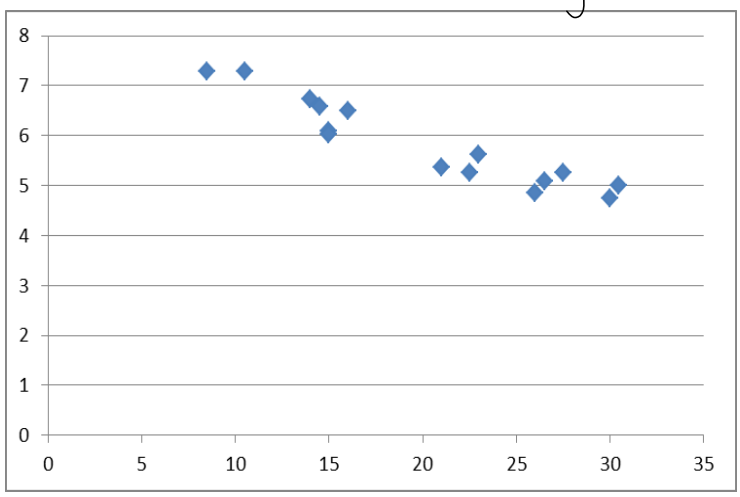
2.3 Coefficient of Determination

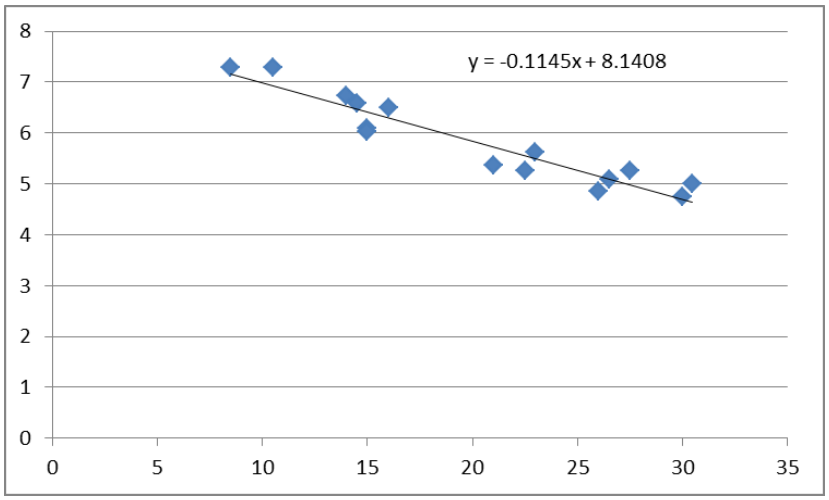
Does a linear model appear appropriate?

x	y
10.5	7.28
23	5.63
27.5	5.26
14.5	6.58
30.5	5.01
14	6.73
21	5.37
8.5	7.28
26	4.85
26.5	5.08
15	6.1
30	4.75
15	6.03
22.5	5.26
16	6.5

Heck yes.

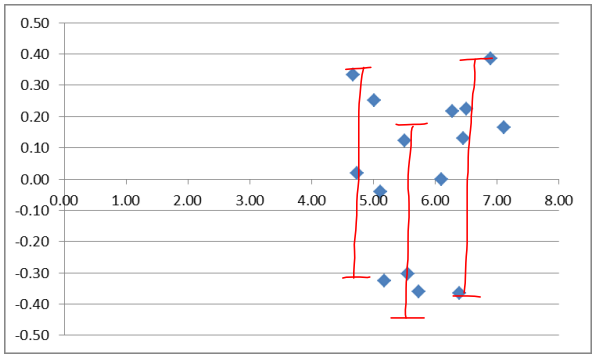


We can fit the Least-Squares Model



We can examine the residual plot.

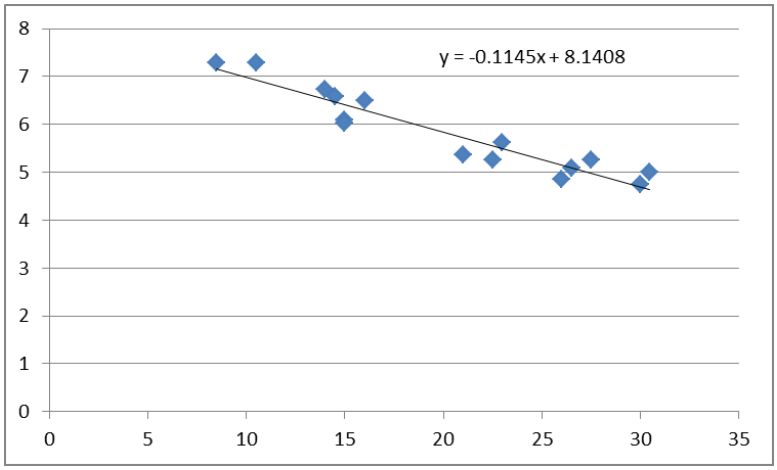
x	y	Fitted Value Yhat	Residual = y-yhat
10.5	7.28	6.89	0.39
23	5.63	5.51	0.12
27.5	5.26	5.01	0.25
14.5	6.58	6.45	0.13
30.5	5.01	4.68	0.33
14	6.73	6.51	0.22
21	5.37	5.73	-0.36
8.5	7.28	7.12	0.16
26	4.85	5.18	-0.33
26.5	5.08	5.12	-0.04
15	6.1	6.10	0.00
30	4.75	4.73	0.02
15	6.03	6.40	-0.37
22.5	5.26	5.56	-0.30
16	6.5	6.28	0.22



When we are satisfied that a linear model is appropriate, we can use it.

Interpolation:

Making inferences within the range of the data.



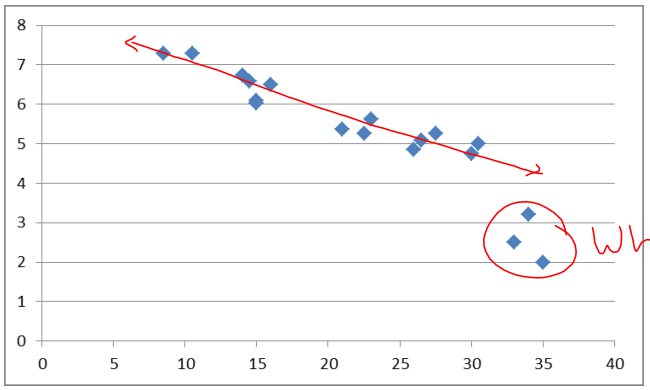
Extrapolation:

Making inferences outside the range of our data.

* do this with caution!

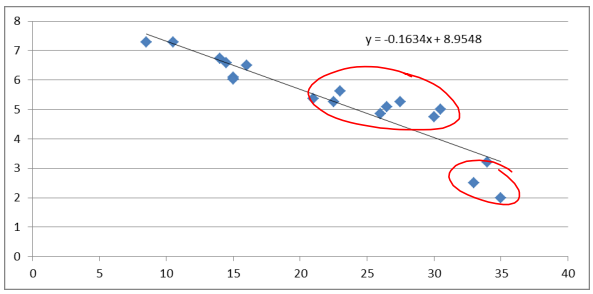
Suppose more data becomes available...

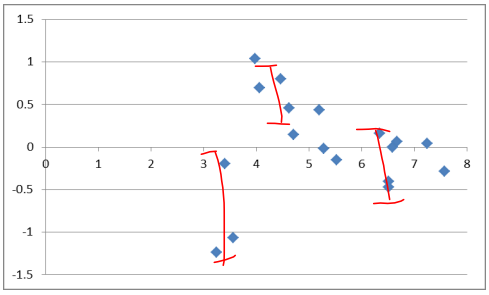
x	y
10.5	7.28
23	5.63
27.5	5.26
14.5	6.58
30.5	5.01
14	6.73
21	5.37
8.5	7.28
26	4.85
26.5	5.08
15	6.1
30	4.75
15	6.03
22.5	5.26
16	6.5
35	2
33	2.5
34	3.2



$r = 0.92$

x	y	Fitted Value Yhat	Residual = y-yhat
10.5	7.28	7.2391	0.0409
23	5.63	5.1966	0.4334
27.5	5.26	4.4613	0.7987
14.5	6.58	6.5855	-0.0055
30.5	5.01	3.9711	1.0389
14	6.73	6.6672	0.0628
21	5.37	5.5234	-0.1534
8.5	7.28	7.5659	-0.2859
26	4.85	4.7064	0.1436
26.5	5.08	4.6247	0.4553
15	6.1	6.5038	-0.4038
30	4.75	4.0528	0.6972
15	6.03	6.5038	-0.4738
22.5	5.26	5.2783	-0.0183
16	6.5	6.3404	0.1596
35	2	3.2358	-1.2358
33	2.5	3.5626	-1.0626
34	3.2	3.3992	-0.1992





Moral:

Do not extrapolate outside the range of the data...
or at least do so very cautiously!

Do not use a Least-Squares Line based on the correlation coefficient alone.

correlation tells us... how close the data points are to a line.

residual plots... how evenly distributed the points are about the line.

Outliers

Mathematical Definition

If we are beyond $Q_3 + 1.5\,IQR$
or below $Q_1 - 1.5\,IQR$ then we have an outlier.

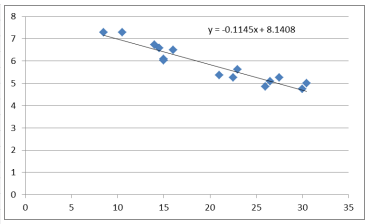
Error? Could be a data entry issue or an equipment malfunction - if so we can... throw it in the trash.

Fit a line with and then without the outlier...influential?

If we have no reason to remove the outlier, we then have to figure out if it is influential.

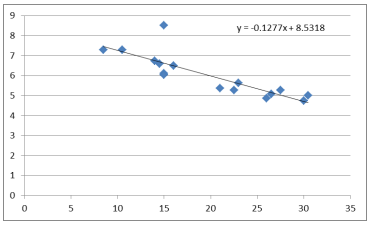
No Outliers – baseline.

x	y
10.5	7.28
23	5.63
27.5	5.26
14.5	6.58
30.5	5.01
14	6.73
21	5.37
8.5	7.28
26	4.85
26.5	5.08
15	6.1
30	4.75
15	6.03
22.5	5.26
16	6.5



Possible Outlier: y-value is unusual

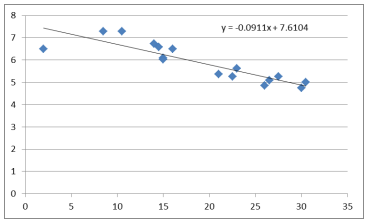
x	y
10.5	7.28
23	5.63
27.5	5.26
14.5	6.58
30.5	5.01
14	6.73
21	5.37
8.5	7.28
26	4.85
26.5	5.08
15	6.1
30	4.75
15	6.03
22.5	5.26
16	6.5
15	8.5



Non-influential outlier

Possible Outlier: x-value unusual

x	y
10.5	7.28
23	5.63
27.5	5.26
14.5	6.58
30.5	5.01
14	6.73
21	5.37
8.5	7.28
26	4.85
26.5	5.08
15	6.1
30	4.75
15	6.03
22.5	5.26
16	6.5
2	6.5



More influential than the previous outlier, though still not a worry.

Least Squares Line is the line that minimizes the sum of squared residuals. $\sum_{i=1}^{n} e_i^2$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

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

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

So...notice the following:

$$r \cdot \frac{Sy}{Sx} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \cdot \frac{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{\beta}_1$$

Goodness of Fit Statistics

Correlation is a goodness of fit statistic for a linear model.

We've taken a closer look at how the correlation coefficient is formed, but how does it measure the "goodness of fit?"

r = Correlation coefficient

R^2 = Coefficient of determination

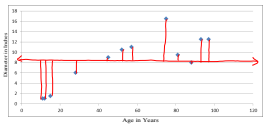
OR

The amount of variation in y that can be accounted for by your model.

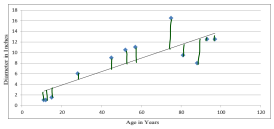
If we do NOT know the age of the tree, what would be our best guess at the tree diameter?

Age (years)	Diameter (inches)
97	12.5
93	12.5
88	8.0
81	9.5
75	16.5
57	11.0
52	10.5
45	9.0
28	6.0
15	1.5
12	1.0
11	1.0

$\bar{y} = 8.25$



$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ = Total Sum of Squares



$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = Sum of squared residuals

OR

Sum of squared Error

If we subtract, then we have an estimate of the REDUCTION in the spread of the points by using the regression line rather than the mean of the diameters.

$$R^2 = \frac{\text{Total SS} - \text{Error SS}}{\text{Total SS}} = 1 - \frac{\text{Error SS}}{\text{Total SS}}$$

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

$r^2 = 0.6893$