IPM of Date Palm Insect Pests and Diseases

Training Course

Statistical Designs and Analysis of IPM data of Date Palm Pests

(Simple and Multiple Regression)

Name: Khaled Al-Shamaa

Date: 28 February 2017

Venue: Muscat, Oman

Citation: Khaled Al-Shamaa (2017). Statistical Designs and Analysis of IPM data of Date Palm Pests (Simple and Multiple Regression). IPM of Date Palm Insect Pests and Diseases Training Course, 28 Feb 2017, Muscat, Oman. BSS/DDG-R, ICARDA, Amman. 30 slides.

ICARDA
Science for Better Livelihoods in Dry Areas

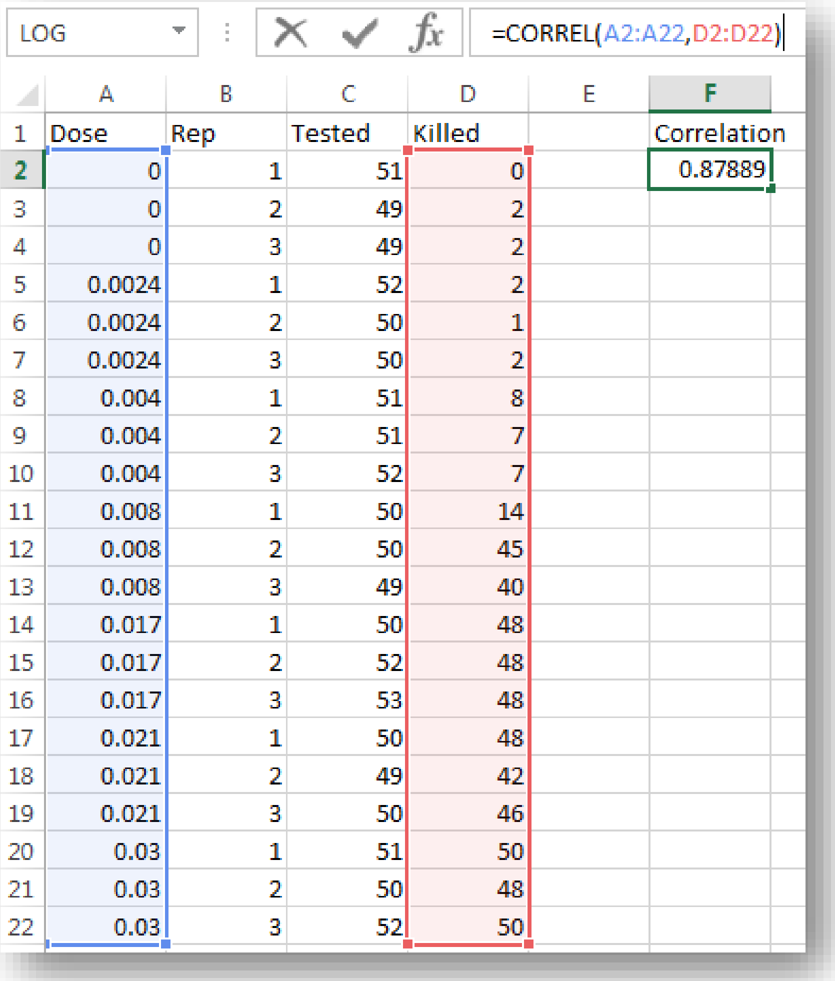
Correlation

- Quantitative variables, linear relationship.
- Correlation does not imply causation.
- Correlation value vary from -1 to +1

  -1 indicates perfect negative correlation, and +1 indicates perfect positive correlation. 0 means no correlation.
- Correlation Significance

  depends on the correlation value and number of observations (test using t-test against 0).

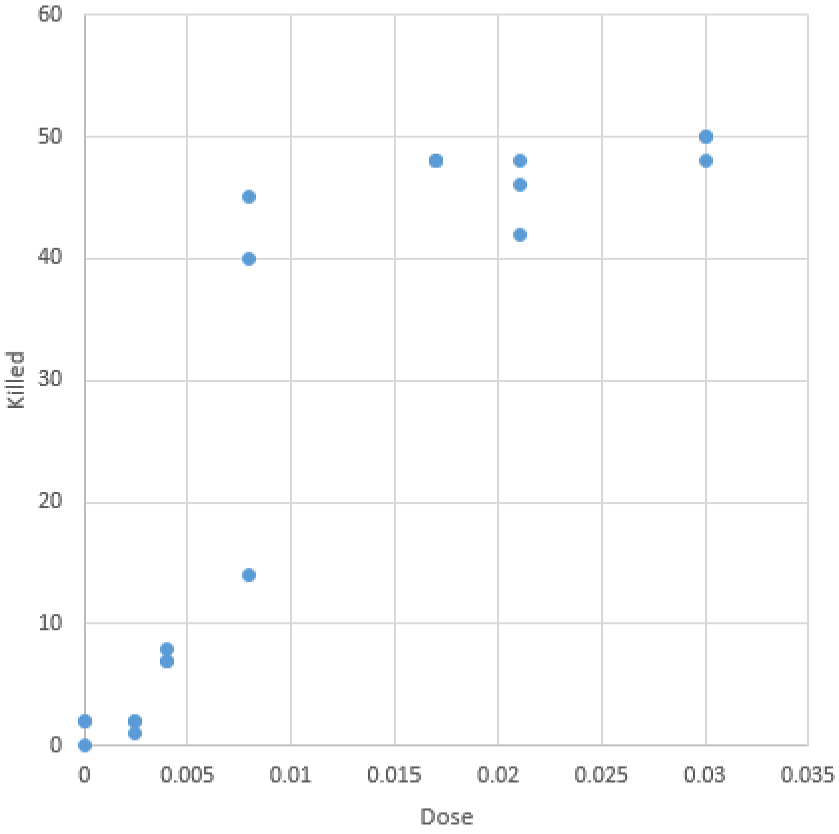
Excel - Correlation



LOG

=CORREL(A2:A22,D2:D22)

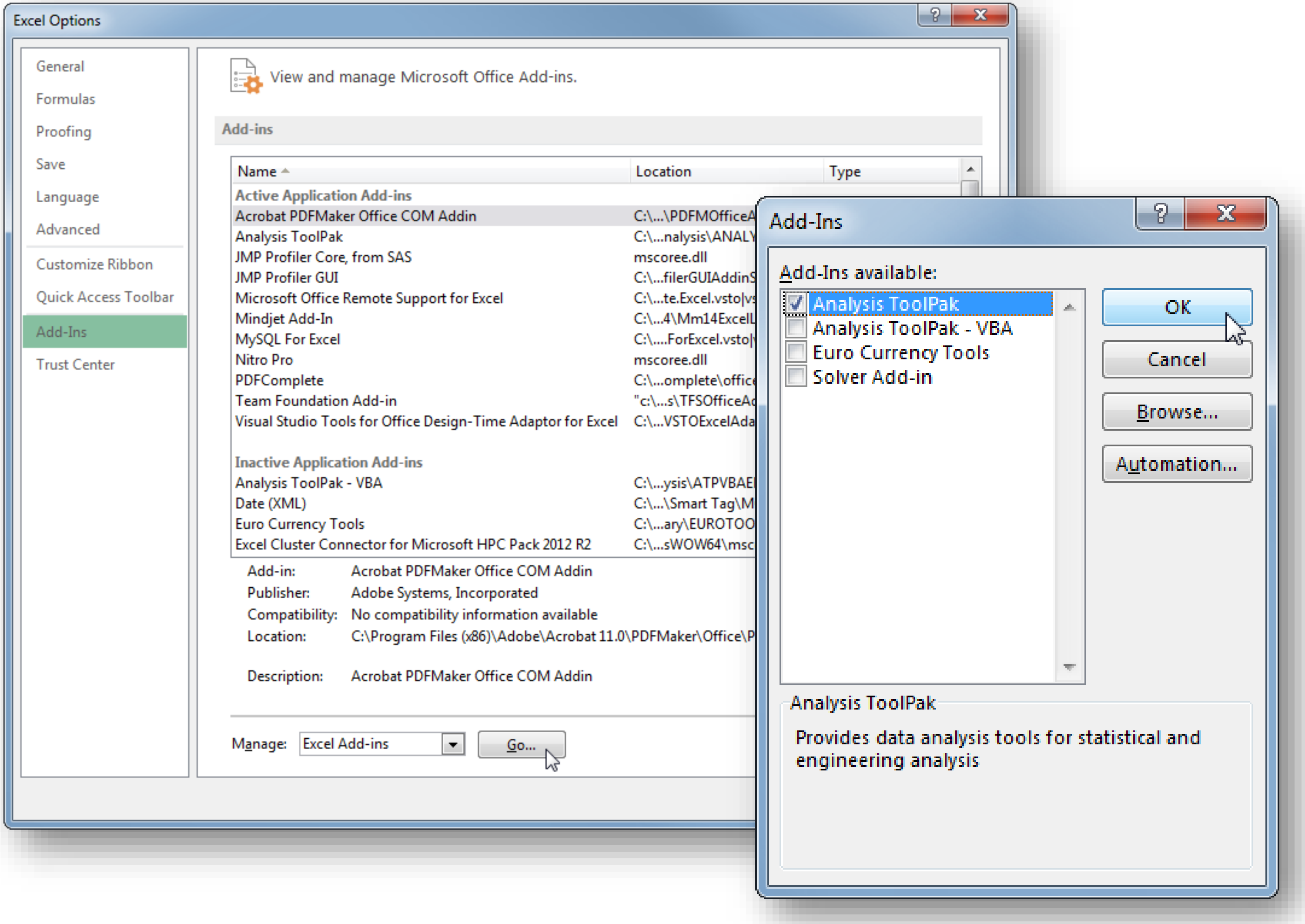
	A	B	C	D	E	F
1	Dose	Rep	Tested	Killed		Correlation
2	0	1	51	0		0.87889
3	0	2	49	2		
4	0	3	49	2		
5	0.0024	1	52	2		
6	0.0024	2	50	1		
7	0.0024	3	50	2		
8	0.004	1	51	8		
9	0.004	2	51	7		
10	0.004	3	52	7		
11	0.008	1	50	14		
12	0.008	2	50	45		
13	0.008	3	49	40		
14	0.017	1	50	48		
15	0.017	2	52	48		
16	0.017	3	53	48		
17	0.021	1	50	48		
18	0.021	2	49	42		
19	0.021	3	50	46		
20	0.03	1	51	50		
21	0.03	2	50	48		
22	0.03	3	52	50		



Treatment: Phosphine, PH3 (mg/l)

Subject: Storage Pests (*R. dominica)*

Excel – Analysis ToolPak



Excel Options

View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
Acrobat PDFMaker Office COM Addin	C:\...\PDFMOfficeA	
Analysis ToolPak	C:\...nalysis\ANALY	
JMP Profiler Core, from SAS	mscoree.dll	
JMP Profiler GUI	C:\...filerGUIAddinS	
Microsoft Office Remote Support for Excel	C:\...te.Excel.vsto\vs	
Mindjet Add-In	C:\...4\Mm14ExcelL	
MySQL For Excel	C:\....ForExcel.vsto\	
Nitro Pro	mscoree.dll	
PDFComplete	C:\...omplete\office	
Team Foundation Add-in	"c:\...s\TFSOfficeAd	
Visual Studio Tools for Office Design-Time Adaptor for Excel	C:\...VSTOExcelAda	
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\...ysis\ATPVBAEI	
Date (XML)	C:\...\Smart Tag\M	
Euro Currency Tools	C:\...ary\EUROTOO	
Excel Cluster Connector for Microsoft HPC Pack 2012 R2	C:\...sWOW64\msc	

Add-in: Acrobat PDFMaker Office COM Addin
Publisher: Adobe Systems, Incorporated
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Adobe\Acrobat 11.0\PDFMaker\Office\P

Description: Acrobat PDFMaker Office COM Addin

Manage: Excel Add-ins

Add-Ins

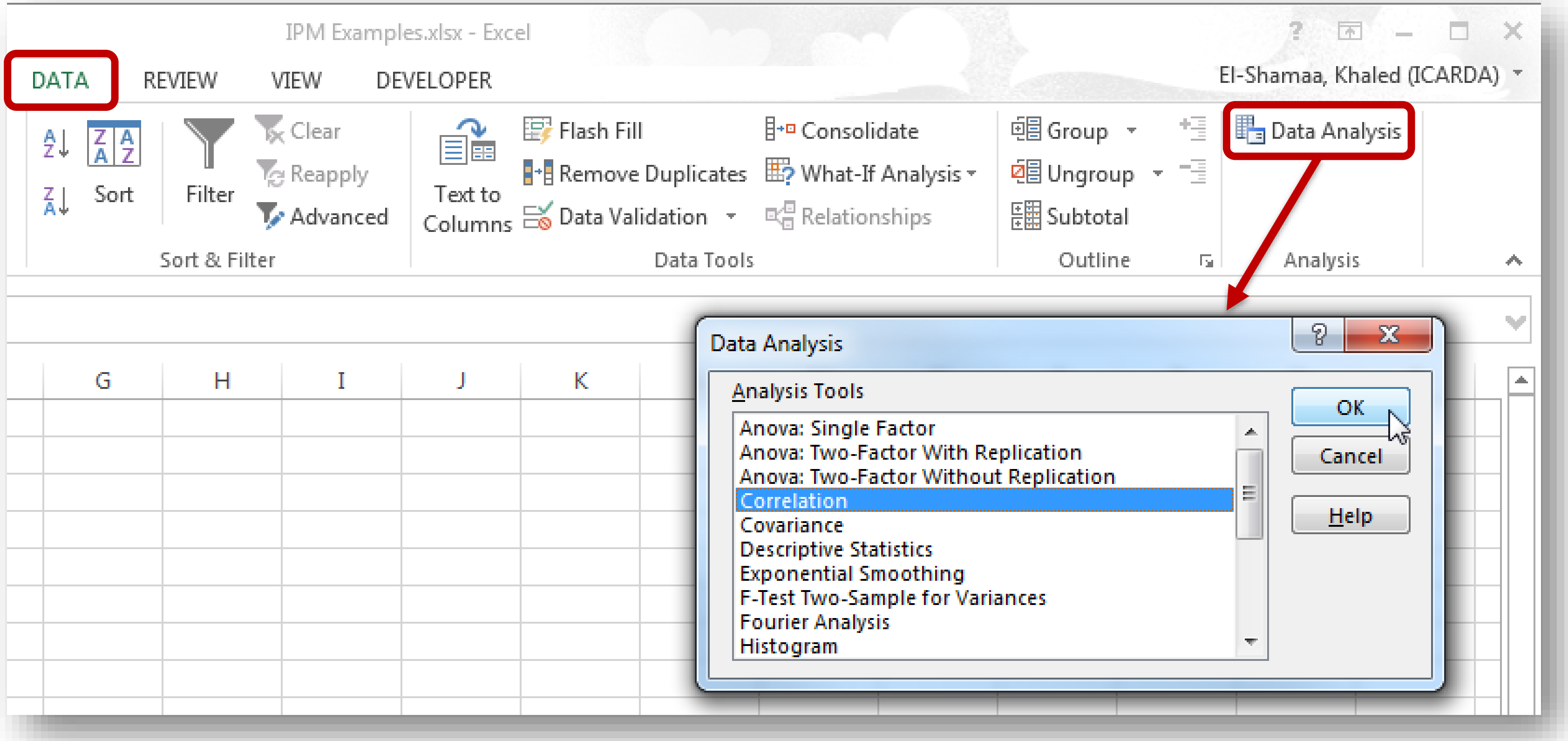
Add-Ins available:

- ☑ Analysis ToolPak
- □ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

Analysis ToolPak

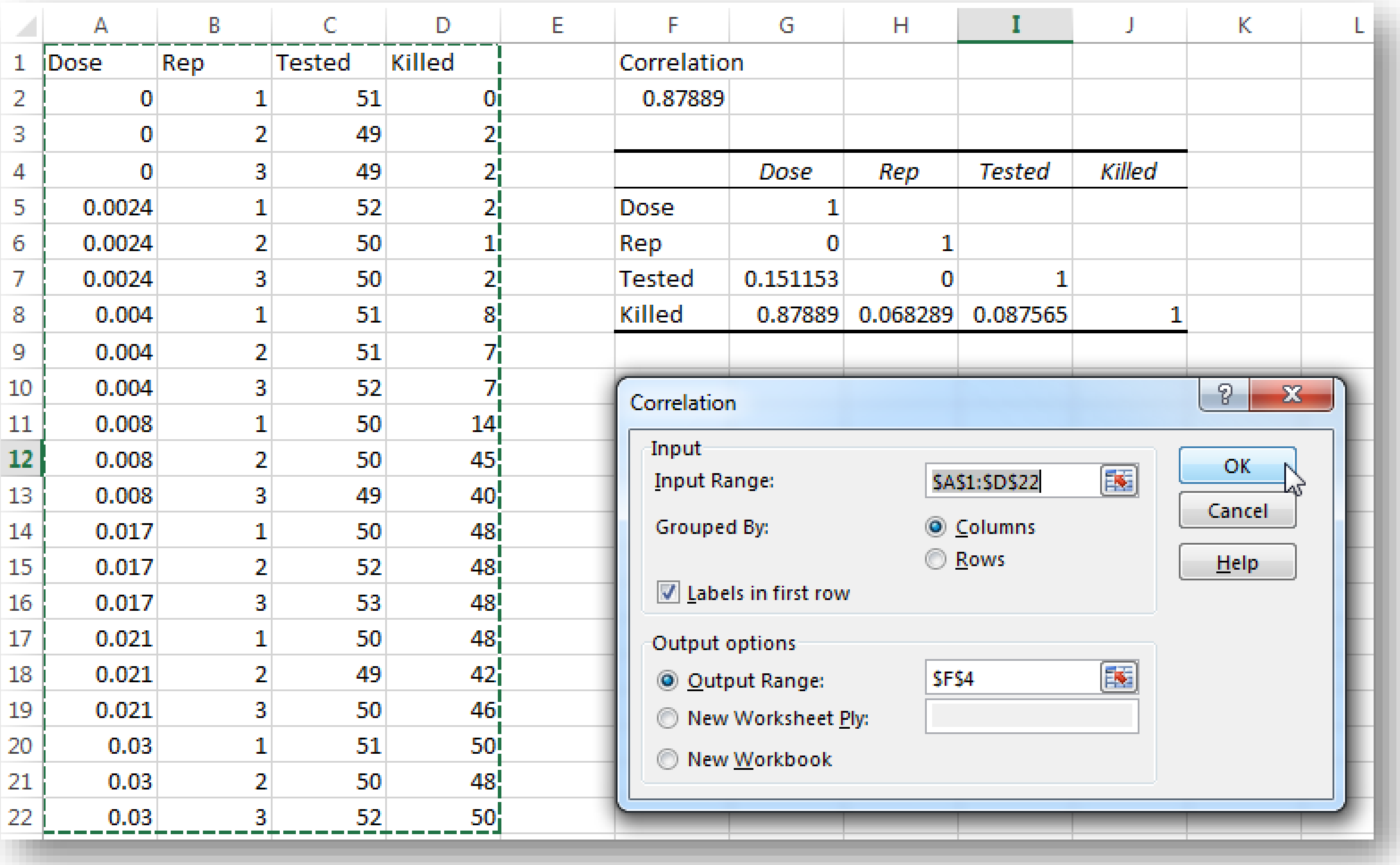
Provides data analysis tools for statistical and engineering analysis

Excel – Analysis ToolPak *(continue)*



Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

Excel (Analysis ToolPak) - Correlation



	Dose	Rep	Tested	Killed
2	0	1	51	0
3	0	2	49	2
4	0	3	49	2
5	0.0024	1	52	2
6	0.0024	2	50	1
7	0.0024	3	50	2
8	0.004	1	51	8
9	0.004	2	51	7
10	0.004	3	52	7
11	0.008	1	50	14
12	0.008	2	50	45
13	0.008	3	49	40
14	0.017	1	50	48
15	0.017	2	52	48
16	0.017	3	53	48
17	0.021	1	50	48
18	0.021	2	49	42
19	0.021	3	50	46
20	0.03	1	51	50
21	0.03	2	50	48
22	0.03	3	52	50

Correlation

0.87889

	Dose	*Rep*	*Tested*	*Killed*
Dose	1			
Rep	0	1		
Tested	0.151153	0	1	
Killed	0.87889	0.068289	0.087565	1

Correlation

Input

Input Range: $A$1:$D$22

Grouped By: ● Columns ○ Rows

☑ Labels in first row

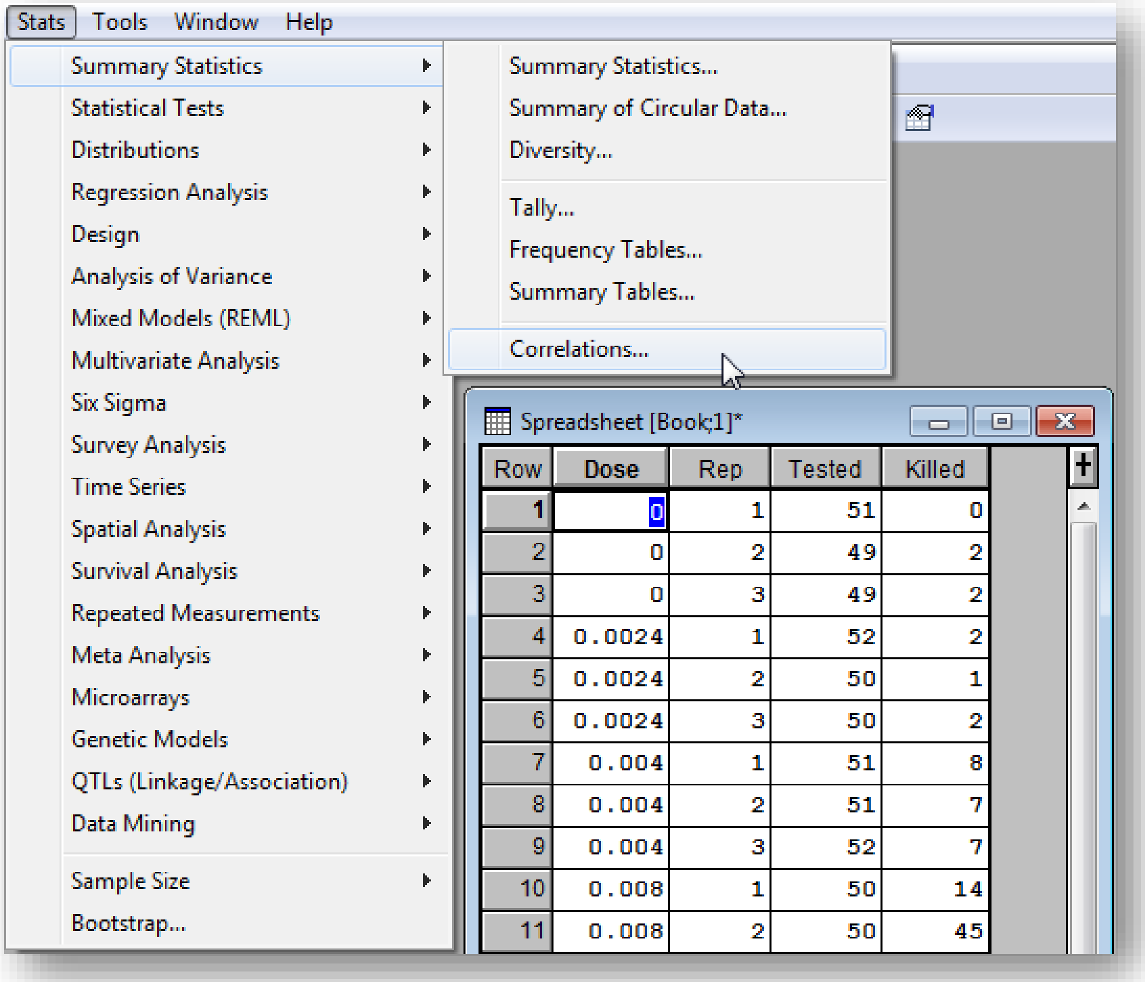
Output options

● Output Range: $F$4

○ New Worksheet Ply:

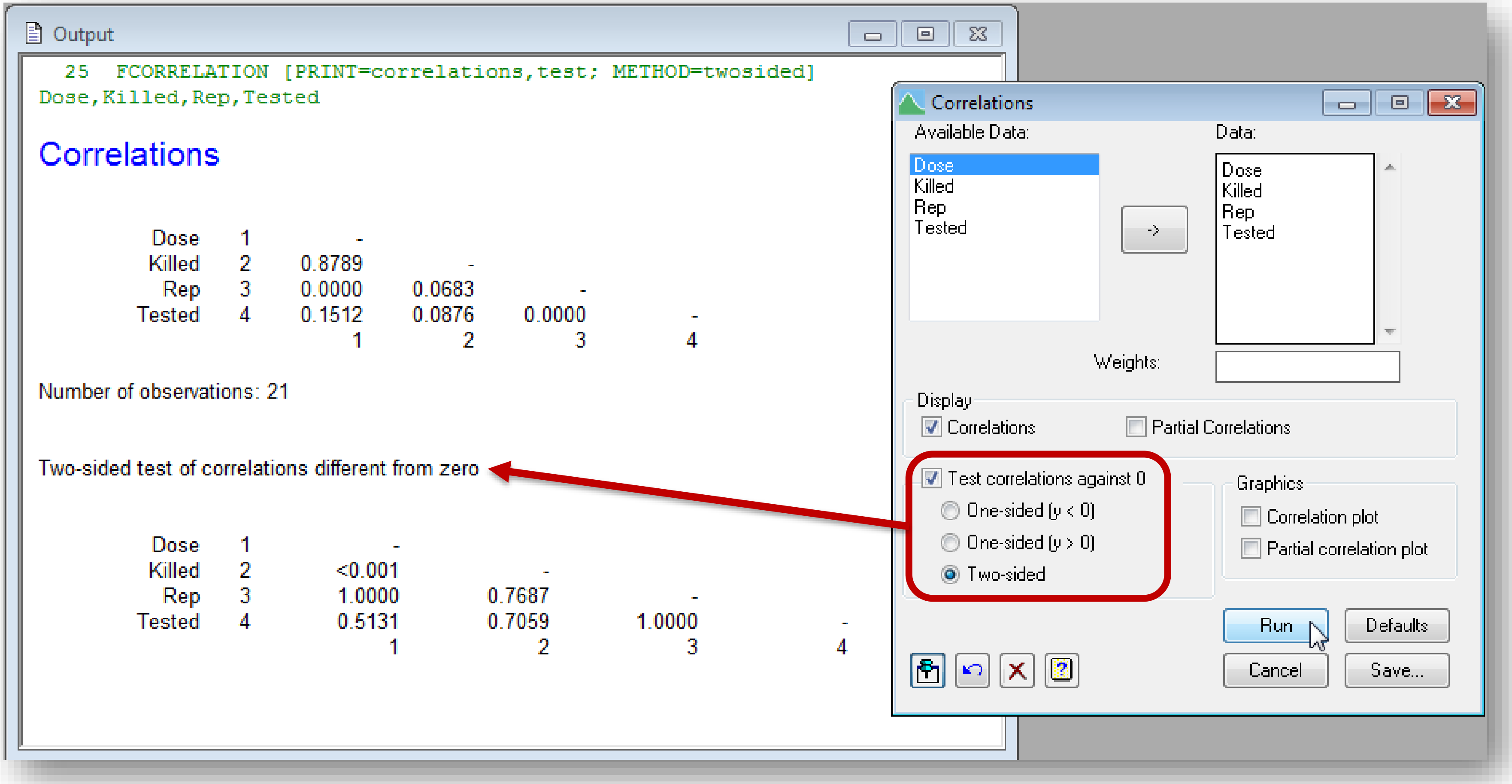
○ New Workbook

GenStat - Correlation



Row	Dose	Rep	Tested	Killed
1	0	1	51	0
2	0	2	49	2
3	0	3	49	2
4	0.0024	1	52	2
5	0.0024	2	50	1
6	0.0024	3	50	2
7	0.004	1	51	8
8	0.004	2	51	7
9	0.004	3	52	7
10	0.008	1	50	14
11	0.008	2	50	45

GenStat – Correlation *(continue)*



```
  25  FCORRELATION [PRINT=correlations,test; METHOD=twosided]
Dose,Killed,Rep,Tested
```

Correlations

```
      Dose    1          -
    Killed    2     0.8789         -
       Rep    3     0.0000    0.0683         -
    Tested    4     0.1512    0.0876    0.0000         -
                         1         2         3         4
```

Number of observations: 21

Two-sided test of correlations different from zero

```
      Dose    1          -
    Killed    2     <0.001          -
       Rep    3     1.0000     0.7687          -
    Tested    4     0.5131     0.7059     1.0000          -
                         1          2          3          4
```

☑ Correlations ☐ Partial Correlations

☑ Test correlations against 0

○ One-sided (ρ < 0)

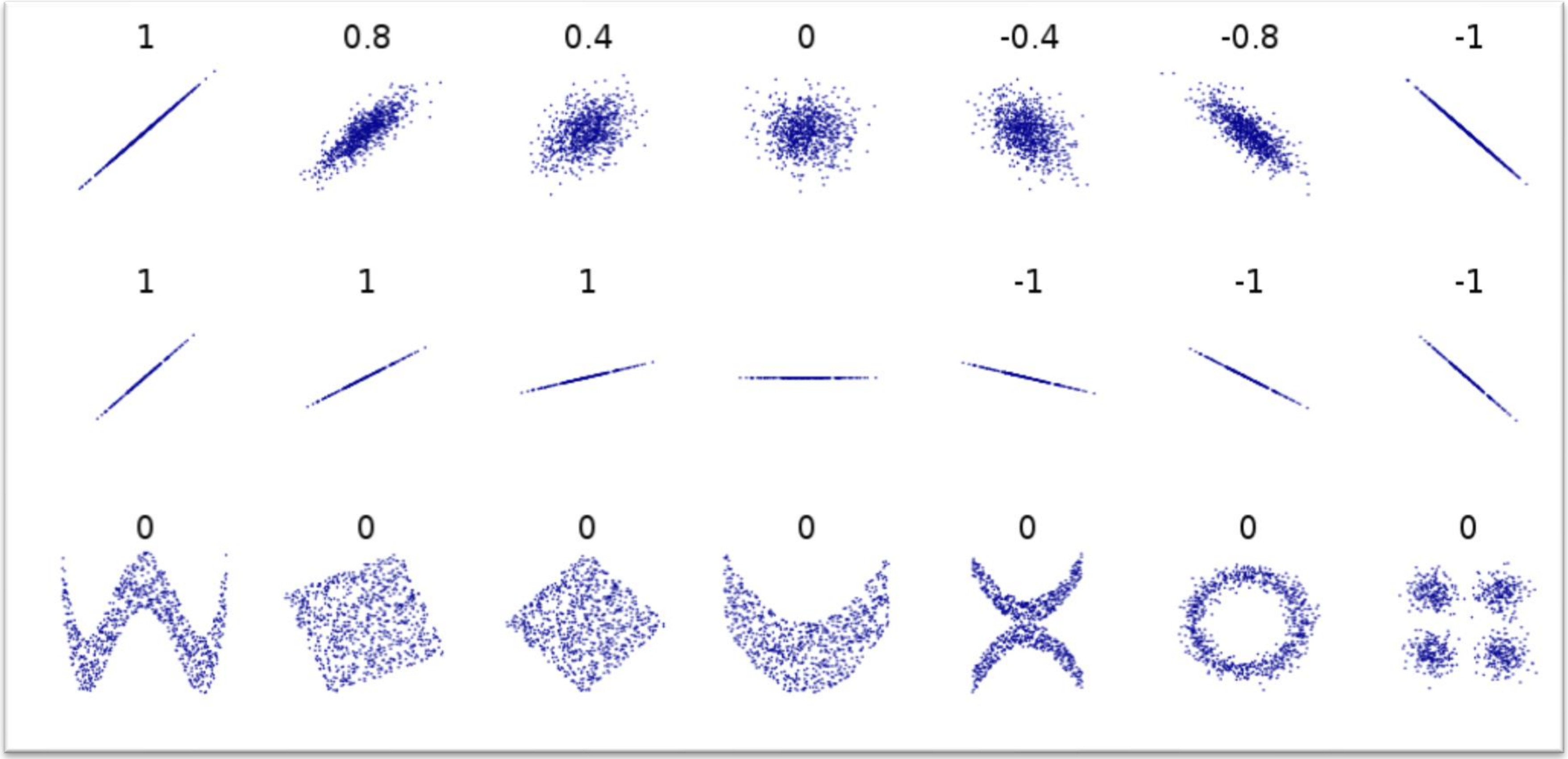
○ One-sided (ρ > 0)

● Two-sided

☐ Correlation plot

☐ Partial correlation plot

Correlation Abuse!



Correlation does not provide any information about the slope of the linear dependency

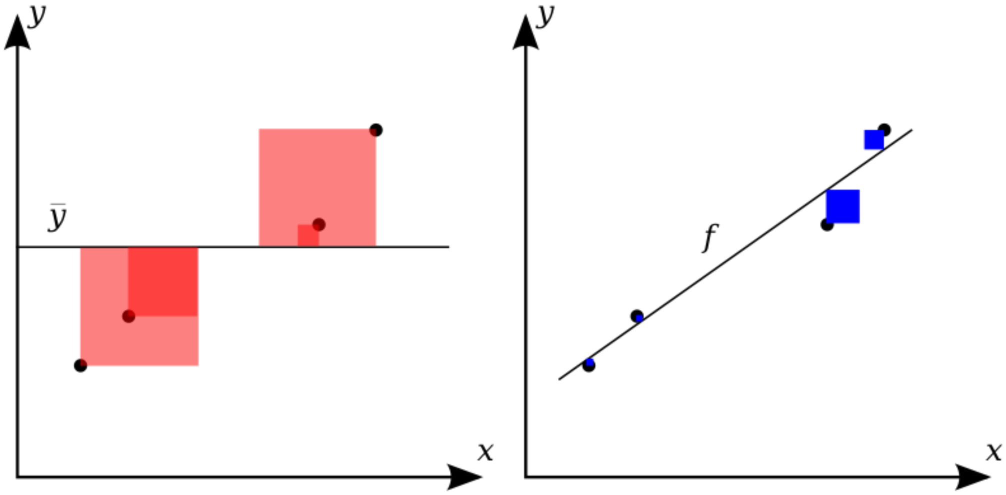
Regression Analysis

- The goal of regression analysis is to use the data on some objects to predict values for another object.
- If X predicts Y it does not mean that X causes Y.
- Accurate prediction depends heavily on measuring the right variables.

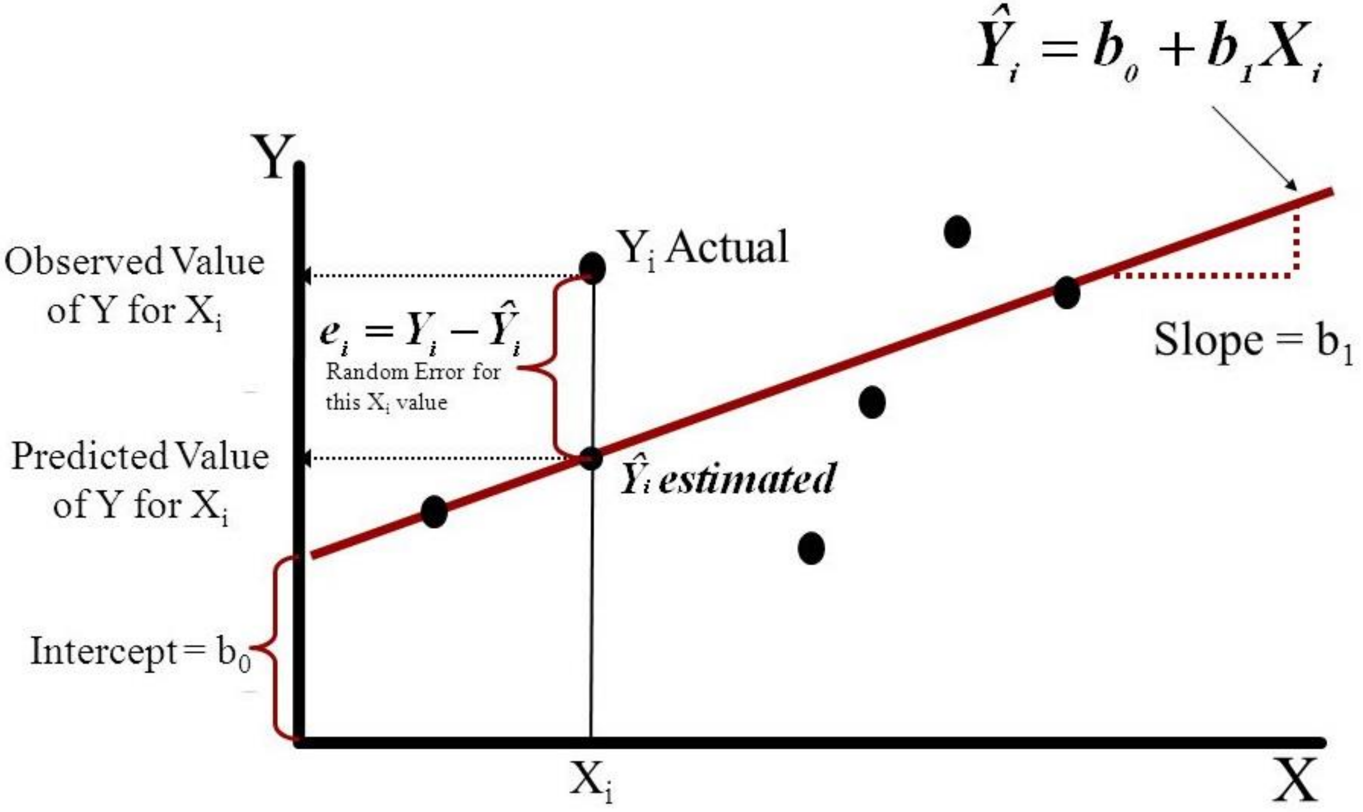
R – Squared (R^2)

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- It is the percentage of the response variable variation that is explained by a linear model.
- Yes! It is squared of Correlation R value.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

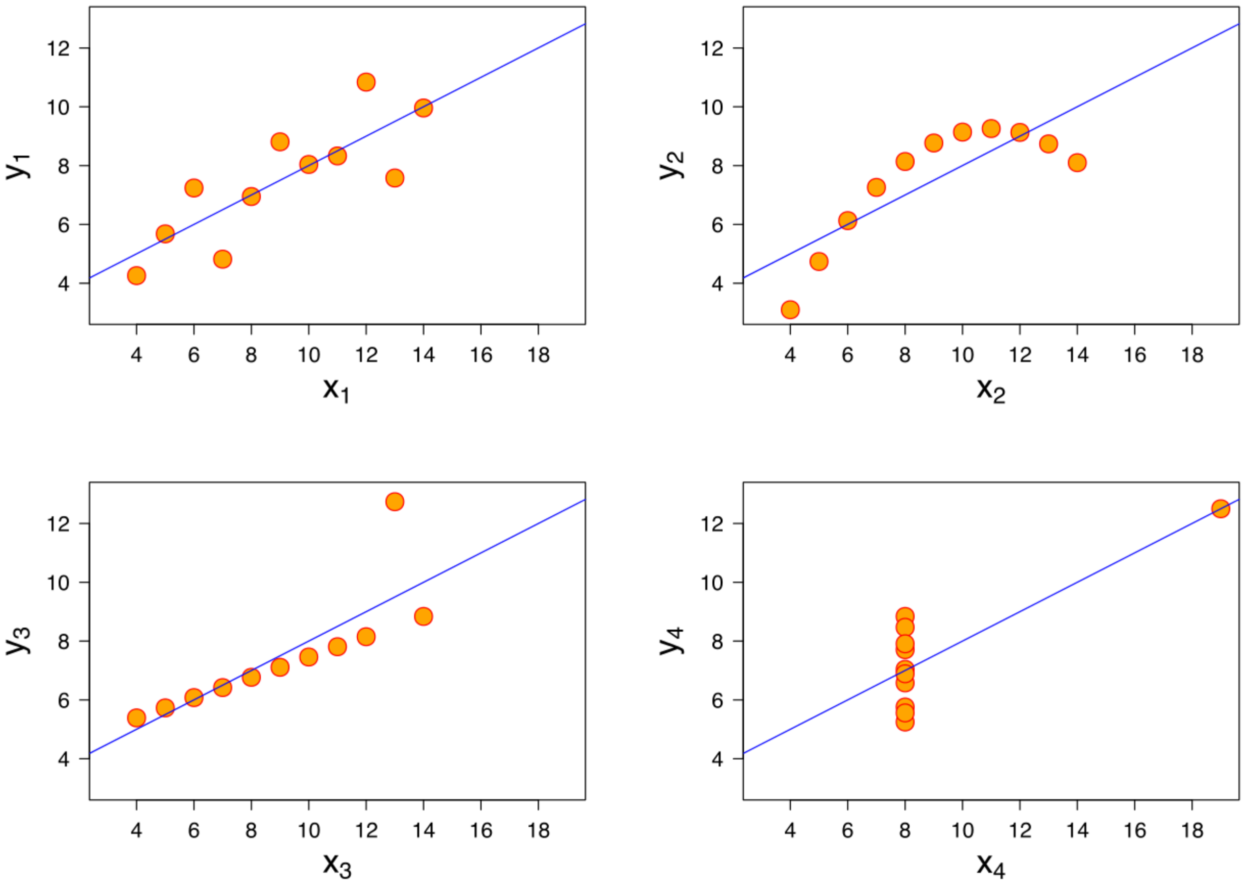


Linear Regression Analysis



$\hat{Y}_i = b_0 + b_1 X_i$
Y
Observed Value of Y for X_i
Y_i Actual
$e_i = Y_i - \hat{Y}_i$
Random Error for this X_i value
Predicted Value of Y for X_i
$\hat{Y}_i$ estimated
Slope = b_1
Intercept = b_0
X_i
X

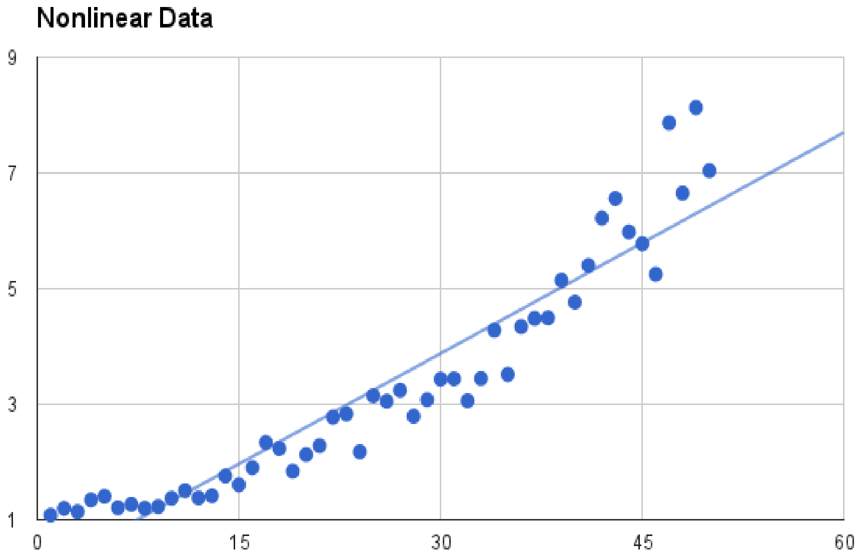
Importance of Graphics!

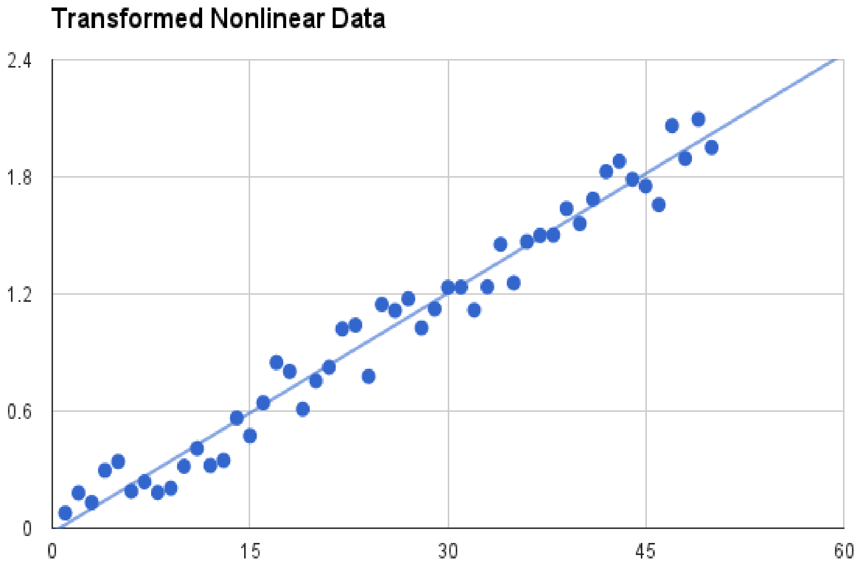


All four sets are identical when examined using simple summary statistics (i.e. mean, variance, correlation, and regression), but vary considerably when graphed

Linear Regression

- When the regression model contains one dependent variable and one independent variable, we call the approach simple linear regression.

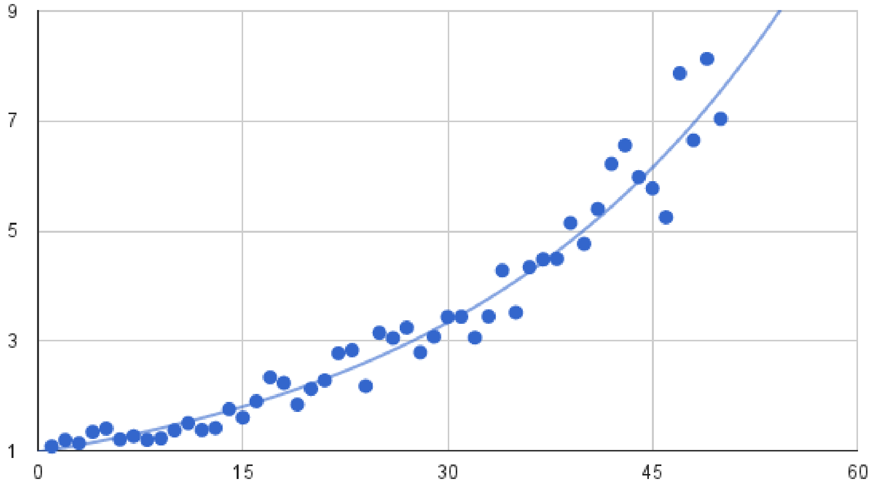




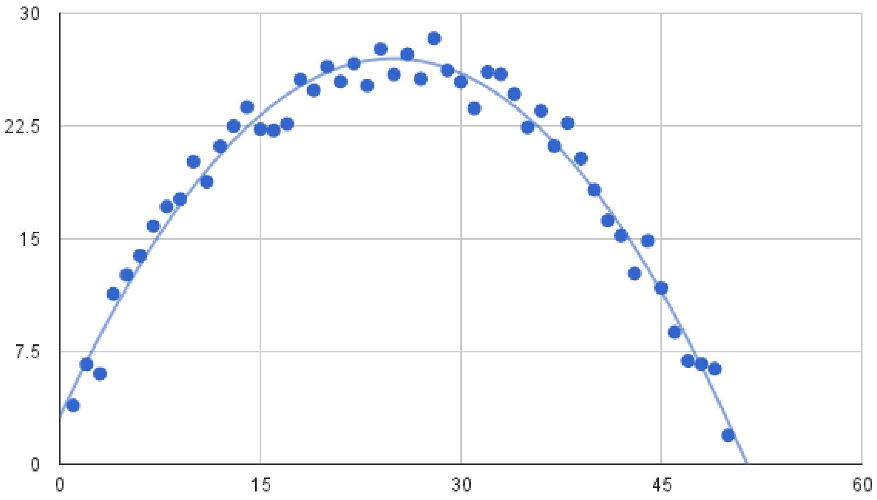
Nonlinear Regression

- When there’s one predictor variable but powers of the variable are included (e.g. X^2, X^3, etc.), we call it polynomial regression (e.g. quadratic or cubic regression).

Nonlinear Data with Exponential Trend

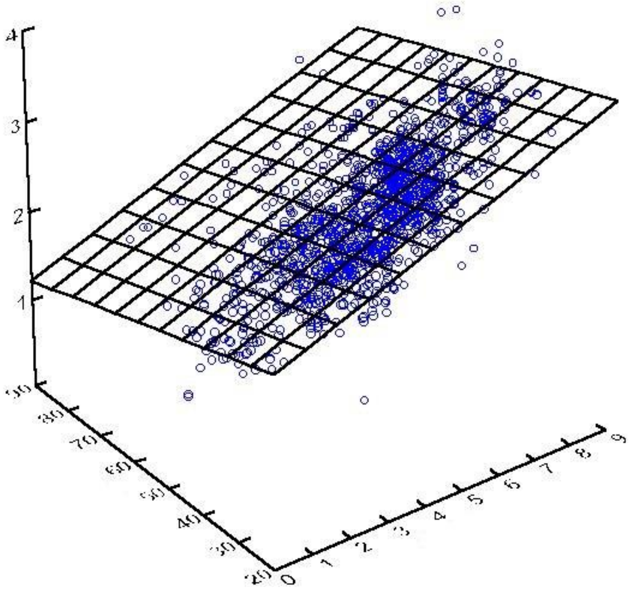


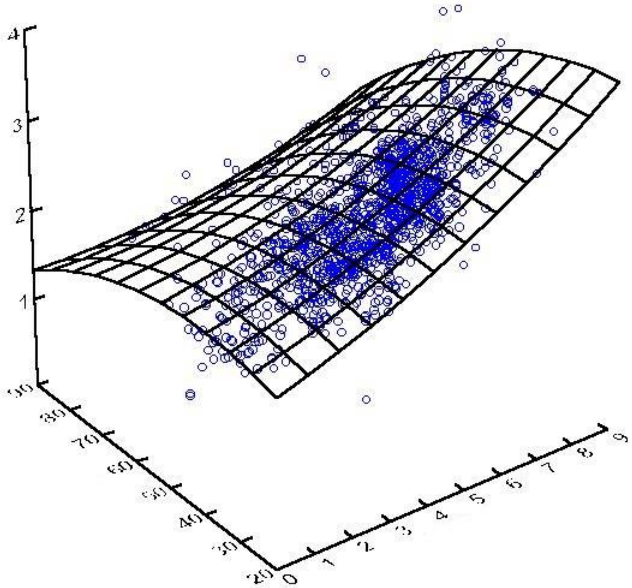
Nonlinear Data



Multiple Regression (Linear & Nonlinear)

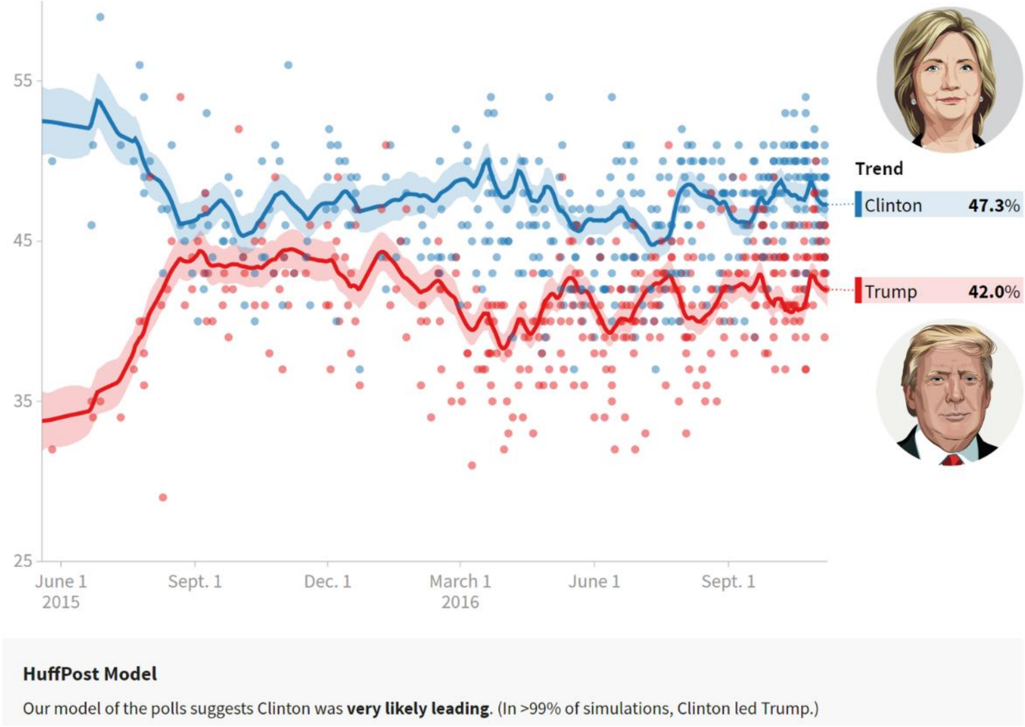
- When there’s more than one predictor variable (e.g. X_1, X_2, X_3, etc.), we call it multiple linear regression.





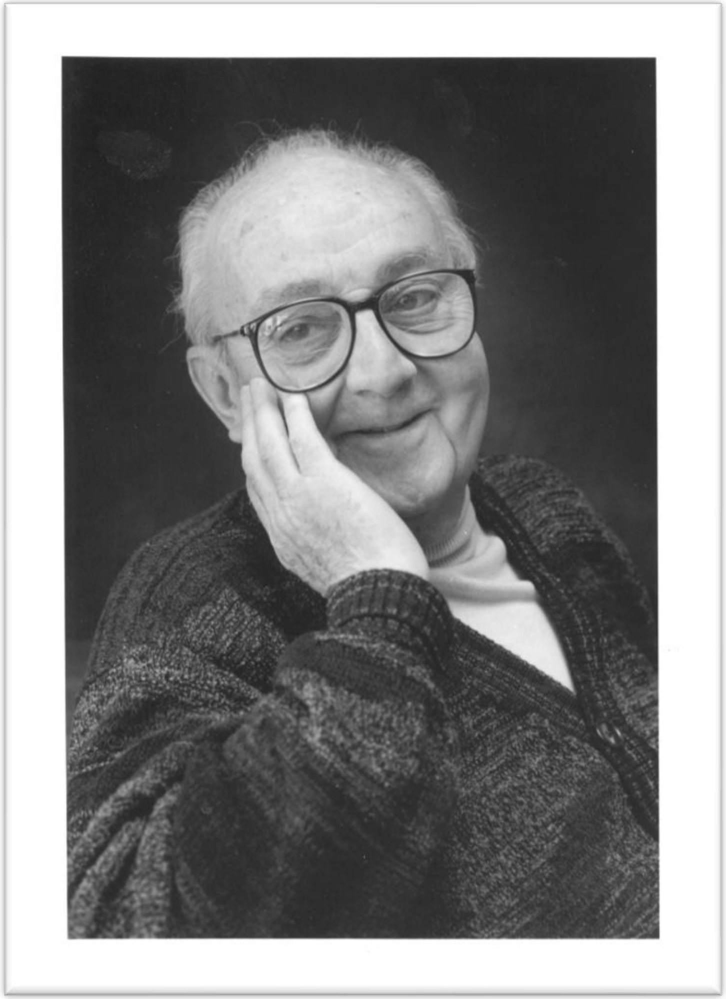
Models Uncertainty

Essentially, all models are wrong, but some are useful



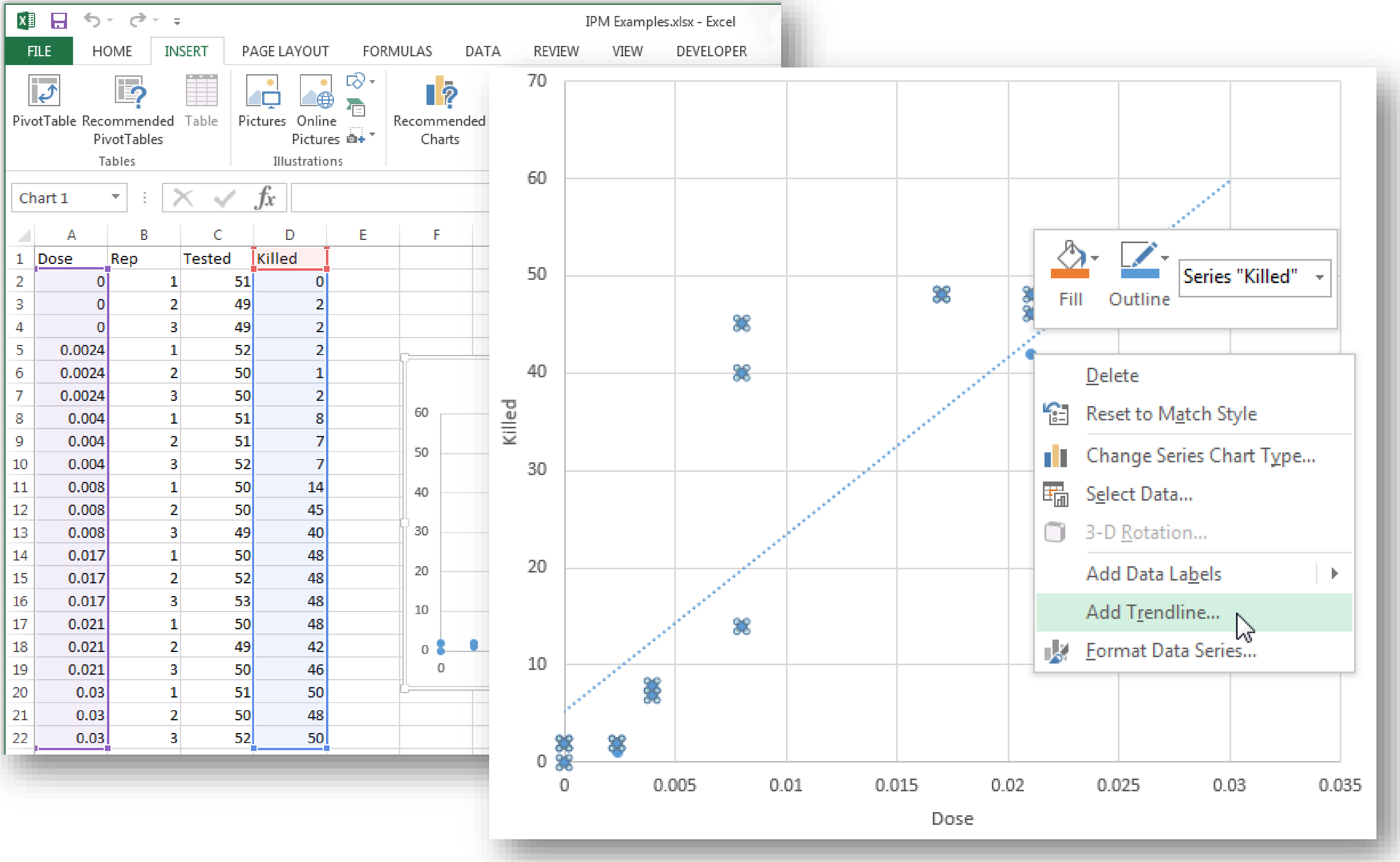
HuffPost Model

Our model of the polls suggests Clinton was **very likely leading**. (In >99% of simulations, Clinton led Trump.)



George E. P. Box

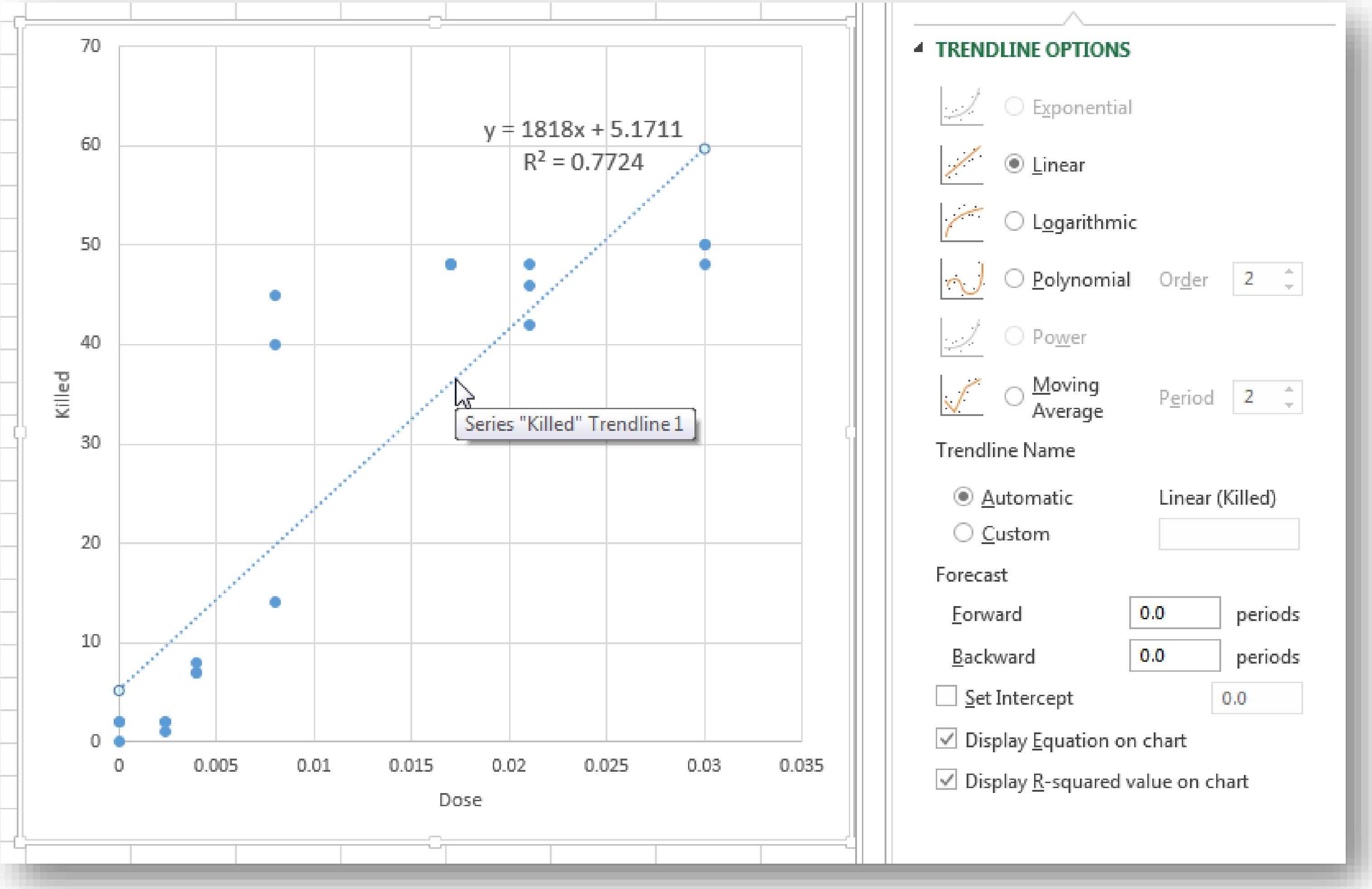
Excel – Scatter Plot & Add Trendline



	A	B	C	D
1	Dose	Rep	Tested	Killed
2	0	1	51	0
3	0	2	49	2
4	0	3	49	2
5	0.0024	1	52	2
6	0.0024	2	50	1
7	0.0024	3	50	2
8	0.004	1	51	8
9	0.004	2	51	7
10	0.004	3	52	7
11	0.008	1	50	14
12	0.008	2	50	45
13	0.008	3	49	40
14	0.017	1	50	48
15	0.017	2	52	48
16	0.017	3	53	48
17	0.021	1	50	48
18	0.021	2	49	42
19	0.021	3	50	46
20	0.03	1	51	50
21	0.03	2	50	48
22	0.03	3	52	50

ICARDA
Science for Better Livelihoods in Dry Areas

Excel – Linear Regression



TRENDLINE OPTIONS

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Killed)

○ Custom

Forecast

Forward 0.0 periods

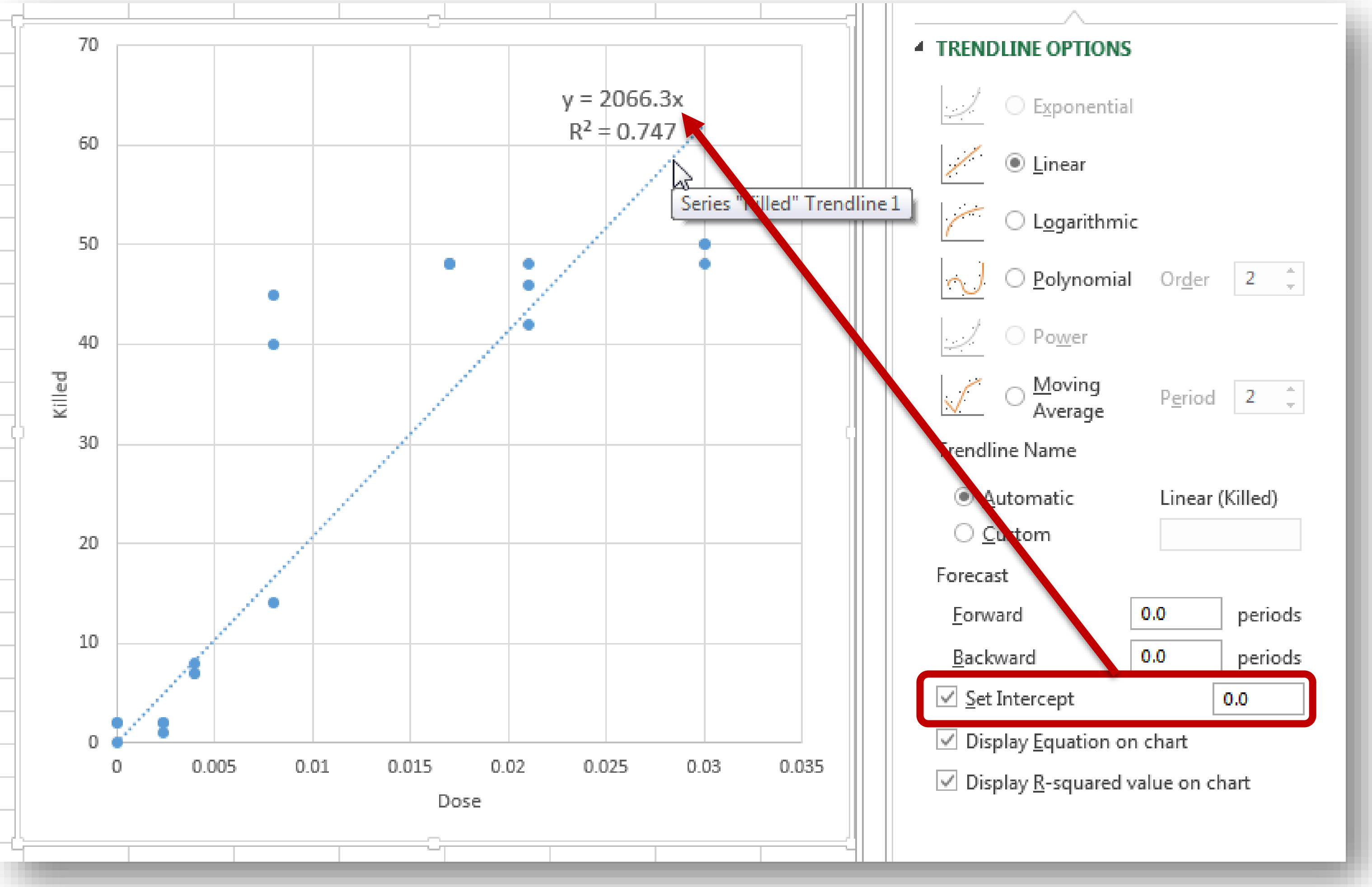
Backward 0.0 periods

□ Set Intercept 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

Excel – Linear Regression, No Intercept



TRENDLINE OPTIONS

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Killed)

○ Custom

Forecast

Forward 0.0 periods

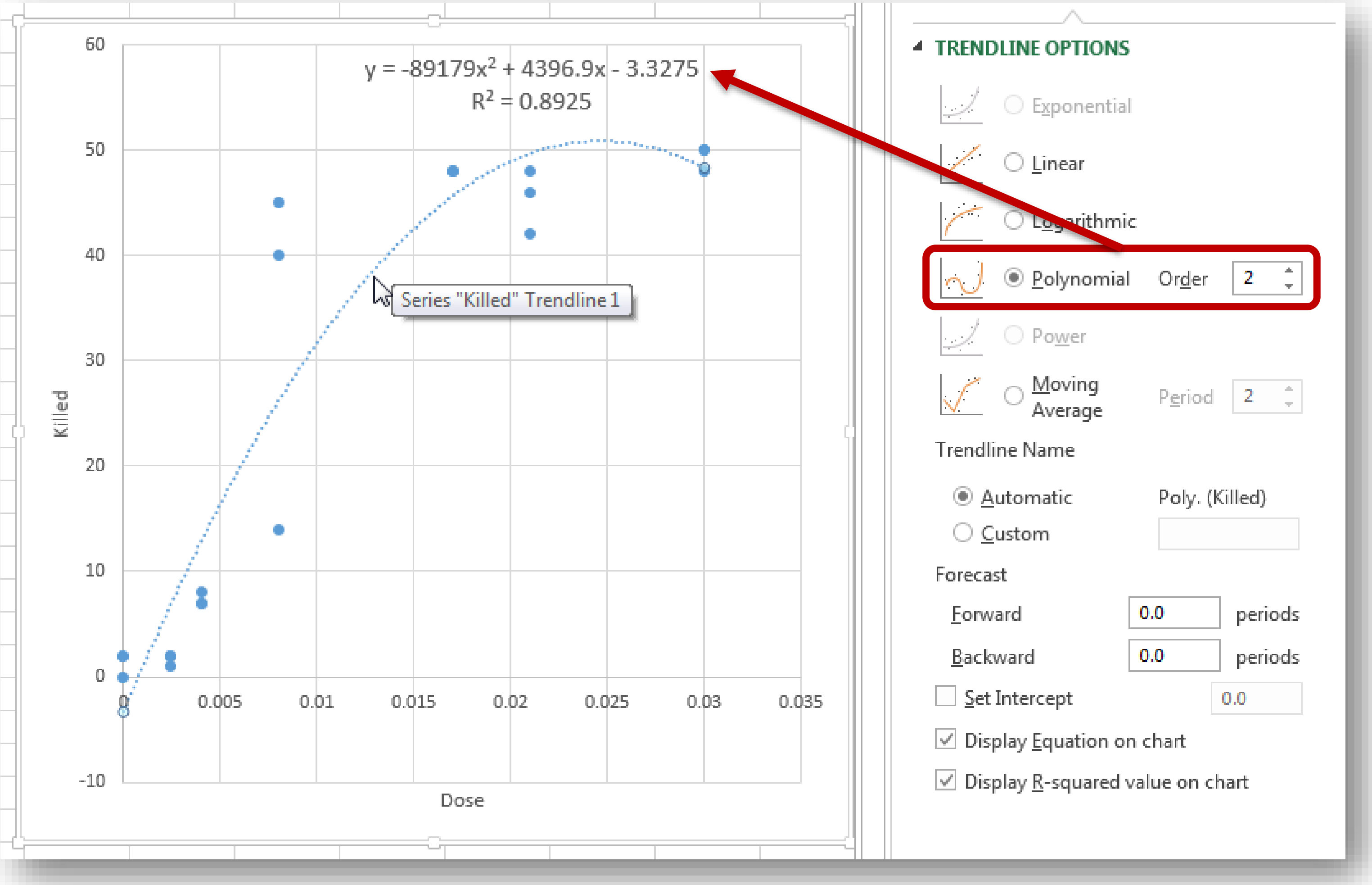
Backward 0.0 periods

☑ Set Intercept 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

Excel - Nonlinear Regression, Polynomial



TRENDLINE OPTIONS

○ Exponential

○ Linear

○ Logarithmic

● Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Poly. (Killed)

○ Custom

Forecast

Forward 0.0 periods

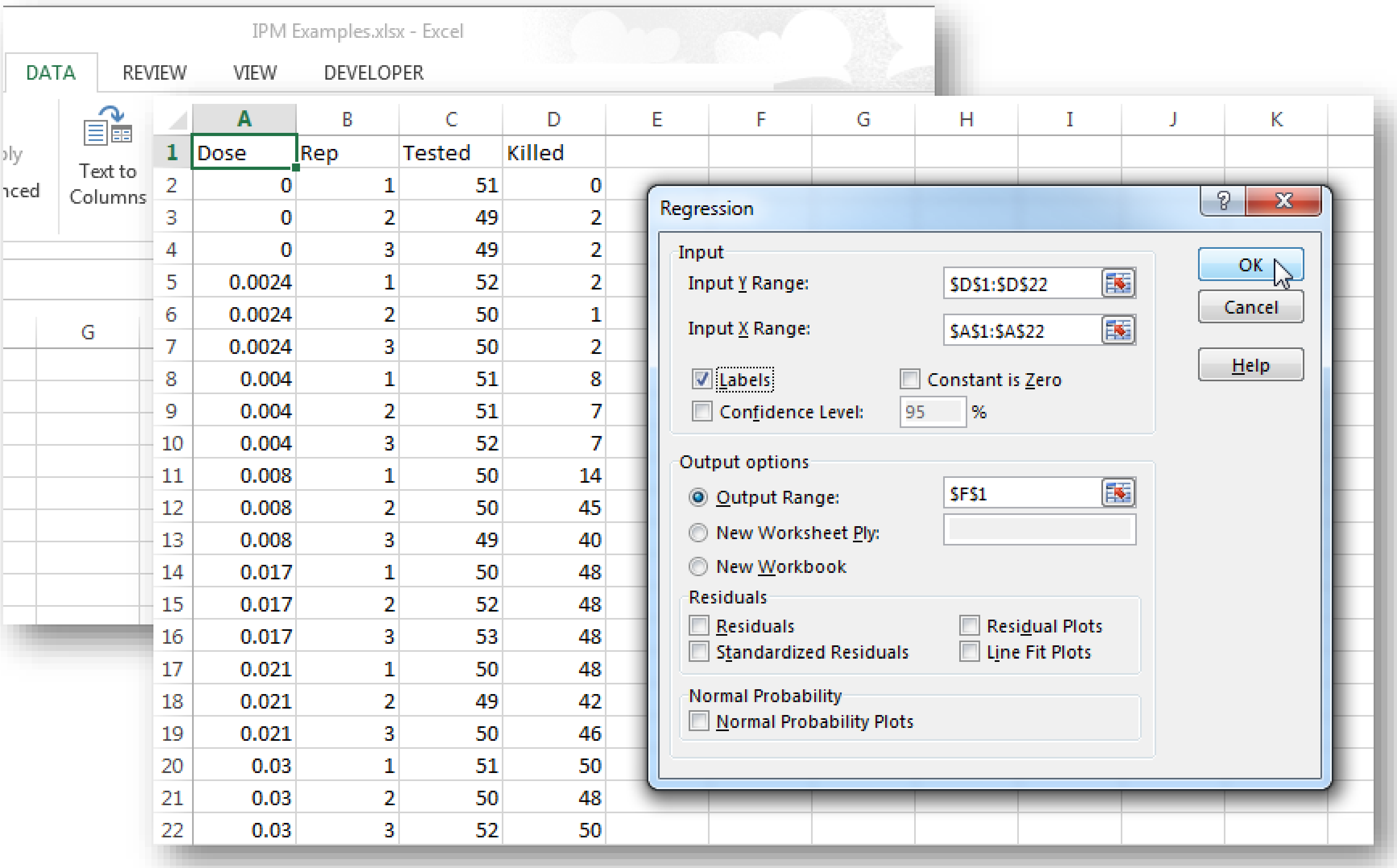
Backward 0.0 periods

□ Set Intercept 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

Excel (Analysis ToolPak) - Regression



	A	B	C	D
1	Dose	Rep	Tested	Killed
2	0	1	51	0
3	0	2	49	2
4	0	3	49	2
5	0.0024	1	52	2
6	0.0024	2	50	1
7	0.0024	3	50	2
8	0.004	1	51	8
9	0.004	2	51	7
10	0.004	3	52	7
11	0.008	1	50	14
12	0.008	2	50	45
13	0.008	3	49	40
14	0.017	1	50	48
15	0.017	2	52	48
16	0.017	3	53	48
17	0.021	1	50	48
18	0.021	2	49	42
19	0.021	3	50	46
20	0.03	1	51	50
21	0.03	2	50	48
22	0.03	3	52	50

Regression

Input

Input Y Range: $D$1:$D$22

Input X Range: $A$1:$A$22

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

Residuals

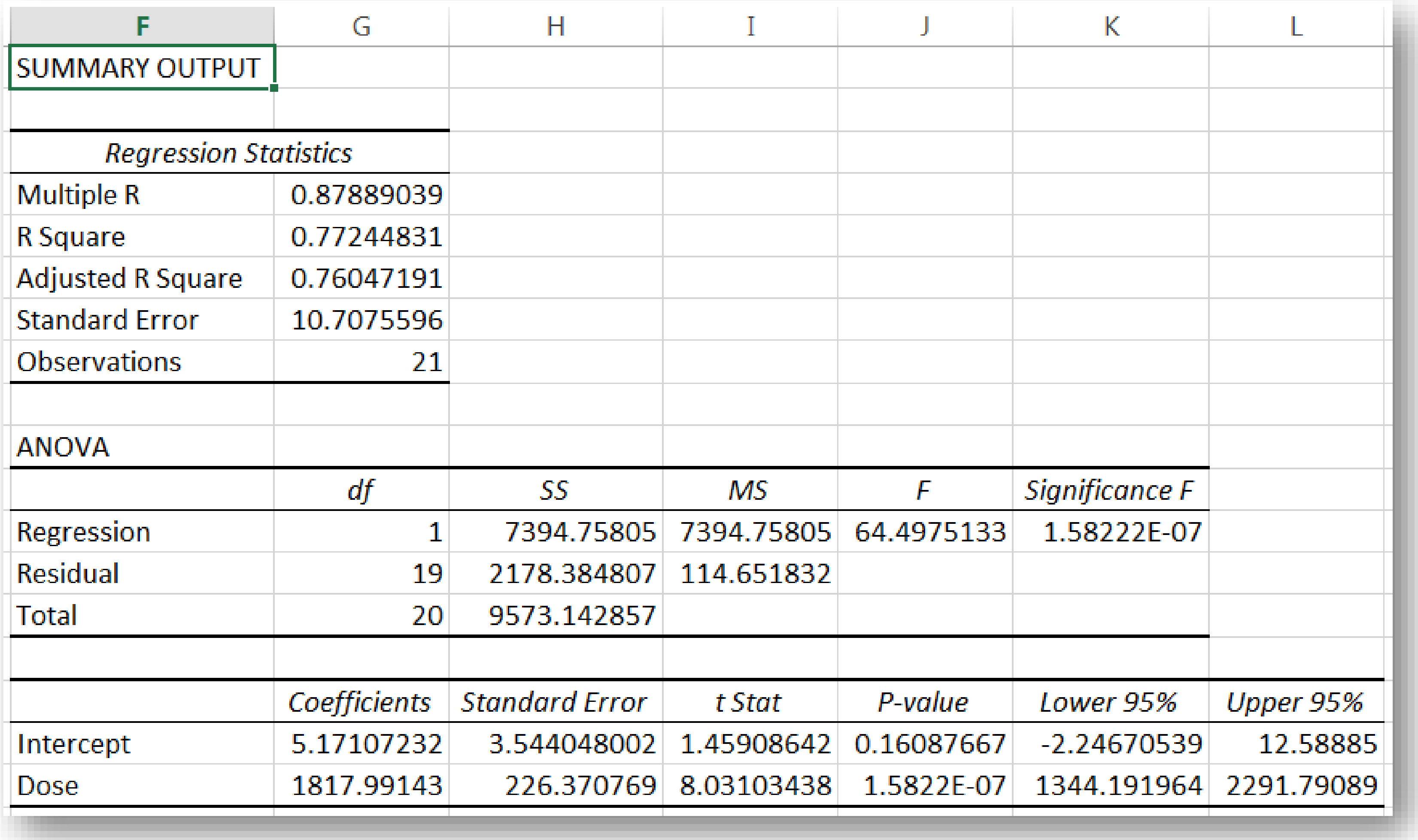
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Excel (Analysis ToolPak) – Regression *(continue)*



SUMMARY OUTPUT

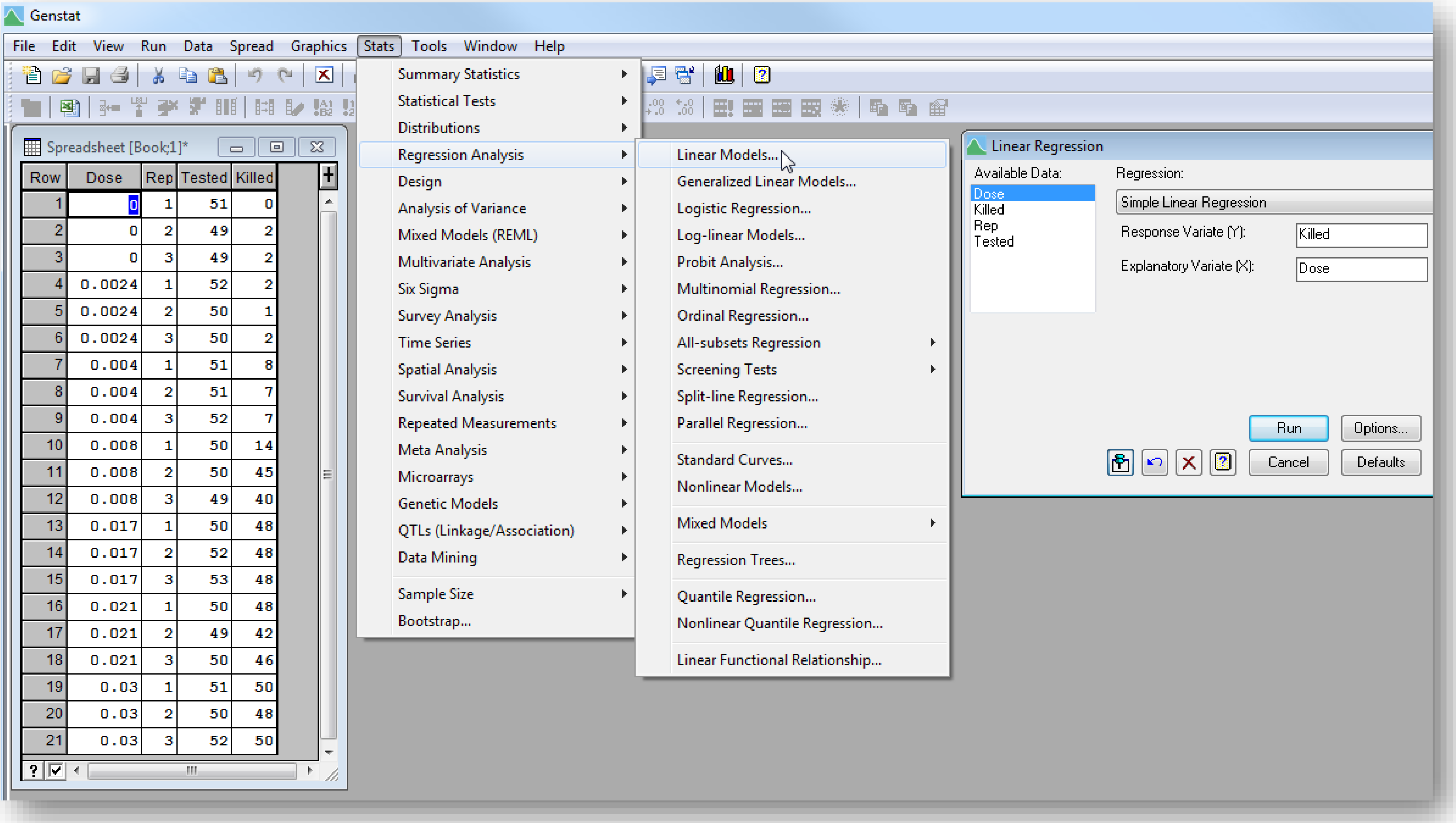
Regression Statistics	
Multiple R	0.87889039
R Square	0.77244831
Adjusted R Square	0.76047191
Standard Error	10.7075596
Observations	21

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7394.75805	7394.75805	64.4975133	1.58222E-07
Residual	19	2178.384807	114.651832		
Total	20	9573.142857			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.17107232	3.544048002	1.45908642	0.16087667	-2.24670539	12.58885
Dose	1817.99143	226.370769	8.03103438	1.5822E-07	1344.191964	2291.79089

GenStat – Simple Linear Regression



Spreadsheet [Book;1]*

Row	Dose	Rep	Tested	Killed
1	0	1	51	0
2	0	2	49	2
3	0	3	49	2
4	0.0024	1	52	2
5	0.0024	2	50	1
6	0.0024	3	50	2
7	0.004	1	51	8
8	0.004	2	51	7
9	0.004	3	52	7
10	0.008	1	50	14
11	0.008	2	50	45
12	0.008	3	49	40
13	0.017	1	50	48
14	0.017	2	52	48
15	0.017	3	53	48
16	0.021	1	50	48
17	0.021	2	49	42
18	0.021	3	50	46
19	0.03	1	51	50
20	0.03	2	50	48
21	0.03	3	52	50

Linear Regression

Available Data:
Dose
Killed
Rep
Tested

Regression: Simple Linear Regression

Response Variate (Y): Killed

Explanatory Variate (X): Dose

GenStat – Simple Linear Regression *(continue)*

```
"Simple Linear Regression"
MODEL Killed
TERMS Dose
FIT [PRINT=model,summary,estimates; CONSTANT=estimate;
```

Regression analysis

```
        Response variate:  Killed
            Fitted terms:  Constant, Dose
```

Summary of analysis

```
Source             d.f.        s.s.        m.s.      v.r.    F pr.
Regression            1       7395.      7394.8     64.50    <.001
Residual             19       2178.       114.7
Total                20       9573.       478.7

Percentage variance accounted for 76.0
Standard error of observations is estimated to be 10.7.
```

Message: the following units have large standardized residuals.

```
            Unit       Response       Residual
              11           45.0           2.43
```

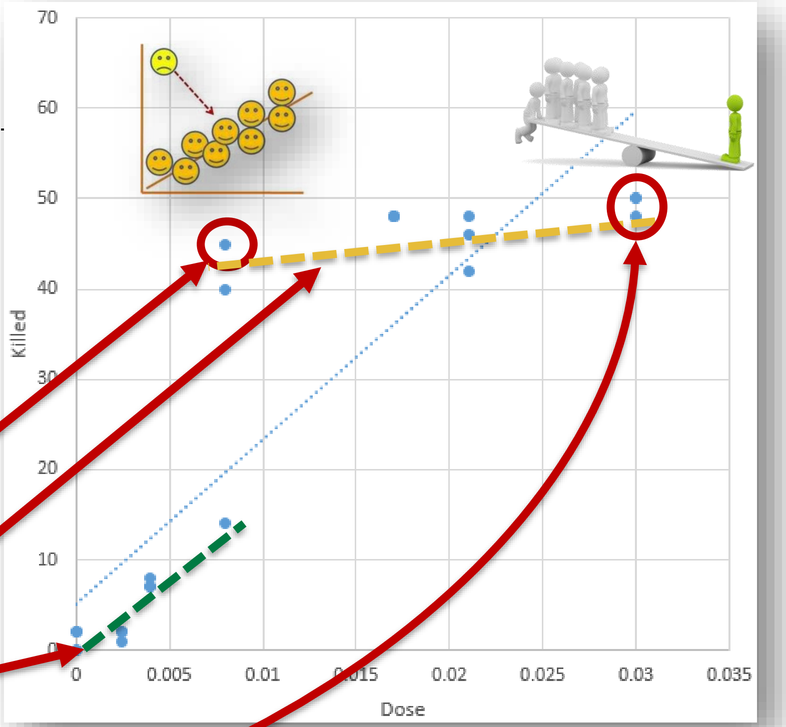
Message: the residuals do not appear to be random; for example, fitted values in the range 5.2 to 19.7 are consistently larger than observed values and fitted values in the range 19.7 to 43.3 are consistently smaller than observed values.

Message: the following units have high leverage.

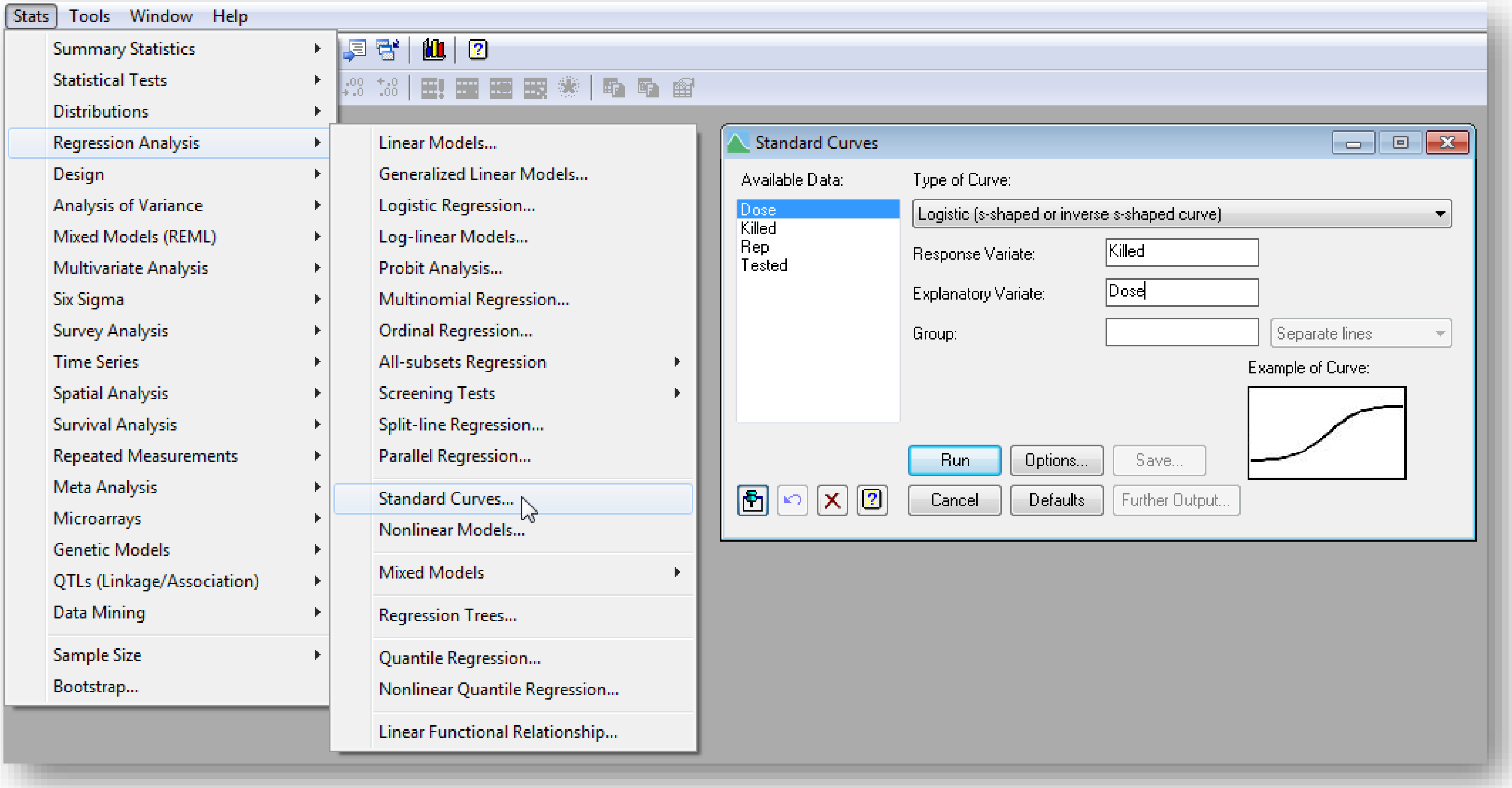
```
            Unit       Response       Leverage
              19           50.0          0.196
              20           48.0          0.196
              21           50.0          0.196
```

Estimates of parameters

```
Parameter          estimate          s.e.       t(19)     t pr.
Constant               5.17          3.54        1.46     0.161
Dose                  1818.          226.        8.03     <.001
```



GenStat – Regression, Standard Curves



Available Data: Dose, Killed, Rep, Tested

Type of Curve: Logistic (s-shaped or inverse s-shaped curve)

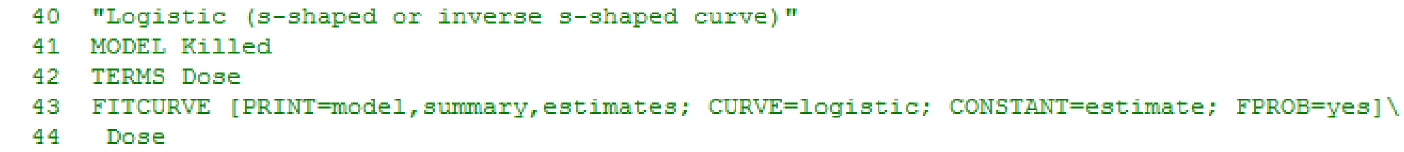
Response Variate: Killed

Explanatory Variate: Dose

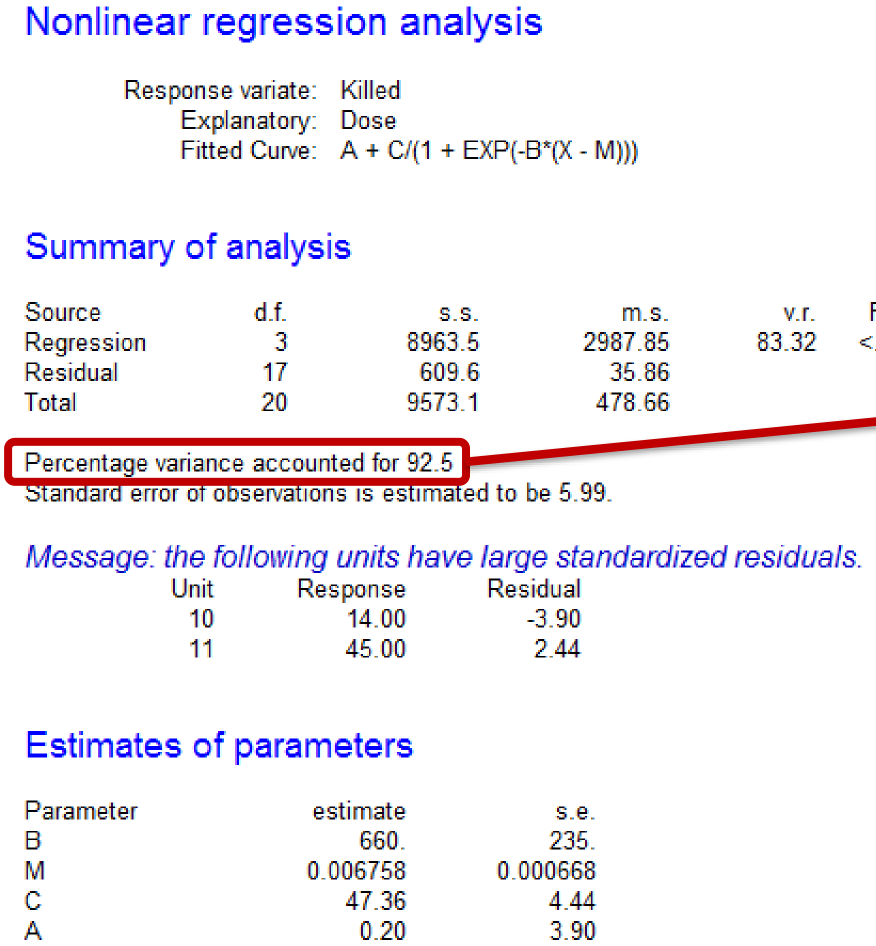
Group:

Separate lines

GenStat – Logistic Regression (S-Shaped Curve)



```
"Logistic (s-shaped or inverse s-shaped curve)"
MODEL Killed
TERMS Dose
FITCURVE [PRINT=model,summary,estimates; CURVE=logistic; CONSTANT=estimate; FPROB=yes]\
 Dose
```



Nonlinear regression analysis

```
    Response variate:  Killed
         Explanatory:  Dose
        Fitted Curve:  A + C/(1 + EXP(-B*(X - M)))
```

Summary of analysis

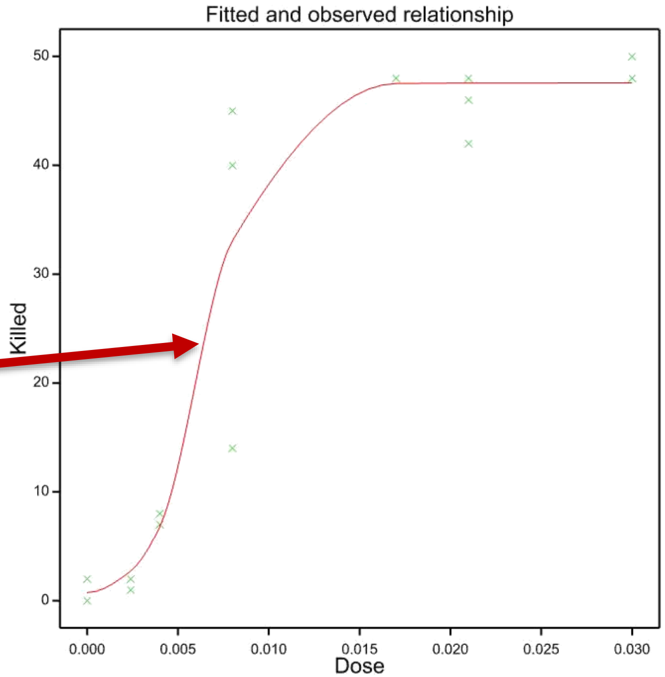
```
Source            d.f.        s.s.        m.s.      v.r.   F pr.
Regression           3      8963.5     2987.85     83.32   <.001
Residual            17       609.6       35.86
Total               20      9573.1      478.66

Percentage variance accounted for 92.5
Standard error of observations is estimated to be 5.99.
```

Message: the following units have large standardized residuals.

```
        Unit      Response      Residual
          10         14.00         -3.90
          11         45.00          2.44
```

Estimates of parameters

```
Parameter           estimate          s.e.
B                       660.          235.
M                   0.006758      0.000668
C                      47.36          4.44
A                       0.20          3.90
```



Thank You

Questions?

Japanese attitude for work:

If one can do it, I can do it. If no one can do it, I must do it.

Middle Eastern attitude for work:

Wallahi... if one can do it, let him do it.
If no one can do it, ya-habibi how can I do it?

Statistical Details (Skip if you'd like)

$$x_1, x_2, \dots, x_n \qquad \sim N(\mu, \sigma^2)$$

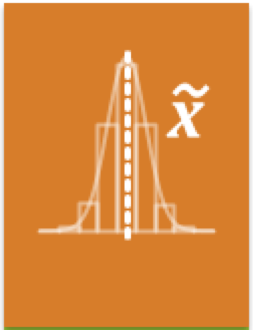
$$\bar{x} = \frac{\sum x_i}{n}$$

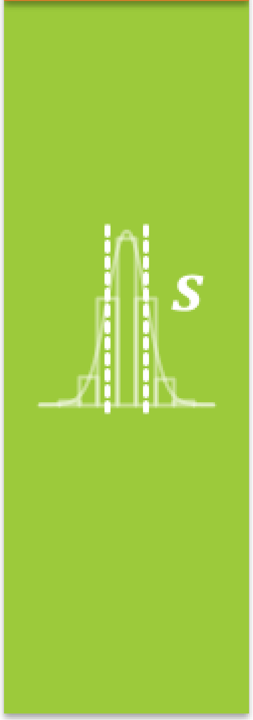
$$Var(x) = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$SD(x) = \sigma = \sqrt{Var(x)}$$

$$Z_i = \frac{x_i - \bar{x}}{SD(x)} \qquad \sim N(0,1)$$

$$t = \frac{\bar{x} - \mu}{SD(x)/\sqrt{n}} \qquad \sim t(n-1)$$





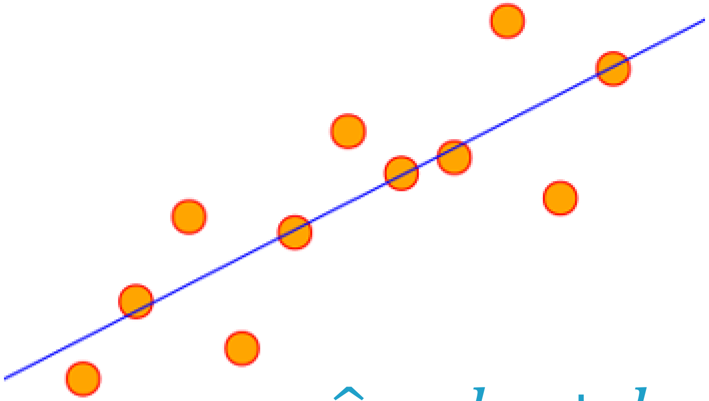
Statistical Details (Skip if you'd like)

$$Cov(x,y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$Cor(x,y) = \frac{Cov(x,y)}{\sigma_x \, \sigma_y}$$

$$b_1 = Cor(x,y)\frac{SD(y)}{SD(x)}$$

$$b_0 = \bar{y} - b_1\bar{x}$$



$$\hat{y} = b_0 + b_1 . x$$