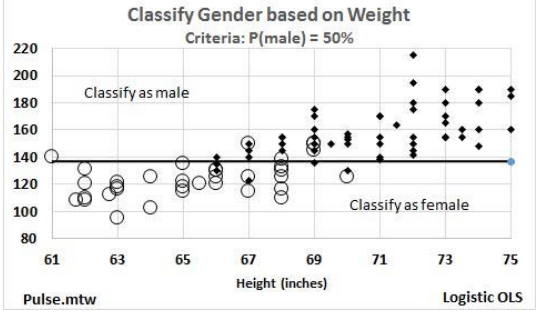
Male if Height > 67.5"; otherwise female

19.6% (18/92) are classified improperly

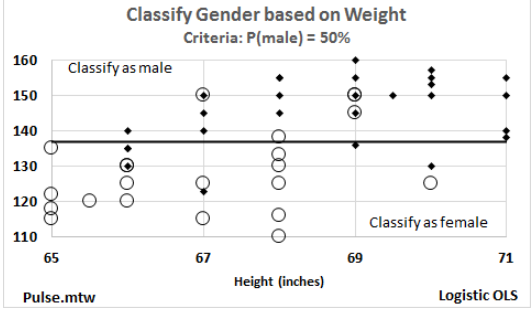
2a. Model Gender on Weight



2b. Predict Sex given Weight Diamond=Male; Circle=Female



2c. Predict Sex given Weight Error Analysis Close-Up



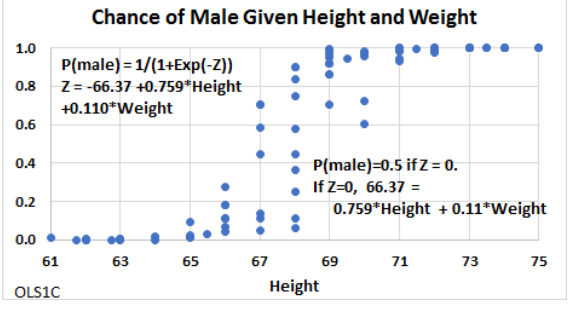
2d. Predict Sex given Weight Error Analysis Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	29	**6**	35
Male	**6**	51	57
All	35	57	92

Male if Weight > 137#; otherwise female

13% (12/92) are classified improperly

3a. Model Gender on Height and Weight



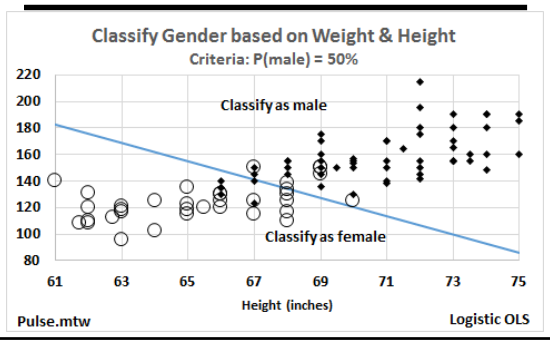
3b. Model Gender on Height and Weight

P(male) = 50%: 66.37 = 0.759*Ht+0.11*Wt
Weight(P50) = (66.37 – 0.759*Height) / 0.11

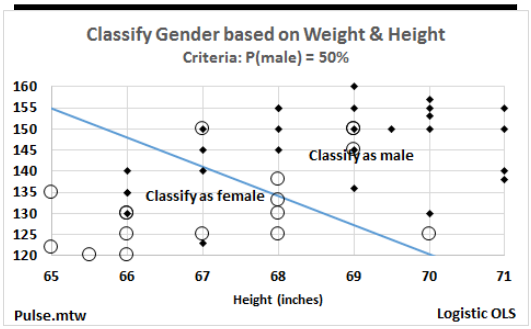
Ht	Weight
60	189
61	182
62	176
63	169
64	162
65	155
66	148
67	141

Ht	Weight
68	134
69	127
70	120
71	113
72	107
73	100
74	93
75	86

3c. Model Gender on Height and Weight



3d. Model Gender on Ht & Wt: Error Close-up



3e. Model Gender on Ht & Wt: Error Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	29	6	35
Male	6	51	57
All	35	57	92

Male if weight > predicted wt given height

13% (12/92) are classified improperly

Summary

Using just height, 19.6% are mis-classified.
Using just weight, 13.0% are misclassified.

Using both height and weight, 13.0% are misclassified.

What is the advantage of using weight instead of height?
34% reduction in error: (13-19.6)/19.6

Disadvantage of using both height & weight vs. weight?
More complex. Can't show in 2D.
Advantage of using both height & weight vs. weight?
Probably better at handling future subjects.

Appendix

Q. Why not just use the average? Mean height or weight?
A. Group average is influenced by the outcome mix.
Logistic regression models the chance of the outcome.
Chance is not influenced by the outcome mix.

==========

Interpreting the coefficients in Logistic Regression: This important topic is beyond this introductory presentation.

Read *The Chicago Guide to "Writing about Multivariate Analysis"* by Jane Miller. See p. 220-243 and 418-431.

Discriminant Analysis using Logistic Regression

by
Milo Schield
Member: International Statistical Institute
US Rep: International Statistical Literacy Project
Director, W. M. Keck Statistical Literacy Project

Slides, output and data at: www.StatLit.org/
pdf/2016-Schield-Logistic-OLS1D-Excel2013-Slides.pdf
pdf/2016-Schield-Logistic-OLS1D-Excel2013-Demo.pdf
Excel/2016-Schield-Logistic-OLS1D-Excel2013-Data.xlsx

Discriminant Analysis: Outcome must be Categorical

Definition: A statistical technique used to classify objects into groups (to predict membership in groups).

Two-Group (Binary) Examples:
Admission to grad, law or medical school
Passing a test (CPA, CMA, etc.)
Toxicity of a substance on insects (causes death in some)

Making a loan; Bankruptcy
Winning an election; Being unemployed
Use of contraceptives; Driving drunk
Pregnancy or divorce; Heart attack or Alzheimer's

Discriminant Analysis Uses Regression

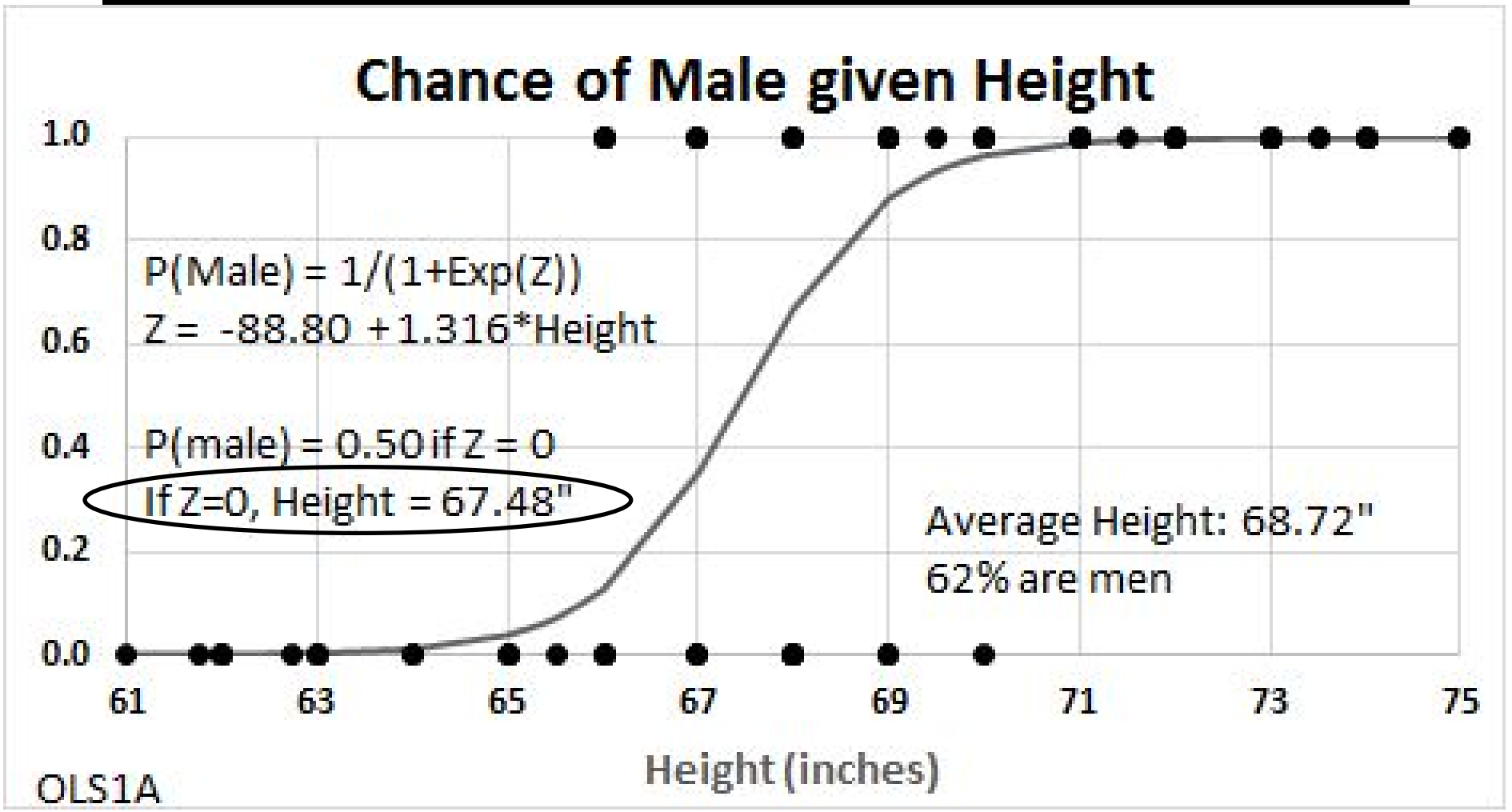
Modelling a binary outcome (loan vs. no-loan) requires logistic regression.

This presentation classifies college students by gender based on their height and weight.

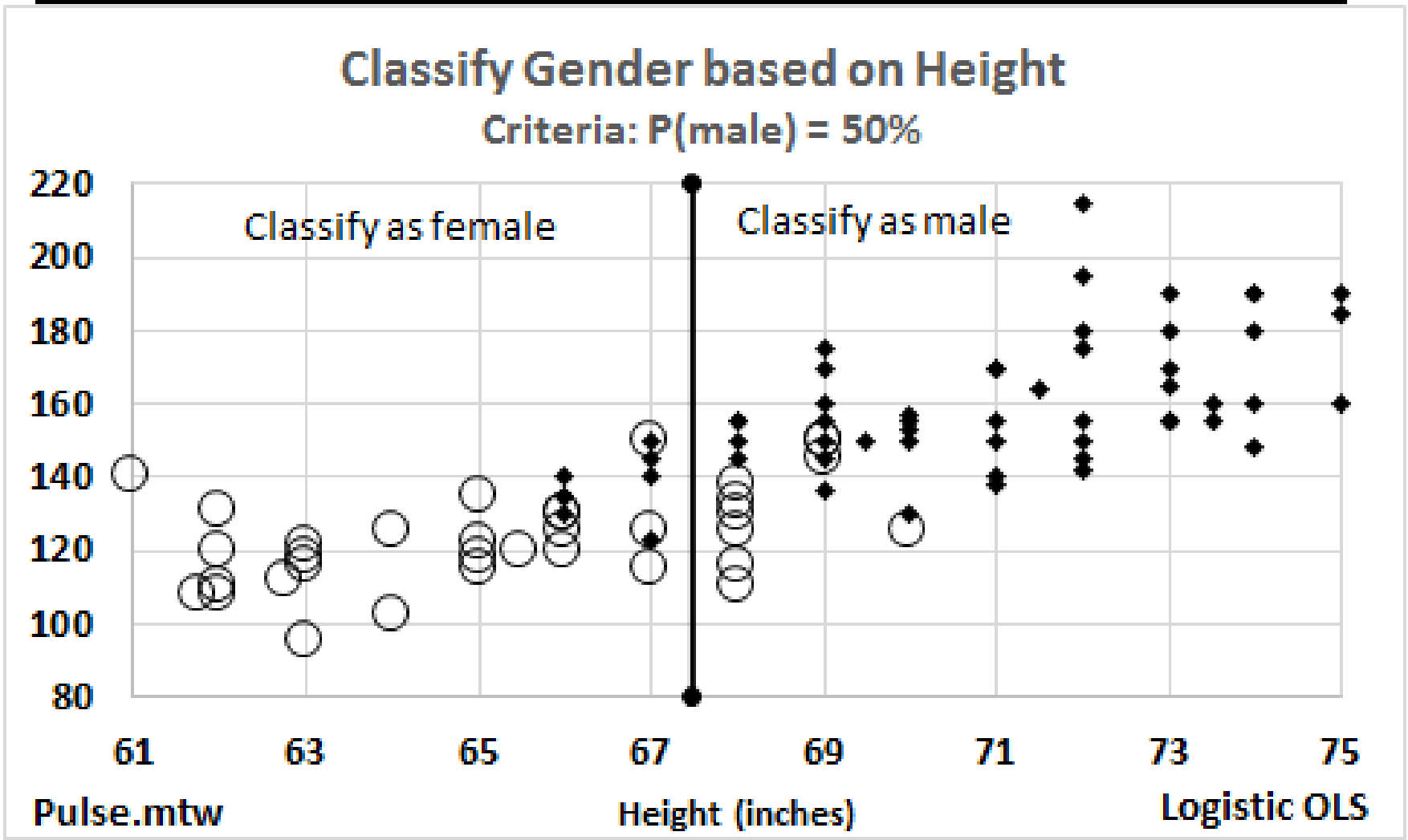
Three logistic models are referenced:

* www.statlit.org/pdf/2015-Schield-Logistic-OLS1A-slides.pdf
* www.statlit.org/pdf/2015-Schield-Logistic-OLS1B-slides.pdf
* www.statlit.org/pdf/2015-Schield-Logistic-OLS1C-slides.pdf

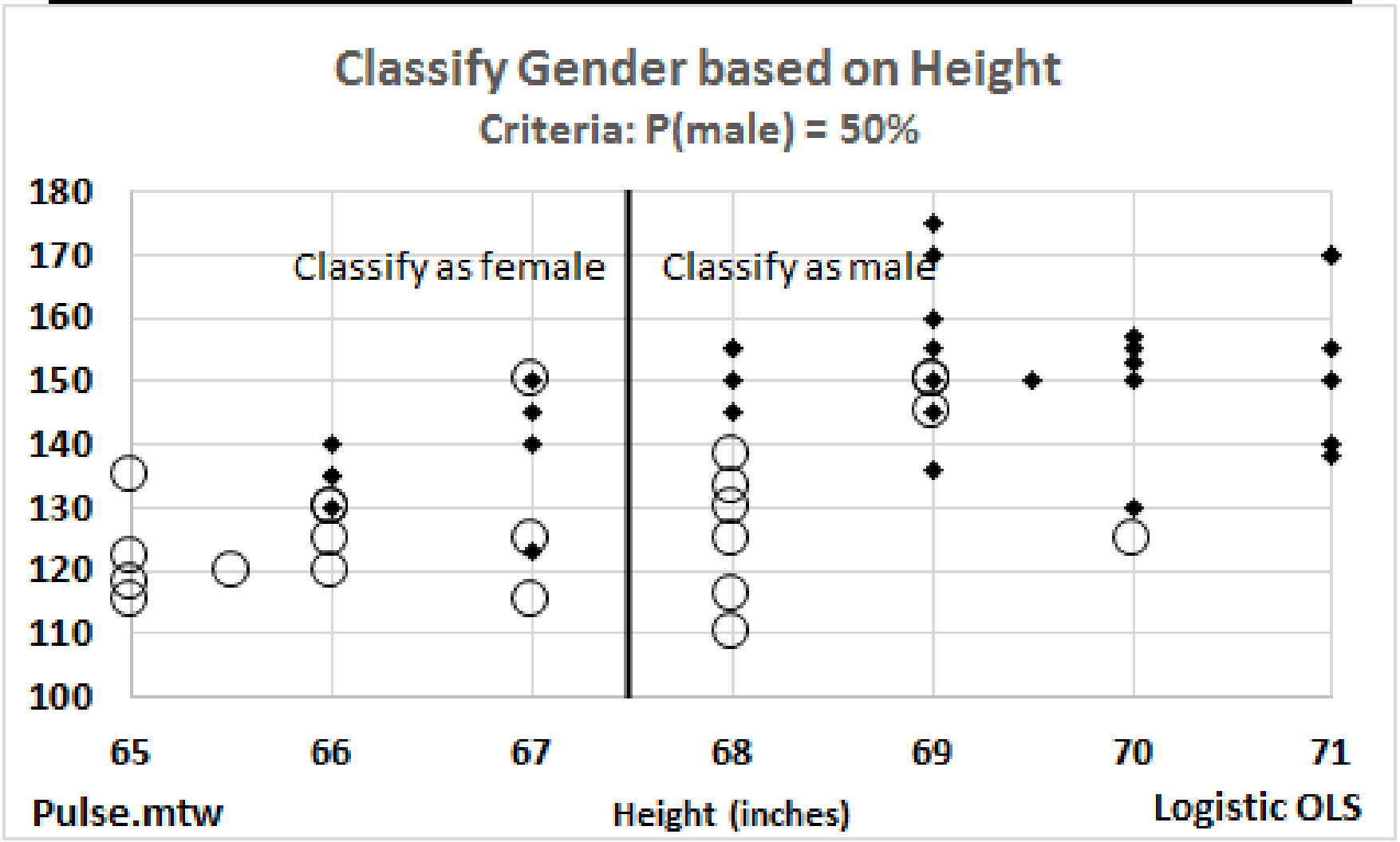
1a: Model gender on Height



1b: Predict Sex given Height Diamond=Male; Circle=Female



1c: Predict Sex given Height: Error Analysis Close-up



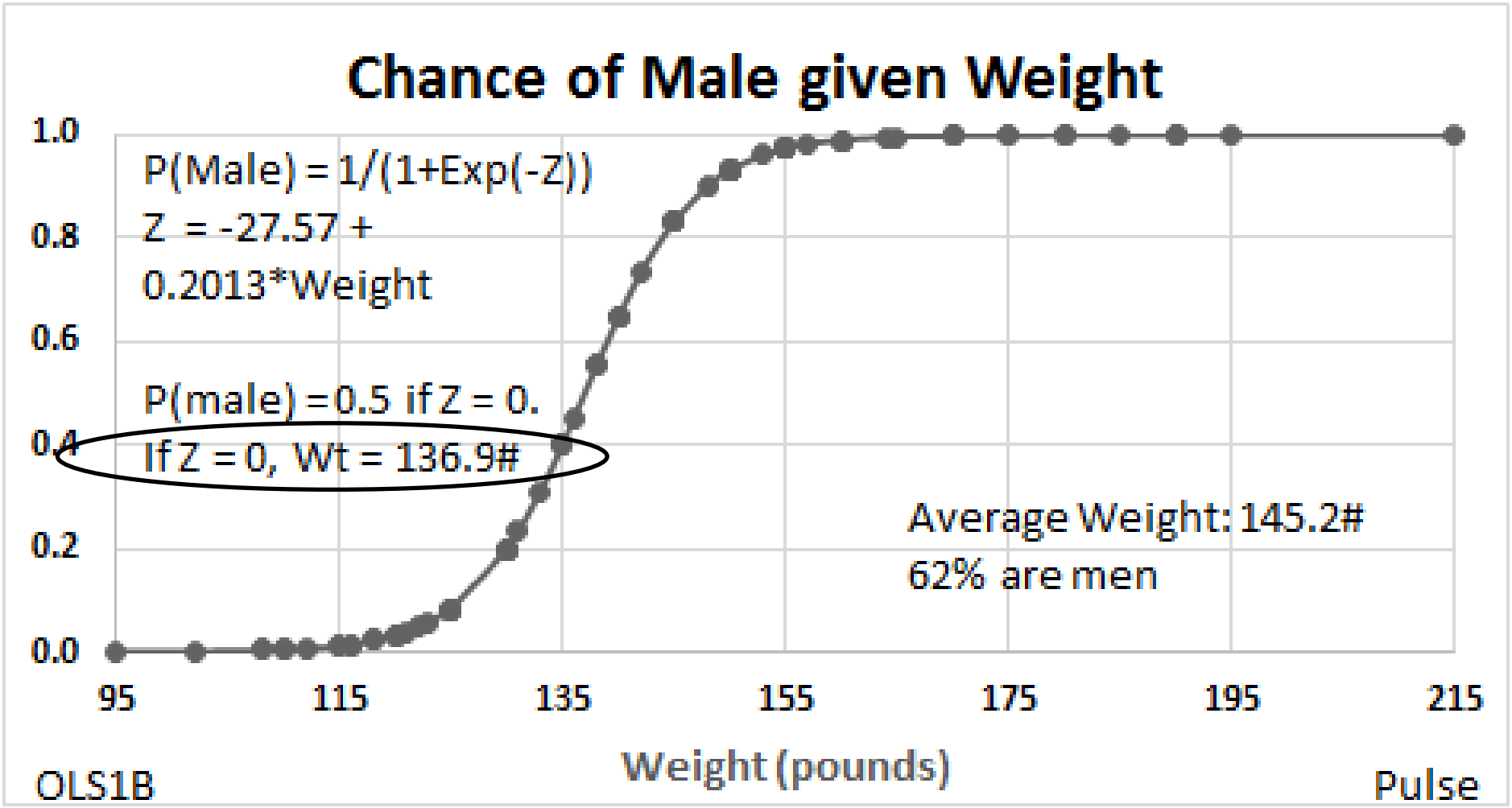
1d. Predict Sex given Height: Error Analysis Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	25	10	35
Male	8	49	57
All	33	59	92

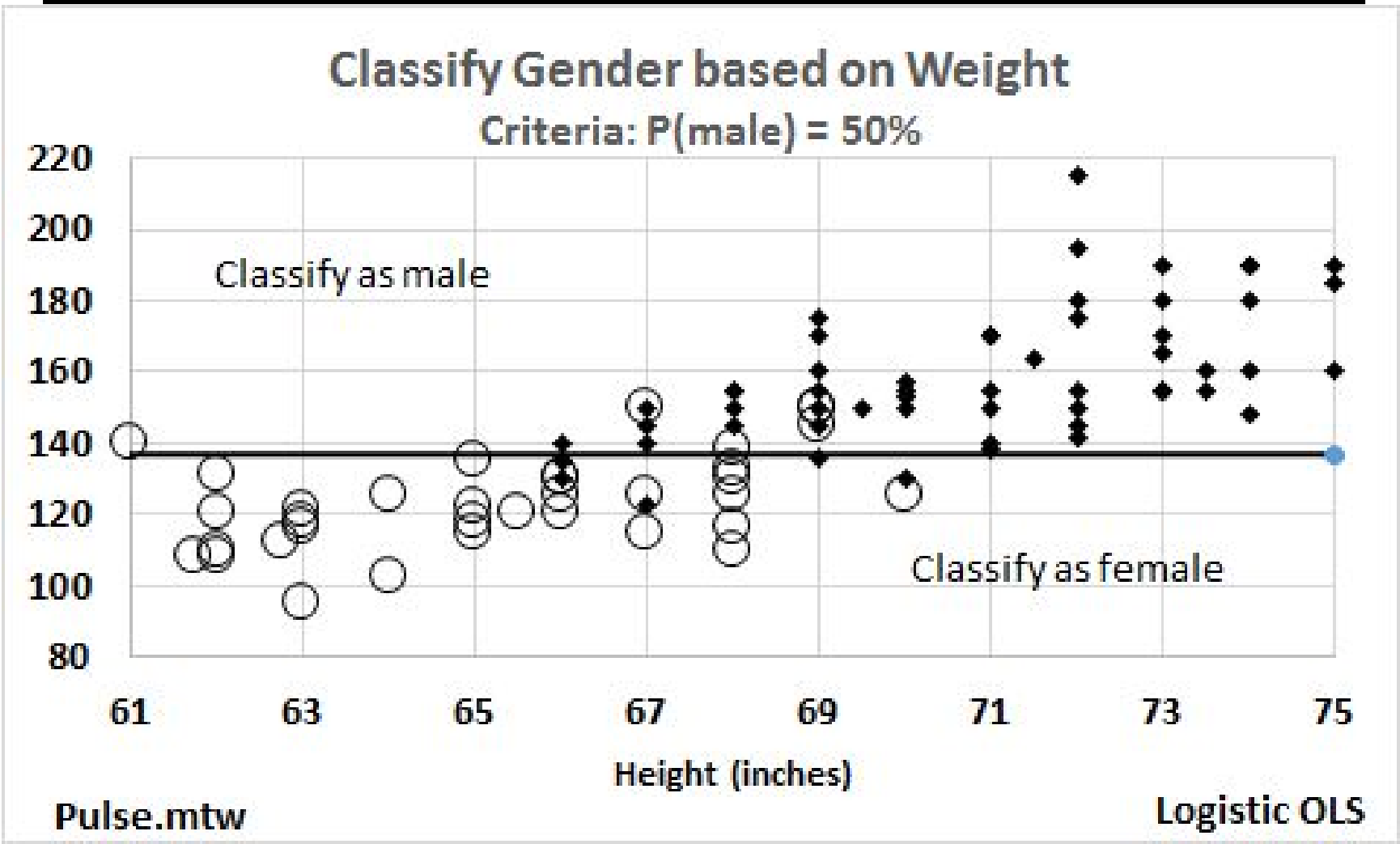
Male if Height > 67.5"; otherwise female

19.6% (18/92) are classified improperly

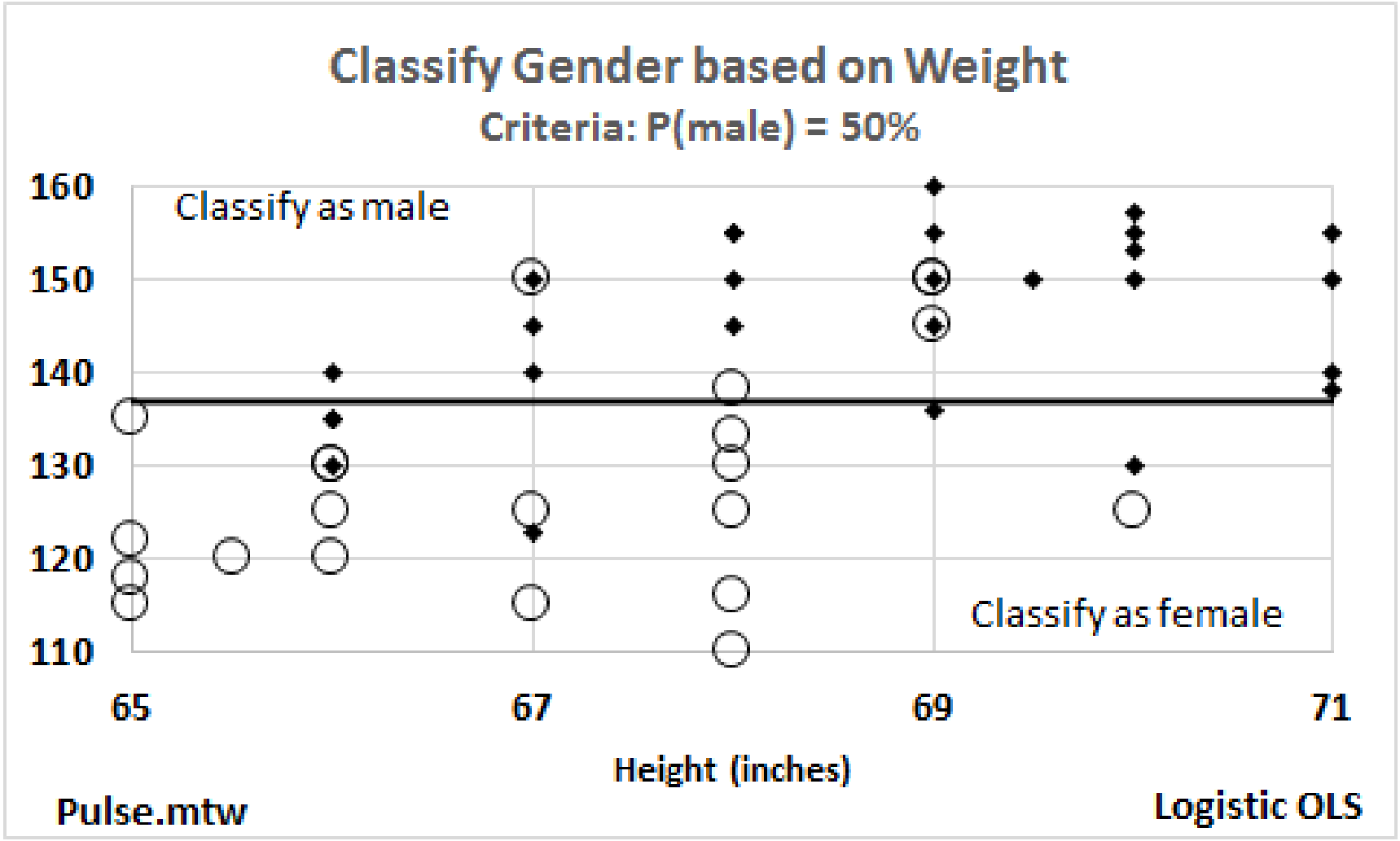
2a. Model Gender on Weight



2b. Predict Sex given Weight Diamond=Male; Circle=Female



2c. Predict Sex given Weight Error Analysis Close-Up



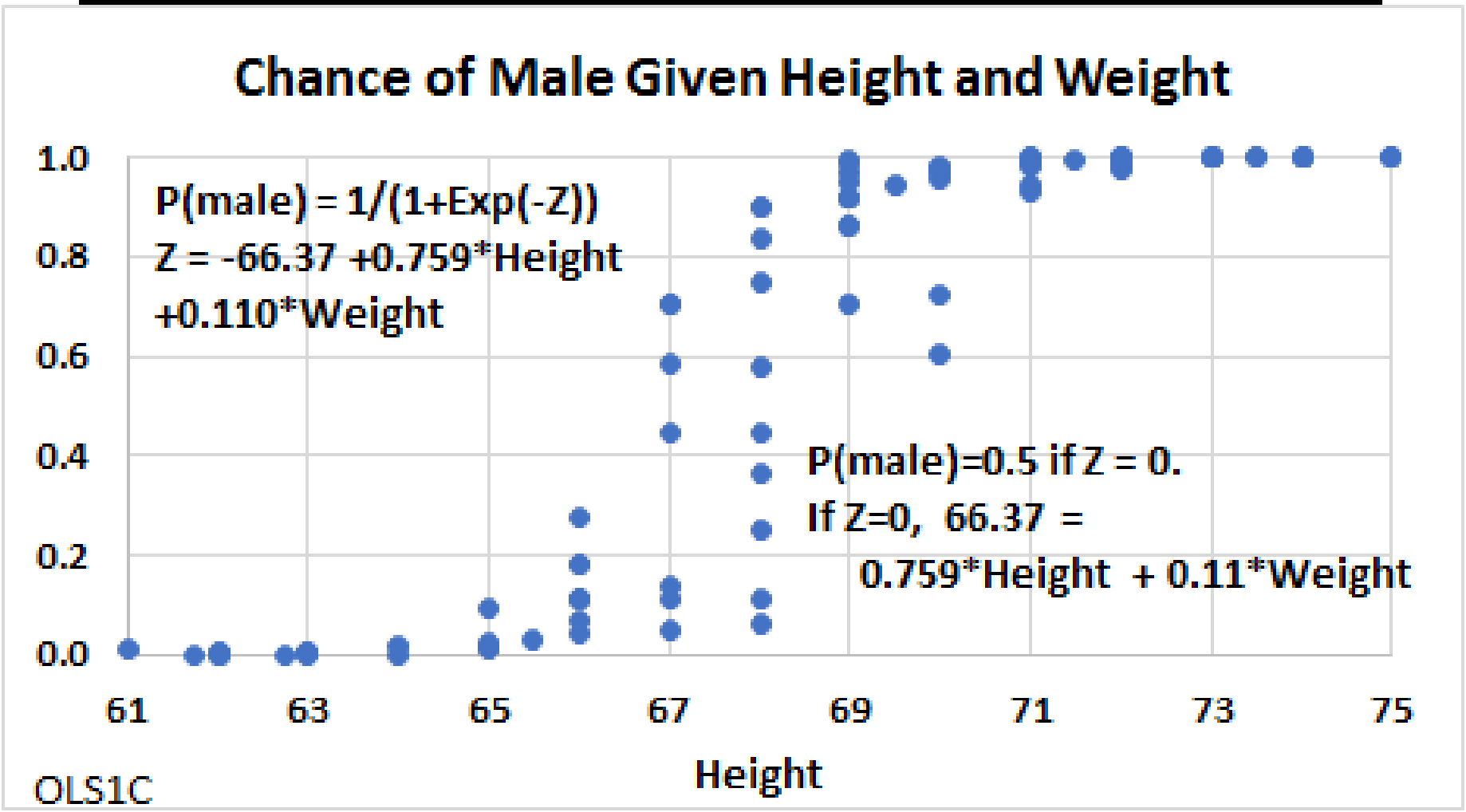
2d. Predict Sex given Weight Error Analysis Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	29	6	35
Male	6	51	57
All	35	57	92

Male if Weight > 137#; otherwise female

13% (12/92) are classified improperly

3a. Model Gender on Height and Weight



3b. Model Gender on Height and Weight

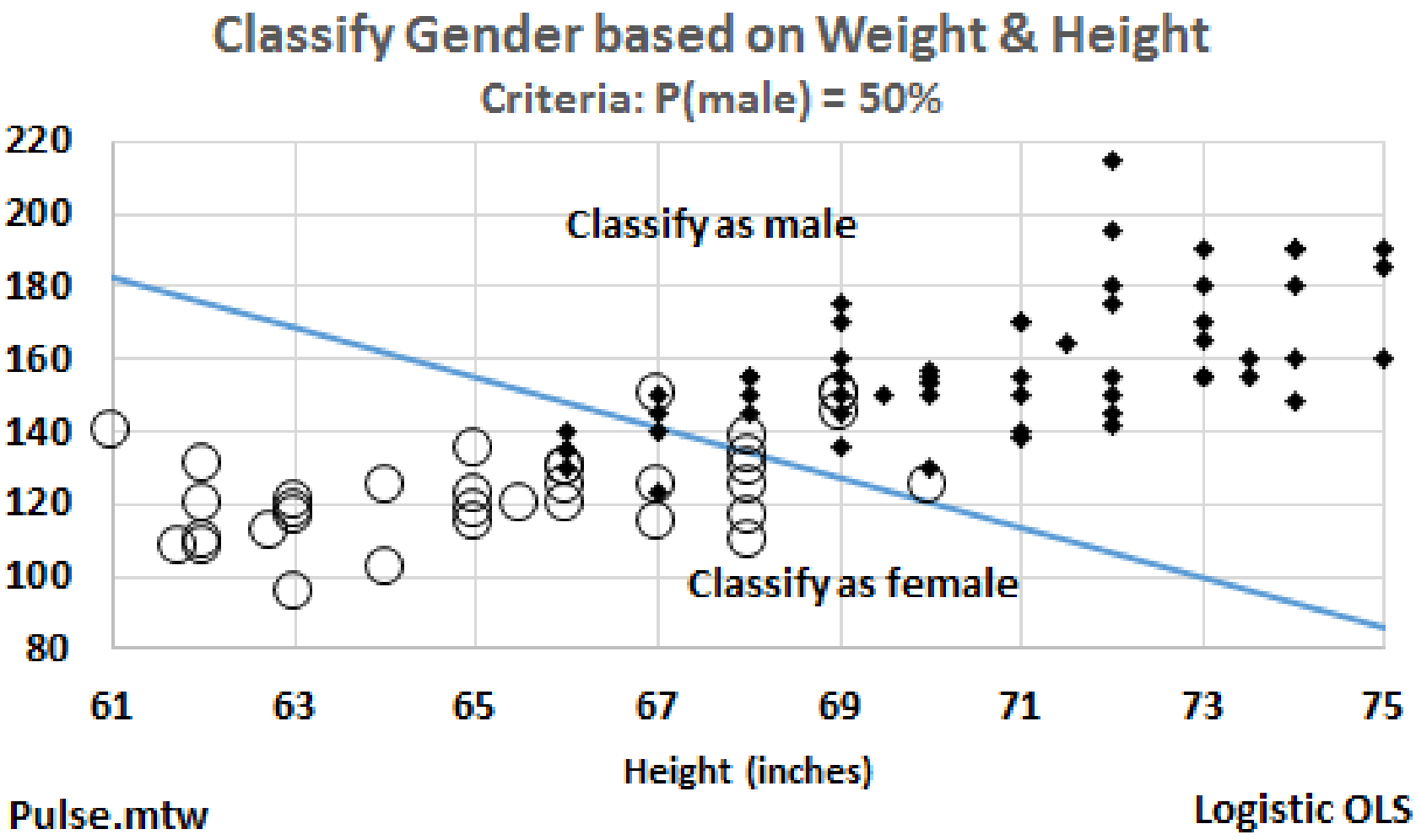
P(male) = 50%: 66.37 = 0.759*Ht+0.11*Wt

Weight(P50) = (66.37 – 0.759*Height) / 0.11

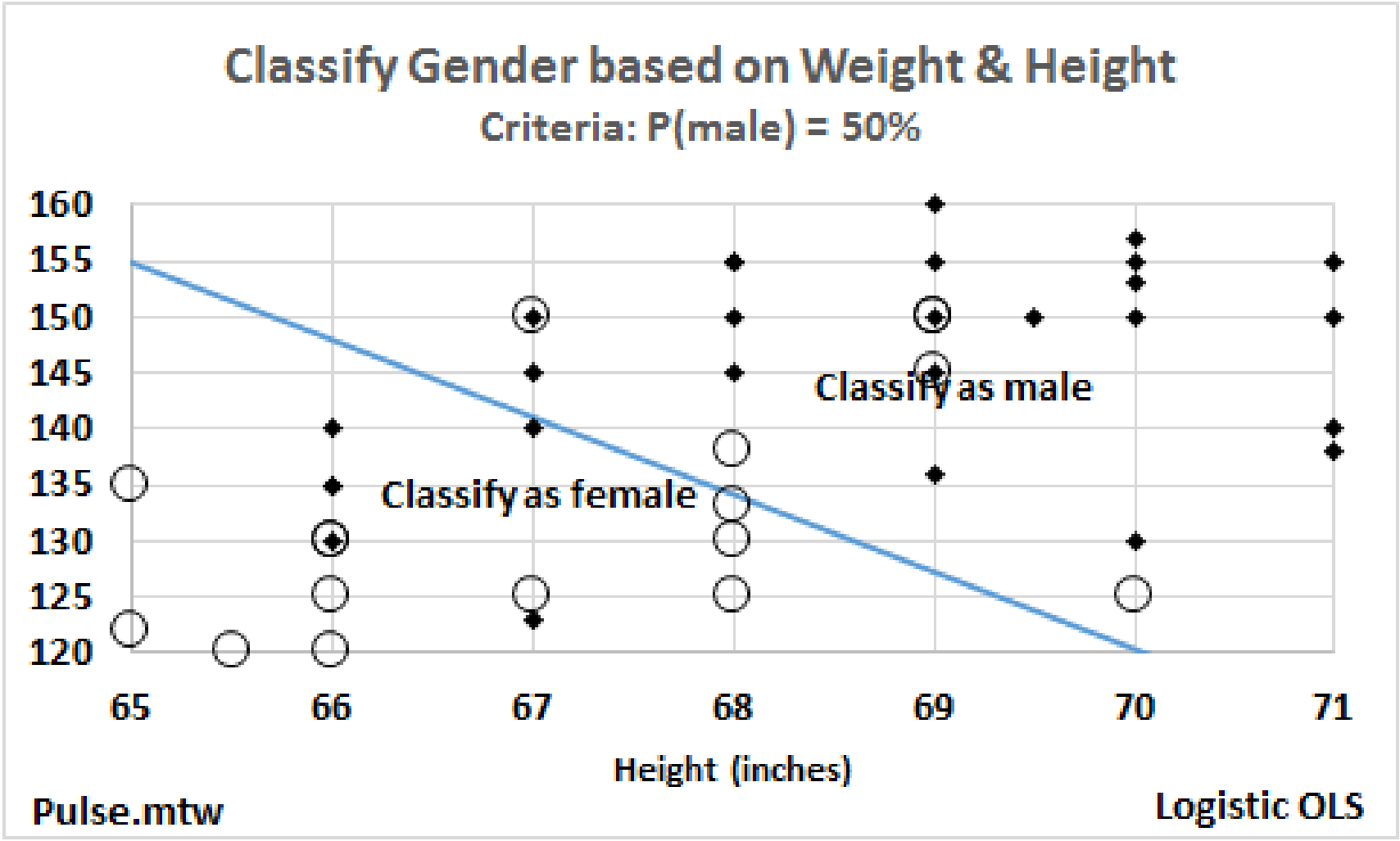
Ht	Weight
60	189
61	182
62	176
63	169
64	162
65	155
66	148
67	141

Ht	Weight
68	134
69	127
70	120
71	113
72	107
73	100
74	93
75	86

3c. Model Gender on Height and Weight



3d. Model Gender on Ht & Wt: Error Close-up



3e. Model Gender on Ht & Wt: Error Summary

	Classified As: P=50%		
Actual	Female	Male	All
Female	29	6	35
Male	6	51	57
All	35	57	92

Male if weight > predicted wt given height

13% (12/92) are classified improperly

Summary

Using just height, 19.6% are mis-classified.
Using just weight, 13.0% are misclassified.

Using both height and weight, 13.0% are misclassified.

What is the advantage of using weight instead of height?
34% reduction in error: (13-19.6)/19.6

Disadvantage of using both height & weight vs. weight?
More complex. Can't show in 2D.

Advantage of using both height & weight vs. weight?
Probably better at handling future subjects.

Appendix

Q. Why not just use the average? Mean height or weight?
A. Group average is influenced by the outcome mix.
Logistic regression models the chance of the outcome.
Chance is not influenced by the outcome mix.

==========

Interpreting the coefficients in Logistic Regression: This important topic is beyond this introductory presentation.

Read *The Chicago Guide to "Writing about Multivariate Analysis"* by Jane Miller. See p. 220-243 and 418-431.