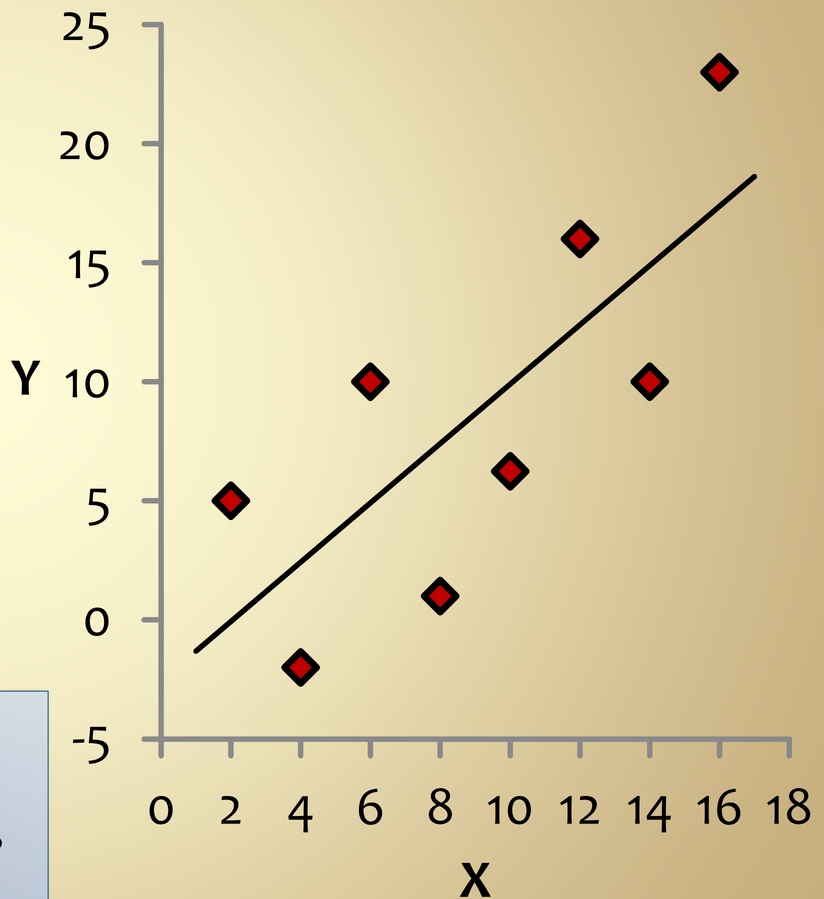
Regression Statistics

Engineers and Regression

Engineers often:

- Regress data to a model
  - Used for assessing theory
  - Used for predicting
  - Empirical or theoretical model
- Use the regression of others
  - Antoine Equation
  - DIPPR



What are uncertainties of regression?

- Do the data fit the model?
- What are the errors in the prediction?
- What are the errors in the parameters?

Linear Regression

- Two classes for a model
  - Linear
  - Non-linear
- "Linear" refers to parameters, not the dependent variable (i.e. written in matrix form- linear algebra)
- You can use the Mathcad function "linfit" on linear equations

Example

$$y = ax^2 + bx + c$$

$$y = \begin{bmatrix} x^2 & x & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

- parameters are linear
- x is not linear

Quiz

Which models can you use linear regression?

1. $$y = ax^2 + bx + c$$
2. $$y = ae^{bx}$$
3. $$y = a + \frac{b}{T} + \frac{c}{T^3} + \frac{d}{T^4} + \frac{e}{T^5}$$
4. $$y = \exp\left(A - \frac{B}{T + C}\right)$$
5. $$y = mx + b$$

Regression- Straight line

- $Y= b_0 + b_1X$
- Two parameters

Practice

- Open Excel Regression example (1^{st} order Example)
- If necessary, add-in the Data Analysis ToolPak
- Regress the data in Columns B and C- select 95 % confidence and residual plot. You may include data labels but you must select label.
- Copy the regressed slope and intercept into m and b cells in column H
- The next slides will help us understand the output

Straight Line Model

$$Y_i = b_0 + b_1 X_i + e_i$$

Intercept

slope

"X" data

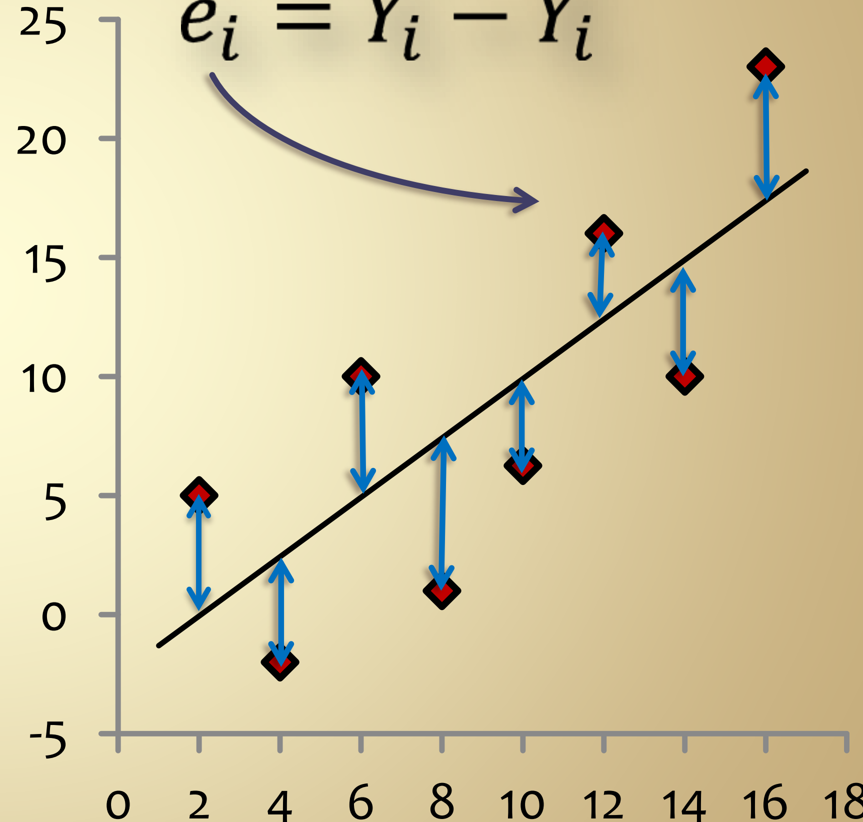
"Y" measured

$$\hat{Y}_i = b_0 + b_1 X_i$$

"Y" predicted

residual (error)

$$e_i = Y_i - \hat{Y}_i$$



Straight Line Model

X_i	Y_i	$\hat{Y}_i$	e_i
1	2.749032178	1.4390884[illegible]3	1.309943694
2	3.719910224	2.362003033	1.357907192
3	0.925995017	3.284917582	-2.35892257
4	2.623482686	4.207832132	-1.58434945
5	6.539797342	5.130746681	1.409050661
6	6.779909177	6.053661231	0.726247946
7	4.946150401	6.976575781	-2.03042538
8	9.674178069	7.89949033	1.774687739
9	7.61959821	8.82240488	-1.20280667
10	7.650020996	9.745319429	-2.09529843
11	11.514	10.66823398	0.845766021
12	13.18285068	11.59114853	1.591702152
13	13.28173635	12.51406308	0.767673275
14	13.60444592	13.43697763	0.16746829
15	12.79535218	14.35989218	-1.56454
16	17.82374778	15.28280673	2.540941056
17	14.55068379	16.20572128	-1.65503748
			42.76608602

"X" data

"Y" data

$$e_i = Y_i - \hat{Y}_i$$

residual (error)

"Y" predicted

$$\hat{Y}_i = b_0 + b_1 X_i$$

$b_0 = 0.923$, $b_1 = 0.516$

sum squared error

$$SS_E = \sum_{i=1}^{n} e_i^2$$

$$MS_E = \hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Number of fitted parameters:
2 for a two-parameter model

The R² Statistic

$$R^2 = \frac{SS\ due\ to\ regression}{(Total\ SS, corrected\ for\ the\ mean\ \bar{Y})}$$

$$= \frac{SS_R}{SS_T}$$

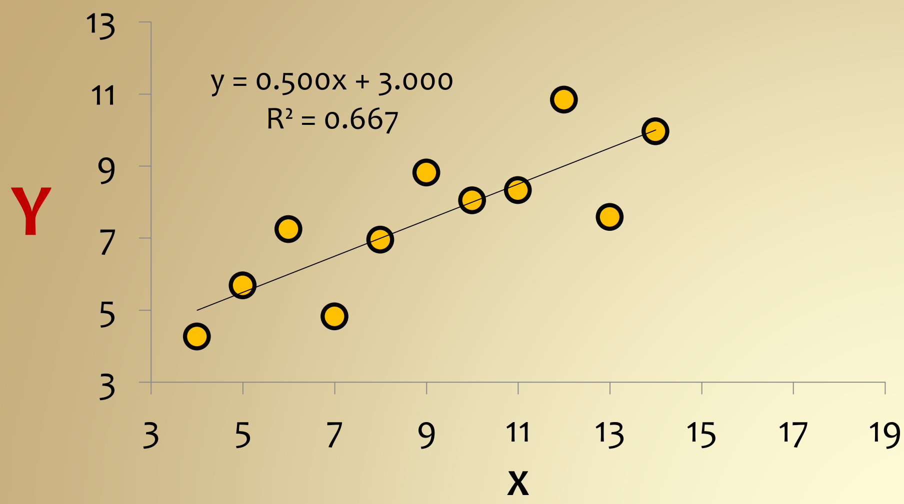
$$= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

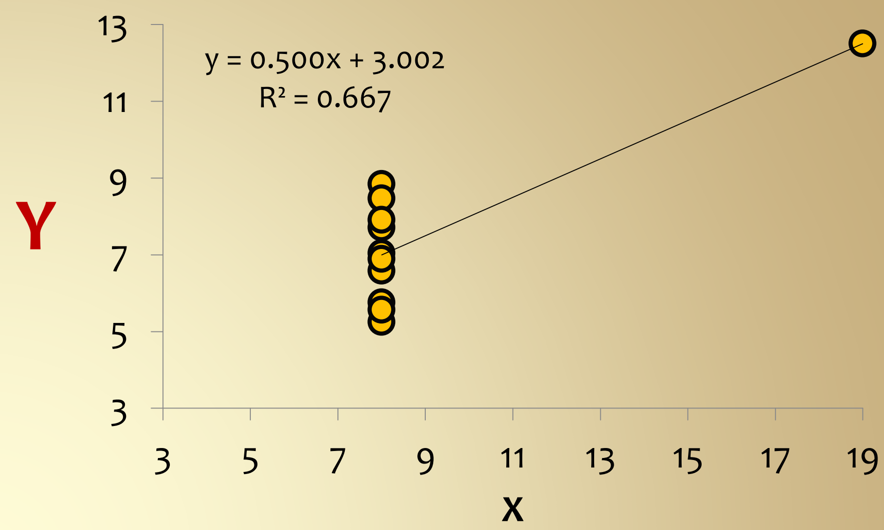
Average measured y value

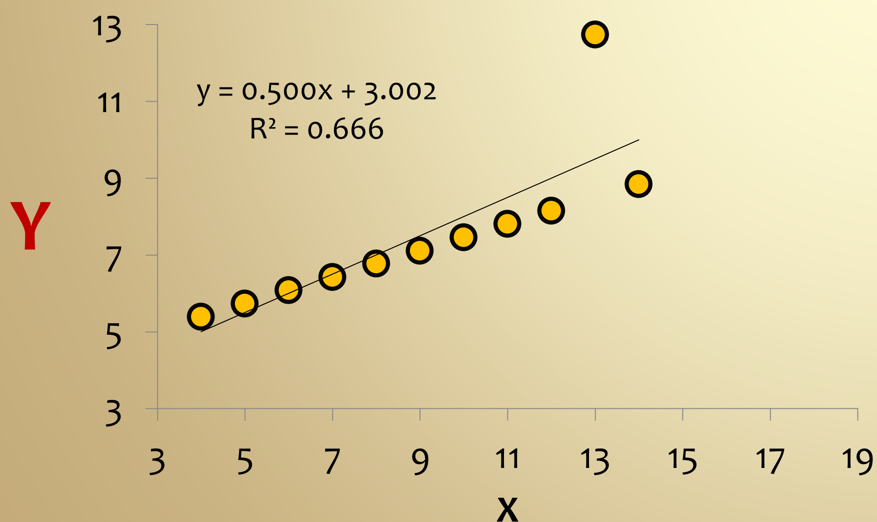
- A *useful* statistic but not definitive
- Tells you how well the data fit the model.
- It does *not* tell you if the model is correct.

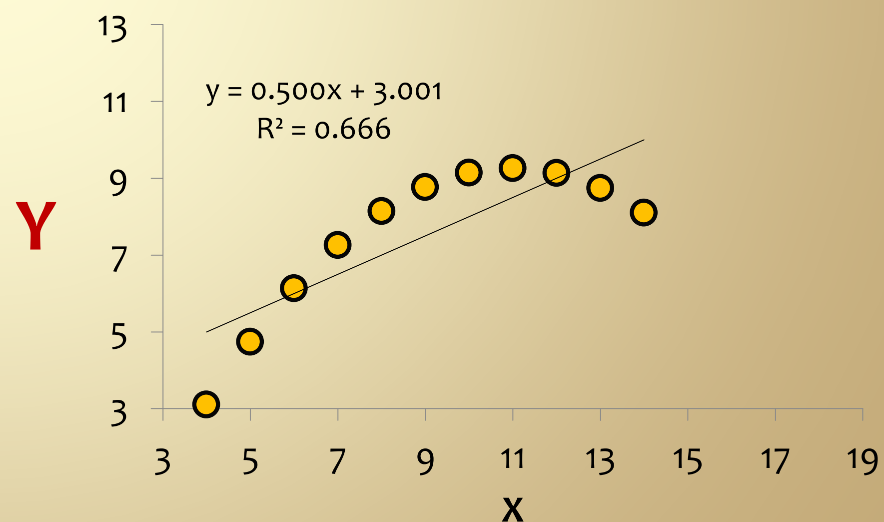
How much of the distribution of the data *about the mean* is described by the model.

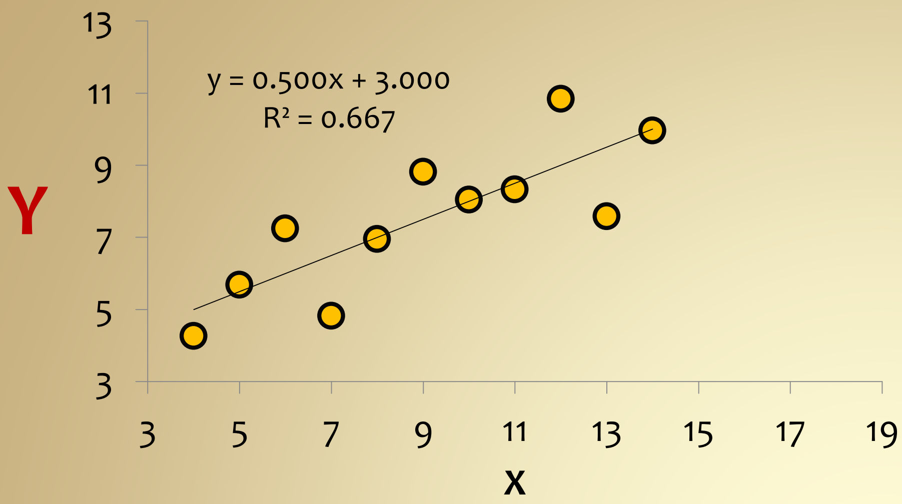
Problems with R^2

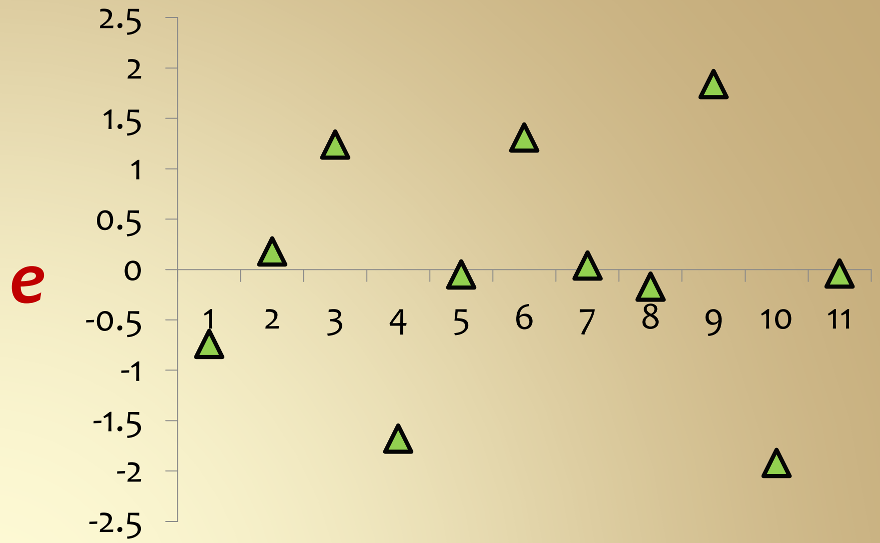




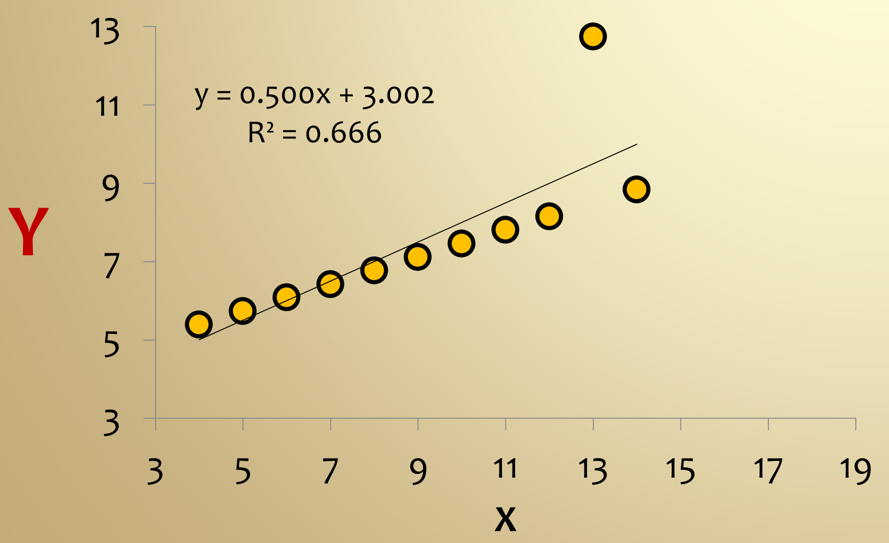


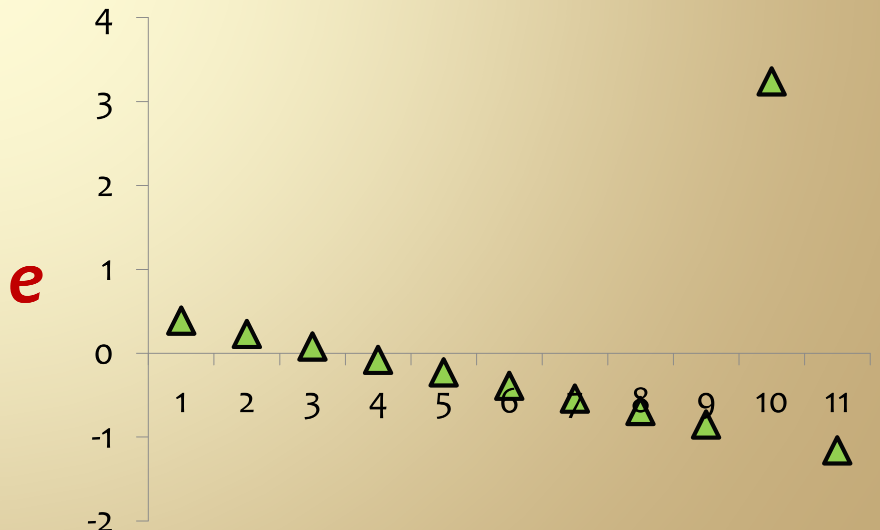


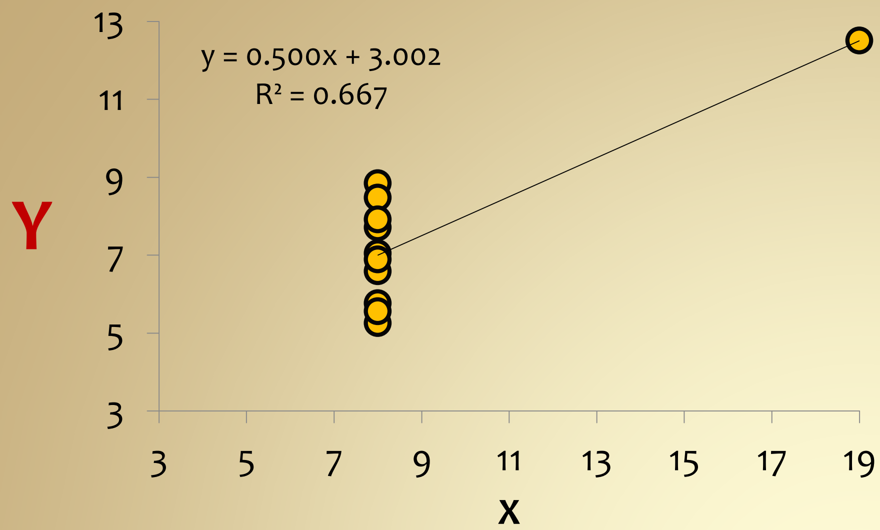


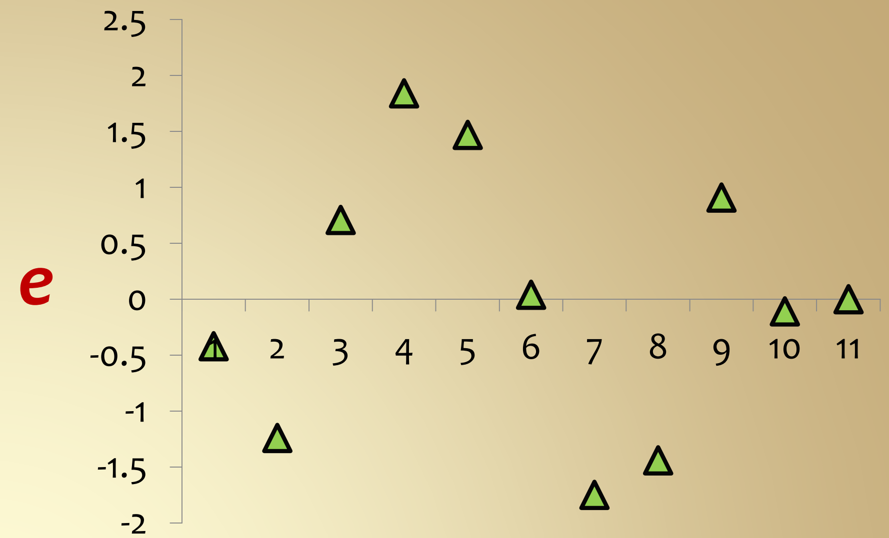


Residuals (e_i) should be normally distributed

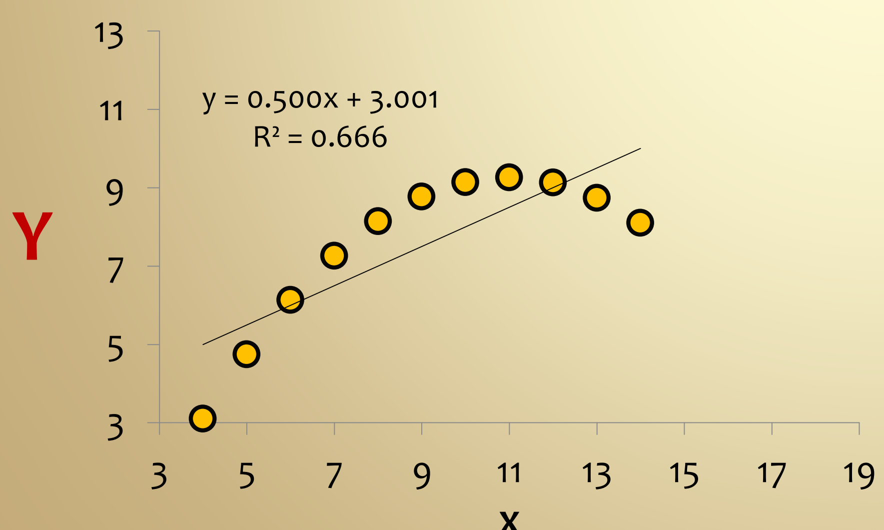


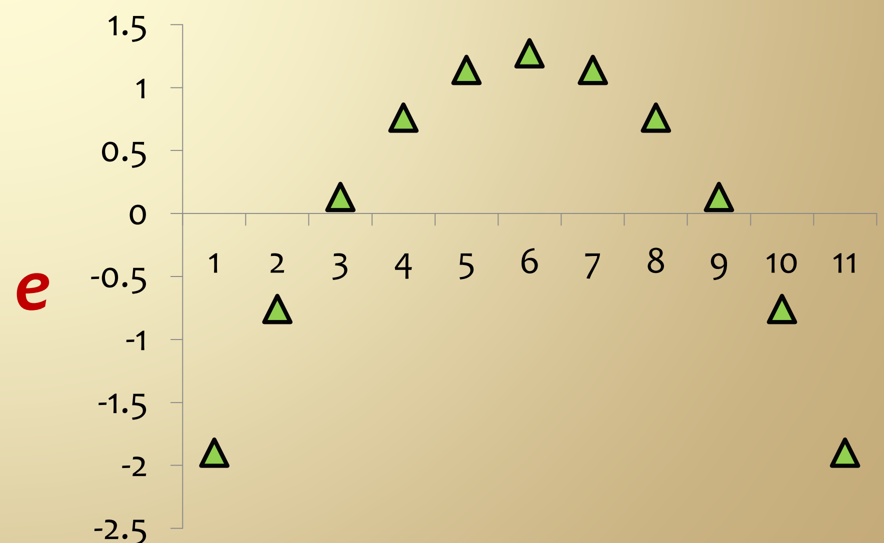






Residuals (e_i) should be normally distributed





Statistics: Slope/Intercept

(1-α)100% Confidence Intervals

$where\ t = f\left(\frac{\alpha}{2}, n - 2\right)$

Number of fitted parameters:
2 for a two-parameter model

intercept $b_0 \pm S_{b_0} t$

$$S_{b_0} = \left(\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

slope $b_1 \pm S_{b_1} t$

$$S_{b_1} = \left(\frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

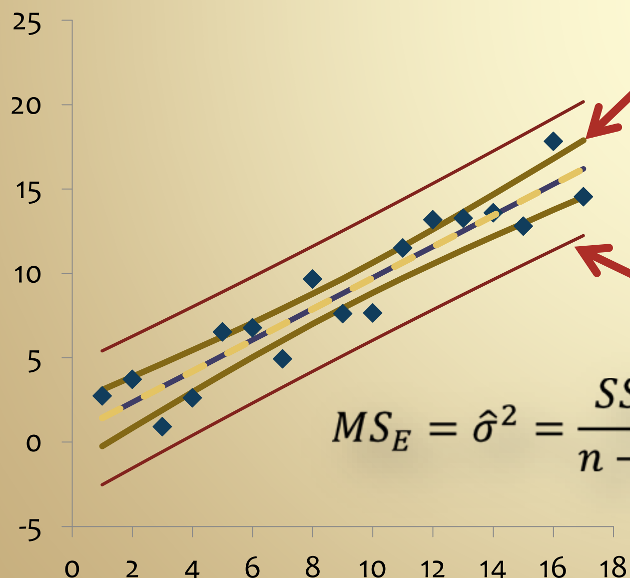
Called standard error in Excel

Statistics: Predicted Variable

(1-α)100% Confidence Intervals

$$Y_p = \hat{Y}_p \pm S_{\hat{Y}} t$$

$$where\ t = f\left(\frac{\alpha}{2}, n - 2\right)$$



$$MS_E = \hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Confidence Interval on the Prediction

Given a specific value for X what is the error in $\hat{Y}_P$?

n → ∞
S_Y → 0

$$S_{\hat{Y}} = \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

Expected Range of Data

If I take more data, where will the data fall? (Where will Y_P be found if measured at given X?)

n → ∞
S_Y → σ

$$S_{\hat{Y}} = \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

Generalized Linear Regression

- Linear regression can be written in matrix form.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

X	Y
21	186
24	214
32	288
47	425
50	455
59	539
68	622
74	675
62	562
50	453
41	370
30	274

Straight Line Model

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 21 \\ 1 & 24 \\ \vdots & \vdots \\ 1 & 41 \\ 1 & 30 \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} 186 \\ 214 \\ \vdots \\ 370 \\ 274 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix}$$

Quadratic Model

$$Y_i = b_0 + b_1 X_i + b_2 X_i^2 + e_i$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 21 & 441 \\ 1 & 24 & 576 \\ \vdots & \vdots & \vdots \\ 1 & 41 & 1681 \\ 1 & 30 & 900 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Excel Practice

1. Complete columns D-H in 1st order example- look at 95% confidence level of data.

2. Complete multiple regression Practice example
   a) Highlight all x columns for x input
   b) Note that x data could represent non-linear data such as T, 1/T, T^2, etc.
   c) Look at regressed parameter associated with x_1 and calculate the 95% confidence. Compare with Excel.
   d) Select residuals and residual plots. Note that there are two types of residual information: 1) each dependent variable has a residual plot where the x-axis is the data and the y-axis is the residual for the data; 2) the summary output gives a residual for y (actual Y – predicted Y). If desired, you can plot as a function of the # of data points.