V1F 2017 ASA 1

Logistic Regression using Excel OLS with Nudge

Milo Schield, Augsburg College

Elected Member: International Statistical Institute

US Rep: International Statistical Literacy Project

VP. National Numeracy Network

JSM Philadelphia

July 31, 2017

www.StatLit.org/pdf/2017-Schield-ASA-Slides.pdf

V1F 2017 ASA 2

Logistic Regression (LR) is Common and Important

Yes/No decisions (binary outcomes) are common in

- Marketing: Predicting whether someone will buy
- Finance: Deciding whether to grant a loan
- Medicine: Determining whether one has a condition
- Epidemiology: Identifying related factors to an outcome

Logistic regression is the most common way of modelling binary outcomes. It is one of the main topics in Stat 200.

It is almost never taught in Stat 100.

But it should be!!!

V1F 2017 ASA 3

Why Isn't Logistic Regression Taught in Intro Course?

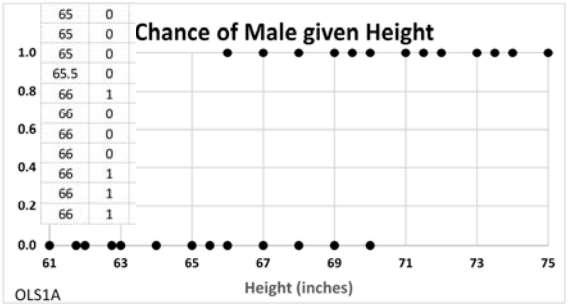
LR isn't taught in Stat 100 for several reasons:

1. Complexity: Maximum likelihood estimation is complex as are odds, log-odds and quality measures.
2. Availability: Not available in Excel or on calculators.
3. Infinity: |Log(Odds)| goes to infinity when p=0 or p=1
4. Non-analytic: Requires trial & error to find best solution.
5. Time: No extra time for extra topics in Intro Statistics.

V1F 2017 ASA 4

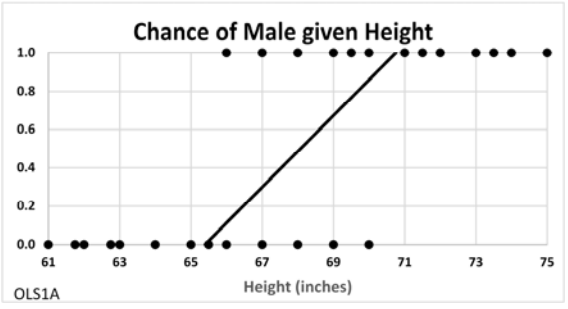
The Data: Height and Gender

65	0
65	0
65	0
65.5	0
66	1
66	0
66	0
66	0
66	1
66	1
66	1



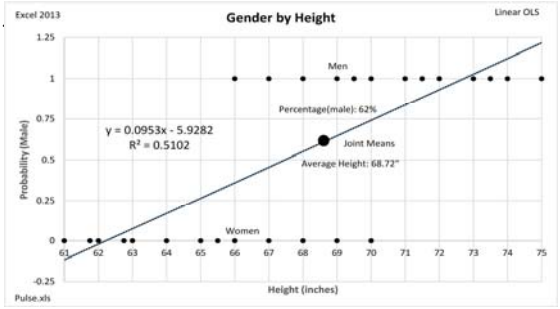
V1F 2017 ASA 5

Simple Model #1: Connect the Mean Heights



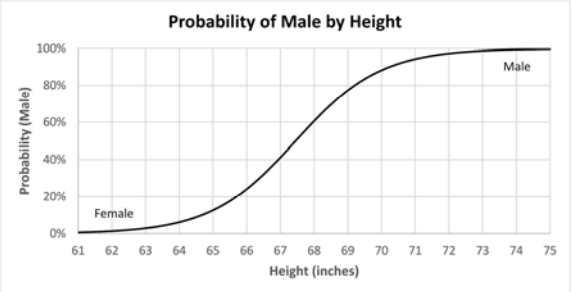
V1F 2017 ASA 6

Simple Model #2: Linear



V1F 2017 ASA 7

Simple Model #3: Logistic Curve



V1F 2017 ASA 8

Simple Solution #4

This simple solution involves two shortcuts:

1. Use the logistic function, but nudge the zero-one data to be epsilon and one minus epsilon. This 'nudge' eliminates the infinities in Ln[Odds(p)].
2. Use the Ordinary Least Squares (OLS) in place of Maximum Likelihood Estimation (MLE). This eliminates the need for industrial-strength software.

Benefits: This allows more attention to the results and to subsequent topics such as confounding and classification.

V1F 2017 ASA 9

Ln[Odds(Nudged Prob)]

A	B	C	D	E	F	G
Predict chance of being male given height. Regress using						
C7	=IF(B7=0, 0.001, 0.999)			E7	=LN(D7)	
D7	=C7/(1-C7)					
Height	Male	Male1	Odds	LN(Odds)	yPred	6
61	0	0.001	0.001	-6.91		7
61.75	0					8
62	0					9

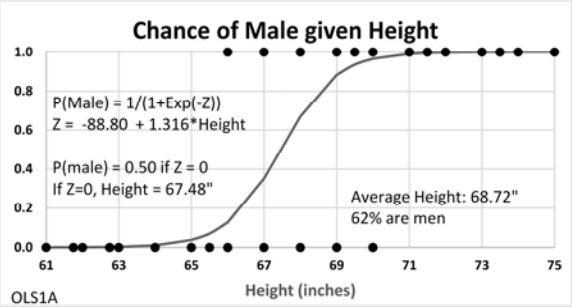
V1F 2017 ASA 10

OLS Results: Regress Gender on Height

17	SUMMARY OUTPUT			
18				
19	*Regression Statistics*			
20	Multiple R	0.7142818		
21	R Square	0.5101985		
22	Adjusted R Square	0.5047563		
23	Standard Error	4.745373		
24	Observations	92		
32		*Coefficients*	*ndard Err*	*t Stat*
33	Intercept	-88.79665	9.354652	-9.49224
34	Height	1.3162354	0.135942	9.682351

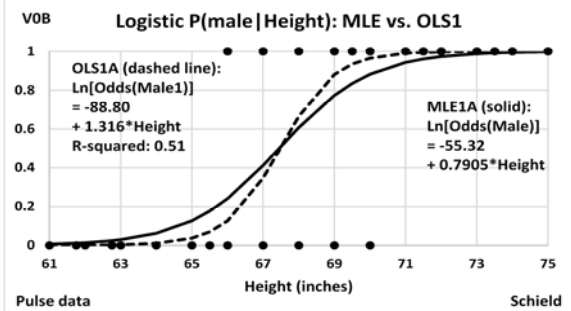
V1F 2017 ASA 11

Translate back to P-space; Plot Probability vs. Height



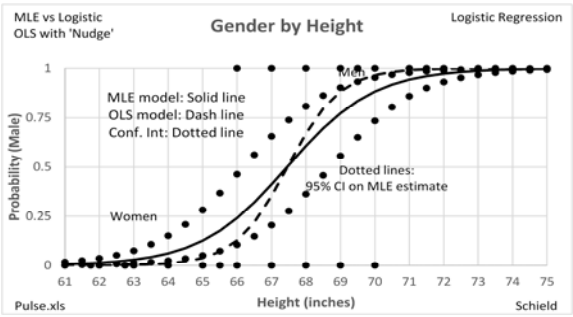
V1F 2017 ASA 12

How close is OLS to MLE? Height: Fairly close…



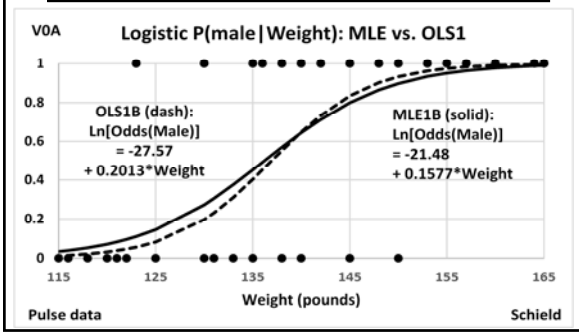
V1F 2017 ASA 13

MLE vs. OLS+Nudge: Significant Difference? No!



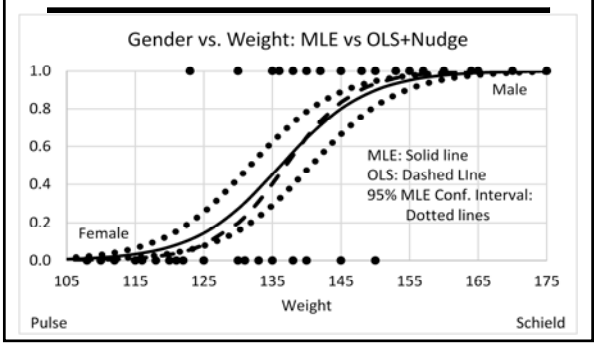
V1F 2017 ASA 14

How close is OLS to MLE? Weight: Fairly close…



V1F 2017 ASA 15

MLE vs. OLS+Nudge: No Significant Difference



V1F 2017 ASA 16

Quotes

"the maximized log likelihood method has always impressed me as an exercise in excessive fine-tuning, reminiscent on some occasions of what Alfred North Whitehead identified as *the fallacy of misplaced concreteness*, and on others of what Freud described as *the narcissism of small differences*."

Comparing exact MLE with OLS regression of Ln[Odds(p)] where p is for grouped data: "The second reason is that in most real-world cases there is *little if any practical difference between the results of the two methods*."

Richard Lowry, Vassar. http://vassarstats.net/logreg1.html

V1F 2017 ASA 17

Recommendation

Those teaching intro statistics needs to think broadly.

Going deeper is good for those who plan to continue on. But almost none of those taking Stat 101 will take Stat 201.

Introducing logistic regression using OLS is simple. The difference between MLE and OLS may not be significant .

Introducing logistic regression in STAT 101 opens the door for other multivariate items such as confounding, classification analysis and discriminant analysis.

V1F 2017 ASA 18

So Why Won't It Be Taught?

OLS is not right in this case.
We don't want to teach our students bad methods.
This OLS+nudge shortcut has a serious lack of rigor.
This is unprofessional; we shouldn't allow it.

Reply:
Lack of rigor vs. rigor mortis?
Can the perfect be the enemy of the good?

What is our goal?
For students to
1. understand some important ideas or
2. be taught correctly even if they don't understand?

V1F 2017 ASA 19

Conclusion

Focus on GAISE 2017 goals.

- Multivariate thinking
- More focus on confounding

See Schield (2016) Offering Stat 102: Social Statistics for Decision Makers.

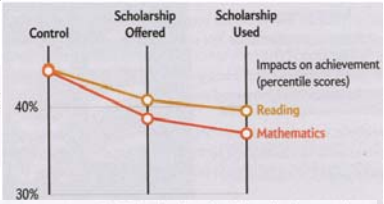
http://www.statlit.org/pdf/2016-Schield-IASE.pdf

V1F 2017 ASA 20

Much More Important Issues Un-Scientific American (2017)

.

Studies show that school vouchers lead to lower math and reading scores.



A recent study of Washington, D.C.'s federally funded voucher program found that math and reading scores among students who used vouchers declined, although the decline in reading scores was not statistically significant.

V1F 2017 ASA 21

Much More Important Issues Un-Scientific American

Three strikes and you are out!

1. Association is not statistically significant
2. Association is not materially significant
3. Author knows that both of these are true, yet puts the association in the headline to the story

Moral: Statistical educators need to put more attention on misuses of statistics in the everyday media. To do less is professional negligence.

V1F 2017 ASA 22

Bibliography

Carlberg, Conrad (2012). Decision Analytics: Microsoft Excel. Que Publishing.

Lowry, R. (2017). E-mail http://vassarstats.net/logreg1.html

Moore, David (2001). Statistical Literacy and Statistical Competence in the New Century. *IASE Proceedings*. http://iase-web.org/documents/papers/sat2001/Moore.pdf

Schield, Milo (2017). Tools at www.StatLit.org/tools.htm

Schield, Milo (2016). Logistic Regression using Minitab and Pulse dataset. http://www.statlit.org/pdf/2016-Minitab-MLE1-Test1.pdf

Logistic Regression using Excel OLS with Nudge

Milo Schield, Augsburg College

Elected Member: International Statistical Institute

US Rep: International Statistical Literacy Project

VP. National Numeracy Network

JSM Philadelphia

July 31, 2017

www.StatLit.org/pdf/2017-Schield-ASA-Slides.pdf

Logistic Regression (LR) is Common and Important

Yes/No decisions (binary outcomes) are common in

- Marketing: Predicting whether someone will buy
- Finance: Deciding whether to grant a loan
- Medicine: Determining whether one has a condition
- Epidemiology: Identifying related factors to an outcome

Logistic regression is the most common way of modelling binary outcomes. It is one of the main topics in Stat 200.

It is almost never taught in Stat 100.

But it should be!!!

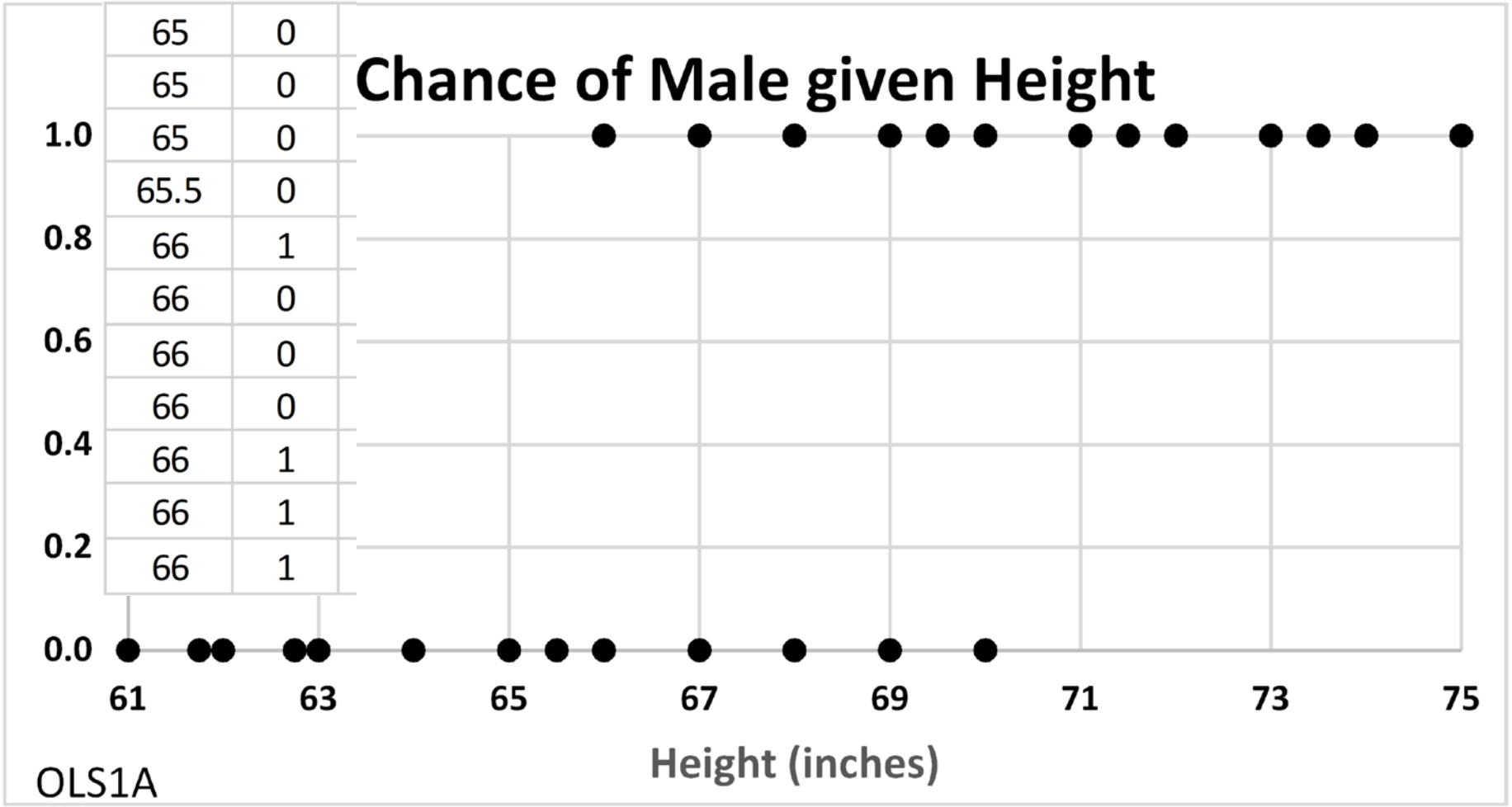
Why Isn't Logistic Regression Taught in Intro Course?

LR isn't taught in Stat 100 for several reasons:

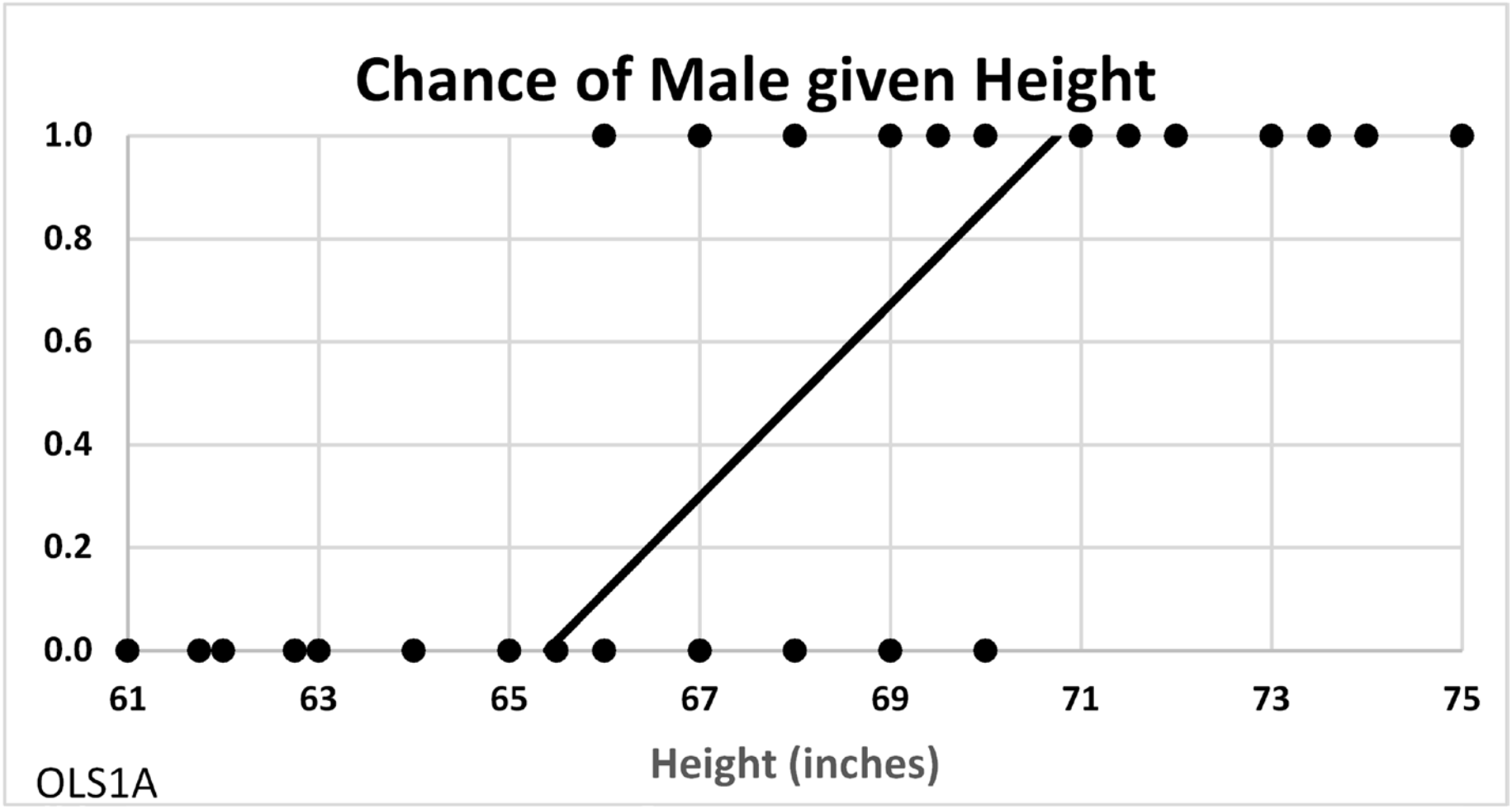
1. Complexity: Maximum likelihood estimation is complex as are odds, log-odds and quality measures.
2. Availability: Not available in Excel or on calculators.
3. Infinity: |Log(Odds)| goes to infinity when p=0 or p=1
4. Non-analytic: Requires trial & error to find best solution.
5. Time: No extra time for extra topics in Intro Statistics.

The Data: Height and Gender

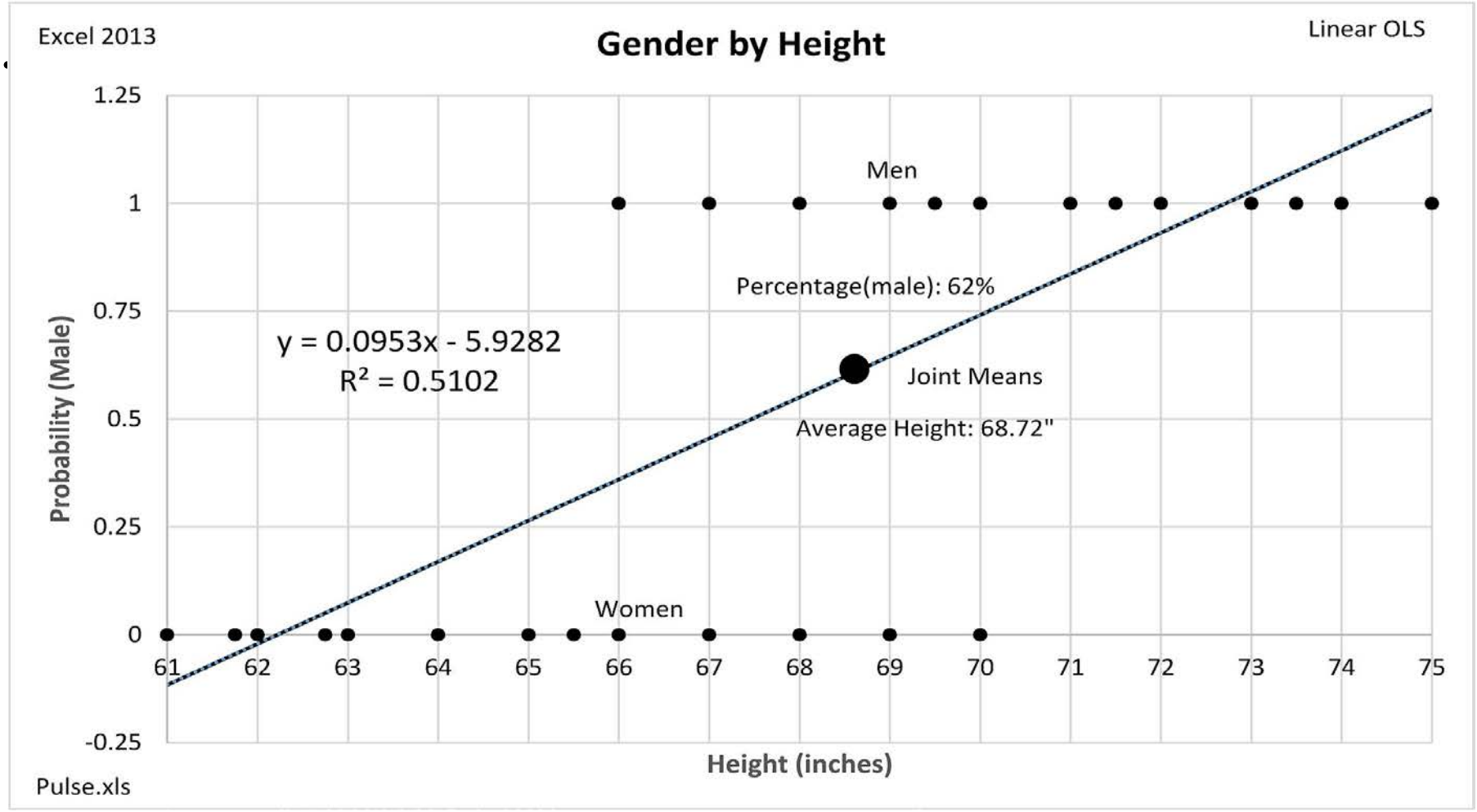
65	0
65	0
65	0
65.5	0
66	1
66	0
66	0
66	0
66	1
66	1
66	1



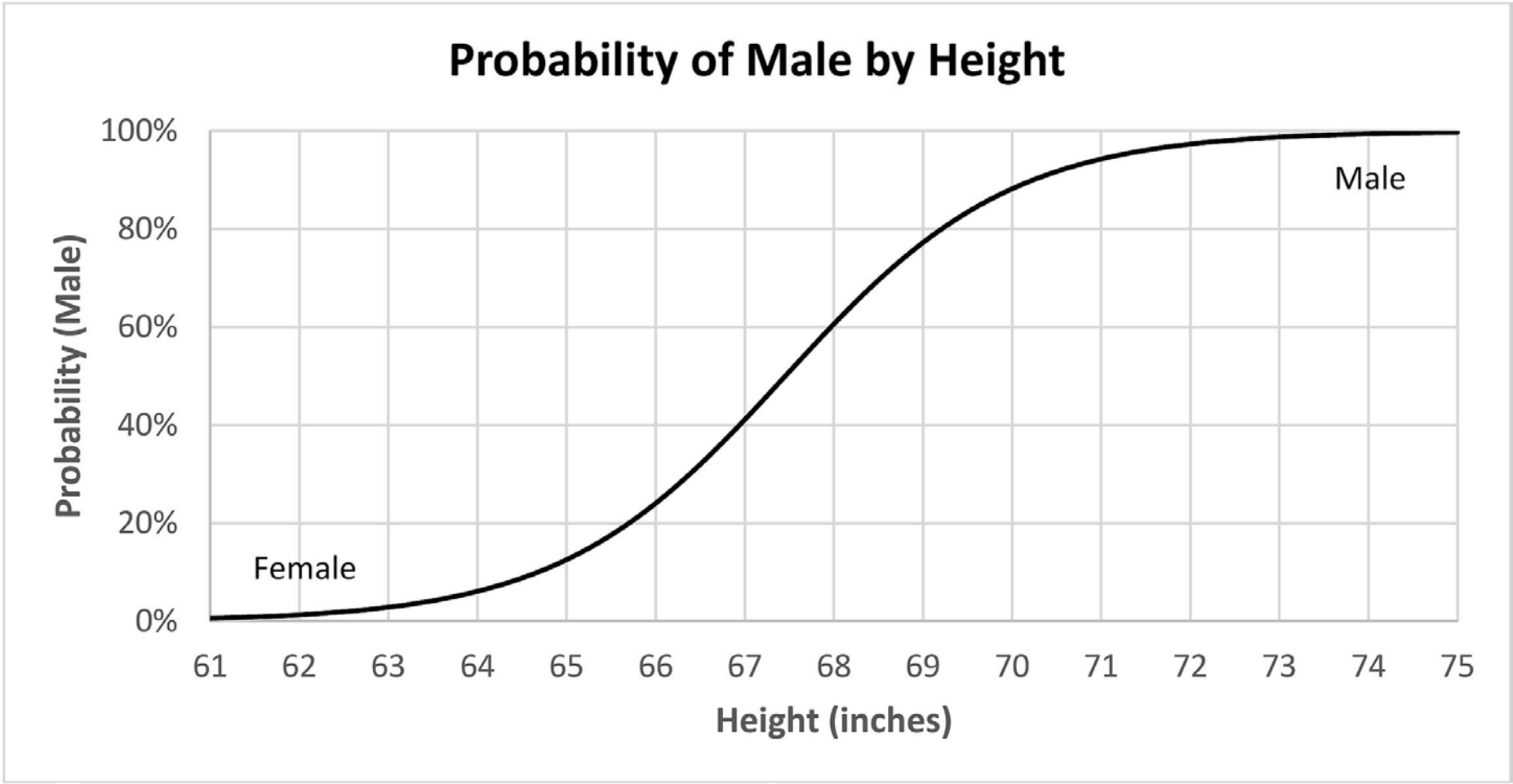
Simple Model #1: Connect the Mean Heights



Simple Model #2: Linear



Simple Model #3: Logistic Curve



Simple Solution #4

This simple solution involves two shortcuts:

1. Use the logistic function, but nudge the zero-one data to be epsilon and one minus epsilon. This 'nudge' eliminates the infinities in Ln[Odds(p)].

2. Use the Ordinary Least Squares (OLS) in place of Maximum Likelihood Estimation (MLE). This eliminates the need for industrial-strength software.

Benefits: This allows more attention to the results and to subsequent topics such as confounding and classification.

Ln[Odds(Nudged Prob)]

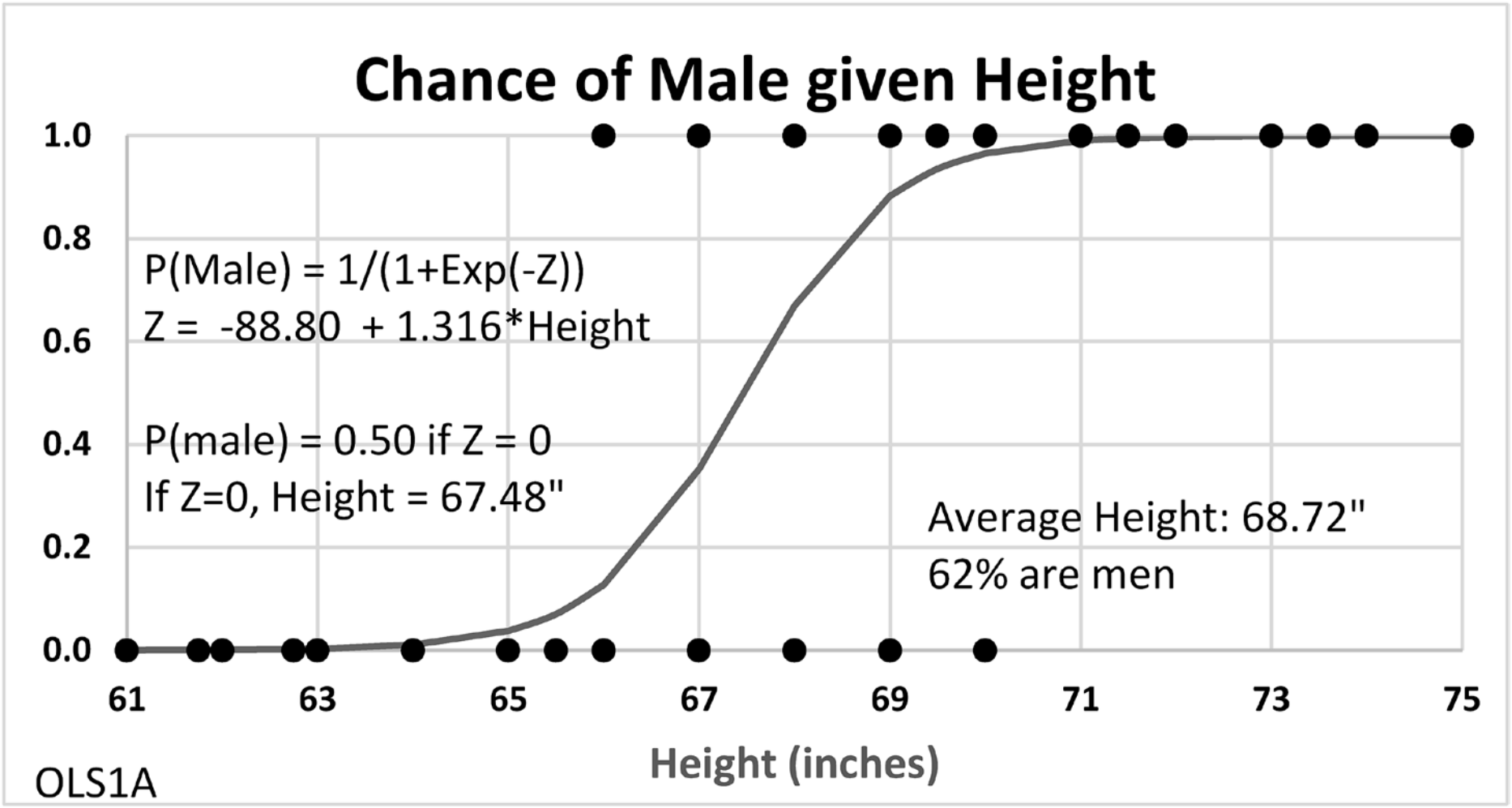
A	B	C	D	E	F	G
Predict chance of being male given height. Regress using						
	C7	=IF(B7=0, 0.001, 0.999)			E7	=LN(D7)
	D7	=C7/(1-C7)				
Height	Male	Male1	Odds	LN(Odds)	yPred	6
61	0	0.001	0.001	-6.91		7
61.75	0					8
62	0					9

OLS Results: Regress Gender on Height

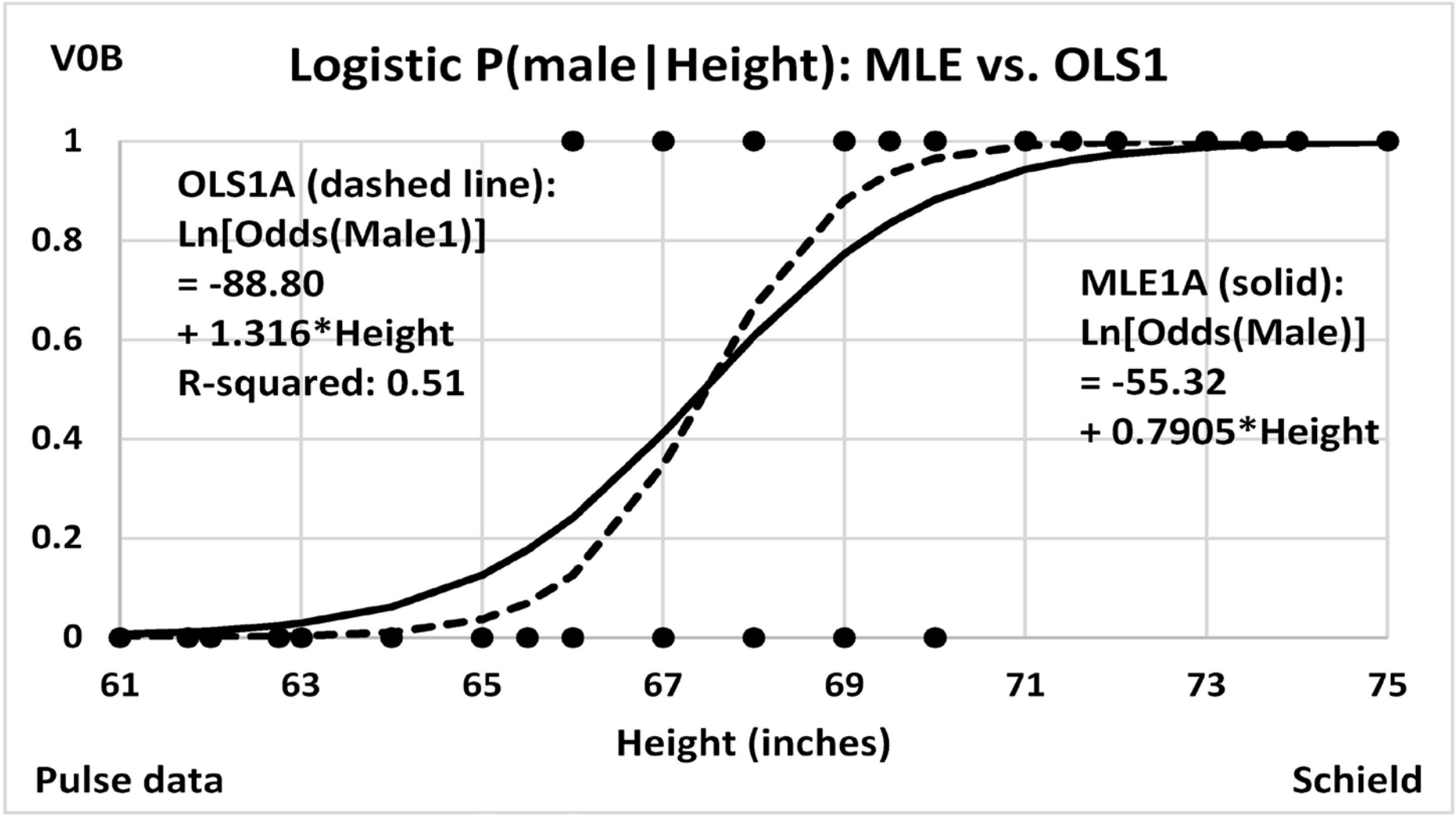
17	SUMMARY OUTPUT	
18		
19	*Regression Statistics*	
20	Multiple R	0.7142818
21	R Square	0.5101985
22	Adjusted R Square	0.5047563
23	Standard Error	4.745373
24	Observations	92

32		*Coefficients*	*ndard Erri*	*t Stat*
33	Intercept	-88.79665	9.354652	-9.49224
34	Height	1.3162354	0.135942	9.682351

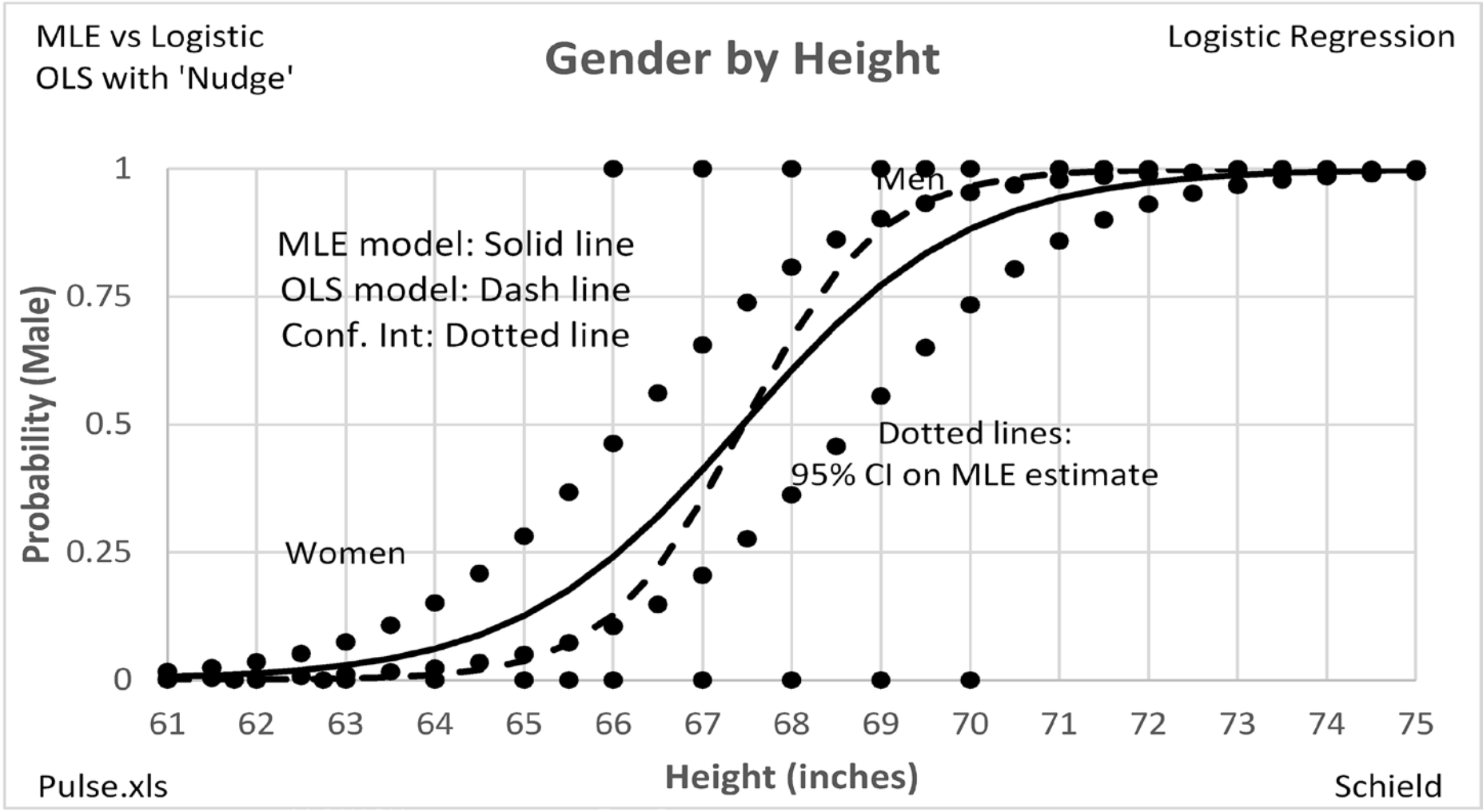
Translate back to P-space; Plot Probability vs. Height



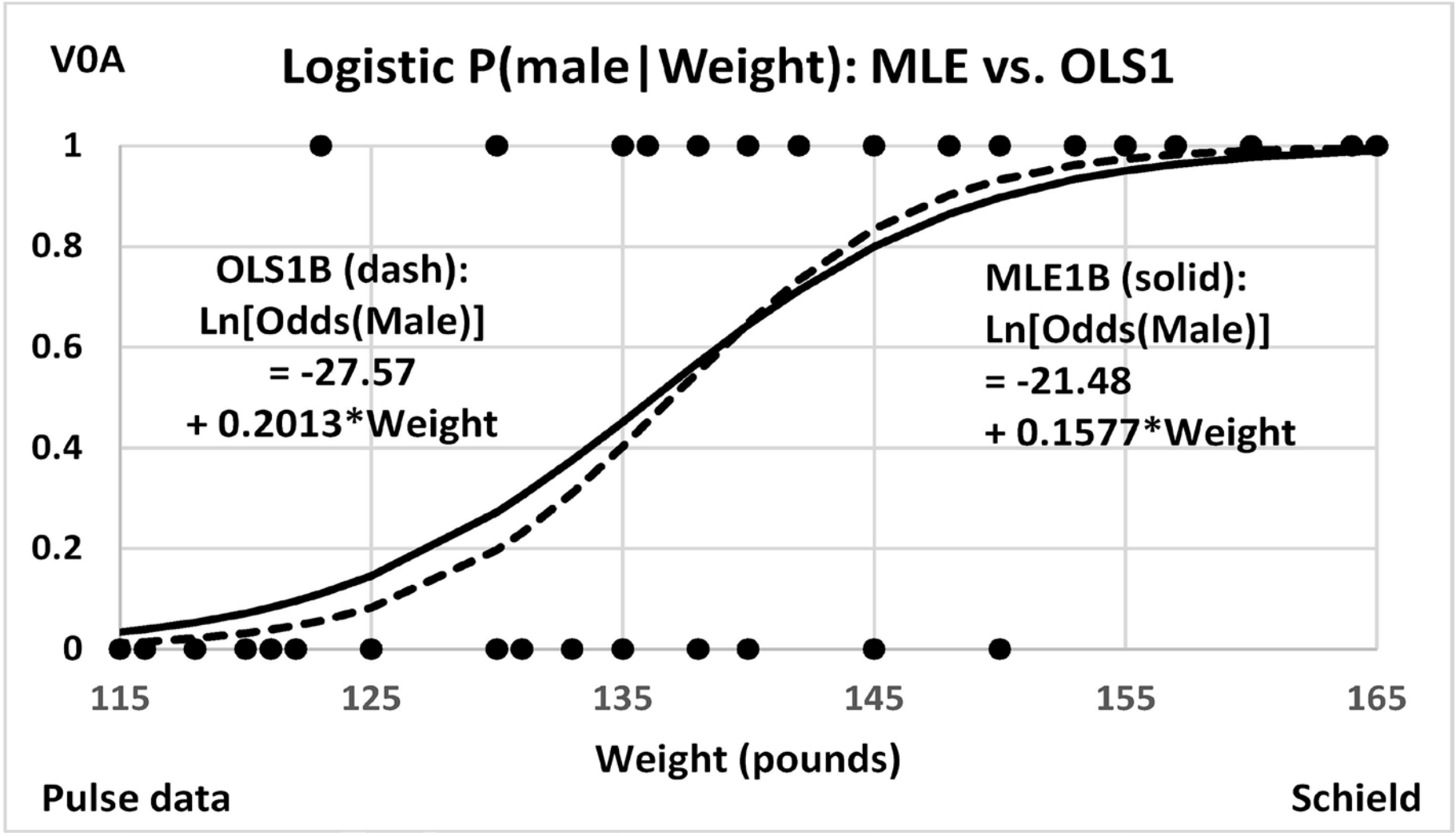
How close is OLS to MLE? Height: Fairly close…



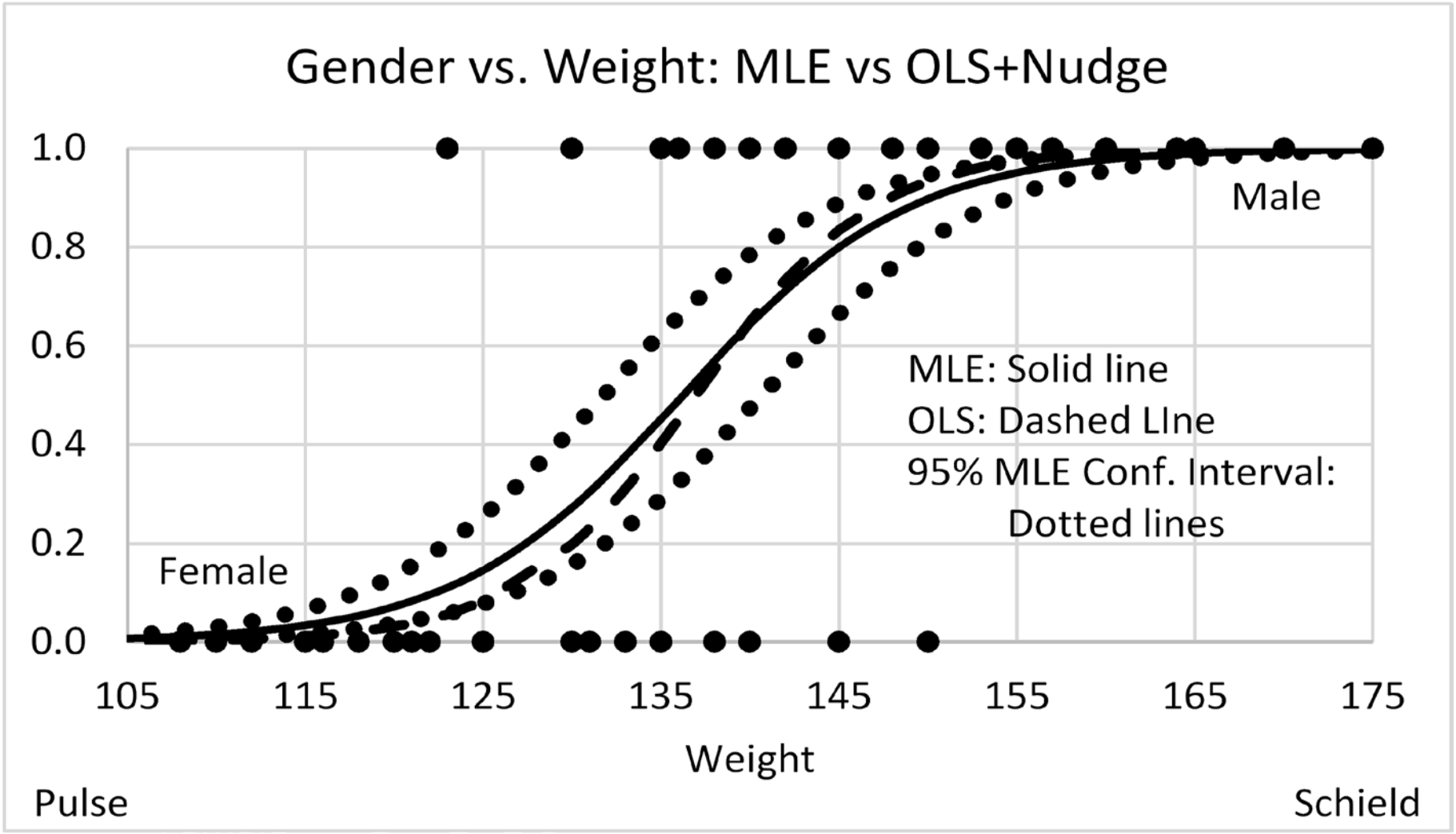
MLE vs. OLS+Nudge: Significant Difference? No!



How close is OLS to MLE? Weight: Fairly close…



MLE vs. OLS+Nudge: No Significant Difference



Quotes

"the maximized log likelihood method has always impressed me as an exercise in excessive fine-tuning, reminiscent on some occasions of what Alfred North Whitehead identified as *the fallacy of misplaced concreteness*, and on others of what Freud described as *the narcissism of small differences*."

Comparing exact MLE with OLS regression of Ln[Odds(p)] where p is for grouped data: "The second reason is that in most real-world cases there is *little if any practical difference between the results of the two methods*."

Richard Lowry, Vassar. http://vassarstats.net/logreg1.html

Recommendation

Those teaching intro statistics needs to think broadly.

Going deeper is good for those who plan to continue on. But almost none of those taking Stat 101 will take Stat 201.

Introducing logistic regression using OLS is simple. The difference between MLE and OLS may not be significant .

Introducing logistic regression in STAT 101 opens the door for other multivariate items such as confounding, classification analysis and discriminant analysis.

So Why Won't It Be Taught?

OLS is not right in this case.
We don't want to teach our students bad methods.
This OLS+nudge shortcut has a serious lack of rigor.
This is unprofessional; we shouldn't allow it.

Reply:
Lack of rigor vs. rigor mortis?
Can the perfect be the enemy of the good?

What is our goal?
For students to
1. understand some important ideas or
2. be taught correctly even if they don't understand?

Conclusion

Focus on GAISE 2017 goals.

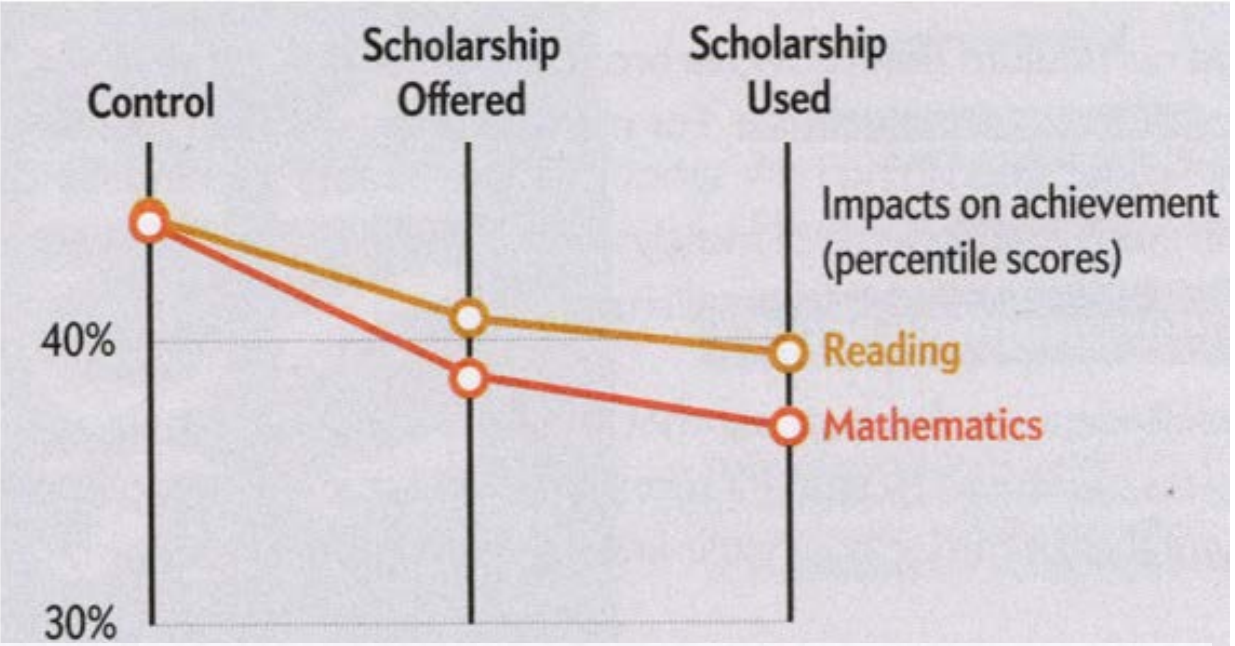
- Multivariate thinking
- More focus on confounding

See Schield (2016) Offering Stat 102: Social Statistics for Decision Makers.

http://www.statlit.org/pdf/2016-Schield-IASE.pdf

Much More Important Issues Un-Scientific American (2017)

- Studies show that school vouchers lead to lower math and reading scores.



A recent study of Washington, D.C.'s federally funded voucher program found that math and reading scores among students who used vouchers declined, although the decline in reading scores was not statistically significant.

Much More Important Issues Un-Scientific American

Three strikes and you are out!

1. Association is not statistically significant
2. Association is not materially significant
3. Author knows that both of these are true,
   yet puts the association in the headline to the story

Moral: Statistical educators need to put more attention on misuses of statistics in the everyday media. To do less is professional negligence.

Bibliography

Carlberg, Conrad (2012). Decision Analytics: Microsoft Excel. Que Publishing.

Lowry, R. (2017). E-mail http://vassarstats.net/logreg1.html

Moore, David (2001). Statistical Literacy and Statistical Competence in the New Century. *IASE Proceedings*. http://iase-web.org/documents/papers/sat2001/Moore.pdf

Schield, Milo (2017). Tools at www.StatLit.org/tools.htm

Schield, Milo (2016). Logistic Regression using Minitab and Pulse dataset. http://www.statlit.org/pdf/2016-Minitab-MLE1-Test1.pdf