V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 1

Logistic Regression: MLE vs. OLS3 in Excel 2013

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

Slides and data at: www.StatLit.org/pdf/2014-Schield-Logistic-MLE-OLS3-Excel2013-slides.pdf
Excel/2014-Schield-Logistic-MLE-OLS3-Excel2013.xlsx

V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 2

Background & Goals

Doing logistic regression properly requires MLE.
Doing MLE in Excel is not easy. See Schield 2014a

Schield has identified three OLS shortcuts:
OLS1: Model Ln(Odds(p)) where p is near 0 or 1
OLS2: Model Ln(Odds(p)) where p is grouped data
OLS3: Use OLS to estimate logistic parameters.

These slides compare OLS3 with MLE.
Schield (2014b) presents the OLS3-based approach.

V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 3

Predict Gender using Height

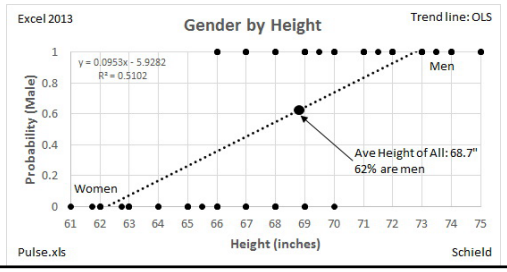
Source: Minitab Pulse dataset

	A	B	C	D	E	F	G	H
1	Pulse1	Pulse2	Height	Weight	Activity	Run?	Smokes?	Male?
2	48	54	68	150	1	0	1	1
3	54	56	69	145	2	0	1	1
4	54	50	69	160	2	0	0	1
5	58	70	72	145	2	1	0	1
6	58	58	66	135	3	0	0	1
7	58	56	67	125	2	0	0	0
8	60	76	71	170	3	1	0	1
9	60	62	71	155	2	0	0	1
10	60	70	71.5	164	2	0	1	1
11	60	66	62	120	2	0	0	0
12	61	70	65.5	120	2	0	0	0
13	62	76	73.5	160	3	1	1	1
14	62	75	72	195	2	1	0	1
15	62	58	72	175	3	1	0	1
16	62	100	66	120	2	1	0	0
17	62	98	62.75	112	2	1	1	0

V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 4

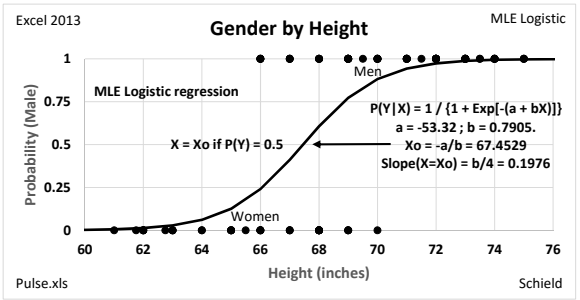
Model Gender by Height (OLS) Must use logistic regression

This trend-line does not satisfy the least-squares assumptions and it goes outside the valid range.



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 5

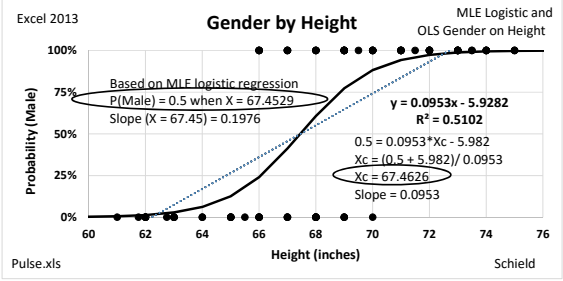
1a: MLE Logistic Regression of Gender by Height



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 6

1b: MLE Logistic Regression vs OLS(Y|X) of Gender on Height

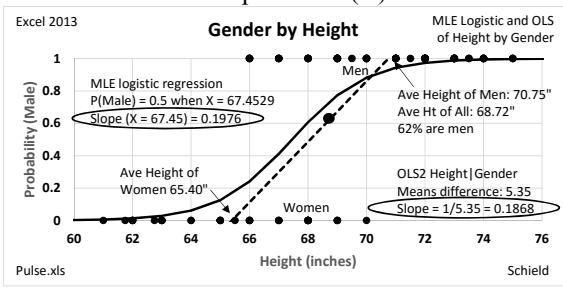
Different slopes but they intersect near P(Y) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 7

1c: MLE Logistic Regression vs OLS(X|Y) of Height on Gender

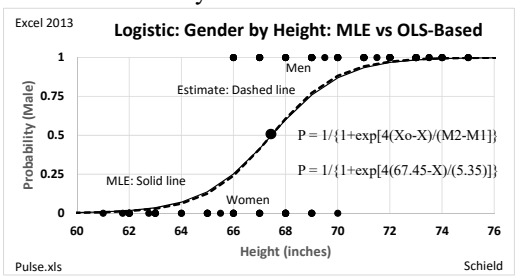
Similar slopes near P(Y) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 8

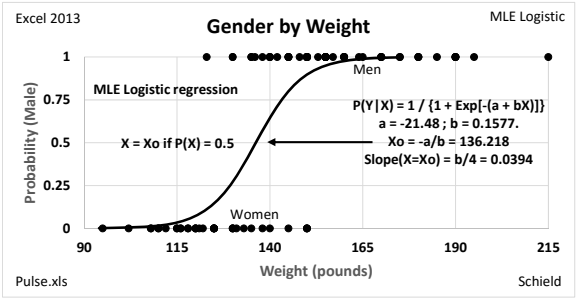
1d: Intersection from OLS(Y|X) Get slope from OLS(X|Y)

Very nice estimate!



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 9

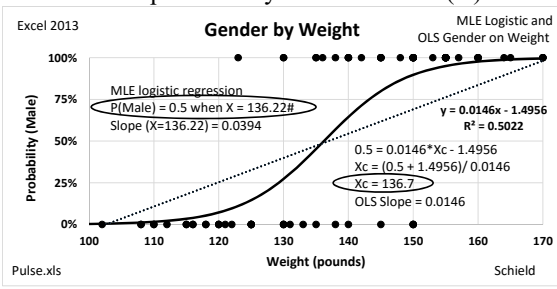
2a: MLE Logistic Regression of Gender by Weight.



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 10

2b: MLE Logistic Regression vs OLS(Y|X) of Gender on Weight

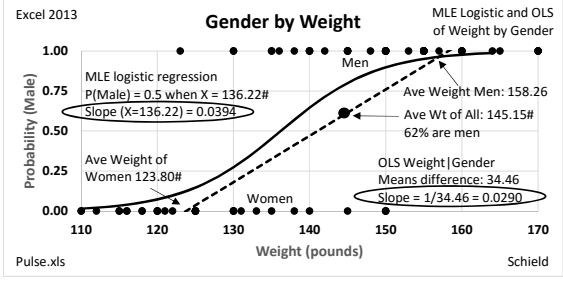
Different slopes but they intersect near P(Y) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 11

2c: MLE Logistic Regression vs OLS(X|Y) of Weight on Gender

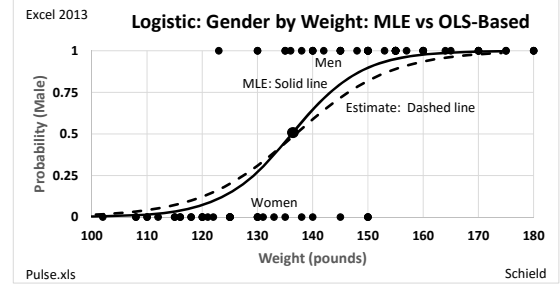
Similar slopes near P(Y) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 12

2d: Intersection from OLS(Y|X) Get slope from OLS(X|Y)

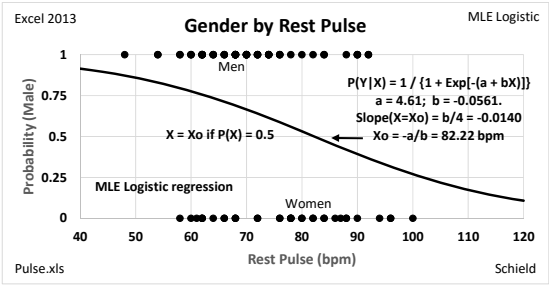
Fairly good estimate



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 13

3a: MLE Logistic Regression of Gender by Rest Pulse

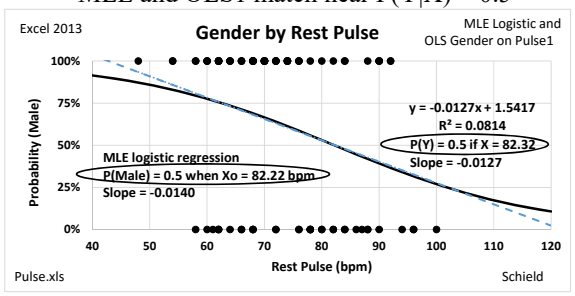
Almost flat! Discriminatory power is weak.



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 14

3b: MLE Logistic Regression vs OLS(Y|X) of Gender on Pulse1

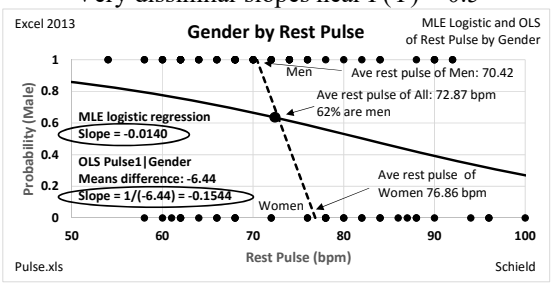
MLE and OLS1 match near P(Y|X) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 15

3c: MLE Logistic Regression vs OLS(X|Y) of Gender on Pulse1

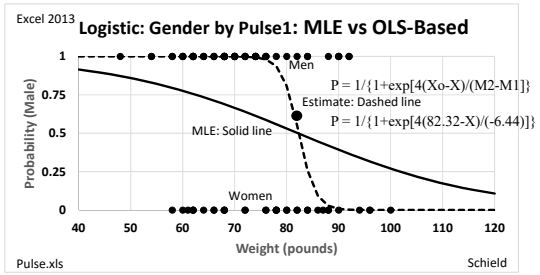
Very dissimilar slopes near P(Y) = 0.5



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 16

3d: Intersection from OLS(Y|X); Get slope from OLS(X|Y)

Very bad estimate



V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 17

Analysis and Conclusion Model Gender:

#1: $R^2 = 51\%$ by Height. "Good estimate"
#2: $R^2 = 50\%$ by Weight. "Fair estimate"
#3: $R^2 = 8\%$ by Rest Pulse. "Bad estimate"

Conclusion #1: Using OLS(Y|X) for Xo and using OLS(X|Y) for the associated slope works fairly well when the overlap is small or moderate: OLS(Y|X). R-squared is high: > 0.5

Conclusion #2: Must use MLE when the overlap is large: OLS(Y|X) R-squared is low (< 0.5). But why bother if the model explains so little?

V0H Schield MLE vs. OLS3-Based Logistic Excel 2013 18

References & Derivation

Schield, Milo (2014a). www.statlit.org/pdf/
Model-Logistic-MLE1A-Excel2013-Slides.pdf
Model-Logistic-MLE1C-Excel2013-Slides.pdf

Schield, Milo (2014b). www.statlit.org/pdf/
2014-Schield-Logistic-OLS3-Excel2013-Slides.pdf

Ln(Odds) = a+bX. P/(1-P) = exp(a+bX)

P = 1 / [1 + exp(-a –bX)]. If P=1/2 at X=Xo, a = -bXo.
P = 1 / {1 + exp[b(Xo – X)]}. b = 4*(dp/dx)|X=Xo.
P = 1 / {1 + exp[4*(Xo-X)*Slope]}
If slope = 1/(M2-M1), P = 1/{1+exp[4(Xo-X)/(M2-M1]}

Logistic Regression: MLE vs. OLS3 in Excel 2013

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

Slides and data at: www.StatLit.org/
pdf/2014-Schield-Logistic-MLE-OLS3-Excel2013-slides.pdf
Excel/2014-Schield-Logistic-MLE-OLS3-Excel2013.xlsx

Background & Goals

Doing logistic regression properly requires MLE.
Doing MLE in Excel is not easy. See Schield 2014a

Schield has identified three OLS shortcuts:
OLS1: Model Ln(Odds(p)) where p is near 0 or 1
OLS2: Model Ln(Odds(p)) where p is grouped data
OLS3: Use OLS to estimate logistic parameters.

These slides compare OLS3 with MLE.
Schield (2014b) presents the OLS3-based approach.

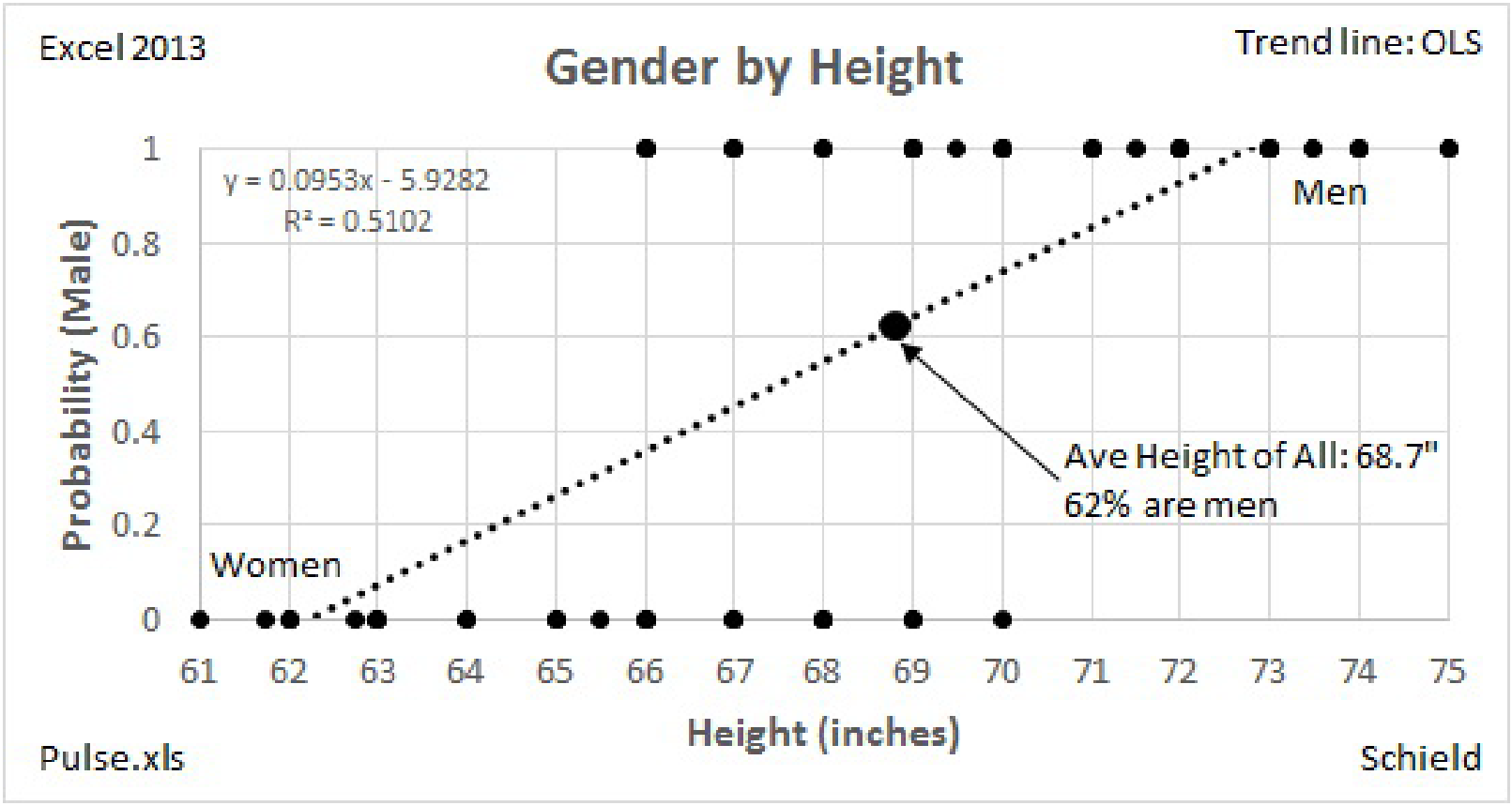
Predict Gender using Height

Source: Minitab Pulse dataset

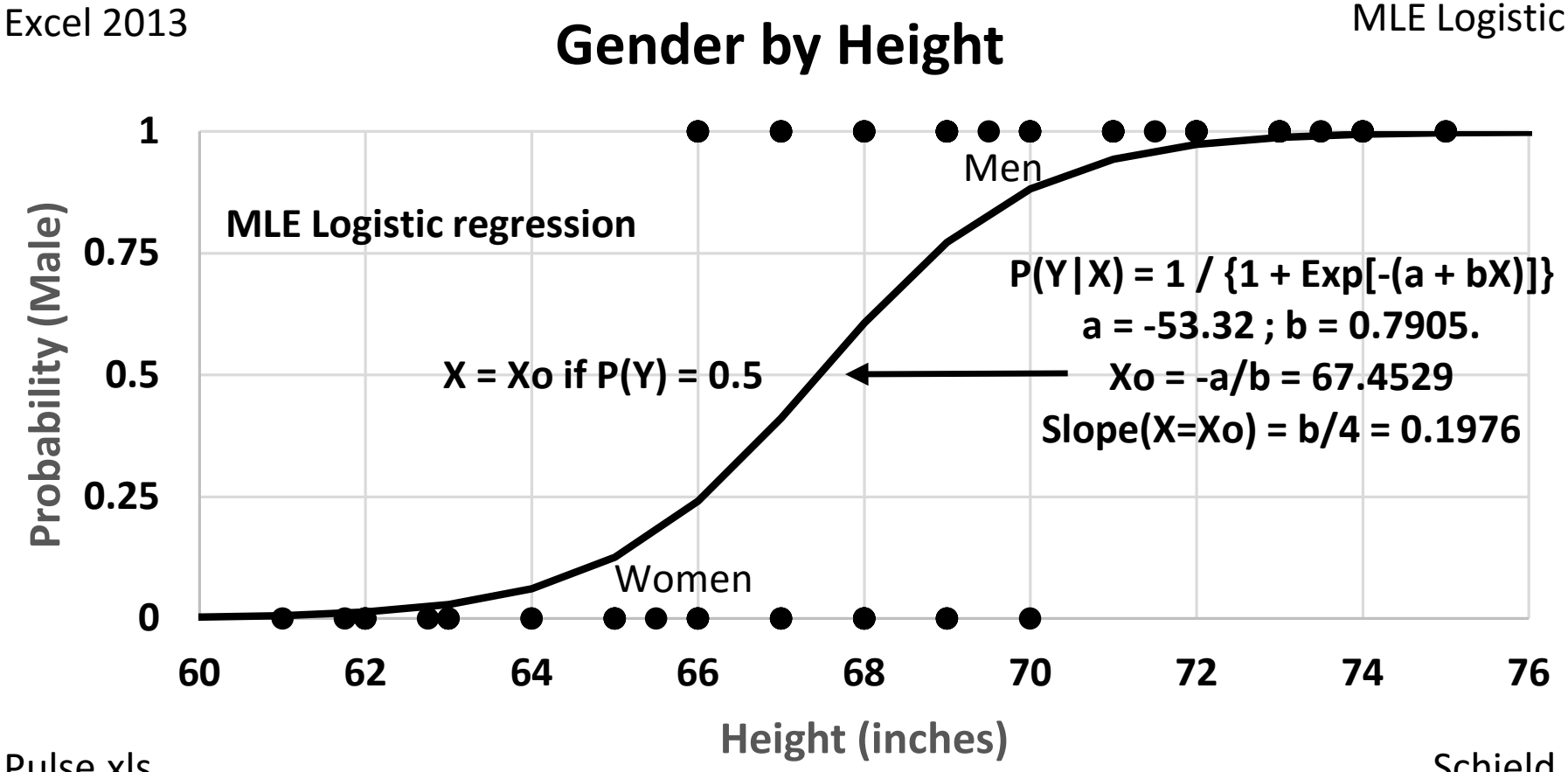
	A	B	C	D	E	F	G	H
1	Pulse1	Pulse2	Height	Weight	Activity	Run?	Smokes?	Male?
2	48	54	68	150	1	0	1	1
3	54	56	69	145	2	0	1	1
4	54	50	69	160	2	0	0	1
5	58	70	72	145	2	1	0	1
6	58	58	66	135	3	0	0	1
7	58	56	67	125	2	0	0	0
8	60	76	71	170	3	1	0	1
9	60	62	71	155	2	0	0	1
10	60	70	71.5	164	2	0	1	1
11	60	66	62	120	2	0	0	0
12	61	70	65.5	120	2	0	0	0
13	62	76	73.5	160	3	1	1	1
14	62	75	72	195	2	1	0	1
15	62	58	72	175	3	1	0	1
16	62	100	66	120	2	1	0	0
17	62	98	62.75	112	2	1	1	0

Model Gender by Height (OLS) Must use logistic regression

This trend-line does not satisfy the least-squares assumptions and it goes outside the valid range.

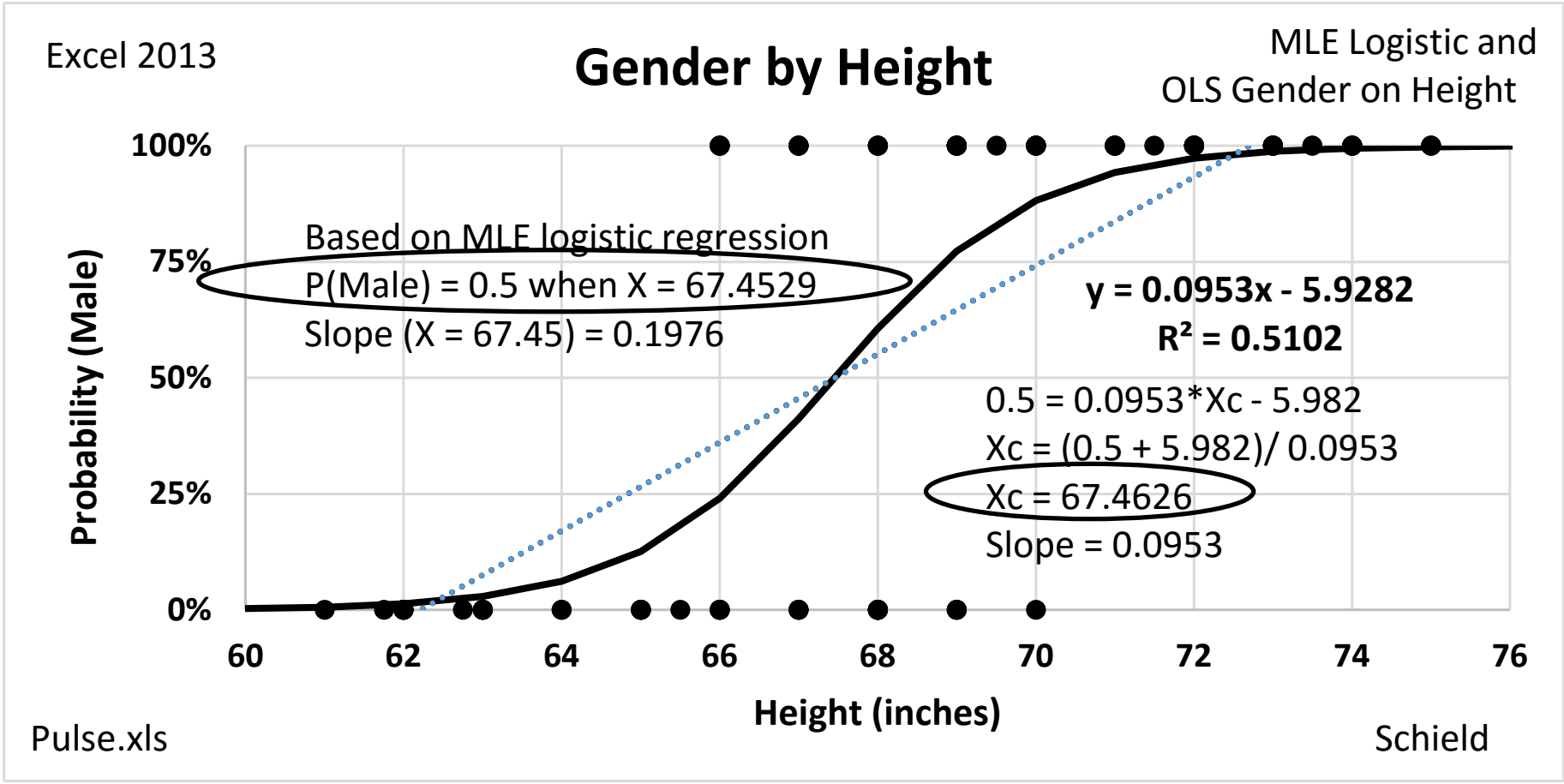


1a: MLE Logistic Regression of Gender by Height



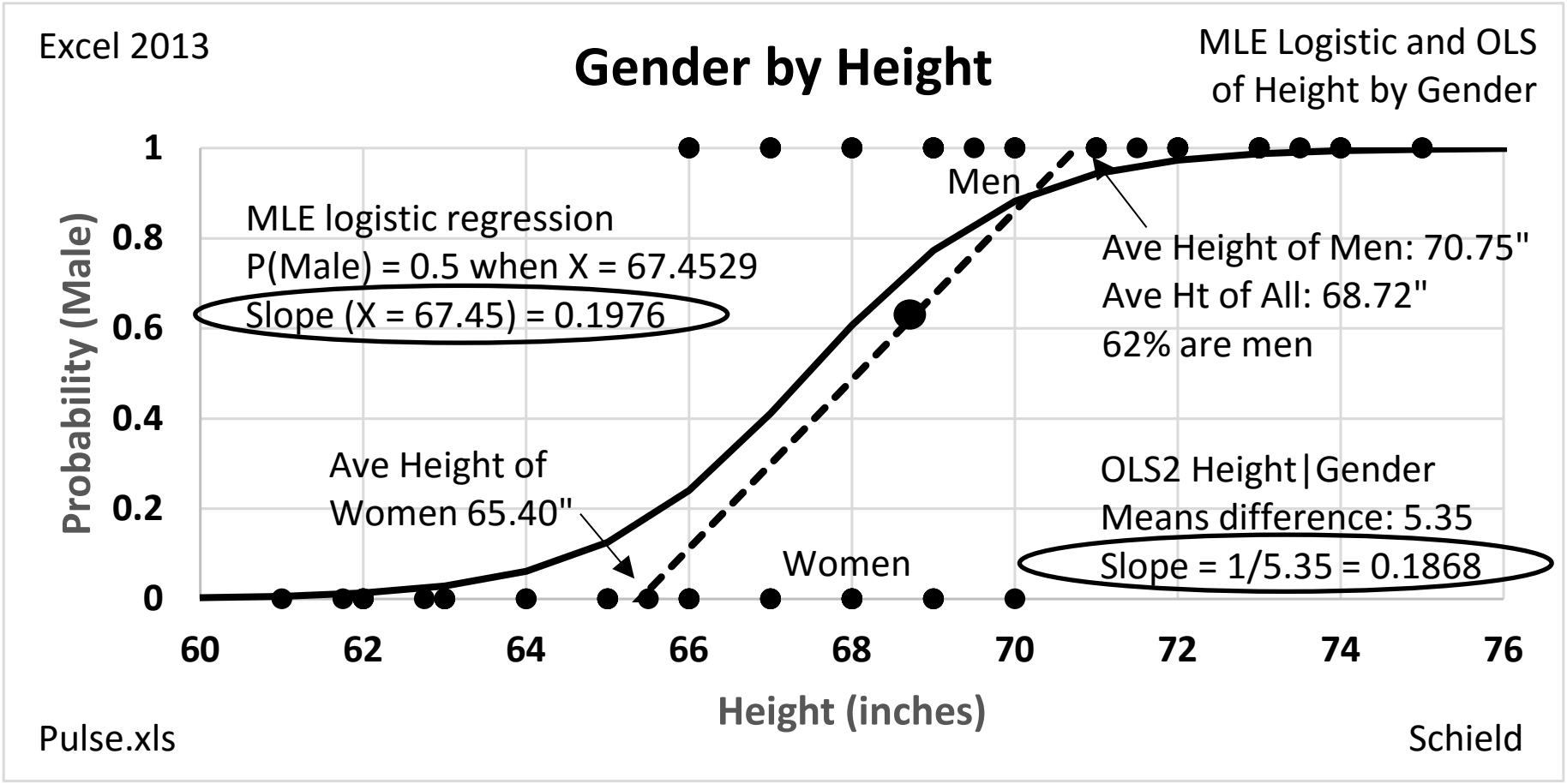
1b: MLE Logistic Regression vs OLS(Y|X) of Gender on Height

Different slopes but they intersect near P(Y) = 0.5



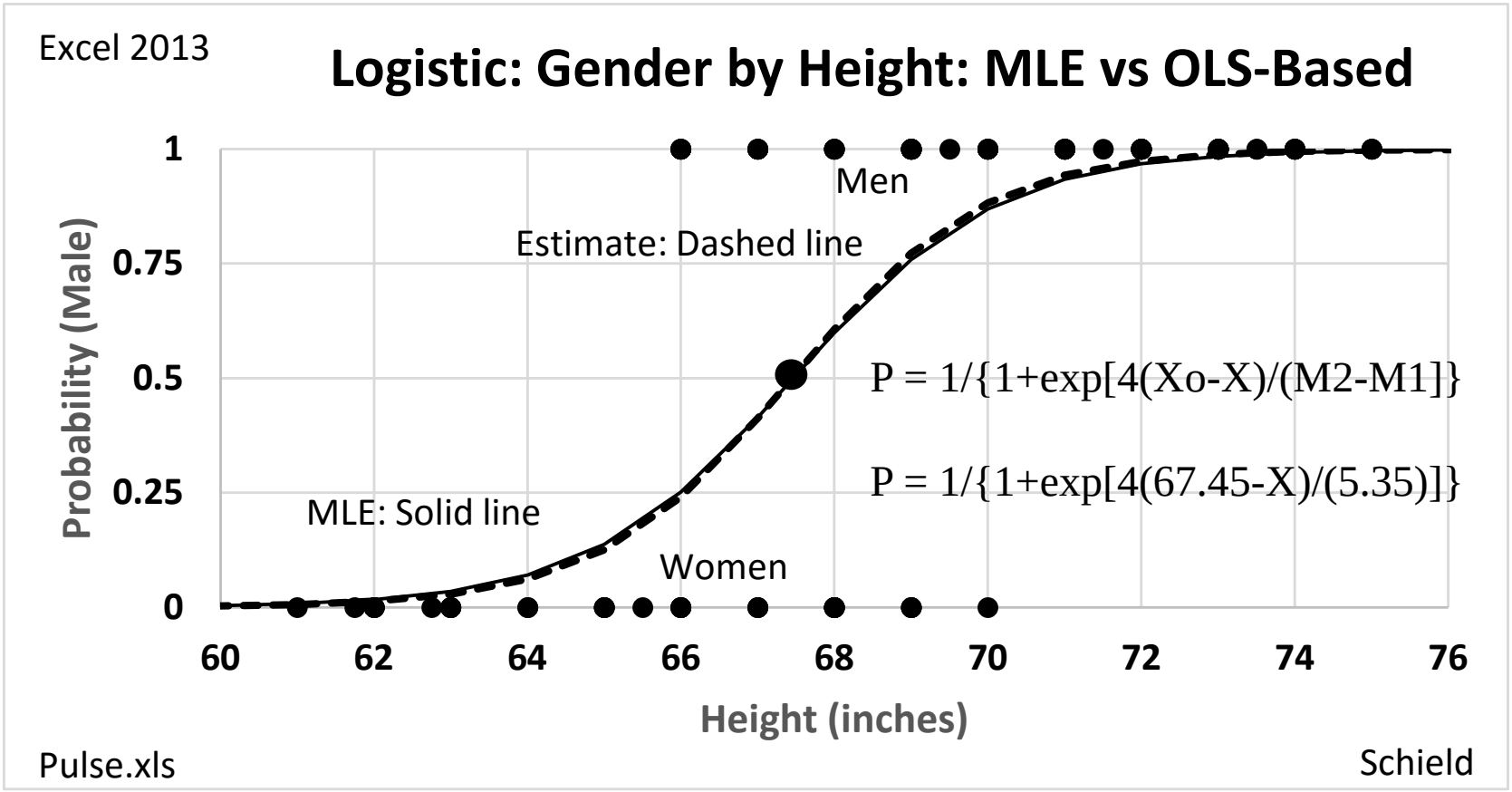
1c: MLE Logistic Regression vs OLS(X|Y) of Height on Gender

Similar slopes near P(Y) = 0.5

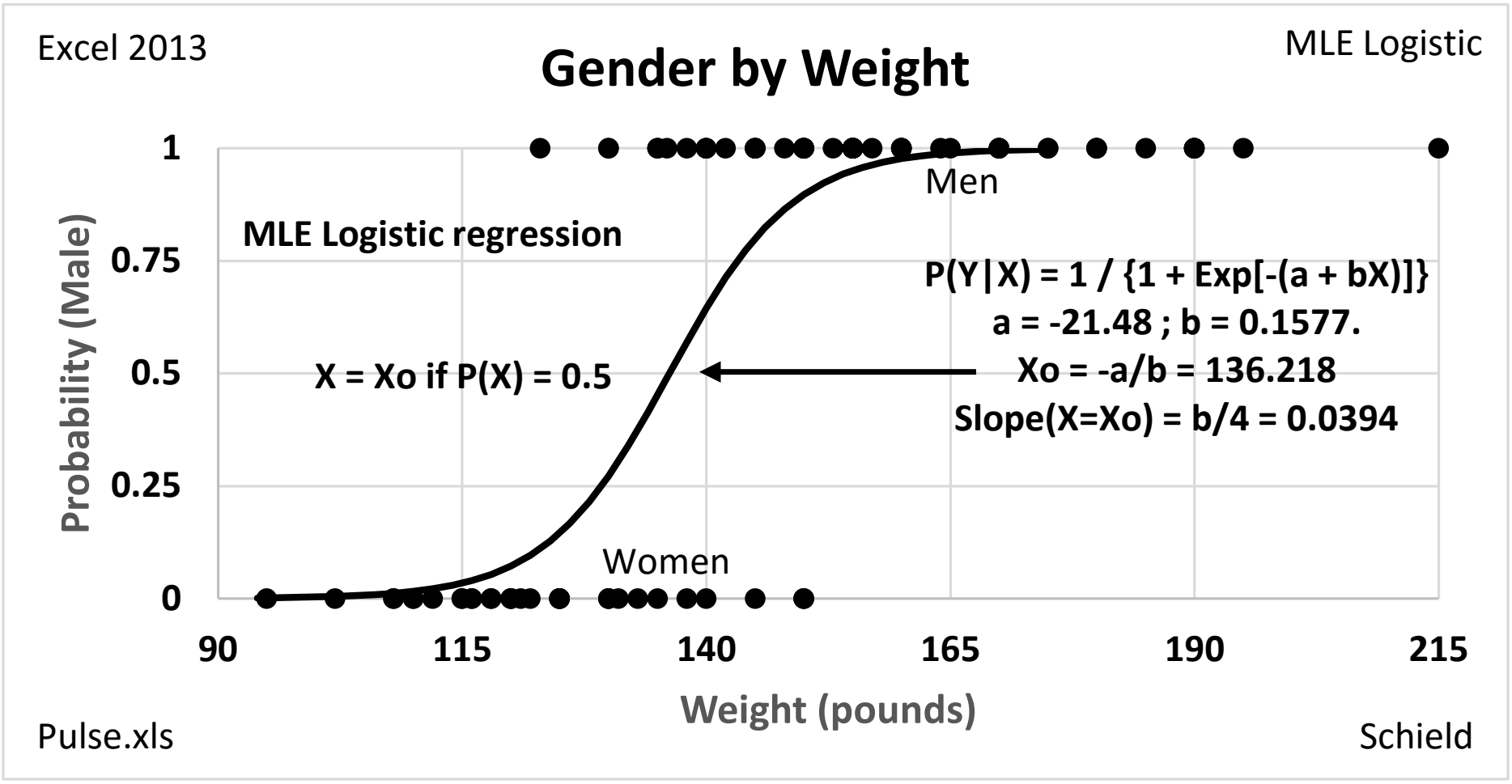


1d: Intersection from OLS(Y|X) Get slope from OLS(X|Y)

Very nice estimate!

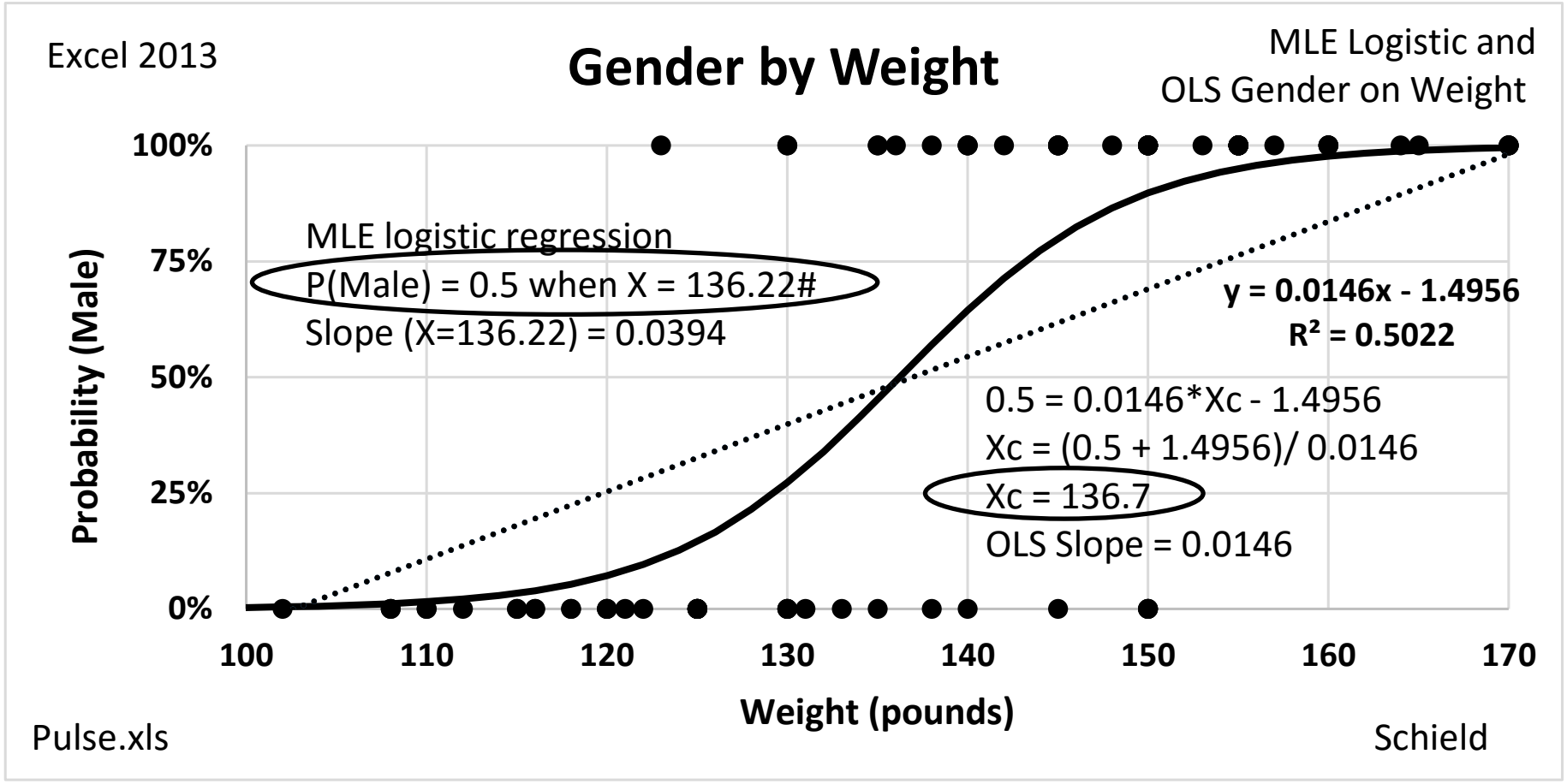


2a: MLE Logistic Regression of Gender by Weight.



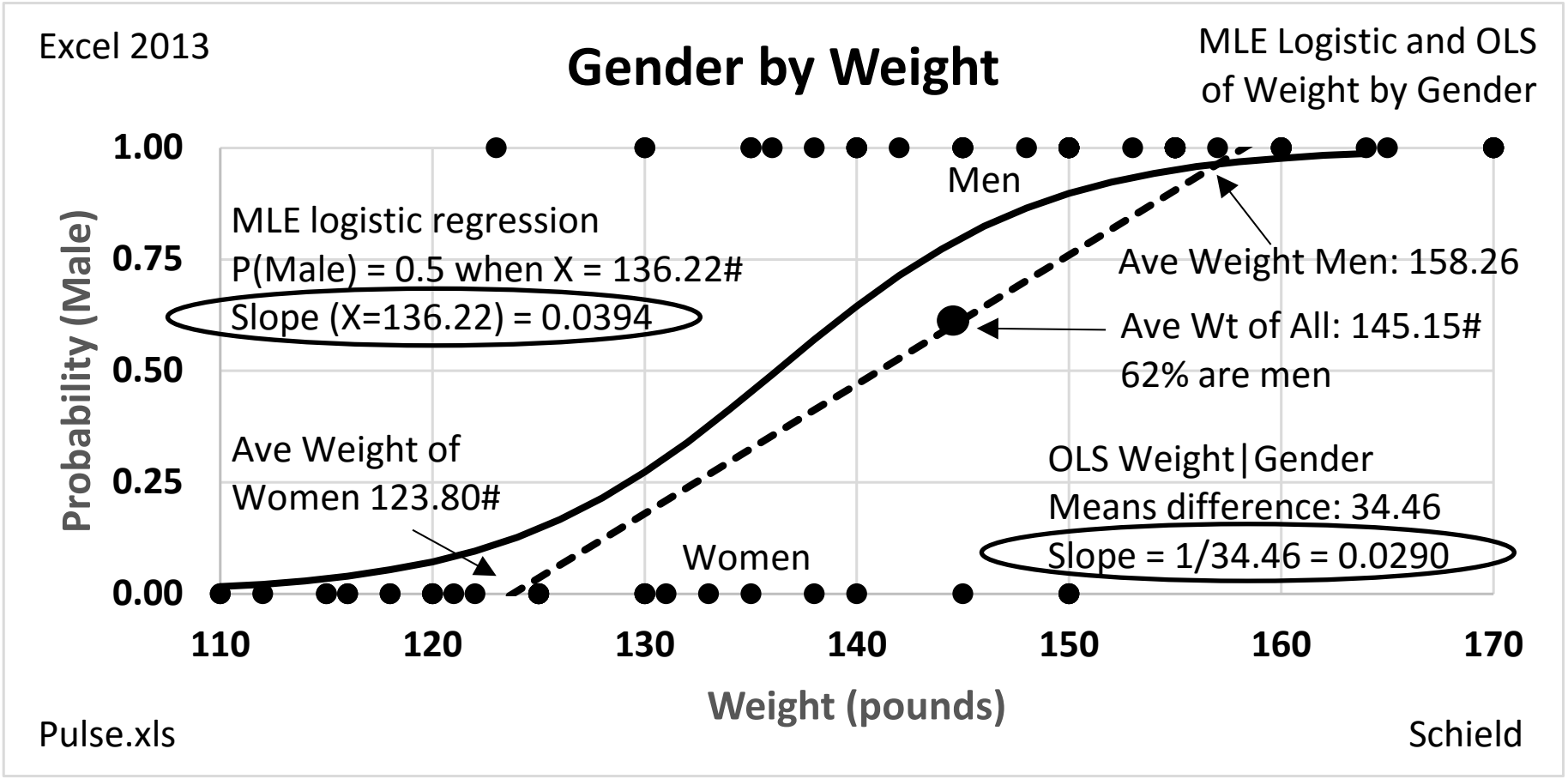
2b: MLE Logistic Regression vs OLS(Y|X) of Gender on Weight

Different slopes but they intersect near P(Y) = 0.5



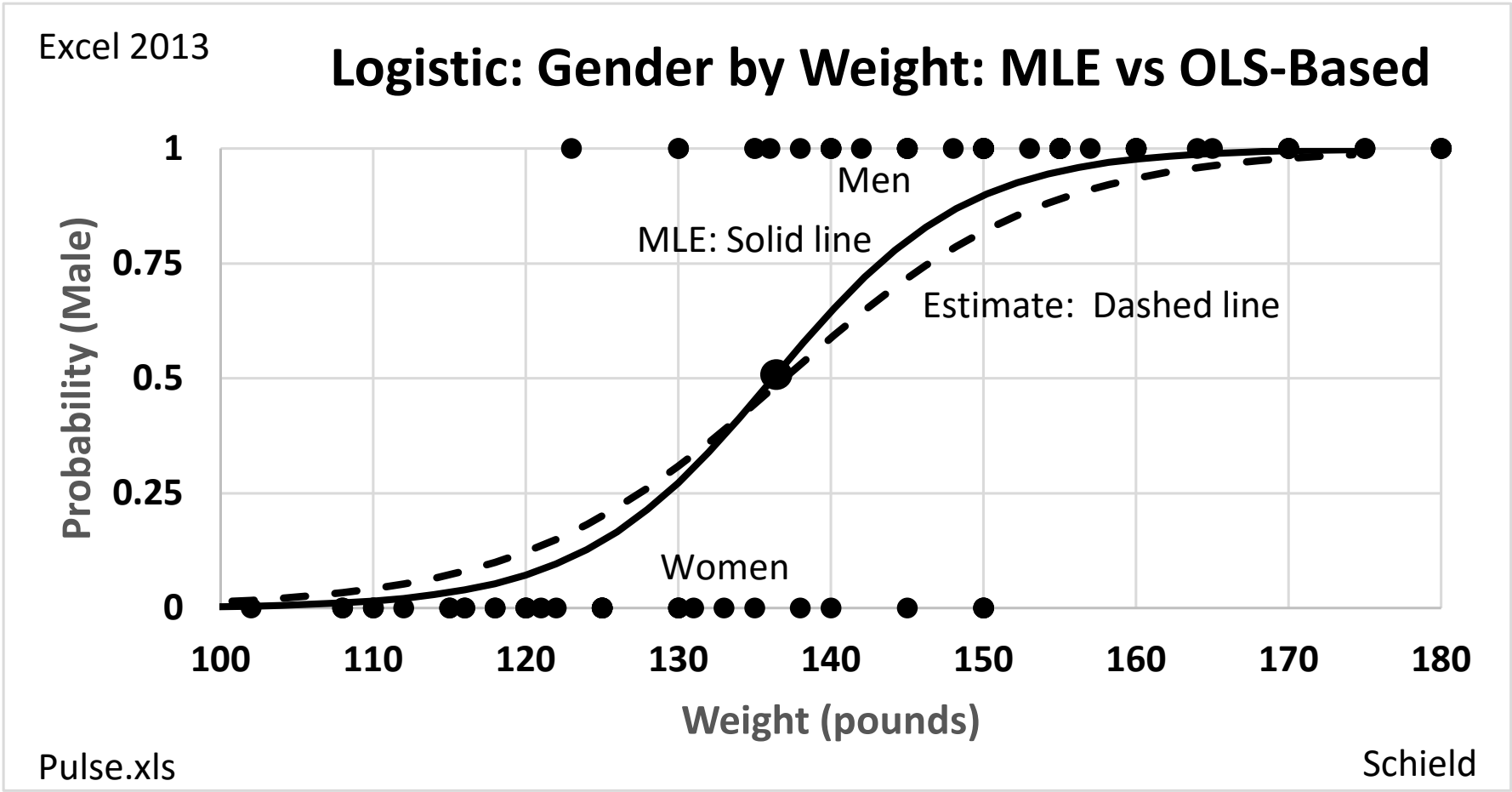
2c: MLE Logistic Regression vs OLS(X|Y) of Weight on Gender

Similar slopes near P(Y) = 0.5



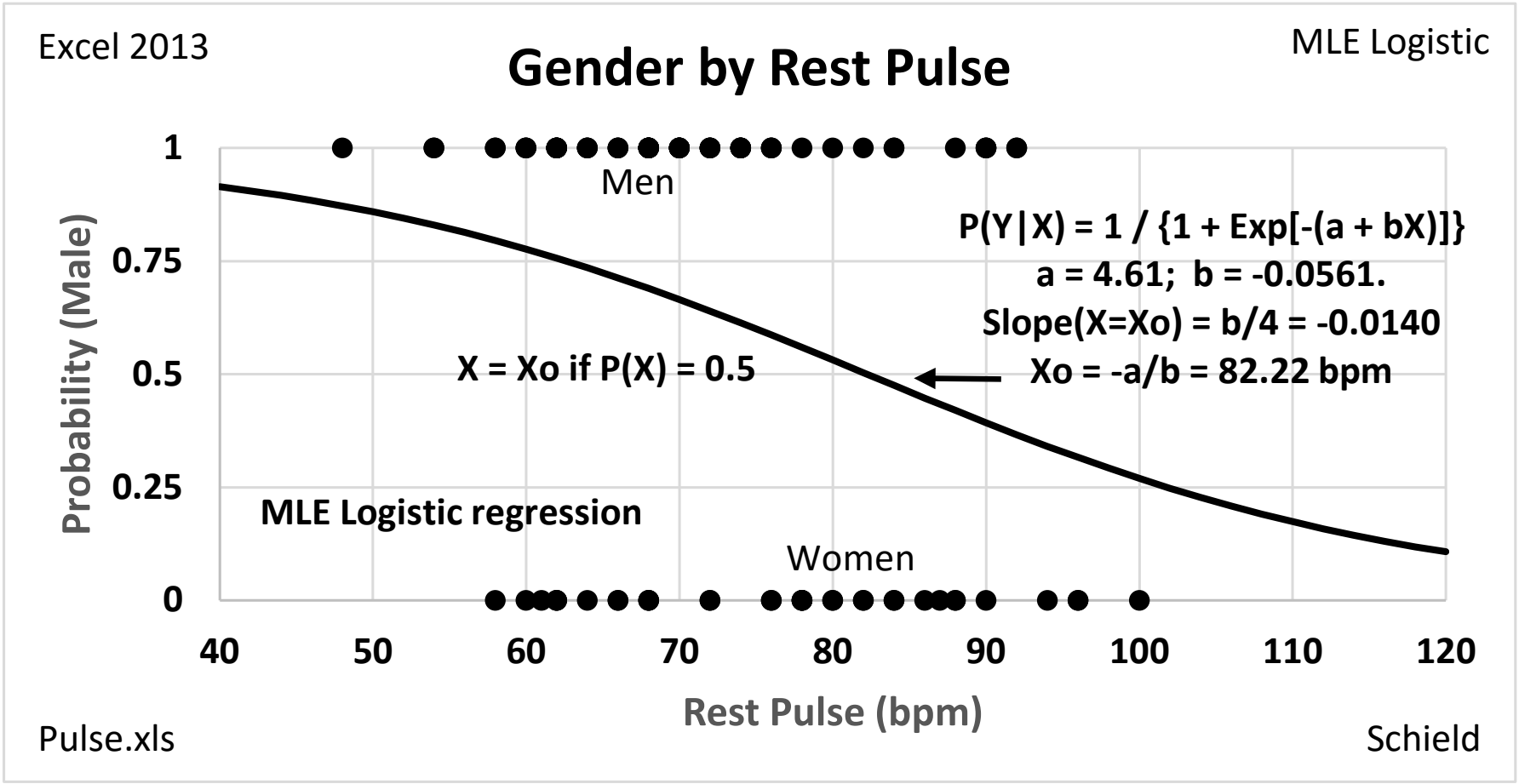
2d: Intersection from OLS(Y|X) Get slope from OLS(X|Y)

Fairly good estimate



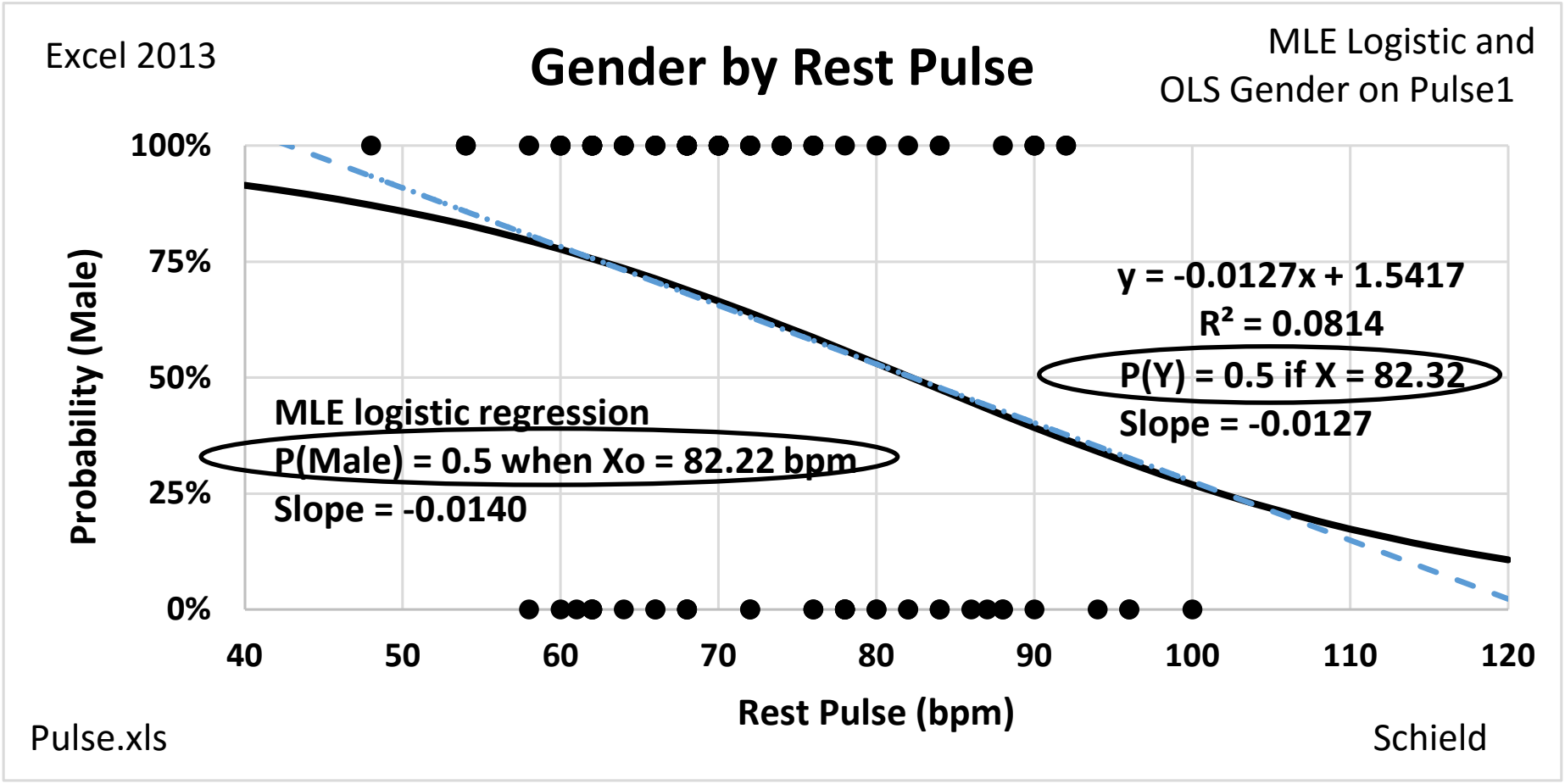
3a: MLE Logistic Regression of Gender by Rest Pulse

Almost flat! Discriminatory power is weak.



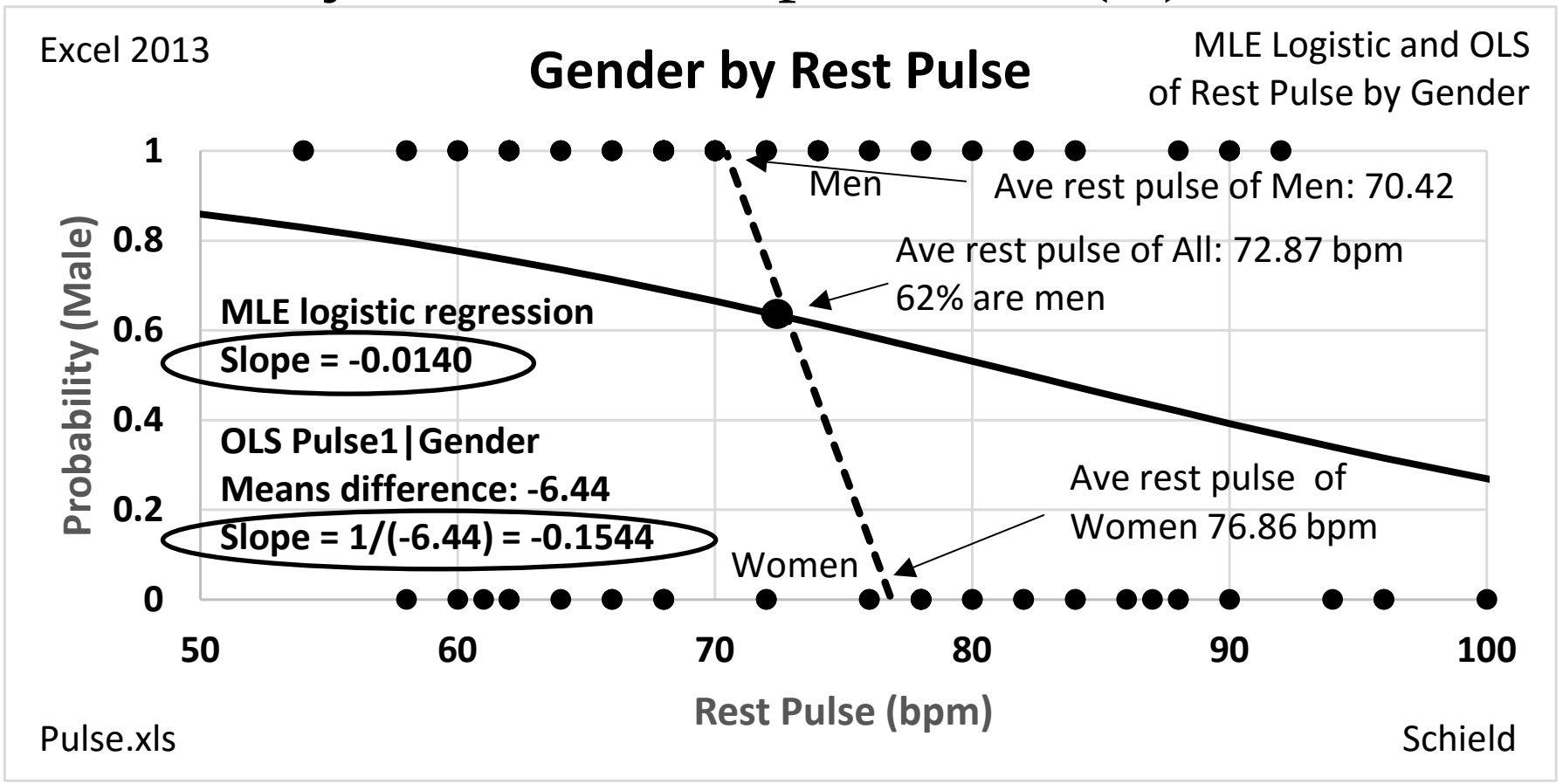
3b: MLE Logistic Regression vs OLS(Y|X) of Gender on Pulse1

MLE and OLS1 match near P(Y|X) = 0.5



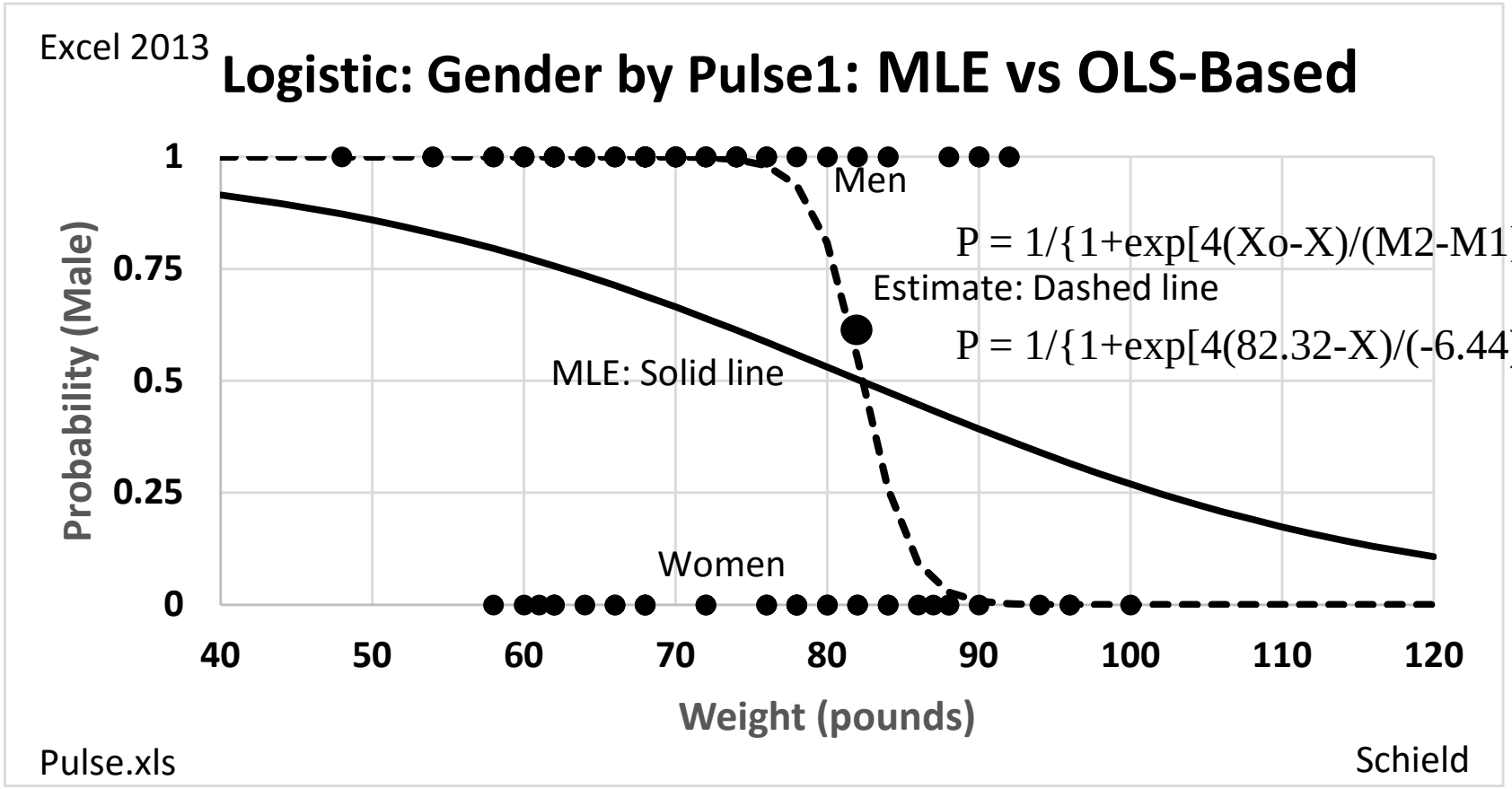
3c: MLE Logistic Regression vs OLS(X|Y) of Gender on Pulse1

Very dissimilar slopes near P(Y) = 0.5



3d: Intersection from OLS(Y|X); Get slope from OLS(X|Y)

Very bad estimate



Analysis and Conclusion Model Gender:

#1: $R^2 = 51\%$ by Height. "Good estimate"

#2: $R^2 = 50\%$ by Weight. "Fair estimate"

#3: $R^2 = 8\%$ by Rest Pulse. "Bad estimate"

Conclusion #1: Using OLS(Y|X) for Xo and using OLS(X|Y) for the associated slope works fairly well when the overlap is small or moderate: OLS(Y|X). R-squared is high: > 0.5

Conclusion #2: Must use MLE when the overlap is large: OLS(Y|X) R-squared is low (< 0.5). But why bother if the model explains so little?

References & Derivation

Schield, Milo (2014a). www.statlit.org/pdf/
Model-Logistic-MLE1A-Excel2013-Slides.pdf
Model-Logistic-MLE1C-Excel2013-Slides.pdf

Schield, Milo (2014b). www.statlit.org/pdf/
2014-Schield-Logistic-OLS3-Excel2013-Slides.pdf

Ln(Odds) = a+bX. P/(1-P) = exp(a+bX)

P = 1 / [1 + exp(-a –bX)]. If P=1/2 at X=Xo, a = -bXo.

P = 1 / {1 + exp[b(Xo – X)]}. b = 4*(dp/dx)|X=Xo.

P = 1 / {1 + exp[4*(Xo-X)*Slope]}

If slope = 1/(M2-M1), P = 1/{1+exp[4(Xo-X)/(M2-M1]}