Logistic Regression using MLE (1A) and 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/2015-Schield-Logistic-MLE1A-Demo.pdf
pdf/2015-Schield-Logistic-MLE1A-Slides.pdf
xls/2015-Schield-Logistic-MLE1A-Data.xlsx

Background & Goals

Modelling a binary outcome (buy/look, payoff/default, go/nogo or male/female) requires logistic regression.

Doing logistic regression in Excel requires Solver. "Since its introduction in .. 1991, … Excel Solver has become the most widely distributed – and almost surely the most widely used – general-purpose optimization modeling system." www.utexas.edu/courses/lasdon/design3.htm

This presentation uses college student data: pulse.xls. This demo models gender (male) based on height.

Goals: Create graph on slide 20.
Determine if slope is statistically significant.

This demo uses Height (col A) to predict Gender (col B)

Column B: 0=Female, 1 = Male (circled)

Ave Heights:

M: 70.75" 62%

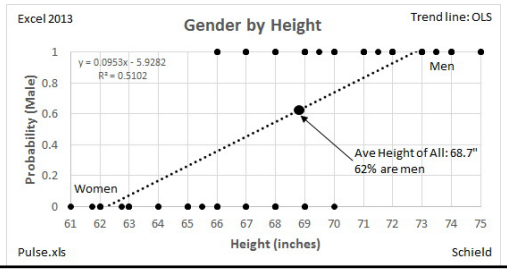
F: 65.3" 38%

Difference: 5.35"

A	B
Height	Male
68	1
69	1
69	1
72	1
66	1
67	0
71	1
71	1
71.5	1
62	0
65.5	0

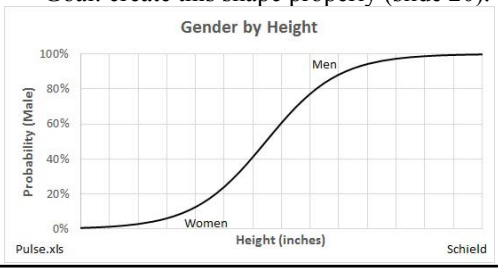
Model Gender by Height. Show Trend, Eq. and Joint Mean.

This invalid trend-line intersects the joint mean.
Insert circle at joint means; insert mean values in textbox.



Linear Trendline is invalid. Intuitive idea of solution

No need to create this graph.
Goal: create this shape properly (slide 20).



Four Step Approach

1) Insert intercept #1 with slope = 0. Record the sum of the errors: the logs of the chance (the likelihood) that the estimate is OK.
2) Solve for intercept & slope using SOLVER; Record the sum of the errors for this model.
3) Test the slope for statistical significance.
4) Generate graphs.

To do: Get data at www.StatLit.org/Excel/ 2015-Schield-Logistic-MLE1A-Excel2013-Data.xlsx

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 7

1a) Get Data; Find Mean(Y). Set Intercept #1 and Slope #1.

#1: Enter formula for E21 and E22.

	D	E	F
19			
20	GENDER & INTERCEPT #1		
21	Male-Pctg	0.62	=AVERAGE(B3:B94)
22	Intercept#1	0.4877	=LN(E21/(1-E21))

#2: Copy value from E22 into D3. Set E3=0.

Row	D	E	F
2	Intercept	Slope	
3	0.4877	0.0000	

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 8

1b) Enter formula for G3:K3. Select G3:K3; pull down to row 94

Row	D	E	F	G	H	I	J	K
2	Intercept	Slope		Logit	Odds	Prob Y=1	Prob OK	Ln-LH-C
3	0.4877	0.0000	See slides 7 and 13	0.49	1.63	0.62	0.62	-0.48
4								
5	Sum LnLk	-0.48	=SUM(K3:K94)					
6	Sum LnLk1		See slide 10					
7	Sum Ln Lk2		See slide 14					
8	Chi-Sq	0.00	=-2*(E6-E7)					
9	P-Value	1	=CHISQ.DIST.RT(E8,1)					
10								
11	FORMULAS & TEXT: Enter, Copy Down							
12	Logit	G3	=D$3+E$3*A3					
13	Odds	H3	=EXP(G3)					
14	Prob Y=1	I3	=H3/(1+H3)					
15	Prob OK	J3	=IF(B3=1,I3,1-I3)					
16	Ln-LH-OK	K3	=LN(J3)					

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 9

1c) Results are as expected. Probability of male = 0.62

Row	D	E	F	G	H	I	J	K
2	Intercept	Slope		Logit	Odds	Prob Y=1	Prob OK	Ln-LH-OK
3	0.4877	0.0000	See slides 7 and 13	0.49	1.63	0.62	0.62	-0.48
4				0.49	1.63	0.62	0.62	-0.48
5	Sum LnLk	-61.11	=SUM(K3:K94)	0.49	1.63	0.62	0.62	-0.48
6	Sum LnLk1		See slide 10	0.49	1.63	0.62	0.62	-0.48
7	Sum Ln Lk2		See slide 14	0.49	1.63	0.62	0.62	-0.48
8	Chi-Sq	0.00	=-2*(E6-E7)	0.49	1.63	0.62	0.38	-0.97
9	P-Value	1	=CHISQ.DIST.RT(E8,1)	0.49	1.63	0.62	0.62	-0.48
10				0.49	1.63	0.62	0.62	-0.48
11	FORMULAS & TEXT: Enter, Copy Down			0.49	1.63	0.62	0.62	-0.48
12	Logit	G3	=D$3+E$3*A3	0.49	1.63	0.62	0.38	-0.97
13	Odds	H3	=EXP(G3)	0.49	1.63	0.62	0.38	-0.97
14	Prob Y=1	I3	=H3/(1+H3)	0.49	1.63	0.62	0.62	-0.48
15	Prob OK	J3	=IF(B3=1,I3,1-I3)	0.49	1.63	0.62	0.62	-0.48
16	Ln-LH-OK	K3	=LN(J3)	0.49	1.63	0.62	0.62	-0.48
17				0.49	1.63	0.62	0.38	-0.97
18				0.49	1.63	0.62	0.38	-0.97

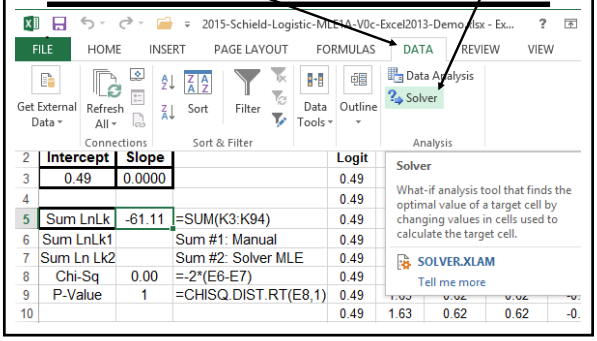
V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 10

1d) Manually: Copy Value of E5 onto E6

Row	D	E	F
2	Intercept	Slope	
3	0.4877	0.0000	See slides 7 and 13
4			
5	Sum LnLk	-61.11	=SUM(K3:K94)
6	Sum LnLk1	-61.11	See slide 10
7	Sum Ln Lk2		See slide 14
8	Chi-Sq	122.22	=-2*(E6-E7)
9	P-Value	2E-28	=CHISQ.DIST.RT(E8,1)
10			

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 11

2a) Solve for Slope and Intercept: From Data menu, select Solver



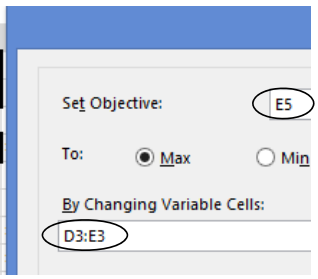
	Intercept	Slope		Logit
3	0.49	0.0000		0.49
4				0.49
5	Sum LnLk	-61.11	=SUM(K3:K94)	0.49
6	Sum LnLk1		Sum #1: Manual	0.49
7	Sum Ln Lk2		Sum #2: Solver MLE	0.49
8	Chi-Sq	0.00	=-2*(E6-E7)	0.49
9	P-Value	1	=CHISQ.DIST.RT(E8,1)	0.49
10				0.49

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 12

2b) Set Solver Parameters. Use GRC Nonlinear. Press Solve

Select Objective Cell (E5) and Variable Cells (D3:E3)

Row	D	E
2	Intercept	Slope
3	0.4877	0.0000
4		
5	Sum LnLk	-61.11
6	Sum LnLk1	-61.11
7	Sum Ln Lk2	
8	Chi-Sq	122.23
9	P-Value	2E-28



V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 13

2c) Results: All constraints & conditions satisfied. Press OK

Row	D	E
2	**Intercept**	**Slope**
3	-53.32	0.7905
4		
5	Sum LnLk	-30.55
6	Sum LnLk1	-61.11
7	Sum Ln Lk2	
8	Chi-Sq	122.23
9	P-Value	2E-28
10		
11	FORMULAS & TEXT	
12	Logit	G3
13	Odds	H3
14	Prob Y=1	I3
15	Prob OK	J3

Solver Results

Solver found a solution. All Constraints and optimality conditions satisfied.

Reports: Answer, Sensitivity, Limits

Keep Solver Solution

Restore Original Values

Return to Solver Parameters Dialog

Outline Rep

OK | Cancel | Save Scen

Solver found a solution. All Constraints and optimality conditions are satisfied.

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 14

2d) Manually: Copy Value of E5 onto E7

Row	D	E	F
2	**Intercept**	**Slope**	
3	-53.3227	0.7905	See slides 7 and 13
4			
5	Sum LnLk	-30.55	=SUM(K3:K94)
6	Sum LnLk1	-61.11	See slide 10
7	Sum Ln Lk2	-30.55	See slide 14
8	Chi-Sq	61.12	=-2*(E6-E7)
9	P-Value	5E-15	=CHISQ.DIST.RT(E8,1)

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 15

3) Hypothesis test: Is non-zero slope statistically significant?

Conduct a right-tail Chi^2 test with 1 degree of freedom.

Sum LnLk	-30.55	=SUM(K3:K94)
Sum LnLk1	-61.11	See slide 10
Sum Ln Lk2	-30.55	See slide 14
Chi-Sq	61.12	=-2*(E6-E7)
P-Value	5E-15	=CHISQ.DIST.RT(E8,1)

Slope is statistically significant: P-value < 0.05

Note: E-15 means the decimal point is 15 places to the left: 0.000 000 000 000 005

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 16

4a) Analyze X axis: Enter formula for V2:V6

U	V	W	X	Y
HEIGHT	**VALUE**	**FORMULA**		
Ht-Average	68.72	=AVERAGE(A3:A94)		
Ht-Max	75.00	=MAX(A3:A94)		
Ht-Min	61.00	=MIN(A3:A94)		
Ht Ave Guy	70.75	=AVERAGEIF(B3:B94,"=1",A3:A94)		
Ht Ave Gal	65.40	=AVERAGEIF(B3:B94,"=0",A3:A94)		

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 17

4b) #1) Set N6 = 61. Enter formula for O6, P6 & Q6

M	N	O	P	Q	R	S
2	**CHART**	N6	Enter X manually		P6	=EXP(O6)
3	**SETUP**	O6	=D$3+E$3*N6		Q6	=P6/(1+P6)
4						
5	**X-Ht**	**Logit**	**Odds**	**Prob Y=1**		
6	61.00	-5.10	0.01	1%		

#2: Set N7 = 61.5; Select O6:Q6. Pull down to O7:Q7 [Row 7]

5	X-Ht	Logit	Odds	Prob Y=1
6	61.00	-5.10	0.01	1%
7	61.50	-4.71	0.01	1%

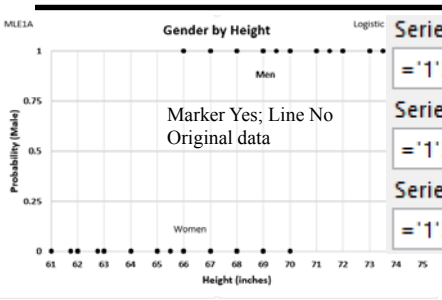
V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 18

4c) Select N6:Q7: Rows 6+7. Drag 2row box to row 34

M	N	O	P	Q
2	**CHART**	N6	Enter X manually	
3	**SETUP**	O6	=D$3+E$3*N6	
4				
5	**X-Ht**	**Logit**	**Odds**	**Prob Y=1**
6	61.00	-5.10	0.01	1%
7	61.50	-4.71	0.01	1%

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 19

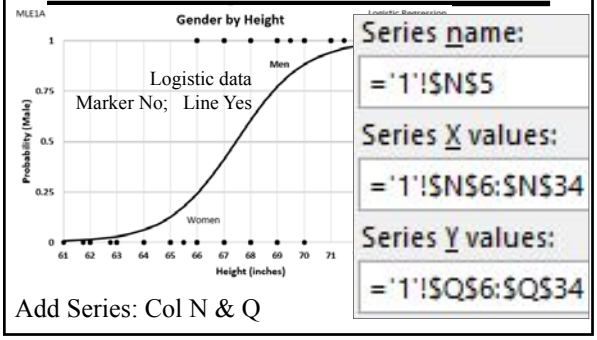
4d) Graph Data on XY Plot: Gender (B) by Height (A)



Original data: Col A & B

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 20

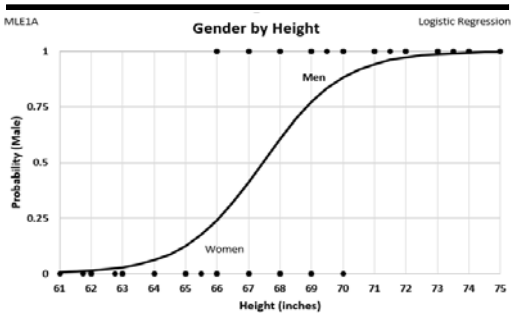
4e) Graph Logistic Regression: Gender (Q) by Height (N)



Add Series: Col N & Q

V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 21

4f) Final Result



V0D 2015 Schield Logistic MLE 1A Excel2013 Slides 22

Acknowledgment and Reference

ACKNOWLEDGMENT:

This presentation closely follows the Carlberg (2012) presentation in Chapter 2: pages 21-52.

These slides present the how – step by step – of logistic regression for a single predictor.

Carlberg (2012) discusses the how and the why.

Schield introduced the shortcut on slides 7 and 8.

REFERENCE:

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

Logistic Regression using MLE (1A) and 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/2015-Schield-Logistic-MLE1A-Demo.pdf
pdf/2015-Schield-Logistic-MLE1A-Slides.pdf
xls/2015-Schield-Logistic-MLE1A-Data.xlsx

Background & Goals

Modelling a binary outcome (buy/look, payoff/default, go/nogo or male/female) requires logistic regression.

Doing logistic regression in Excel requires Solver. "Since its introduction in .. 1991, … Excel Solver has become the most widely distributed – and almost surely the most widely used – general-purpose optimization modeling system." www.utexas.edu/courses/lasdon/design3.htm

This presentation uses college student data: pulse.xls. This demo models gender (male) based on height.

Goals: Create graph on slide 20.
Determine if slope is statistically significant.

This demo uses Height (col A) to predict Gender (col B)

Column B: 0=Female, 1 = Male (circled)

Ave Heights:

M: 70.75” 62%

F: 65.3” 38%

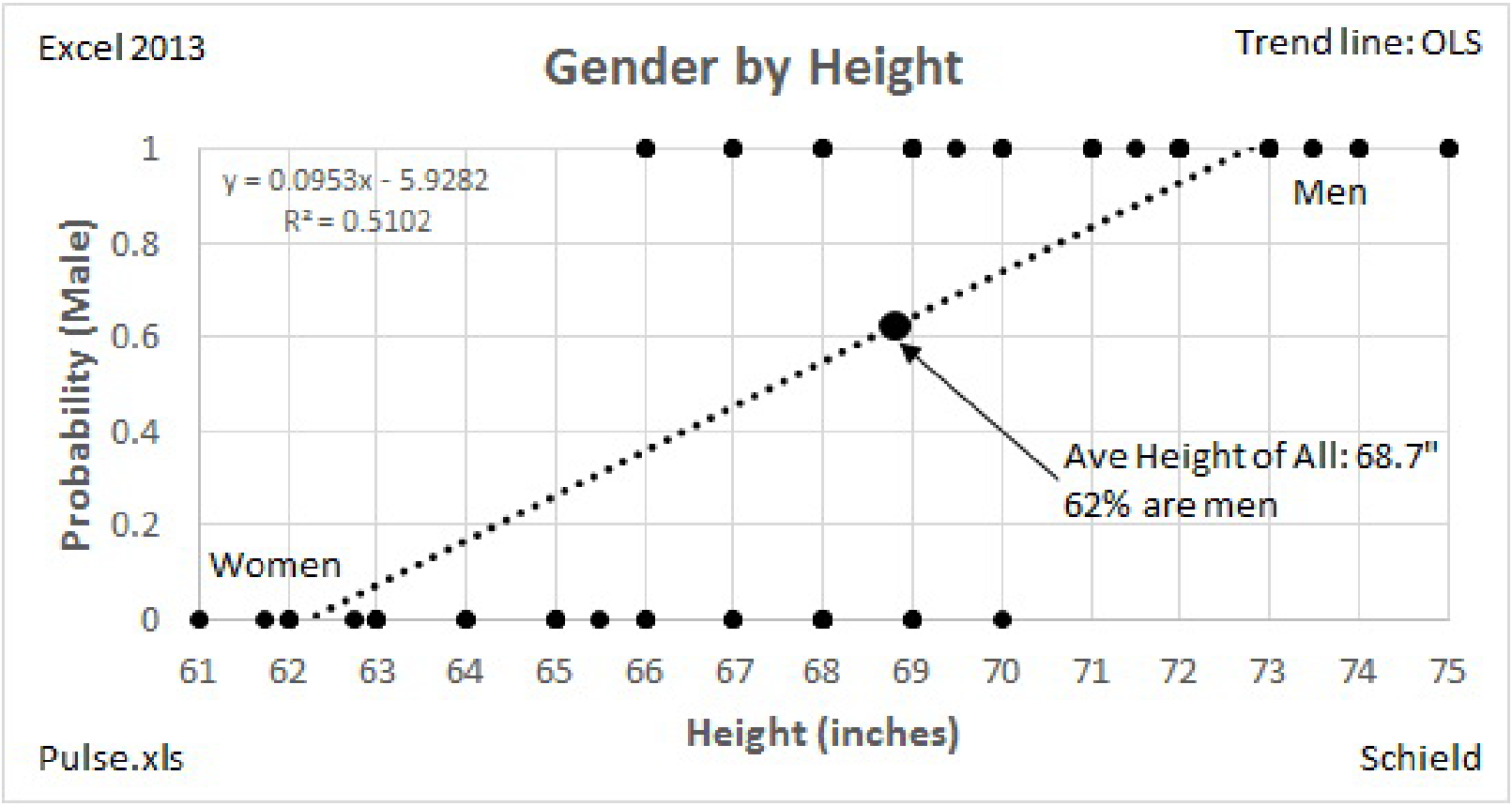
Difference:

5.35”

A	B
Height	Male
68	1
69	1
69	1
72	1
66	1
67	0
71	1
71	1
71.5	1
62	0
65.5	0

Model Gender by Height. Show Trend, Eq. and Joint Mean.

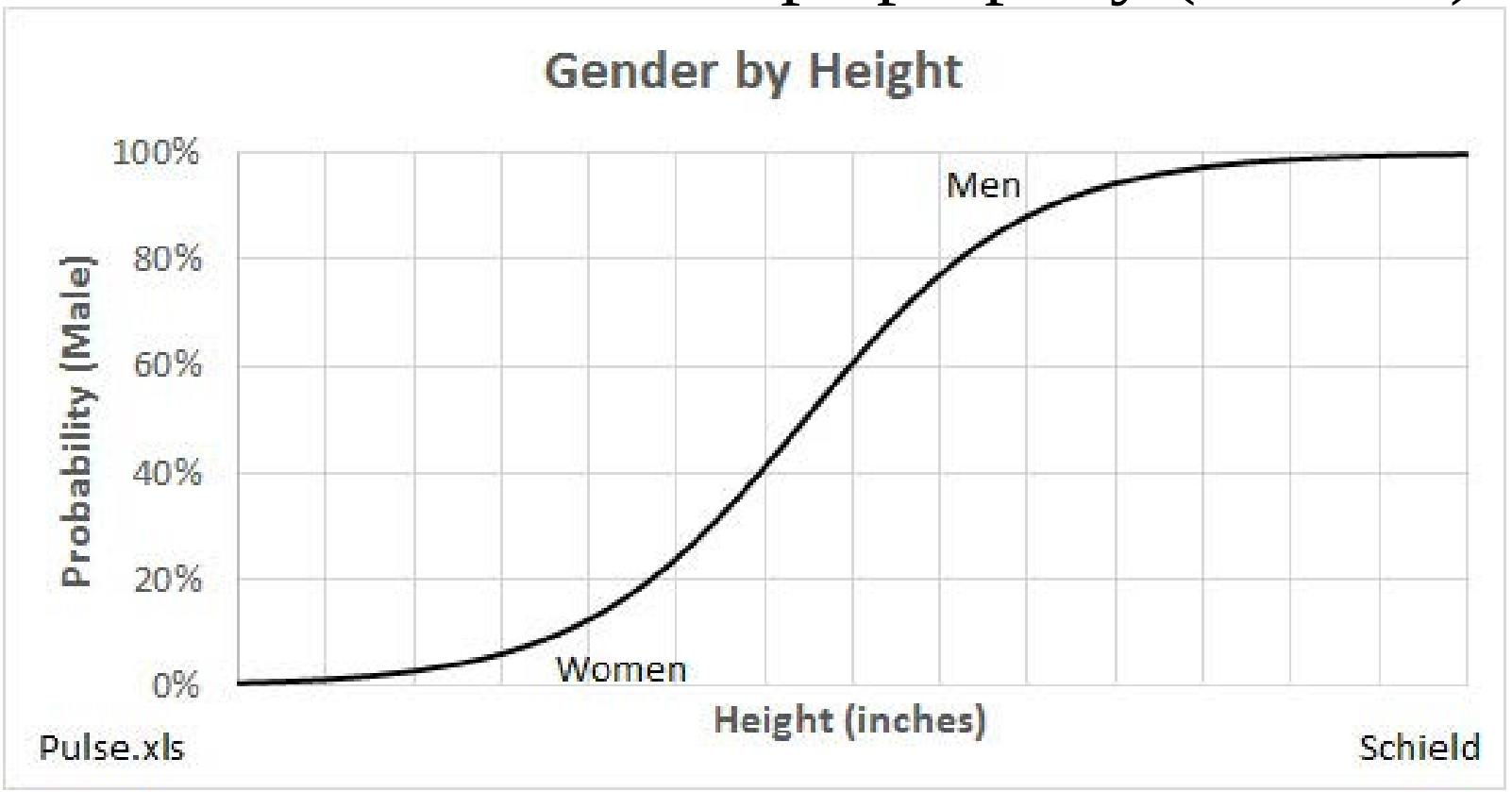
This invalid trend-line intersects the joint mean.
Insert circle at joint means; insert mean values in textbox.



Linear Trendline is invalid. Intuitive idea of solution

No need to create this graph.

Goal: create this shape properly (slide 20).



Four Step Approach

1) Insert intercept #1 with slope = 0. Record the sum of the errors: the logs of the chance (the likelihood) that the estimate is OK.
2) Solve for intercept & slope using SOLVER; Record the sum of the errors for this model.
3) Test the slope for statistical significance.
4) Generate graphs.

To do: Get data at www.StatLit.org/Excel/ 2015-Schield-Logistic-MLE1A-Excel2013-Data.xlsx

1a) Get Data; Find Mean(Y). Set Intercept #1 and Slope #1.

#1: Enter formula for E21 and E22.

19	D	E	F
20	**GENDER & INTERCEPT #1**		
21	Male-Pctg	0.62	=AVERAGE(B3:B94)
22	Intercept#1	0.4877	=LN(E21/(1-E21))

#2: Copy value from E22 into D3. Set E3=0.

Row	D	E	F
2	**Intercept**	**Slope**	
3	0.4877	0.0000	

1b) Enter formula for G3:K3. Select G3:K3; pull down to row 94

Row	D	E	F	G	H	I	J	K
2	**Intercept**	**Slope**		**Logit**	**Odds**	**Prob Y=1**	**Prob OK**	**Ln-LH-C**
3	0.4877	0.0000	See slides 7 and 13	0.49	1.63	0.62	0.62	-0.48
4								
5	Sum LnLk	-0.48	=SUM(K3:K94)					
6	Sum LnLk1		See slide 10					
7	Sum Ln Lk2		See slide 14					
8	Chi-Sq	0.00	=-2*(E6-E7)					
9	P-Value	1	=CHISQ.DIST.RT(E8,1)					
10								
11	**FORMULAS & TEXT: Enter, Copy Down**							
12	Logit	G3	=D$3+E$3*A3					
13	Odds	H3	=EXP(G3)					
14	Prob Y=1	I3	=H3/(1+H3)					
15	Prob OK	J3	=IF(B3=1,I3,1-I3)					
16	Ln-LH-OK	K3	=LN(J3)					

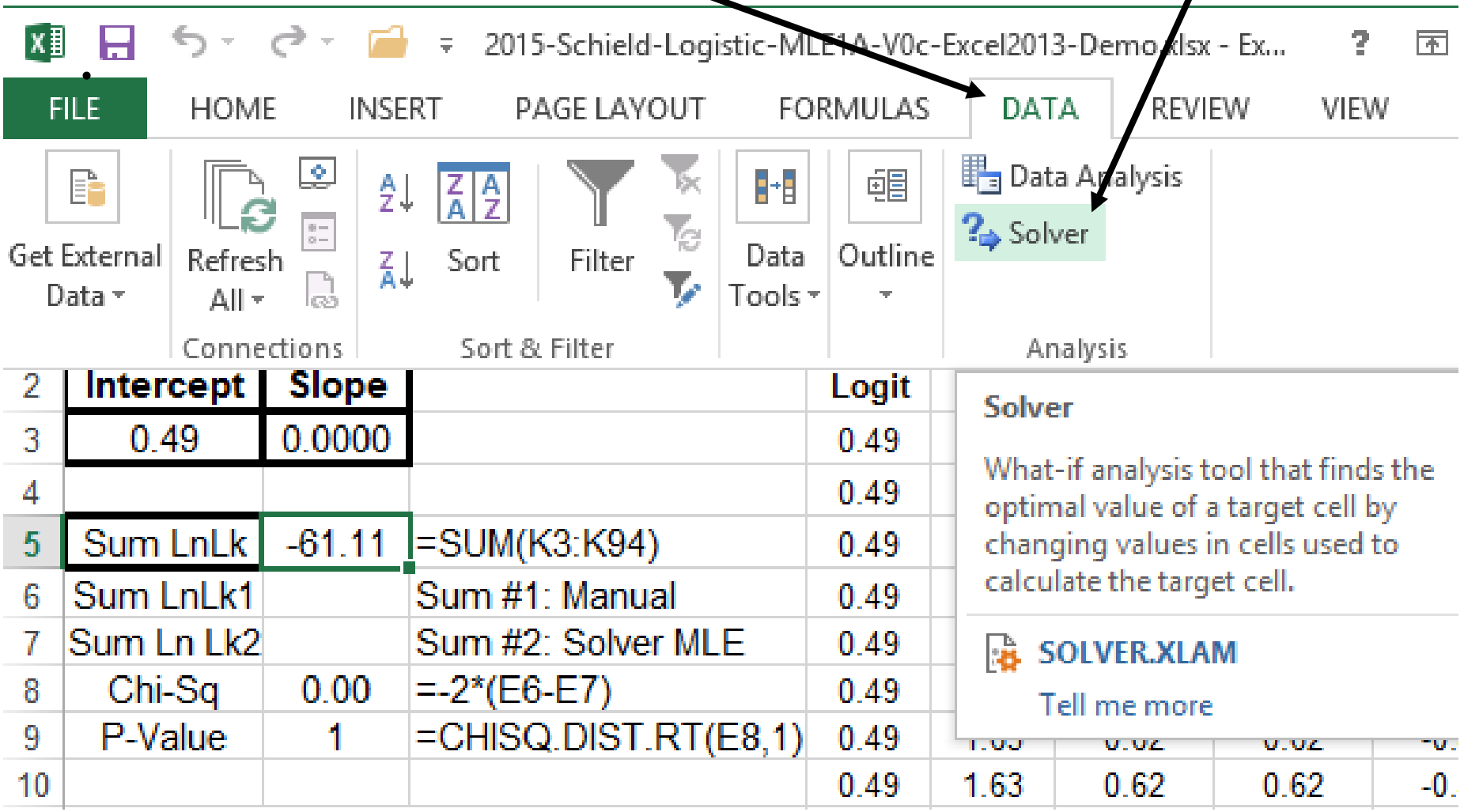
1c) Results are as expected. Probability of male = 0.62

Row	D	E	F	G	H	I	J	K
2	**Intercept**	**Slope**		**Logit**	**Odds**	**Prob Y=1**	**Prob OK**	**Ln-LH-OK**
3	0.4877	0.0000	See slides 7 and 13	0.49	1.63	0.62	0.62	-0.48
4				0.49	1.63	0.62	0.62	-0.48
5	Sum LnLk	-61.11	=SUM(K3:K94)	0.49	1.63	0.62	0.62	-0.48
6	Sum LnLk1		See slide 10	0.49	1.63	0.62	0.62	-0.48
7	Sum Ln Lk2		See slide 14	0.49	1.63	0.62	0.62	-0.48
8	Chi-Sq	0.00	=-2*(E6-E7)	0.49	1.63	0.62	0.38	-0.97
9	P-Value	1	=CHISQ.DIST.RT(E8,1)	0.49	1.63	0.62	0.62	-0.48
10				0.49	1.63	0.62	0.62	-0.48
11	**FORMULAS & TEXT: Enter, Copy Down**			0.49	1.63	0.62	0.62	-0.48
12	Logit	G3	=D$3+E$3*A3	0.49	1.63	0.62	0.38	-0.97
13	Odds	H3	=EXP(G3)	0.49	1.63	0.62	0.38	-0.97
14	Prob Y=1	I3	=H3/(1+H3)	0.49	1.63	0.62	0.62	-0.48
15	Prob OK	J3	=IF(B3=1,I3,1-I3)	0.49	1.63	0.62	0.62	-0.48
16	Ln-LH-OK	K3	=LN(J3)	0.49	1.63	0.62	0.62	-0.48
17				0.49	1.63	0.62	0.38	-0.97
18				0.49	1.63	0.62	0.38	-0.97

1d) Manually: Copy Value of E5 onto E6

Row	D	E	F
2	**Intercept**	**Slope**	
3	0.4877	0.0000	See slides 7 and 13
4			
5	Sum LnLk	-61.11	=SUM(K3:K94)
6	Sum LnLk1	-61.11	See slide 10
7	Sum Ln Lk2		See slide 14
8	Chi-Sq	122.22	=-2*(E6-E7)
9	P-Value	2E-28	=CHISQ.DIST.RT(E8,1)
10			

2a) Solve for Slope and Intercept: From Data menu, select Solver

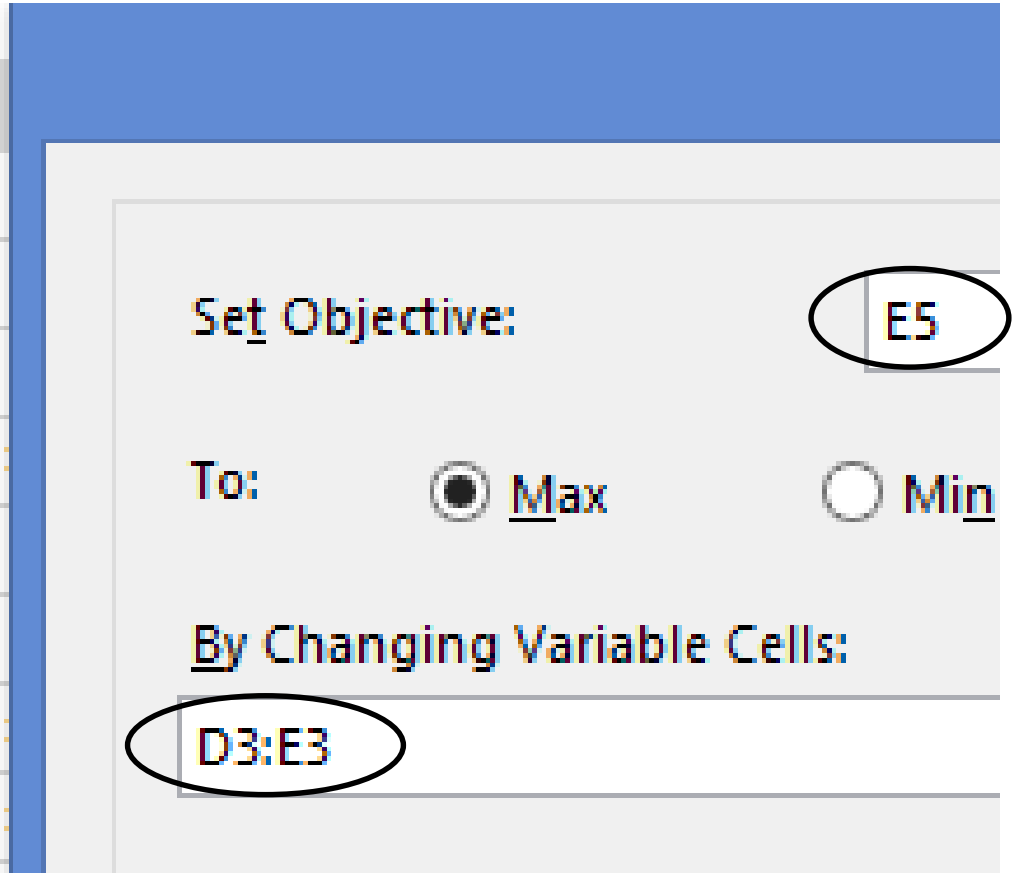


2	Intercept	Slope		Logit				
3	0.49	0.0000		0.49				
4				0.49				
5	Sum LnLk	-61.11	=SUM(K3:K94)	0.49				
6	Sum LnLk1		Sum #1: Manual	0.49				
7	Sum Ln Lk2		Sum #2: Solver MLE	0.49				
8	Chi-Sq	0.00	=-2*(E6-E7)	0.49				
9	P-Value	1	=CHISQ.DIST.RT(E8,1)	0.49				
10				0.49	1.63	0.62	0.62	-0.

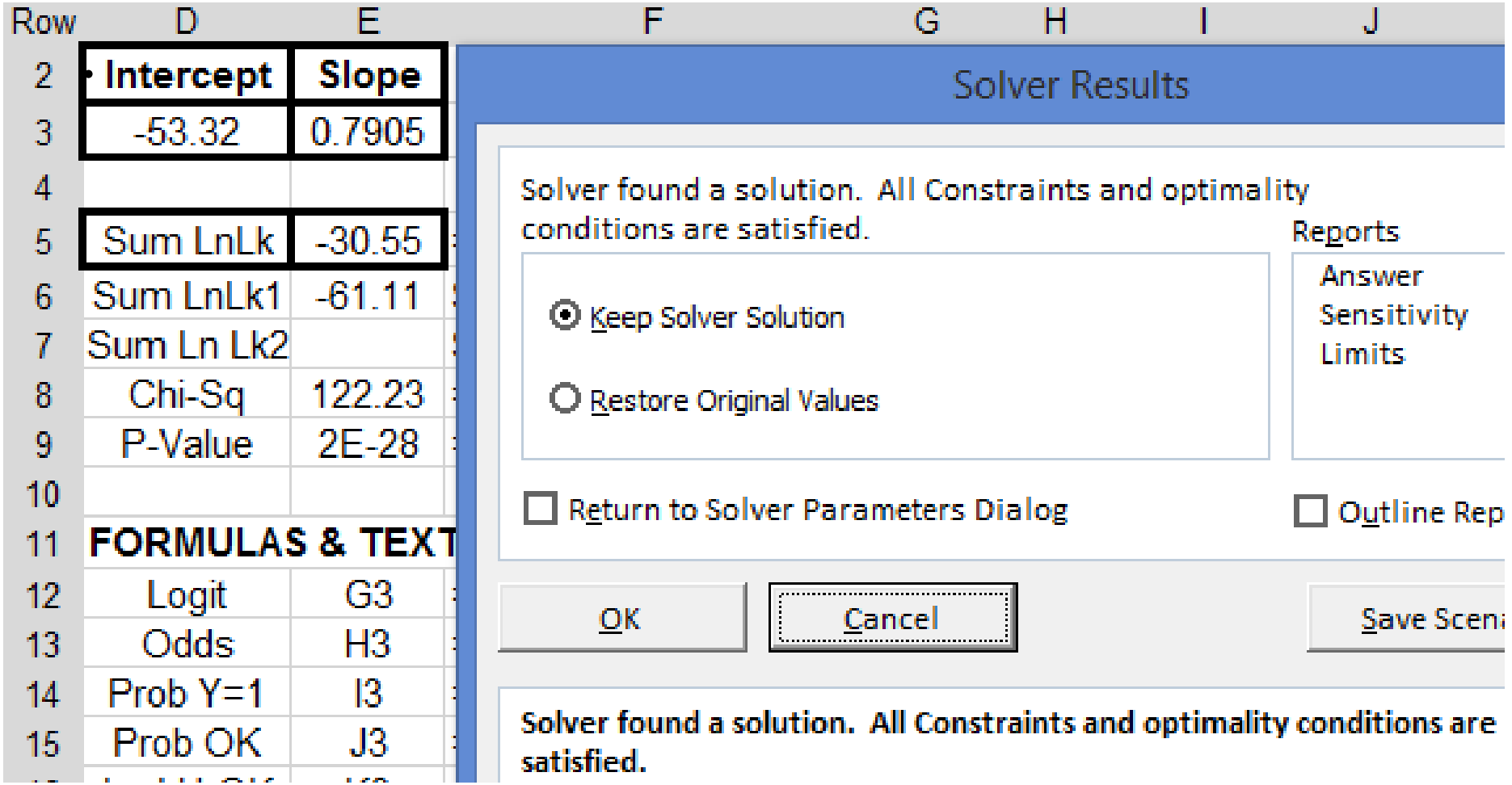
2b) Set Solver Parameters. Use GRC Nonlinear. Press Solve

Select Objective Cell (E5) and Variable Cells (D3:E3)

Row	D	E
2	**Intercept**	**Slope**
3	0.4877	0.0000
4		
5	Sum LnLk	-61.11
6	Sum LnLk1	-61.11
7	Sum Ln Lk2	
8	Chi-Sq	122.23
9	P-Value	2E-28



2c) Results: All constraints & conditions satisfied. Press OK



2d) Manually: Copy Value of E5 onto E7

Row	D	E	F
2	**Intercept**	**Slope**	
3	-53.3227	0.7905	See slides 7 and 13
4			
5	Sum LnLk	-30.55	=SUM(K3:K94)
6	Sum LnLk1	-61.11	See slide 10
7	Sum Ln Lk2	-30.55	See slide 14
8	Chi-Sq	61.12	=-2*(E6-E7)
9	P-Value	5E-15	=CHISQ.DIST.RT(E8,1)

3) Hypothesis test: Is non-zero slope statistically significant?

Conduct a right-tail Chi^2 test with 1 degree of freedom.

Sum LnLk	-30.55	=SUM(K3:K94)
Sum LnLk1	-61.11	See slide 10
Sum Ln Lk2	-30.55	See slide 14
Chi-Sq	61.12	=-2*(E6-E7)
P-Value	5E-15	=CHISQ.DIST.RT(E8,1)

Slope is statistically significant: P-value < 0.05

Note: E-15 means the decimal point is 15 places to the left: 0.000 000 000 000 005

4a) Analyze X axis: Enter formula for V2:V6

U	V	W	X	Y
HEIGHT	VALUE	FORMULA		
Ht-Average	68.72	=AVERAGE(A3:A94)		
Ht-Max	75.00	=MAX(A3:A94)		
Ht-Min	61.00	=MIN(A3:A94)		
Ht Ave Guy	70.75	=AVERAGEIF(B3:B94,"=1",A3:A94)		
Ht Ave Gal	65.40	=AVERAGEIF(B3:B94,"=0",A3:A94)		

4b) #1) Set N6 = 61.
Enter formula for O6, P6 & Q6

M	N	O	P	Q	R	S
2	**CHART**	N6	Enter X manually		P6	=EXP(O6)
3	**SETUP**	O6	=D$3+E$3*N6		Q6	=P6/(1+P6)
4						
5	**X-Ht**	**Logit**	**Odds**	**Prob Y=1**		
6	61.00	-5.10	0.01	1%		

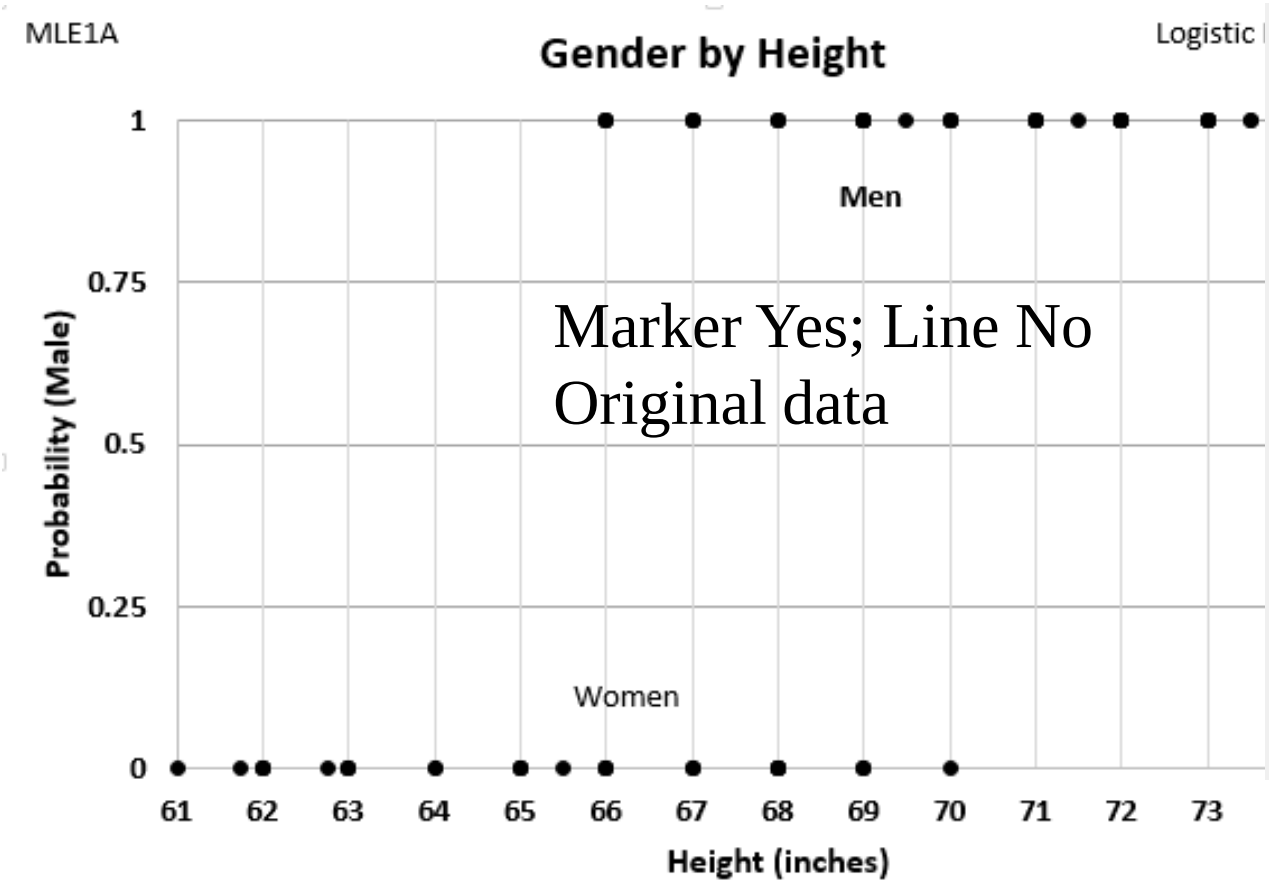
#2: Set N7 = 61.5; Select O6:Q6.
Pull down to O7:Q7 [Row 7]

5	X-Ht	Logit	Odds	Prob Y=1
6	61.00	-5.10	0.01	1%
7	61.50	-4.71	0.01	1%

4c) Select N6:Q7: Rows 6+7. Drag 2row box to row 34

M	N	O	P	Q
2	**CHART**	N6	Enter X manually	
3	**SETUP**	O6	=D$3+E$3*N6	
4				
5	**X-Ht**	**Logit**	**Odds**	**Prob Y=1**
6	61.00	-5.10	0.01	1%
7	61.50	-4.71	0.01	1%

4d) Graph Data on XY Plot: Gender (B) by Height (A)



Marker Yes; Line No
Original data

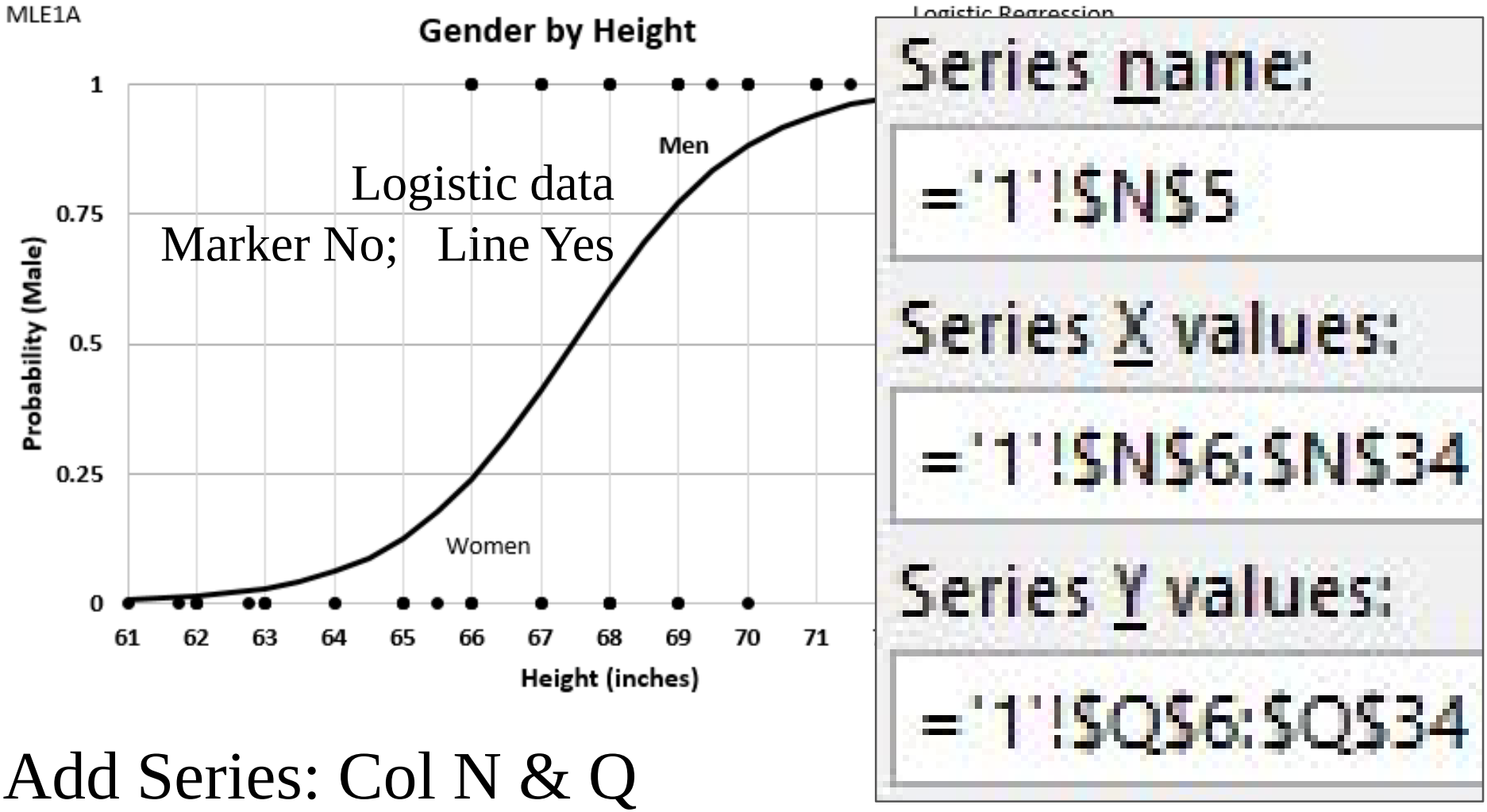
Series name:
='1'!$B$2

Series X values:
='1'!$A$3:$A$94

Series Y values:
='1'!$B$3:$B$94

Original data: Col A & B

4e) Graph Logistic Regression: Gender (Q) by Height (N)



Add Series: Col N & Q

4f) Final Result

Acknowledgment and Reference

ACKNOWLEDGMENT:

This presentation closely follows the Carlberg (2012) presentation in Chapter 2: pages 21-52.

These slides present the how – step by step – of logistic regression for a single predictor.

Carlberg (2012) discusses the how and the why.

Schield introduced the shortcut on slides 7 and 8.

REFERENCE:

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