Curve fitting with least squares

Fitting functions to data

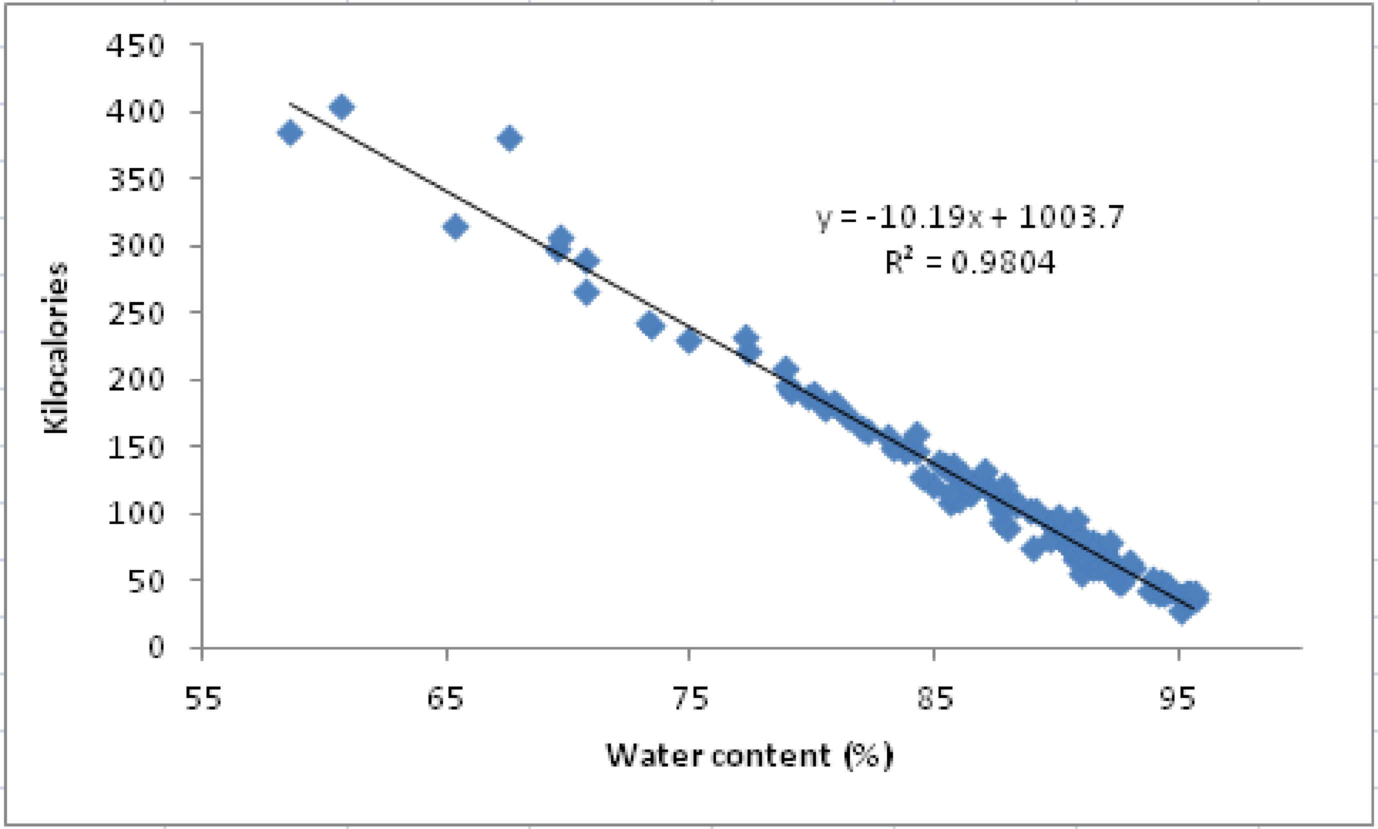
Fitting functions to data

- Common way to analyze data
- Two useful purposes
  - Assess the relationship between variables, obtain a predictive function
  - Obtain estimates of parameters
- We will focus today on “least squares” approaches
  - Least squares criterion: The line of best fit to the data minimizes the squared deviations between the data and the line

Simple linear regression

- Used to assess the straight-line relationship between two numeric variables
- Two variables
  - Independent, or predictor
  - Dependent, or response
- The independent is treated as the cause of change in the dependent
- Deviation from the line is treated as random variation, and only in the response variable

Regression of caloric content on percent water for various foods



Linear functions are easy to solve analytically

- There are formulas for slope and intercept:

$$\hat{y} = a + b\,x$$

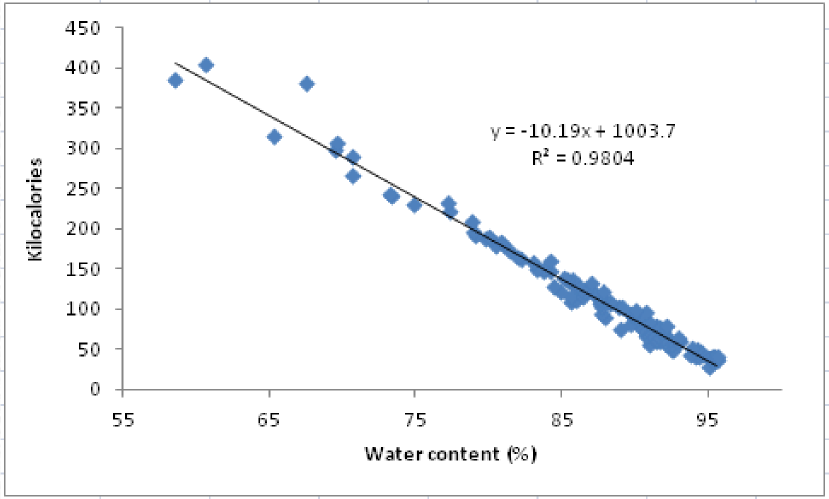
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}$$

- Formulas also available for standard errors of the estimates
- But, some equations can't be so easily solved analytically
- Instead, we can use numerical approaches to fit the line, and obtain standard errors

Least squares

- Want the best fit line – how do we know we have it?
- Least squares criterion: the best fit line minimizes the squared deviations between the line and the data
  - Sum of squared deviations between the y-data and the line is the “residual sums of squares”
  - Sum of squared deviations of y data from y mean is the “total sums of squares”
  - Variation accounted for by the line is “explained” or “model sums of squares”
- r^2 = coefficient of determination
  - Explained sums of squares / total sums of squares
  - (Total SS – residual SS)/total SS

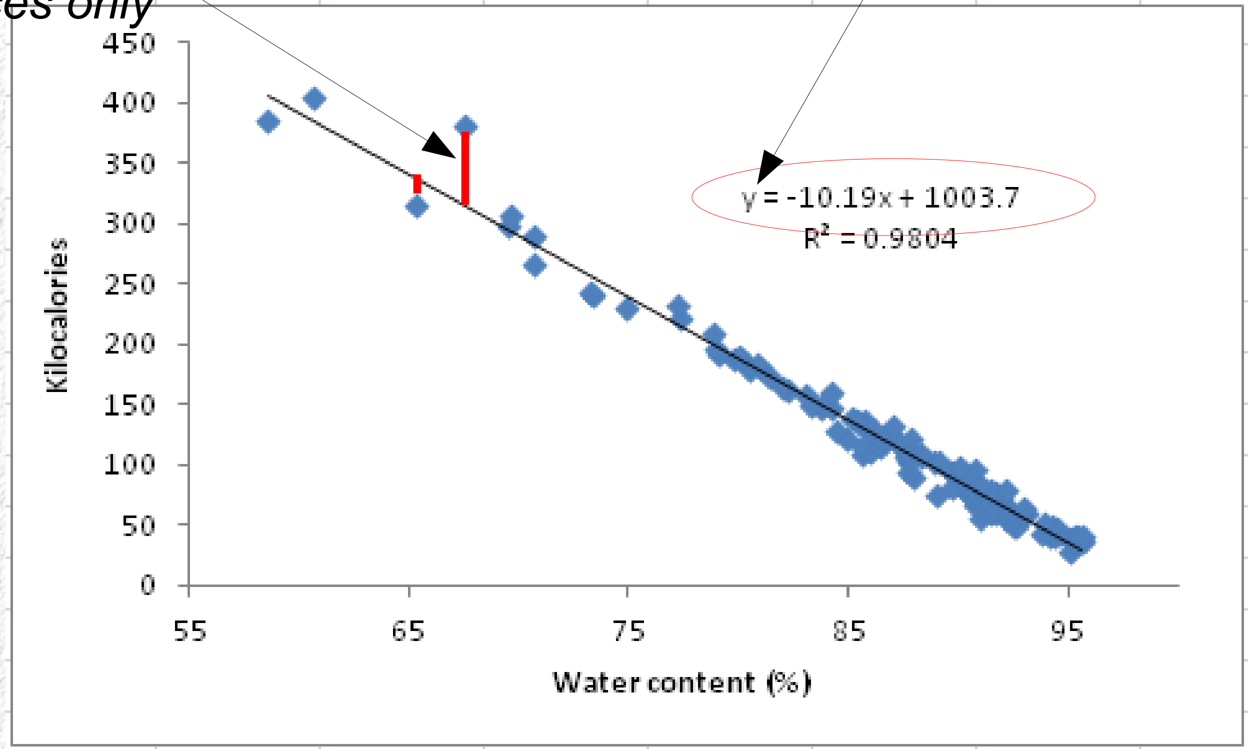


Residuals

Residual = observed – predicted value

Predicted value = average of y expected for a given value of x

Vertical differences only



Numerically fitting data to a function

- Start with a set of x and y data
- Use a function that predicts y from x, using any (reasonable) starting values for the unknown parameters (slope and intercept)
- Calculate the residuals, square them, sum them up
- Use Solver change the slope and intercept parameters until the sum of squared residuals is as small as possible

In Excel

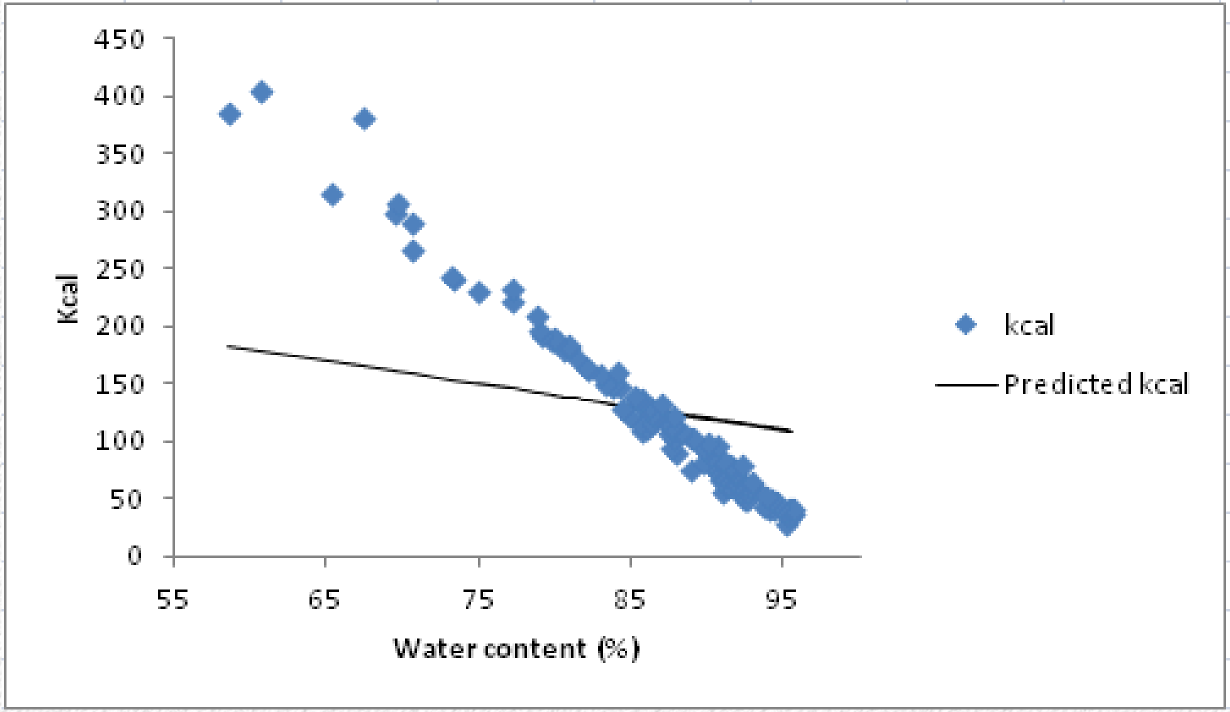
Predicted from straight line formula using initial parameters in B118 and B119

Squared deviations between observed and predicted kcal's

	A	B	C	D	E
1	food	water	kcal	Predicted kcal	Squared deviations
2	watercress	95.11	28	=B$118*B2+B$119	=(C2-D2)^2
3	pak-choi cabbage	95.32	34	109.36	5679.1
4	iceberg lettuce	95.64	36	108.72	5288.2
5	white gourd	95.54	36	108.92	5244.7
6	green leaf lettuce	95.07	38	109.86	5163.8
7	cucumber	95.23	40	109.54	4835.8
8	radish	95.27	41	109.46	4686.8
9	nopales	94.12	41	111.76	5007.0
113	plantains	65.28	316	169.44	21479.6
114	soybeans	67.5	381	165	46440.9
115	garlic	58.58	386	182.84	41274.4
116	prairie turnips	60.69	405	178.62	51248.4
117					
118	Slope	-2		Sum Sq.	452408.5
119	Intercept	300			
120					

Sum of squared deviations – minimize with Solver

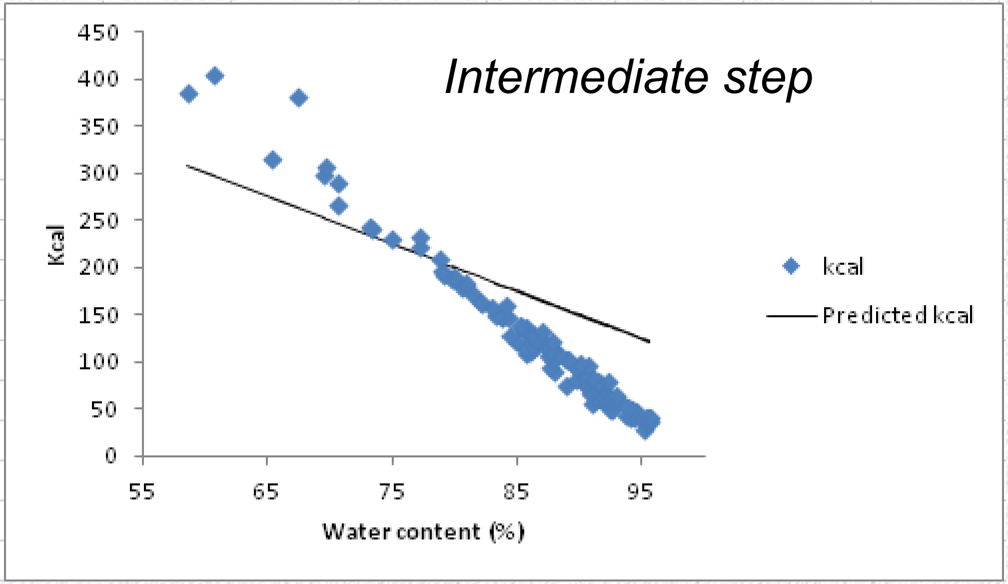
Close enough to start...

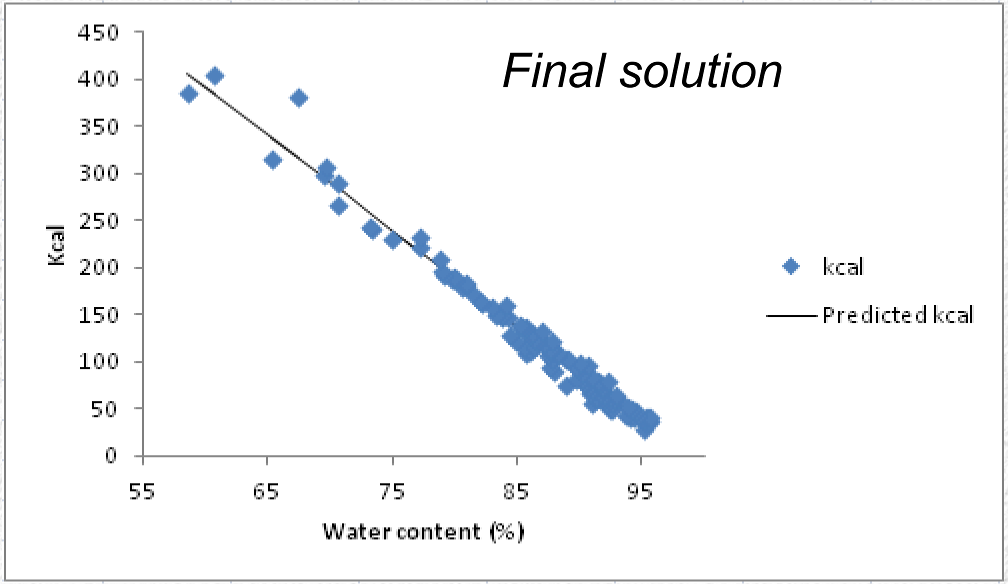


Slope = -2

Intercept = 300

As Solver changes slope and intercept...





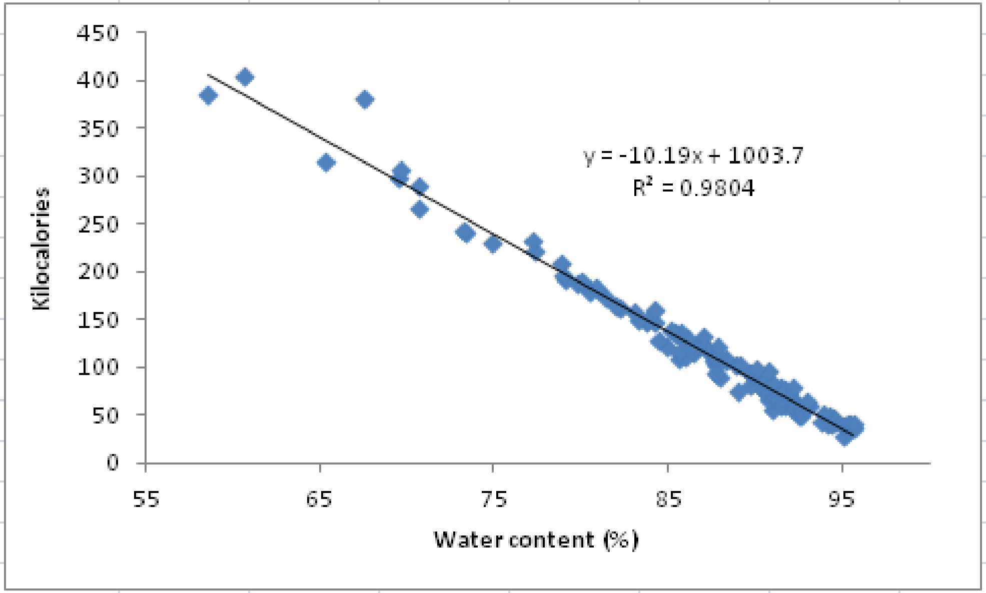
Slope = -5

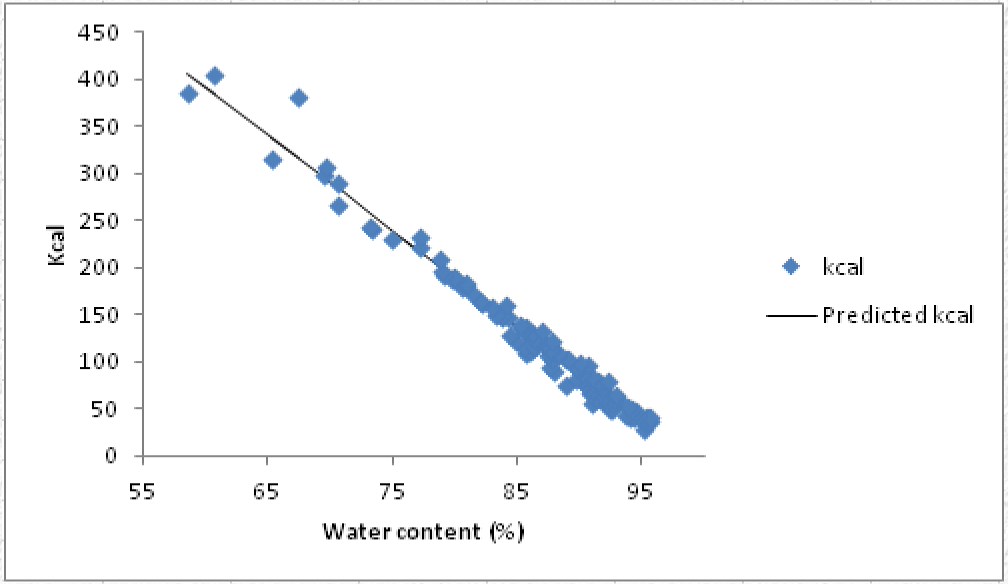
Intercept = 600

Slope = -10.19

Intercept = 1003.7

Match between analytical and numeric solutions





Slope = -10.19

Intercept = 1003.7

Very close agreement!

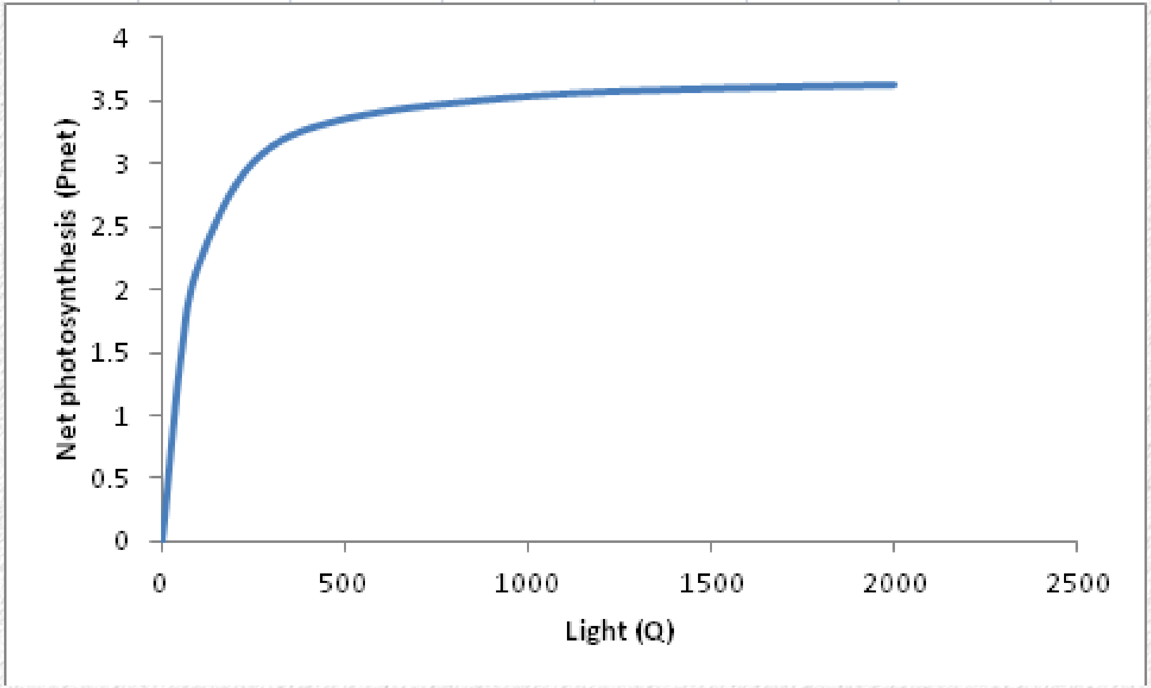
A trickier problem

- Sometimes we can’t directly measure what we want to know
- If we know how the quantity we want to know is related to the things we can measure, we can:
  - Use a function that shows the relationship
  - Fit the function to the data we can measure
  - Use the parameters from the best fit line as estimates of the quantities we are interested in
- Example: photosynthesis data

Photosynthesis measurements

- Portable photosynthesis systems are used to measure photosynthesis in living leaves in the field
- Light levels are set by the machine to different levels, and gas exchange is measured in response
- Gas exchange rates are converted by the system to net photosynthesis rates (P_{net})
- The relationship between light and photosynthesis is non-linear

Net photosynthesis as a function of light intensity



A model of photosynthesis

- A mechanistic model that explains the relationship between light intensity and net photosynthesis is:

$$P_{net} = \frac{\Phi Q + P_{marea} - \sqrt{(\Phi Q + P_{marea})^2 - 4\theta \Phi Q P_{marea}}}{2\theta}$$

Data – reported from the photosynthesis system

P_{net} = net photosynthesis

Q = light intensity

To be estimated

Φ = Phi = Maximum quantum yield (CO_2 molecules fixed per photon)

P_{marea} = maximum area-based rate of net photosynthesis (CO_2 m^2s^{-1})

θ = convexity of the curve (dimensionless shape constant)

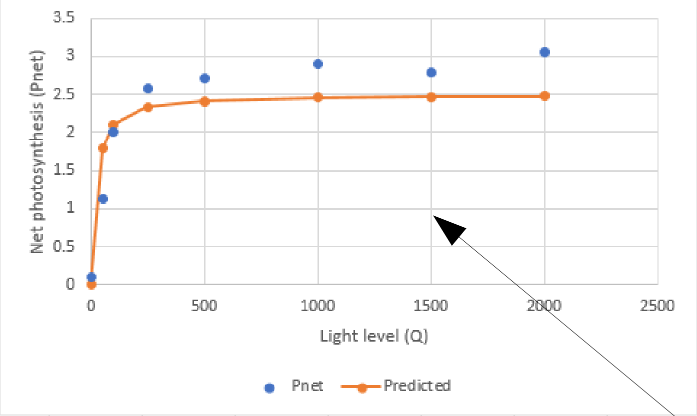
In Excel - setup

The data

Predicted P_{net} from equation

	A	B	C	D
1	Q	Pnet	Predicted	(Obs - Predicted)^2
2	0	0.109075	0	0.011897356
3	50	1.130192	1.79564	0.442820399
4	100	2.000815	2.106496	0.011168429
5	250	2.583246	2.332056	0.063096309
6	500	2.719568	2.414259	0.093213628
7	1000	2.899914	2.456689	0.196448631
8	1500	2.788039	2.471028	0.100495874
9	2000	3.059445	2.478235	0.337805601
10				
11	Theta	0.3	SumSq	
12	Phi	0.1	1.256946	
13	Pmarea	2.5		

Squared deviations

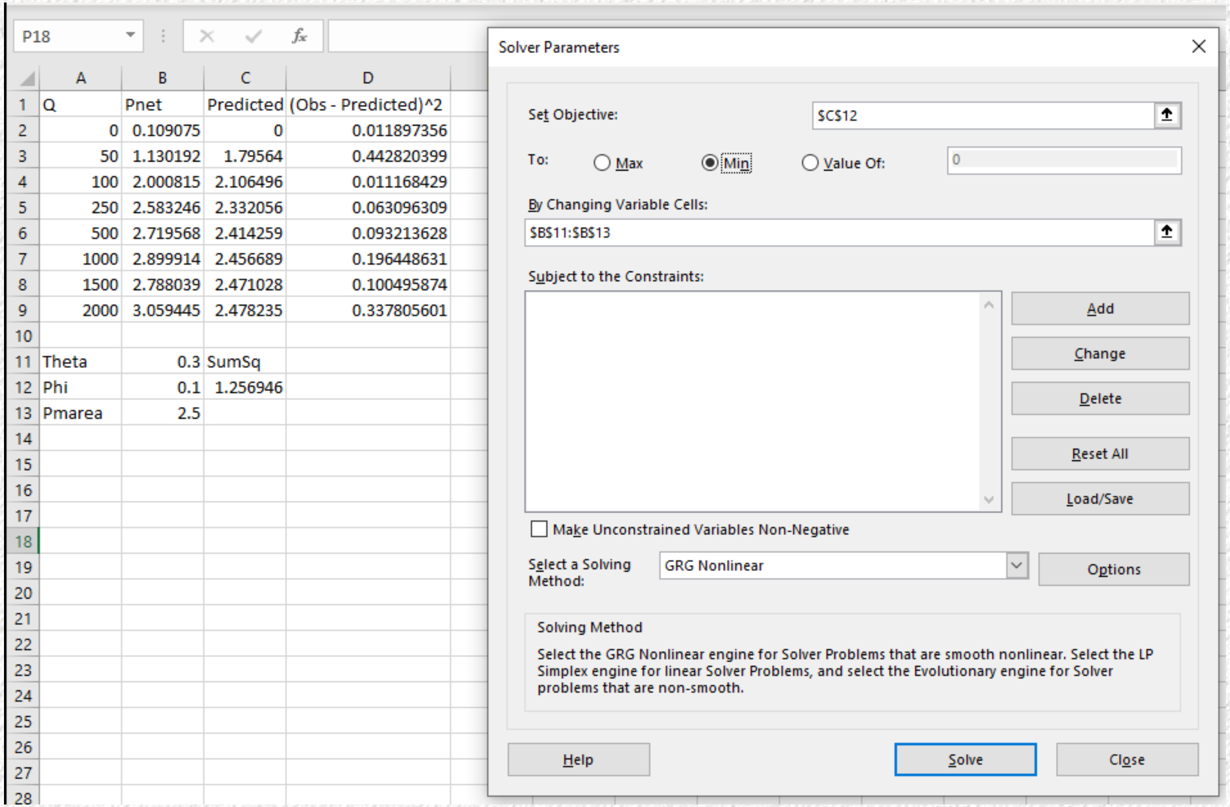


Graph of observed and predicted (based on starting values of parameters)

Sum of squared deviations

Parameters to estimate (initial guesses entered)

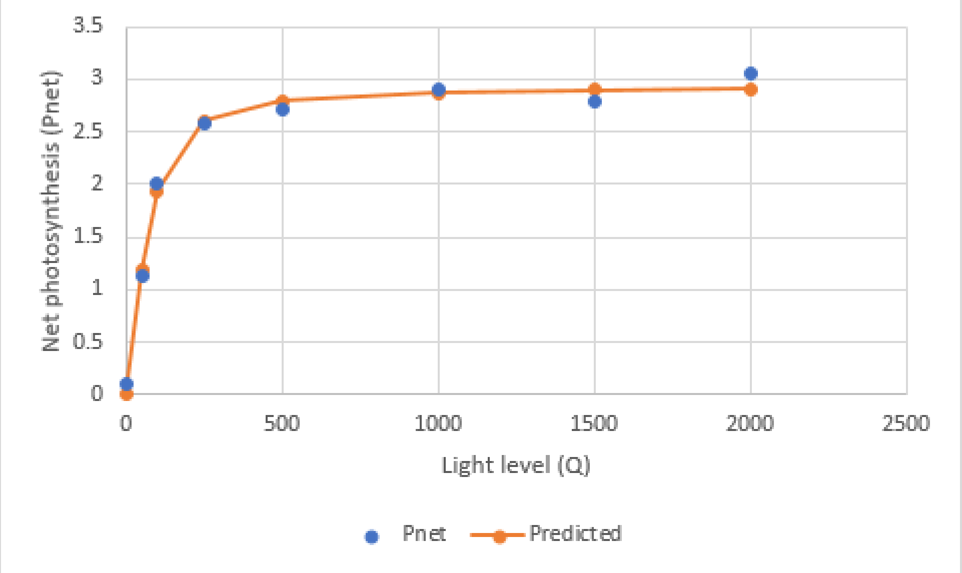
Solver settings



	A	B	C	D
1	Q	Pnet	Predicted	(Obs - Predicted)^2
2	0	0.109075	0	0.011897356
3	50	1.130192	1.79564	0.442820399
4	100	2.000815	2.106496	0.011168429
5	250	2.583246	2.332056	0.063096309
6	500	2.719568	2.414259	0.093213628
7	1000	2.899914	2.456689	0.196448631
8	1500	2.788039	2.471028	0.100495874
9	2000	3.059445	2.478235	0.337805601
10				
11	Theta	0.3	SumSq	
12	Phi	0.1	1.256946	
13	Pmarea	2.5		

Solver's solution

	A	B	C	D
1	Q	Pnet	Predicted	(Obs - Predicted)^2
2	0	0.109075	0	0.011897356
3	50	1.130192	1.18214	0.002698579
4	100	2.000815	1.933932	0.004473283
5	250	2.583246	2.606909	0.000559929
6	500	2.719568	2.793659	0.005489489
7	1000	2.899914	2.872731	0.000738939
8	1500	2.788039	2.896997	0.011871802
9	2000	3.059445	2.90875	0.022708838
10				
11	Theta	0.797057	SumSq	
12	Phi	0.026865	0.060438	
13	Pmarea	2.942537		



Today...

- We will set up the worksheet and find the estimates for the unknown parameters