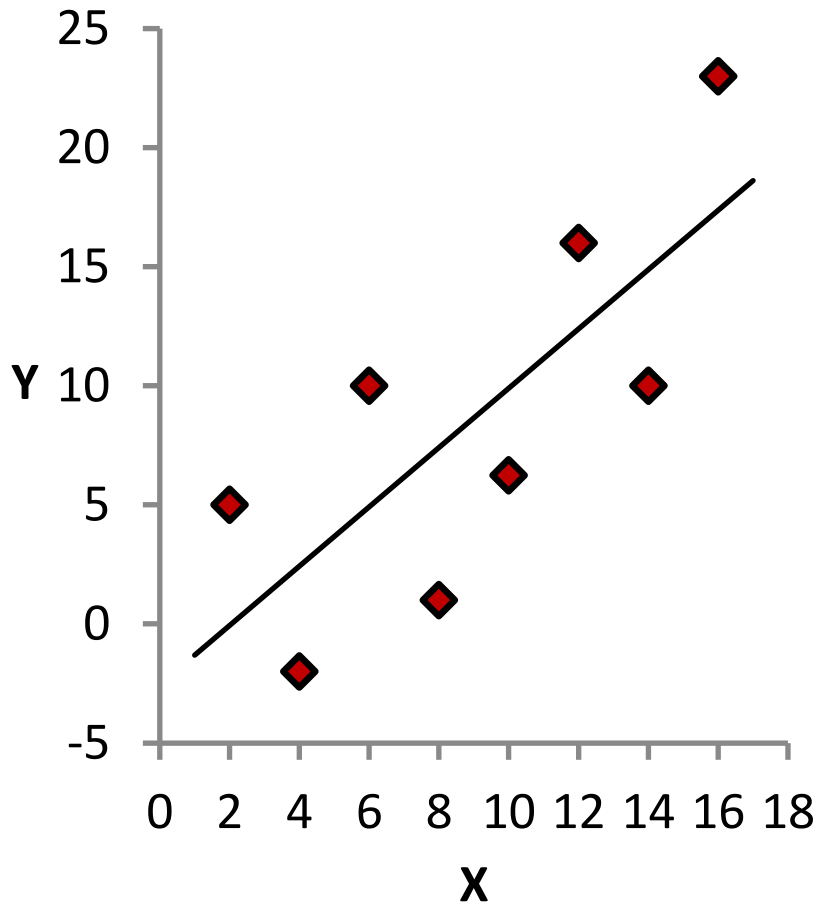
Linear Regression Statistics

Ch En 479

Engineers and Regression

- Regress data
  - Analysis
  - Fit to theory
  - Data reduction
- Use the regression of others
  - Antoine Equation
  - DIPPR



We need to be able to report uncertainties associated with regression.

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

Linear Regression

- Two types of regressions
  - Linear
  - Non-linear
- "Linear" refers to the parameters
  - Only one parameter per term
  - Parameter is not inside a function (sin, log, etc.)
- *Sensitivity coefficients* of linear models contain no model parameters.

Quiz

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. $y = mx + b$

Linear Regression

- Two types of regressions
  - Linear
  - Non-linear
- "Linear" refers to the parameters
  - Only one parameter per term
  - Parameter is not inside a function (sin, log, etc.)
- *Sensitivity coefficients* of linear models contain no model parameters.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\left(\frac{\partial y}{\partial \beta_0} = 1\right)$$

$$\left(\frac{\partial y}{\partial \beta_1} = x\right)$$

$$\left(\frac{\partial y}{\partial \beta_2} = x^2\right)$$

Straight Line Model

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

slope

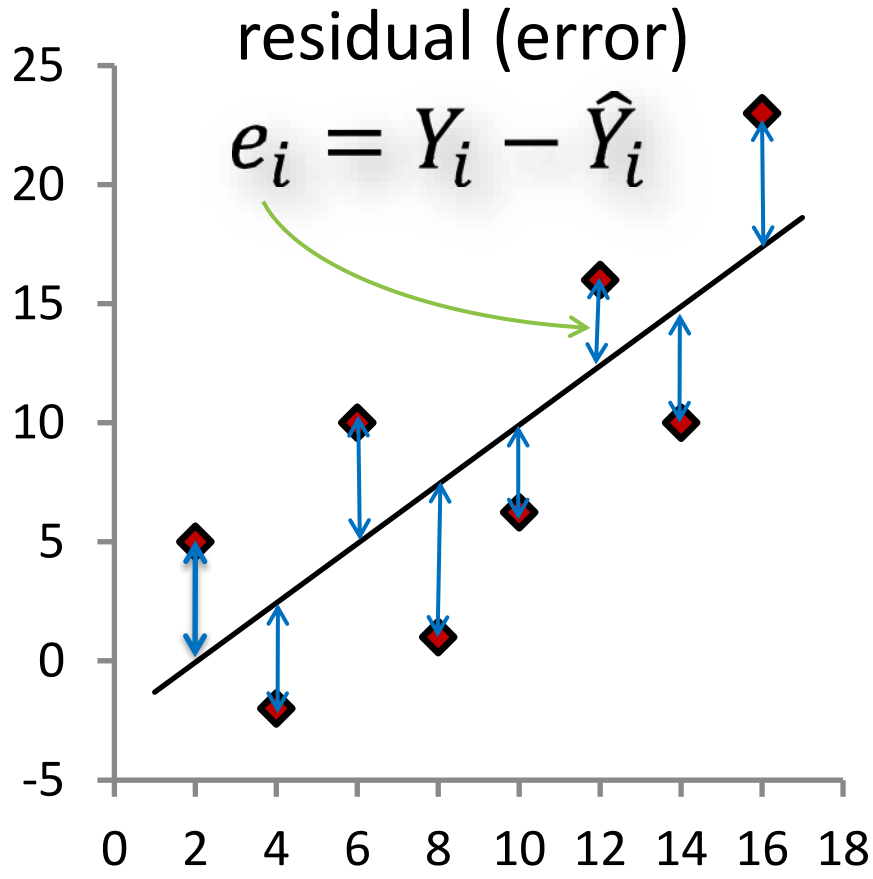
Intercept

"X" data

"Y" Measured

"Y" Predicted

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



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

intercept	0.92291455	slope	0.516173934

X_i	Y_i	$\hat{Y}_i$	e_i
1	2.749032178	1.439088483	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	SS_E

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

residual (error)

"X" data

"Y" data

"Y" predicted

sum squared error

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

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

mean squared error

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

The R^2 Statistic

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

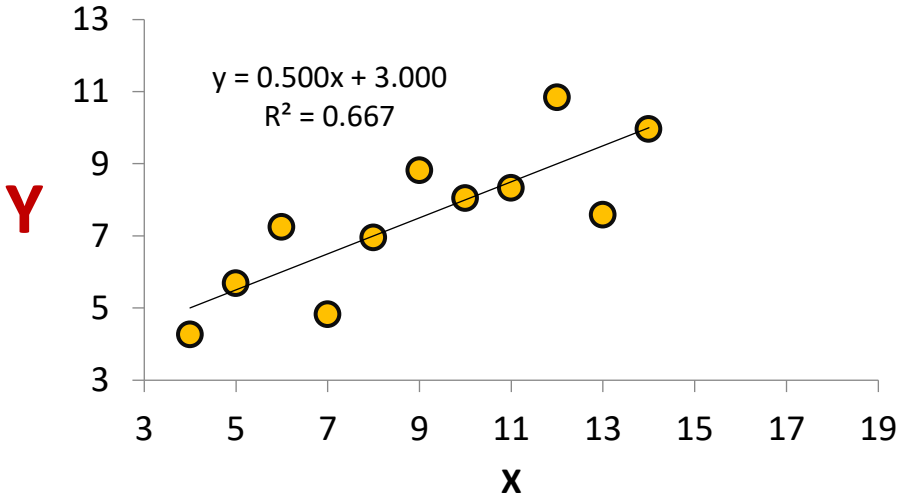
$$= \frac{SS_R}{SS_T}$$

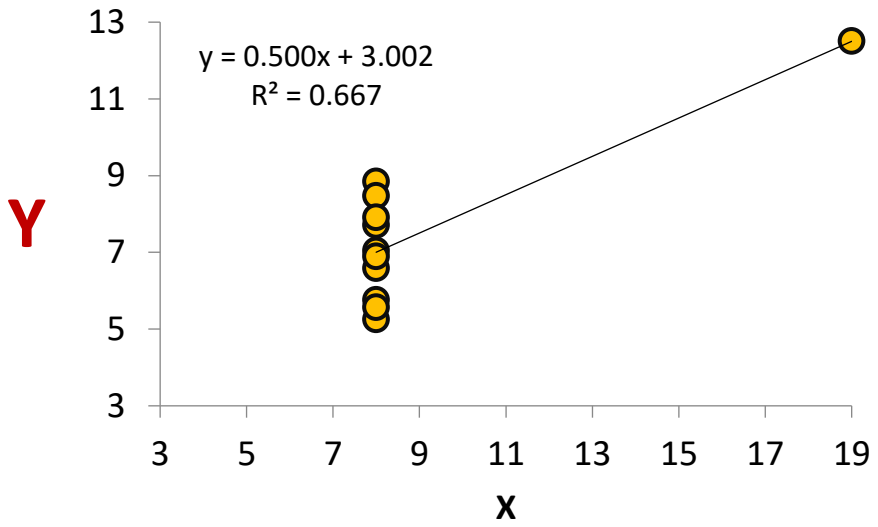
$$= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

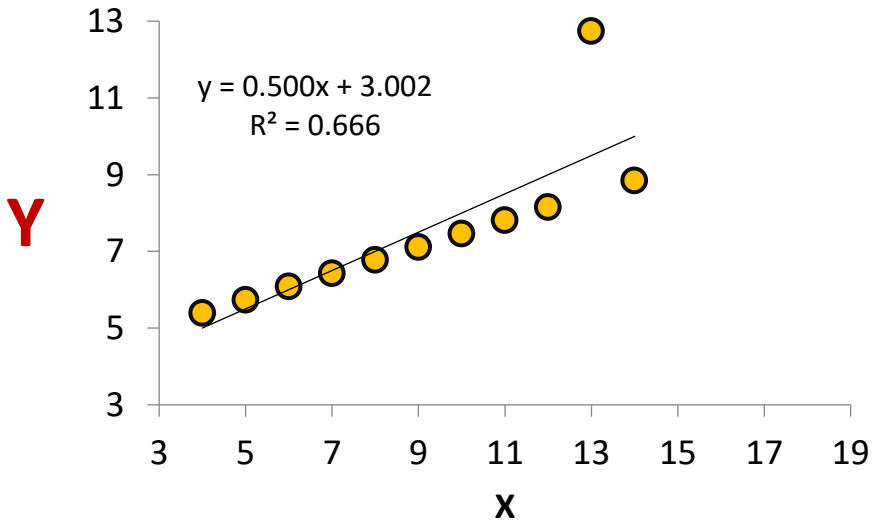
- A *useful* statistic but not definitive
- Tells you how well the data fit the model.
- It does *not* tell you if the model is correct.

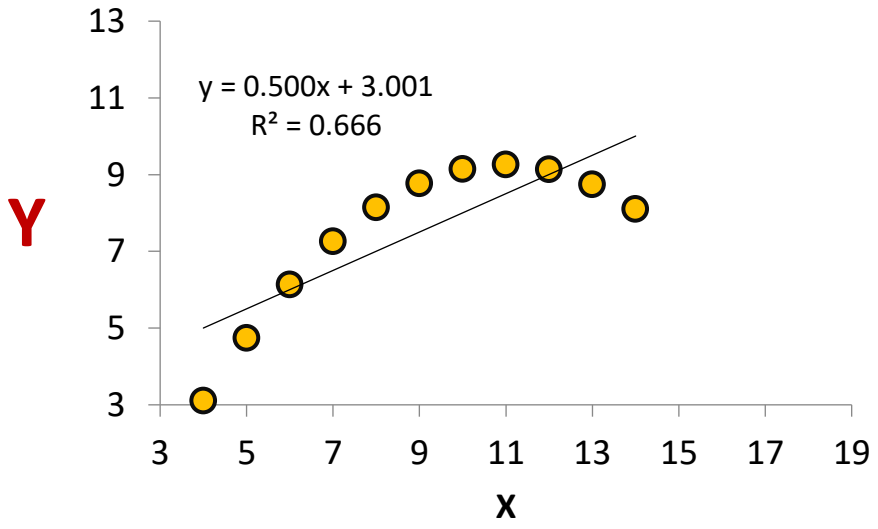
How much of the distribution of the data *about the mean* is described by the model.

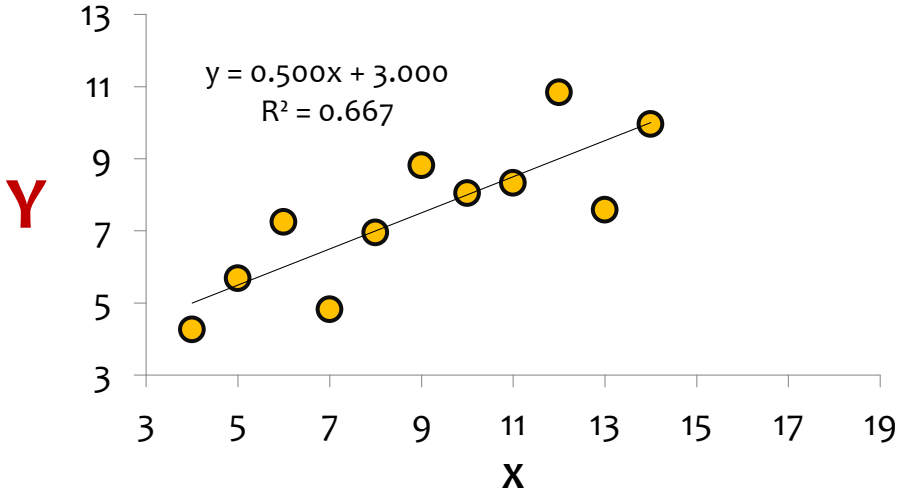
Problems with R^2

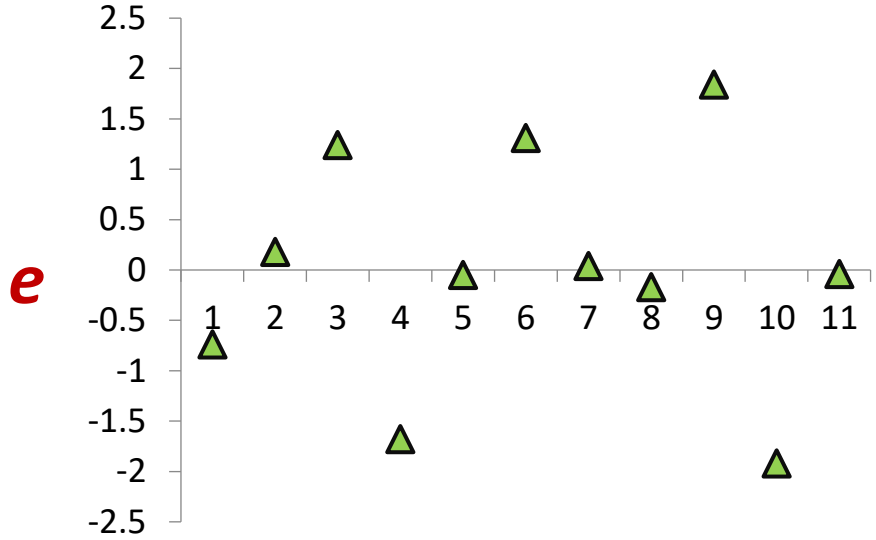




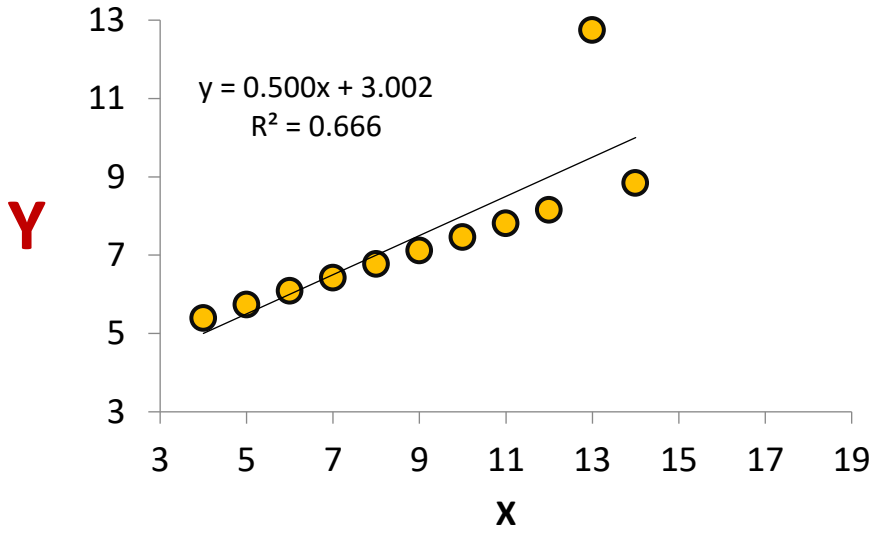


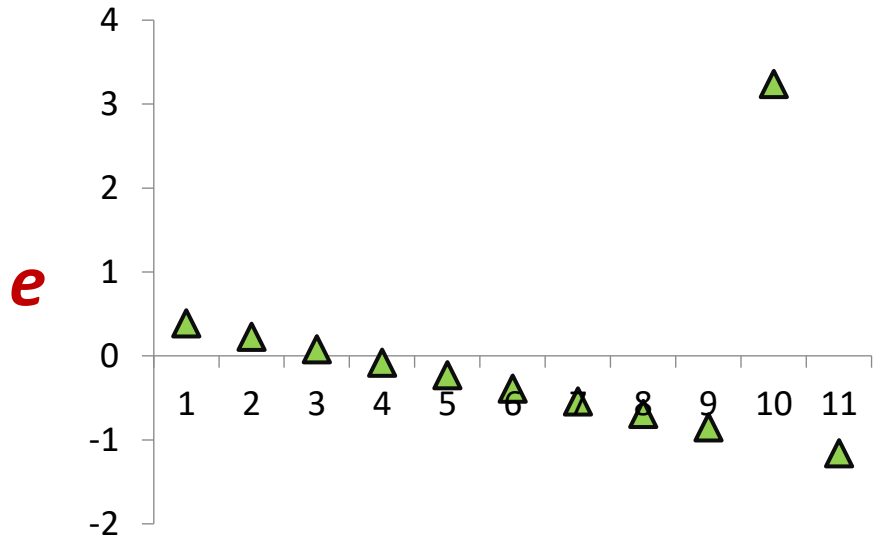


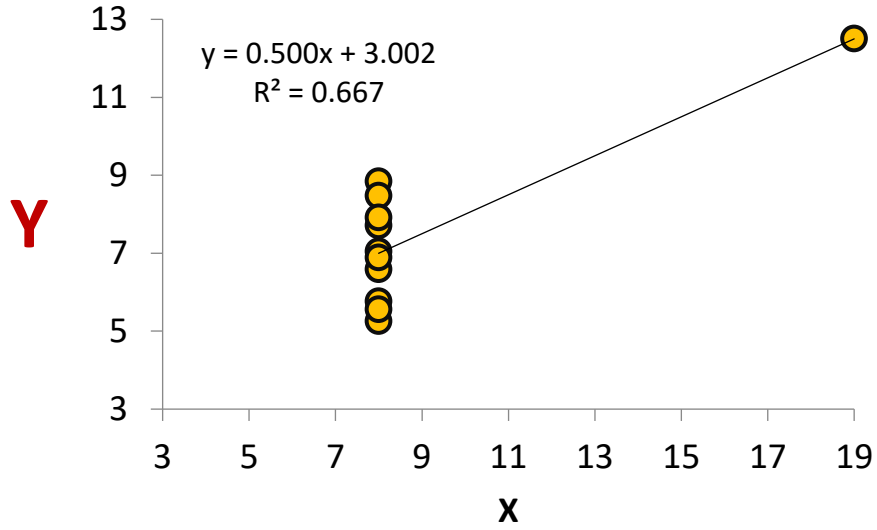


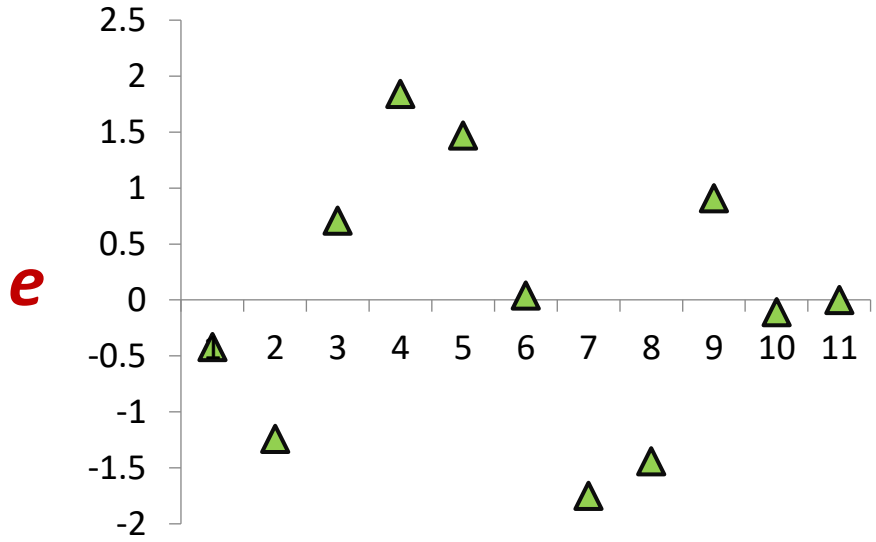


Residuals (e_i) should be randomly distributed.

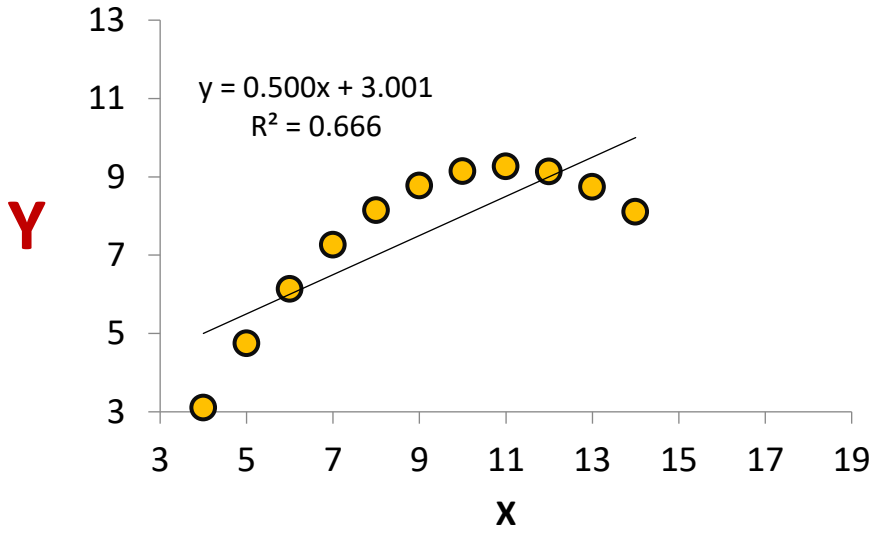


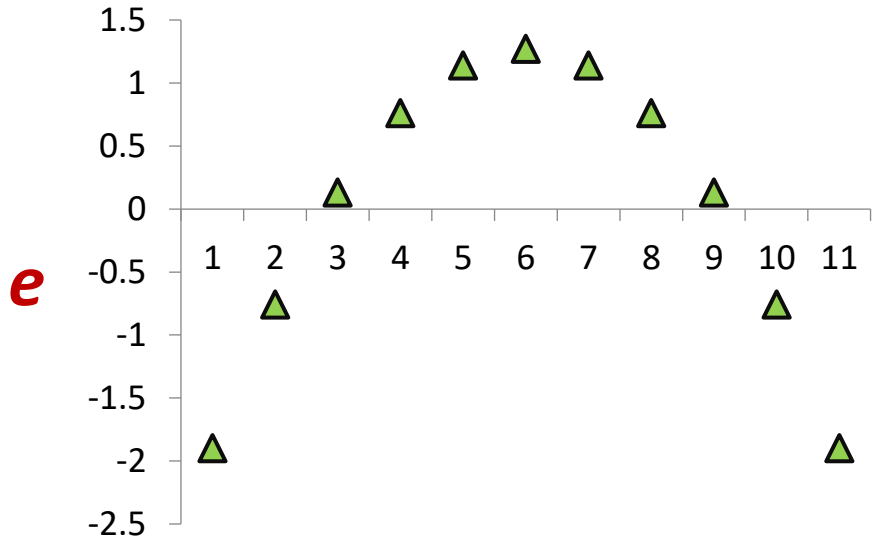






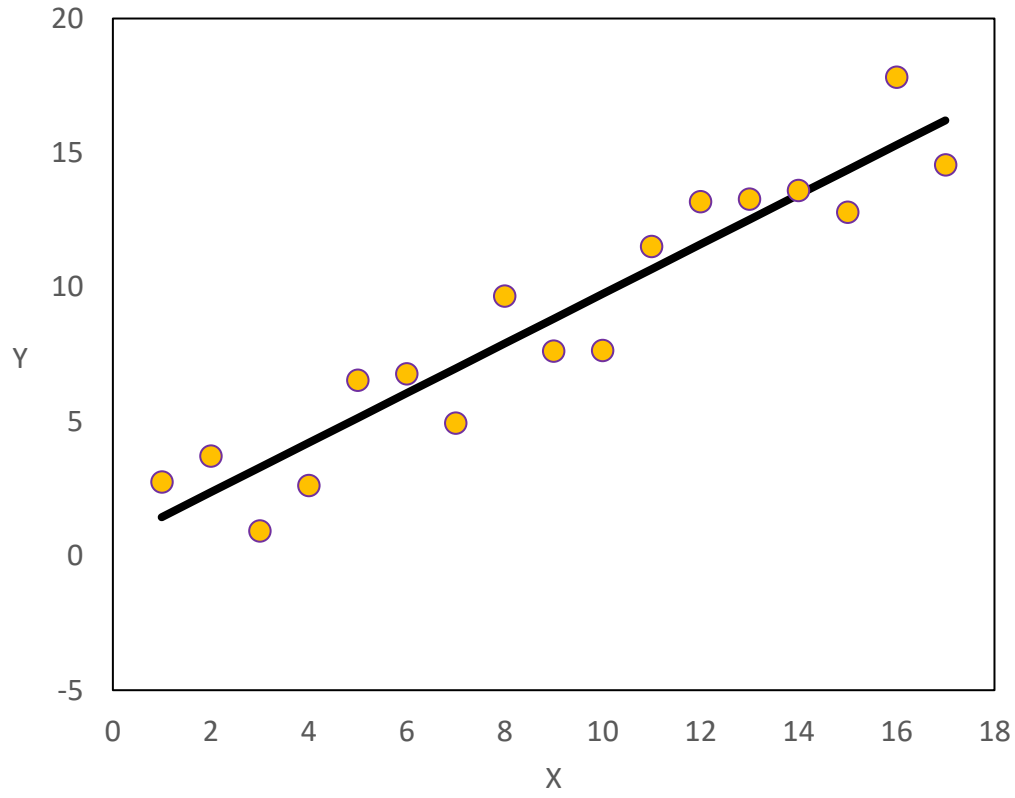
Residuals (e_i) should be randomly distributed.





Errors with the Regression

- What is the error on the *slope* and *intercept* (the parameters).
- What *error in Y* is expected from using the equation?
- Where should *future data* be found?



Statistics on the Slope/Intercept

The slope and intercept from fitting are *estimates* of the *true* slope and intercept.

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

(1-α)100% CI

$$b_0 \pm S_{b_0} t_{n-2,1-\frac{\alpha}{2}}$$

(slope)

$$b_1 \pm S_{b_1} t_{n-2,1-\frac{\alpha}{2}}$$

(intercept)

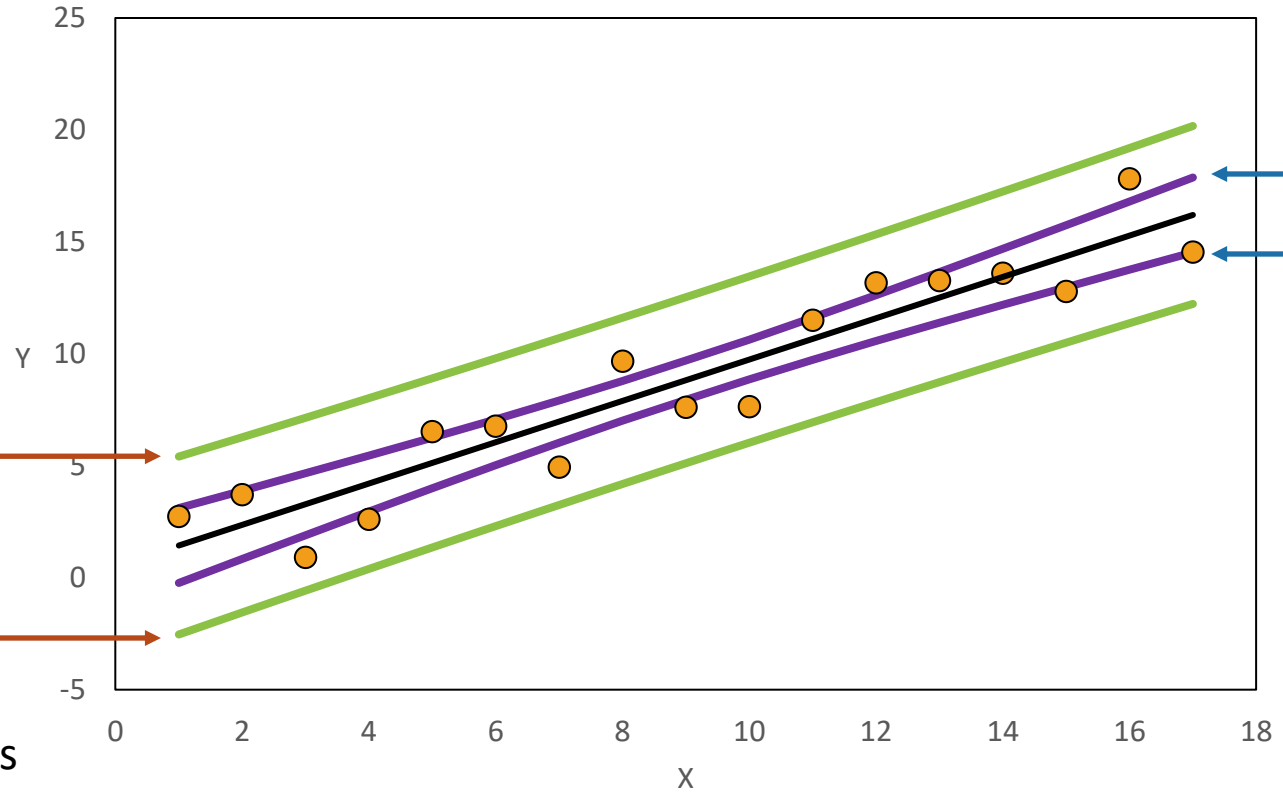
Standard Errors

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

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

The “Bands”

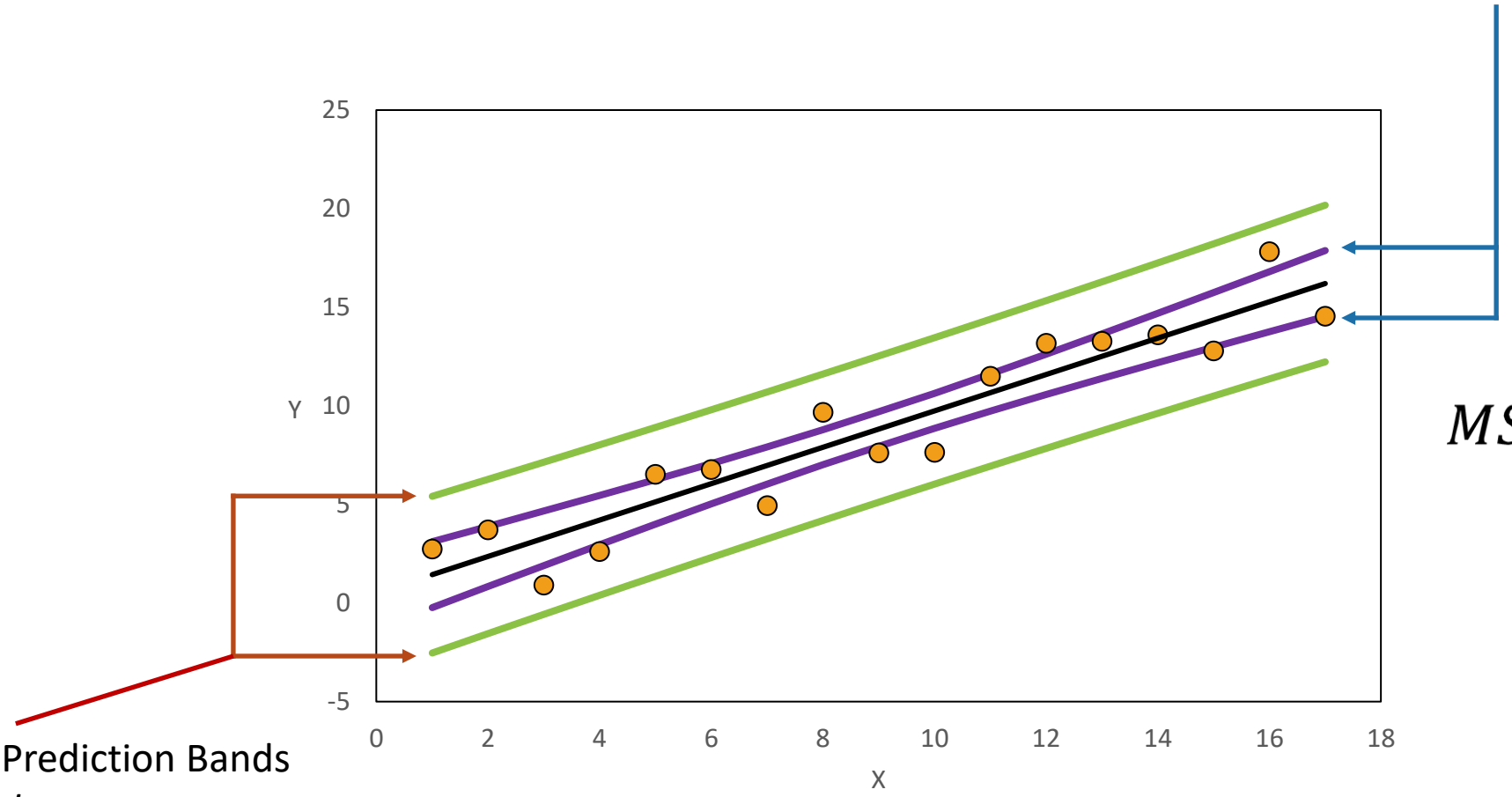
- Mean Prediction Bands
- *Inside Bands*
  - Error in $\hat{Y}$ (predicted Y)
  - Interval narrows with increasing N_{points}



- Single Point Prediction Bands
- *Outside Bands*
  - Where should a new point be found?
  - Little narrowing with increasing N_{points}

$$\hat{Y}_0 \pm S_{\hat{Y}} t_{n-2,1-\frac{\alpha}{2}} \qquad S_{\hat{Y}} = \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

- Mean Prediction Bands
- *Inside Bands*
  - Error in $\hat{Y}$ (predicted Y)
  - Interval narrows with increasing N_{points}



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

- Single Point Prediction Bands
- *Outside Bands*
  - Where should a new point be found?
  - Little narrowing with increasing N_{points}

$$Y_0 = \hat{Y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma}$$

Example Using Excel

Matrix Form of Linear Regression

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

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

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

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

Quadratic Model

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 + e_i$$

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

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

Statistics with Matrices

$$SS_E = \boldsymbol{Y^T Y} - \boldsymbol{b^T X^T Y} \qquad MS_E = \hat{\sigma}^2 = \frac{SS_E}{n-p}$$

Parameter Confidence Intervals

$$b_i \pm S_{b_i} t_{n-p,1-\frac{\alpha}{2}}$$

Standard Error of b_i is the square root of the i-th diagonal term of the matrix

$$\left(\boldsymbol{X^T X}\right)^{-1} \hat{\sigma}^2$$

Mean Prediction Bands

$$\hat{Y}_0 \pm S_{\hat{Y}} t_{n-p,1-\frac{\alpha}{2}}$$

$$S_{\hat{Y}} = \left(\boldsymbol{X_0^T}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X_0}\right)^{0.5} \hat{\sigma}$$