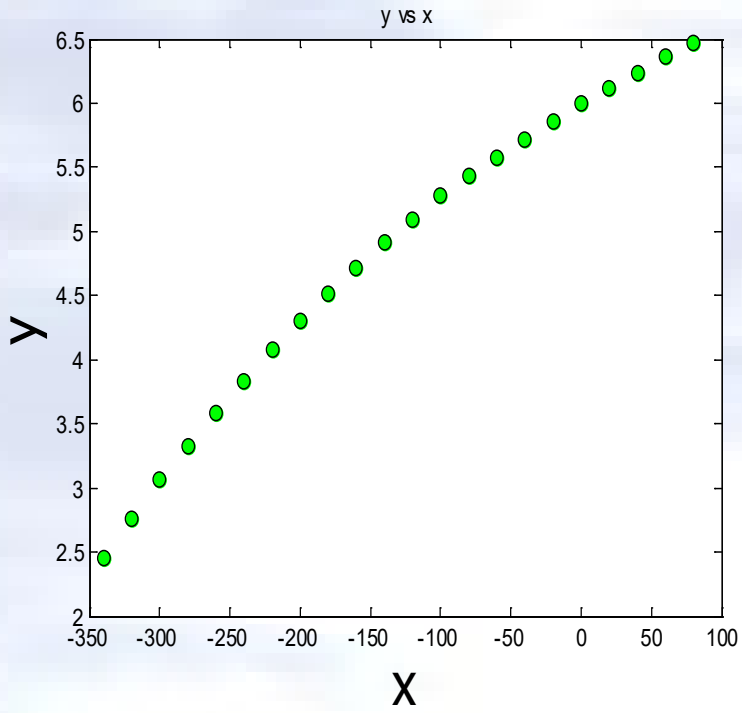
6.05
Adequacy of Linear Regression Models

Data



Therm exp coeff vs temperature

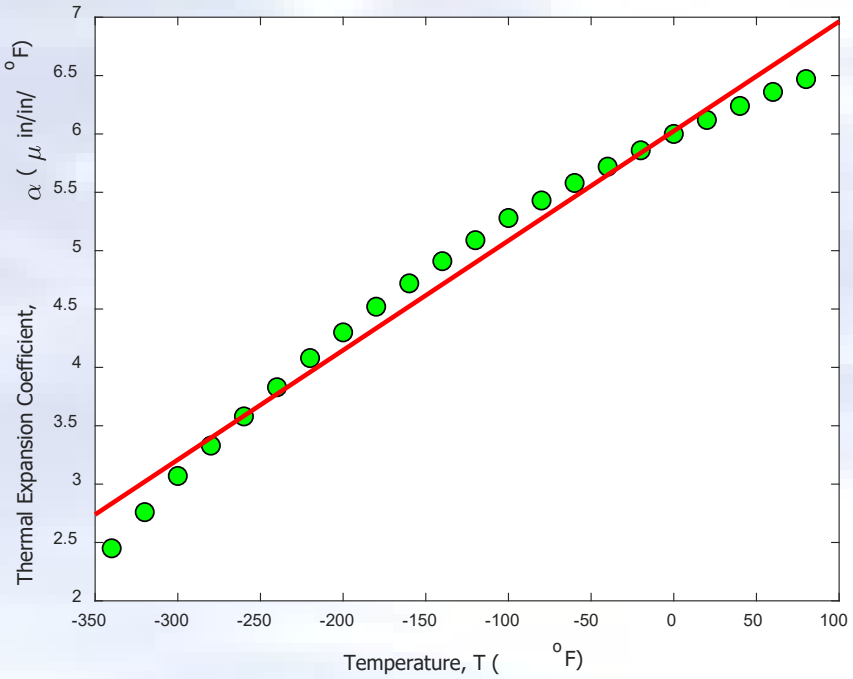
T	α
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.2
-60	5.58
-80	5.43
-100	5.28
-120	5.09

T	α
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

T is in ^oF

α is in $\mu in/in/\,^oF$

Is this adequate?



Straight Line Model

Quality of Fitted Data

- **Does the model describe the data adequately?**
- **How well does the model predict the response variable predictably?**

Linear Regression Models

- **Limit our discussion to adequacy of straight-line regression models**

Four checks

1. Does the model look like it explains the data?
2. Do 95%of the residuals fall with ±2 standard error of estimate?
3. Is the coefficient of determination acceptable?
4. Does the model meet the assumption of random errors?

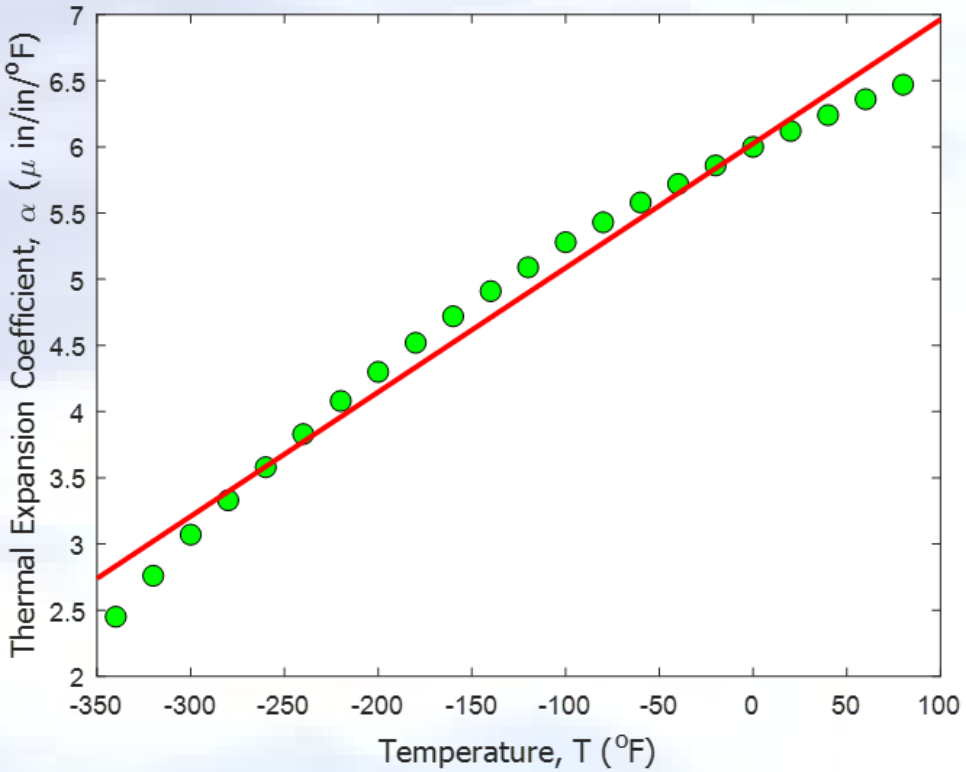
Therm exp coeff vs temperature (22 data points)

T	α
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.2
-60	5.58
-80	5.43
-100	5.28
-120	5.09

T	α
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

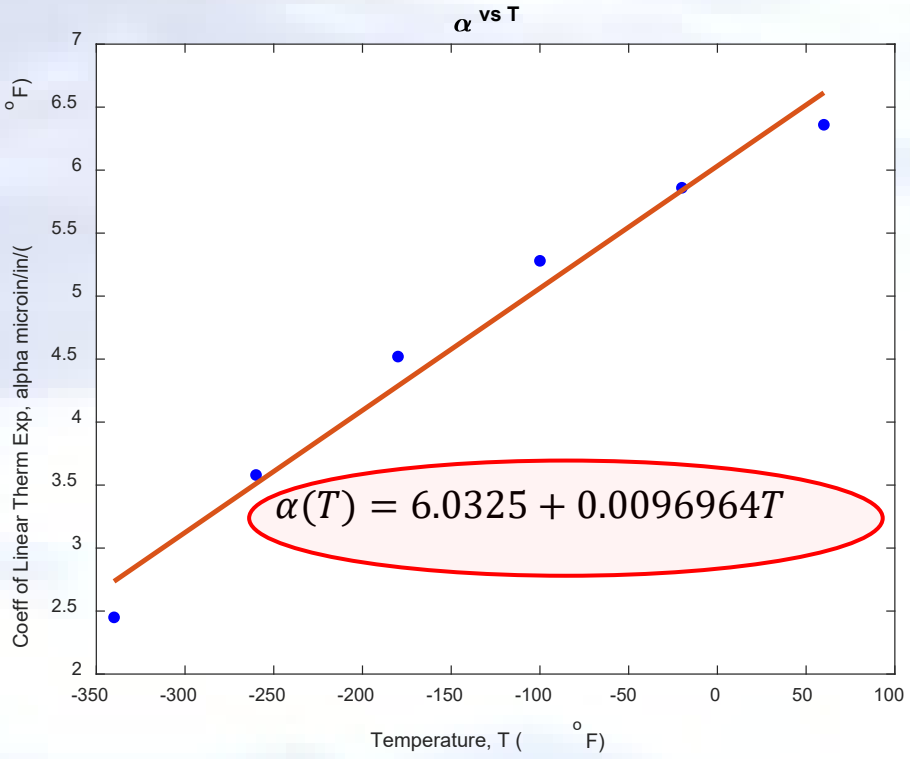
T is in ^oF

α is in $\mu in/in/^oF$



Therm exp coeff vs temperature (6 data points)

T_i	α_i
-340	2.45
-260	3.58
-180	4.52
-100	5.28
-20	5.86
60	6.36



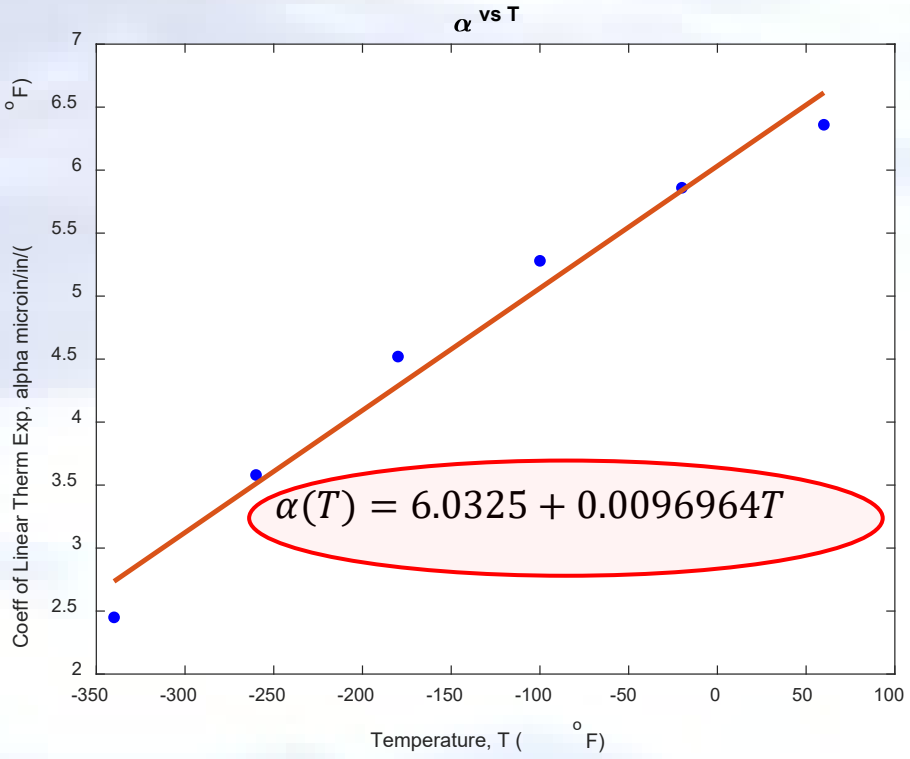
Four checks (with 6 data points)

1. Does the model look like it explains the data?
2. Do 95%of the residuals fall with ±2 standard error of estimate?
3. Is the coefficient of determination acceptable?
4. Does the model meet the assumption of random errors?

Check 1: Does the model look like it explains the data?

Data and Model

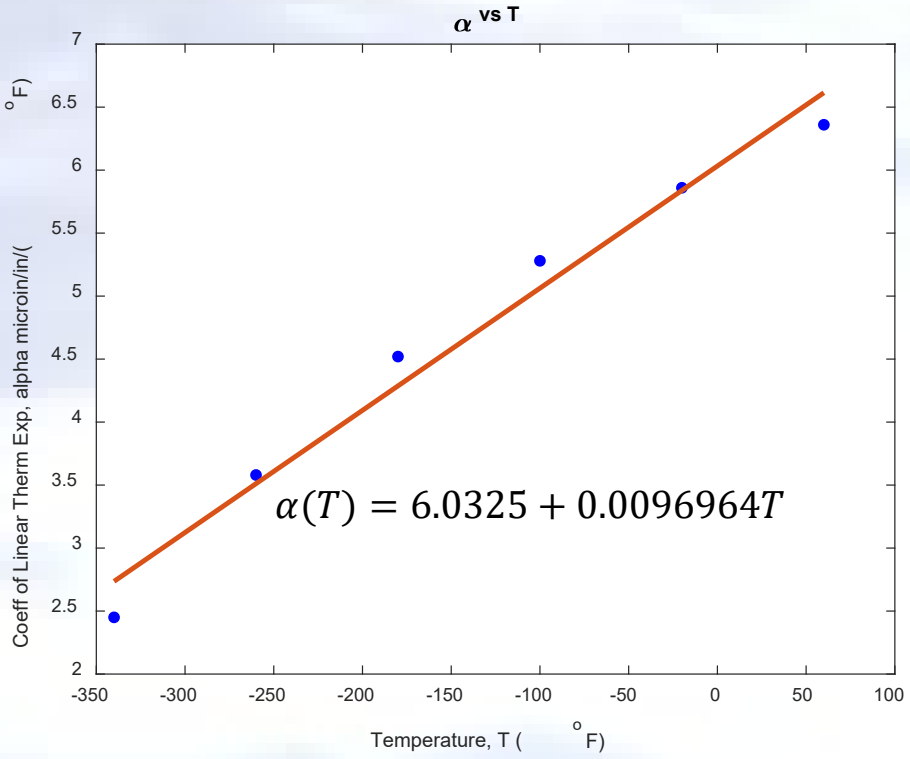
T_i	α_i
-340	2.45
-260	3.58
-180	4.52
-100	5.28
-20	5.86
60	6.36



Check 2. Do 95%of the residuals fall within ±2 standard error of estimate?

Data and Model

T_i	α_i
-340	2.45
-260	3.58
-180	4.52
-100	5.28
-20	5.86
60	6.36



Standard error of estimate

$$S_r = \sum_{i=1}^{n} (\alpha_i - a_0 - a_1 T_i)^2$$

$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$

Standard Error of Estimate

$$\alpha(T) = 6.0325 + 0.0096964T$$

T_i	α_i	$a_0 + a_1 T_i$	$\alpha_i - a_0 - a_1 T_i$
-340	2.45	2.7357	-0.28571
-260	3.58	3.5114	0.068571
-180	4.52	4.2871	0.23286
-100	5.28	5.0629	0.21714
-20	5.86	5.8386	0.021429
60	6.36	6.6143	-0.25429

Standard Error of Estimate

$$S_r = 0.25283$$

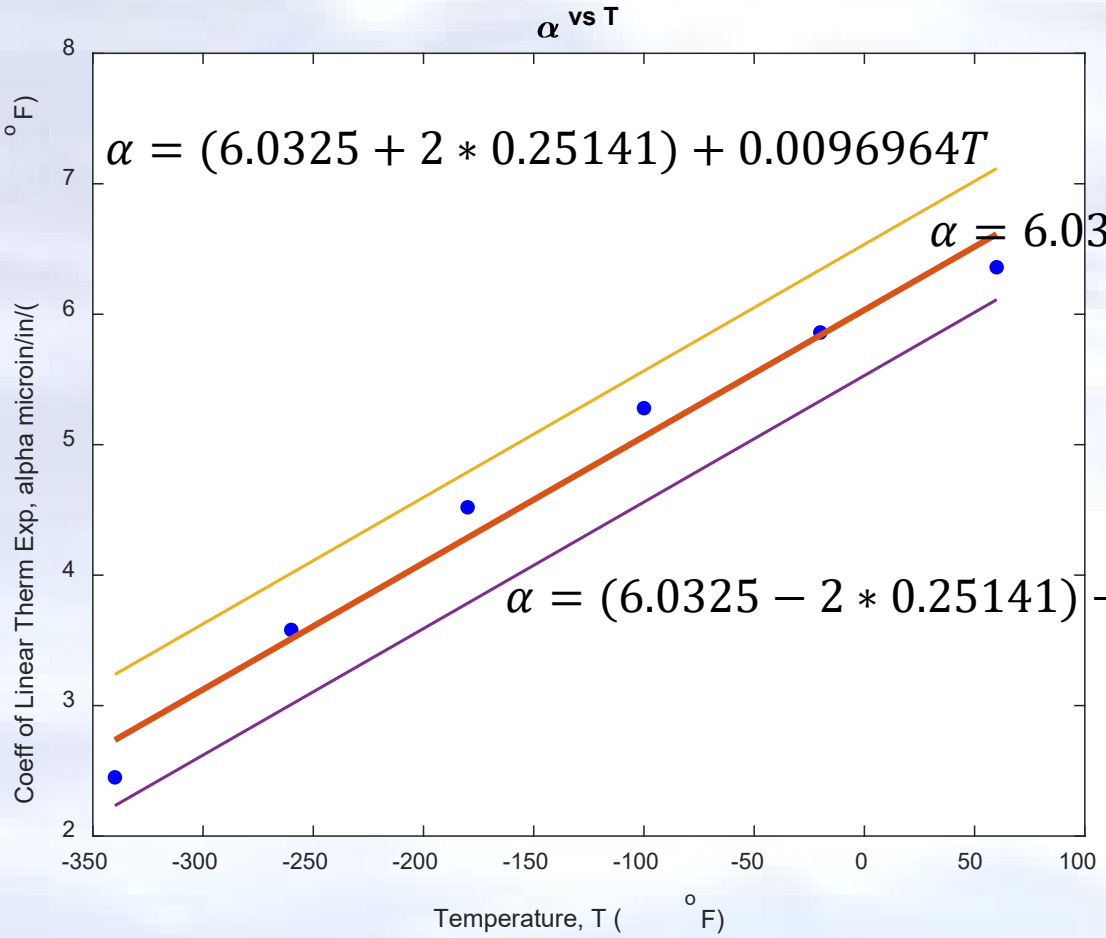
$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$

$$= \sqrt{\frac{0.25283}{6-2}}$$

$$= 0.25141$$

T_i	α_i	$a_0 + a_1 T_i$	$\alpha_i - a_0 - a_1 T_i$
-340	2.45	2.7357	-0.28571
-260	3.58	3.5114	0.068571
-180	4.52	4.2871	0.23286
-100	5.28	5.0629	0.21714
-20	5.86	5.8386	0.021429
60	6.36	6.6143	-0.25429

Standard Error of Estimate



Scaled Residuals

$$\text{Scaled Residual} = \frac{\text{Residual}}{\text{Standard Error of Estimate}}$$

$$\text{Scaled Residual} = \frac{\alpha_i - a_0 - a_1 T_i}{s_{\alpha/T}}$$

95% of the scaled residuals need to be in [-2,2]

Scaled Residuals

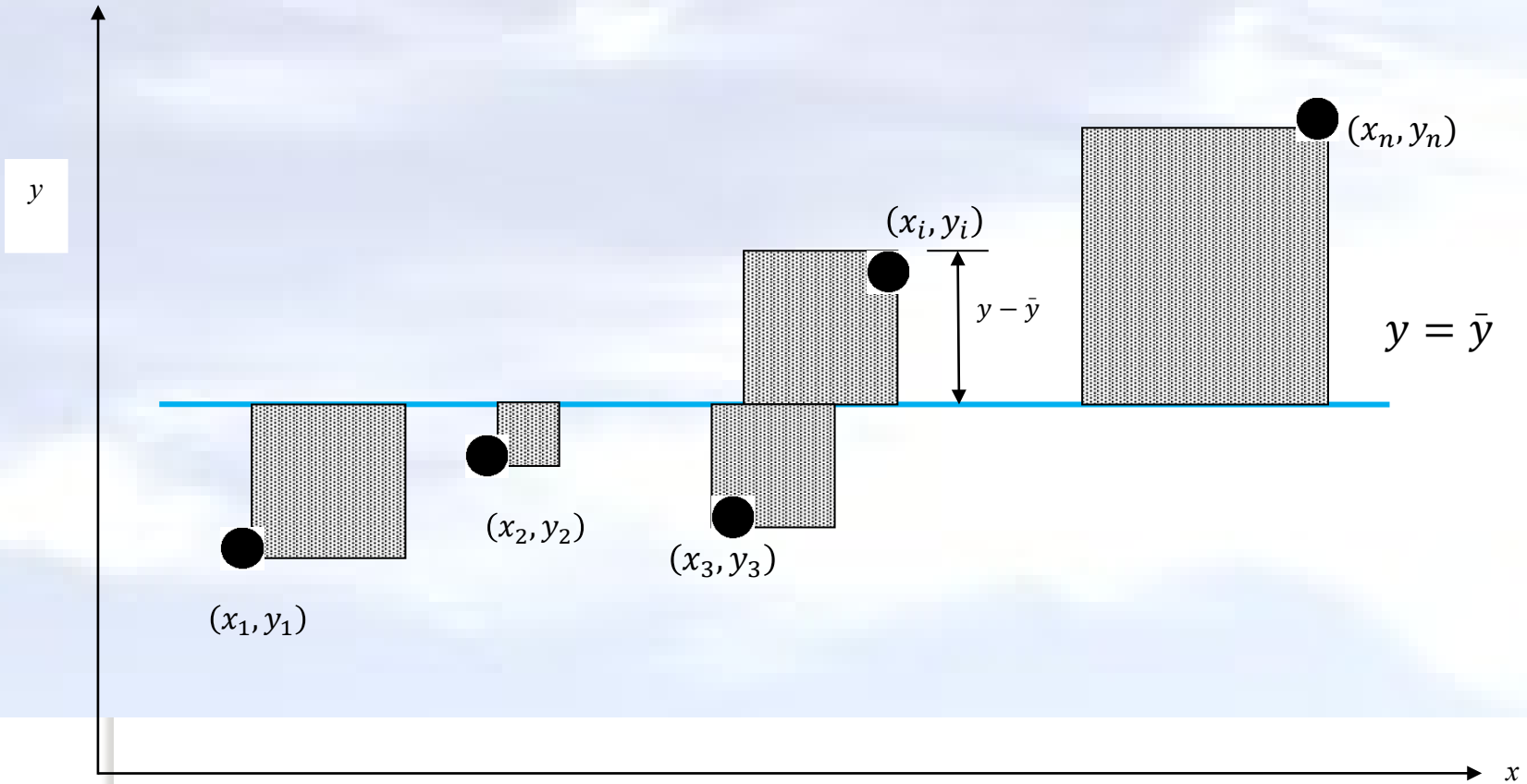
$$s_{\alpha/T} = 0.25141$$

T_i	α_i	Residual	Scaled Residual
-340	2.45	-0.28571	-1.1364
-260	3.58	0.068571	0.27275
-180	4.52	0.23286	0.92622
-100	5.28	0.21714	0.86369
-20	5.86	0.021429	0.085235
60	6.36	-0.25429	-1.0115

3. Is the coefficient of determination acceptable?

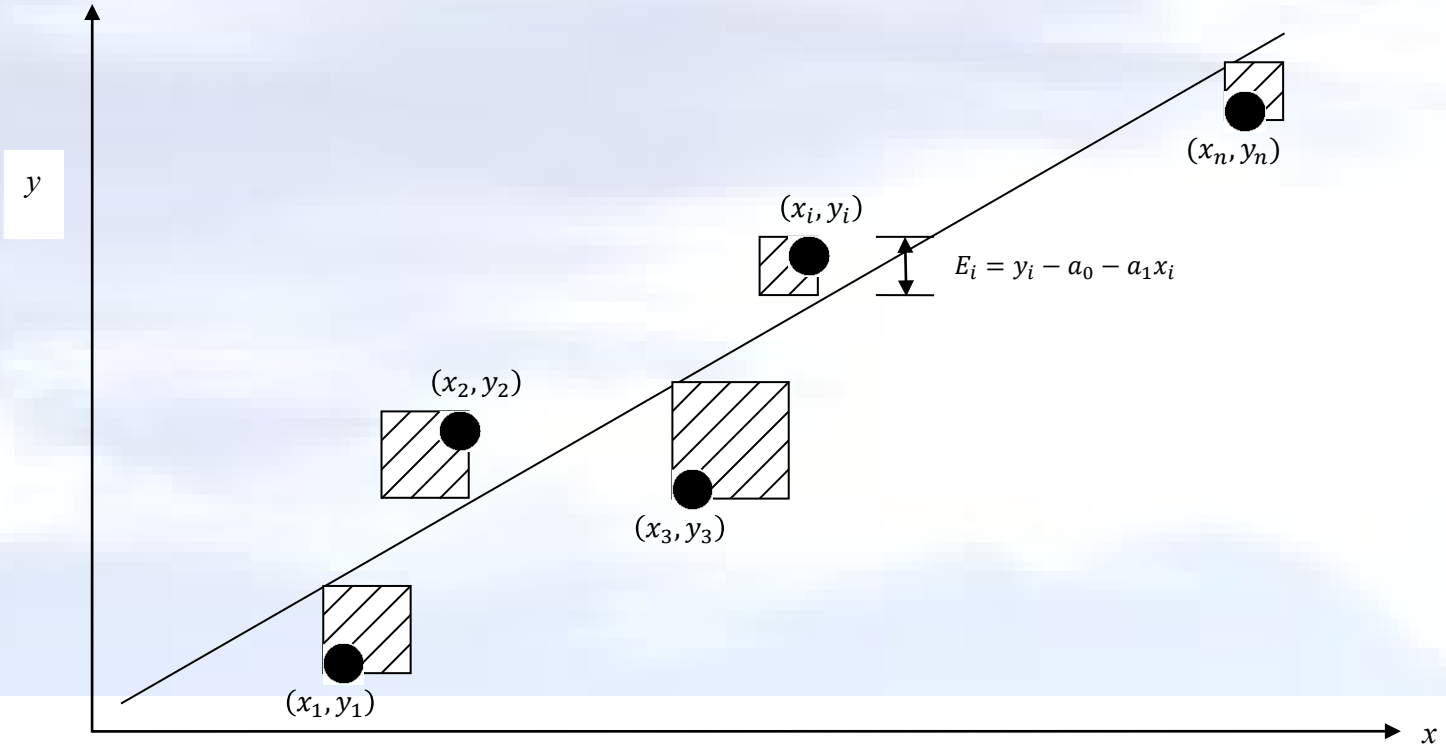
Sum of square of residuals between data and mean

$$S_t = \sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2$$



Sum of square of residuals between observed and predicted

$$S_r = \sum_{i=1}^{n} (\alpha_i - a_0 - a_1 T_i)^2$$



Coefficient of determination

$$S_t = \sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2$$

$$S_r = \sum_{i=1}^{n} (\alpha_i - a_0 - a_1 T_i)^2$$

$$r^2 = \frac{S_t - S_r}{S_t}$$

Calculation of S_t

T_i	α_i	$\alpha_i - \bar{\alpha}$
-340	2.45	-2.2250
-260	3.58	-1.0950
-180	4.52	-0.15500
-100	5.28	0.60500
-20	5.86	1.1850
60	6.36	1.6850

$\bar{\alpha} = 4.6750$

$S_t = 10.783$

Calculation of S_r

$$\alpha(T) = 6.0325 + 0.0096964T$$

T_i	α_i	$a_0 + a_1 T_i$	$\alpha_i - a_0 - a_1 T_i$
-340	2.45	2.7357	-0.28571
-260	3.58	3.5114	0.068571
-180	4.52	4.2871	0.23286
-100	5.28	5.0629	0.21714
-20	5.86	5.8386	0.021429
60	6.36	6.6143	-0.25429

$$S_r = 0.25283$$

Coefficient of determination

$$r^2 = \frac{S_t - S_r}{S_t}$$

$$= \frac{10.783 - 0.25283}{10.783}$$

$$= 0.97655$$

Limits of Coefficient of Determination

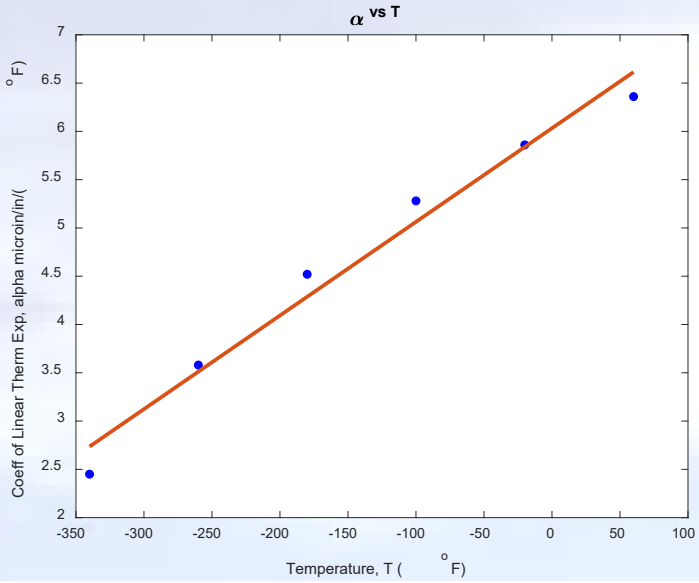
$$r^2 = \frac{S_t - S_r}{S_t}$$

$$0 \leq r^2 \leq 1$$

Correlation coefficient

$$r = \sqrt{\frac{S_t - S_r}{S_t}}$$
$$= \sqrt{0.97655}$$
$$= 0.98820$$

How do you know if r is positive or negative ?



What does a particular value of |r| mean?

0.8 to 1.0 - Very strong relationship

0.6 to 0.8 - Strong relationship

0.4 to 0.6 - Moderate relationship

0.2 to 0.4 - Weak relationship

0.0 to 0.2 - Weak or no relationship

Four checks (with many data points)

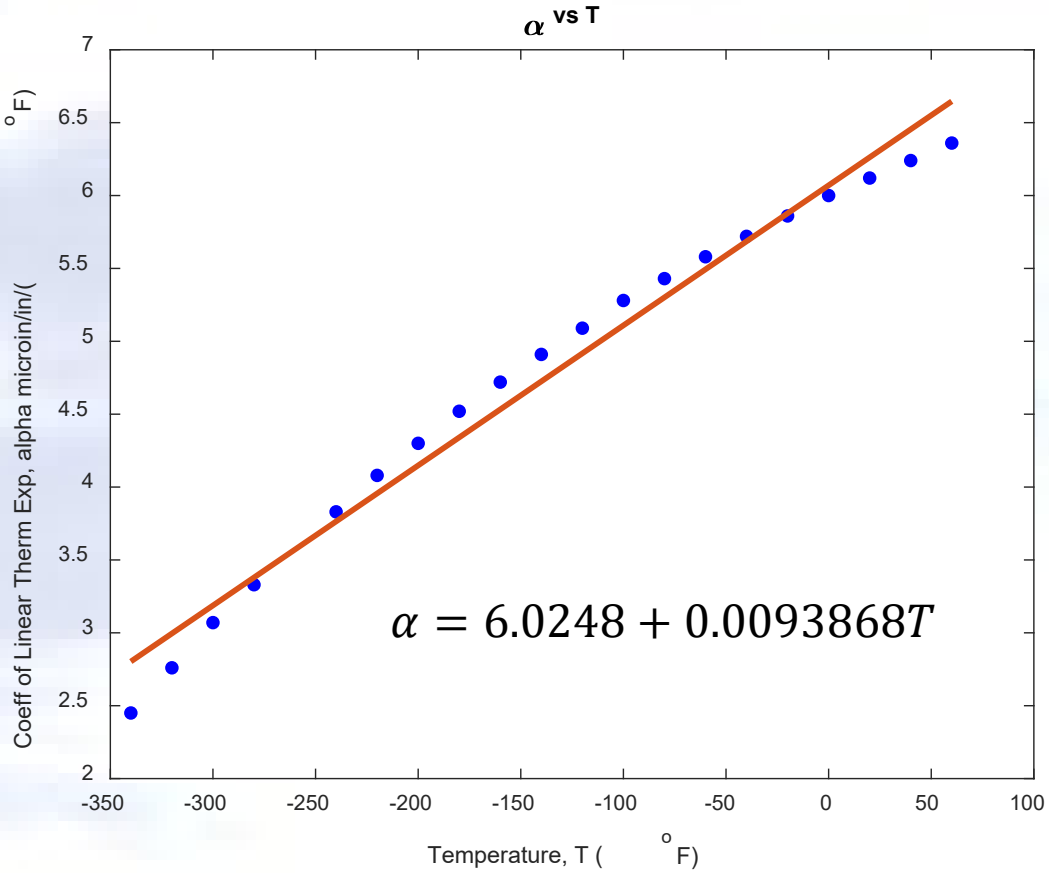
1. Does the model look like it explains the data?
2. Do 95%of the residuals fall within ±2 of standard error of estimate?
3. Is the coefficient of determination acceptable?
4. Does the model meet the assumption of random errors?

Check 1: Does the model look like it explains the data?

Check 1:Plot Model and Data

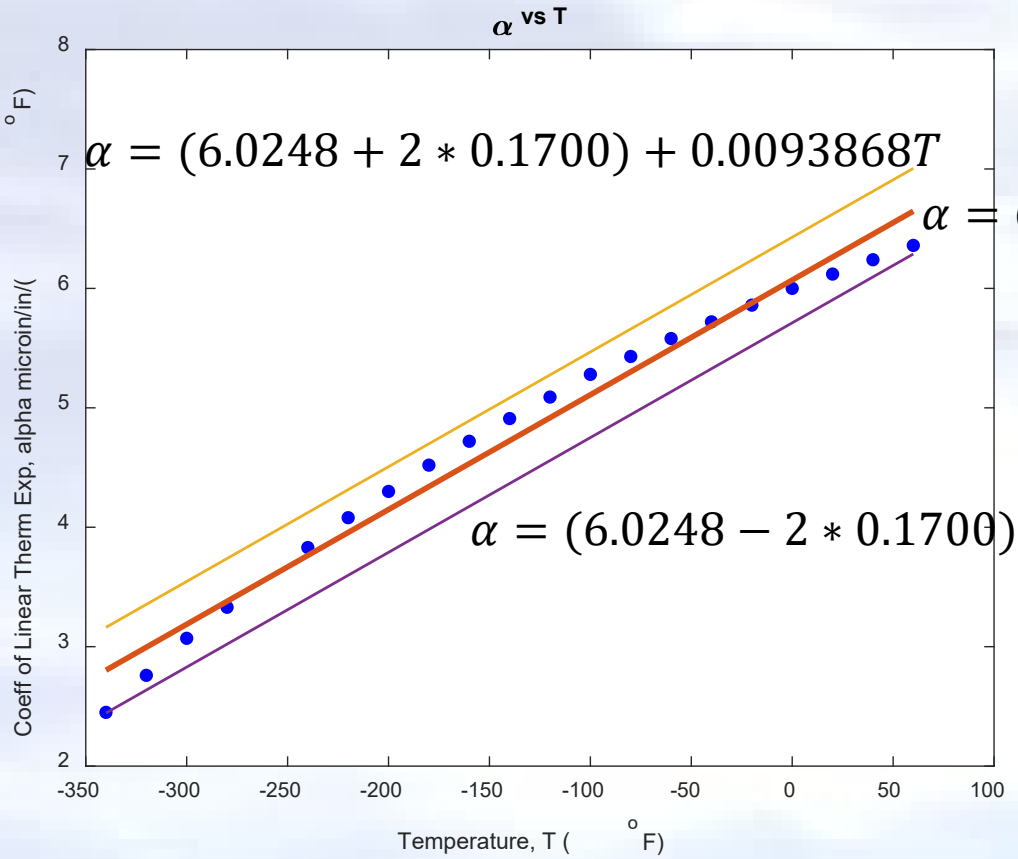
T	α
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.2
-60	5.58
-80	5.43
-100	5.28
-120	5.09

T	α
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45



Check 2. Do 95%of the residuals fall within ±2 standard error of estimate?

Check 2: Using Standard Error of Estimate



$\alpha = (6.0248 + 2 * 0.1700) + 0.0093868T$

$\alpha = 6.0248 + 0.0093868T$

$\alpha = (6.0248 - 2 * 0.1700) + 0.0093868T$

$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$

$s_{\alpha/T}$ =0.1700

3. Is the coefficient of determination acceptable?

Check 3: Using Coefficient of Determination

$$r^2 = \frac{S_t - S_r}{S_t}$$

$$= \frac{27.614 - 0.5785}{27.614}$$

$$= 0.9791$$

Check 4. Does the model meet the assumption of random errors?

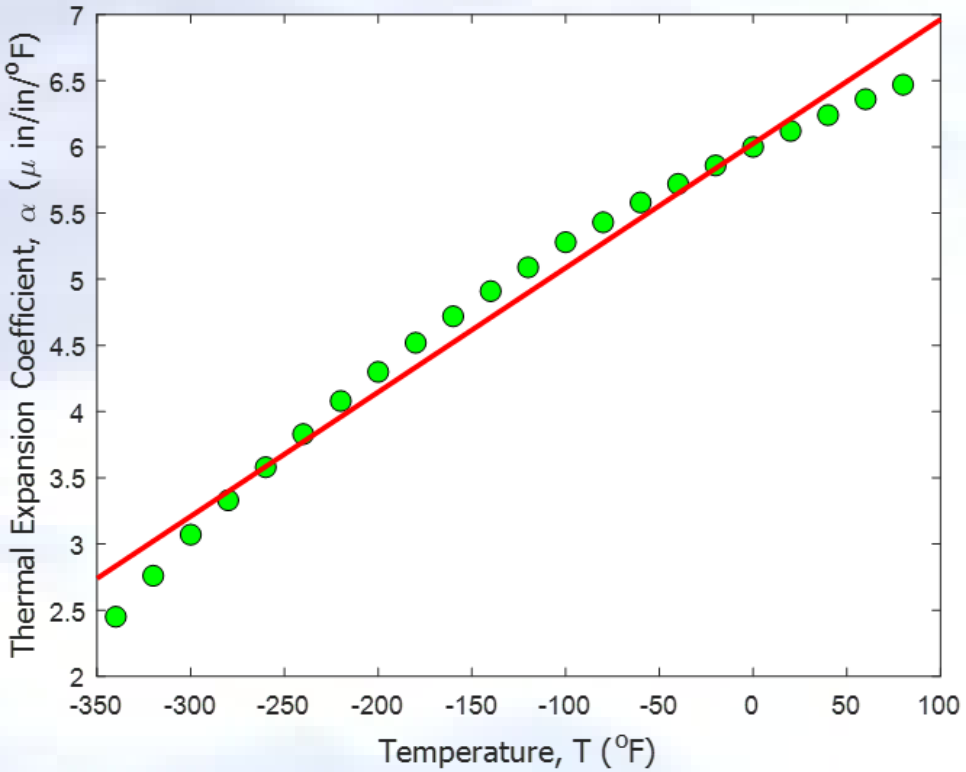
Therm exp coeff vs temperature

T	α
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.2
-60	5.58
-80	5.43
-100	5.28
-120	5.09

T	α
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

T is in ^oF

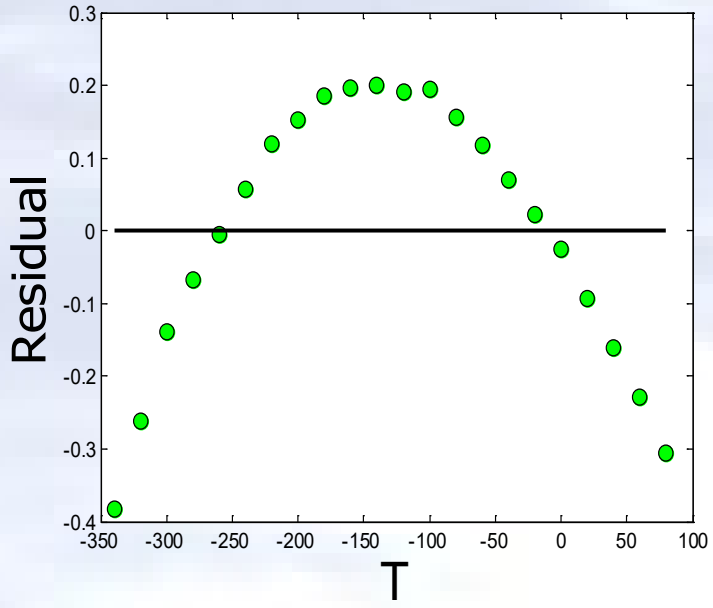
α is in $\mu in/in/^oF$



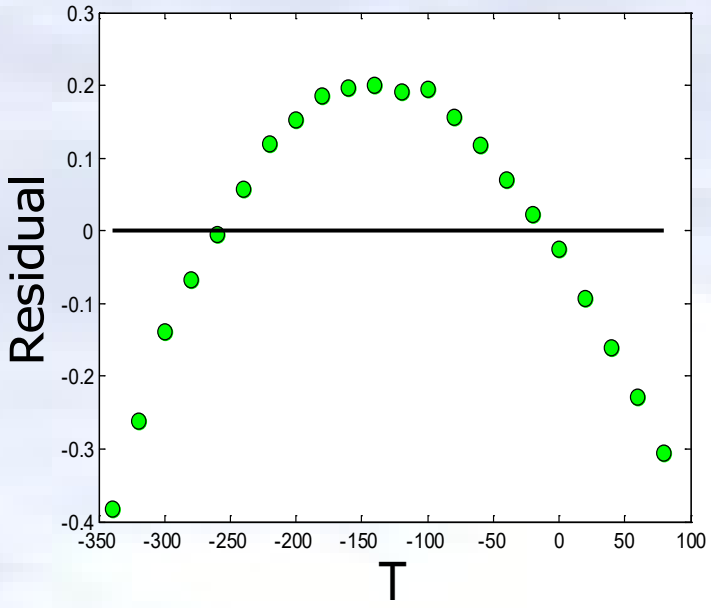
Model meets assumption of random errors

- Residuals are negative as well as positive
- Variation of residuals as a function of the independent variable is random
- Residuals follow a normal distribution
- ~~There is no autocorrelation between the data points.~~

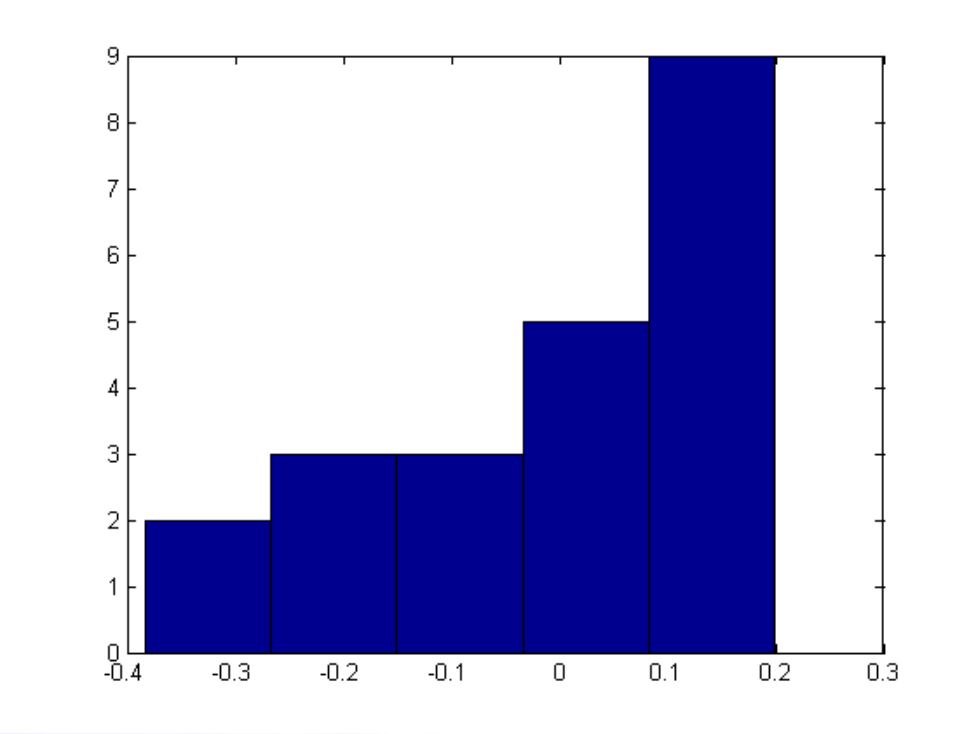
Are residuals negative and positive?



Is variation of residuals as a function of independent variable random?

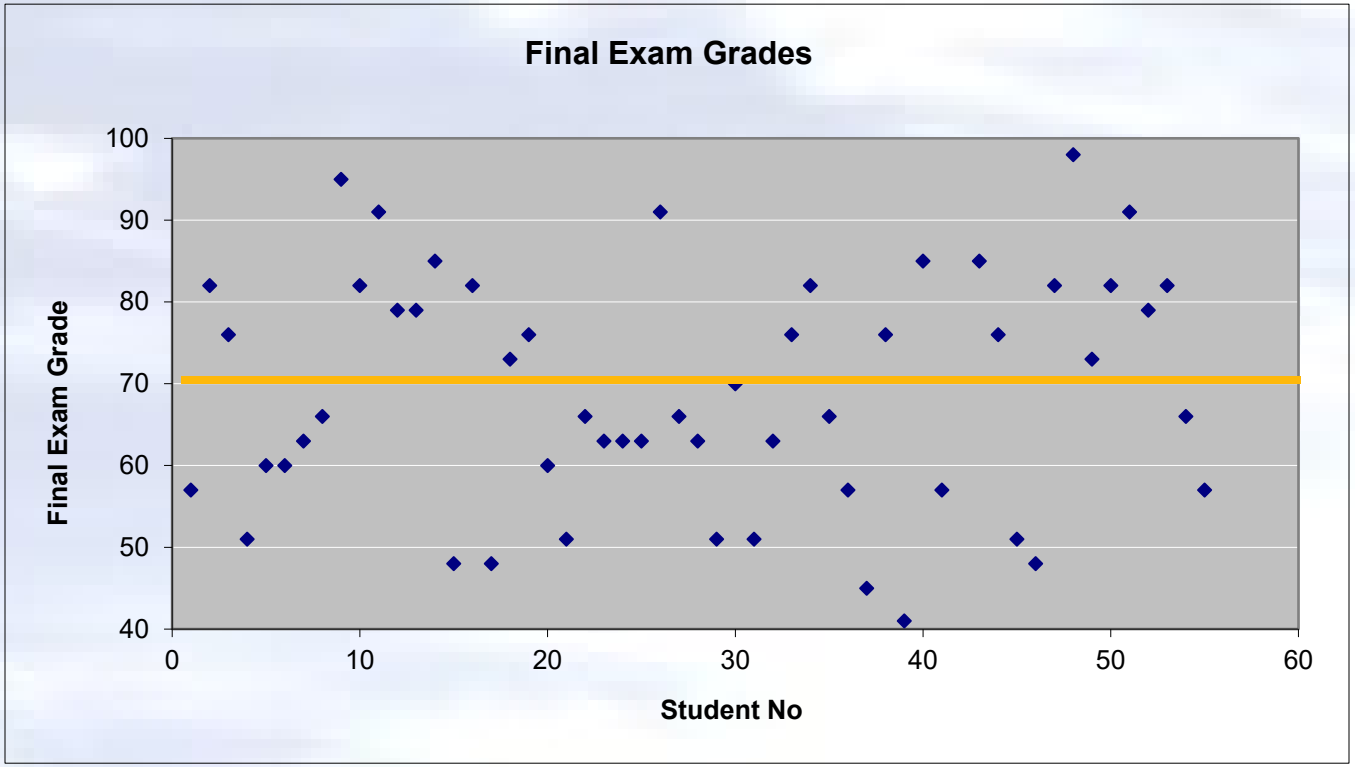


Do the residuals follow normal distribution?

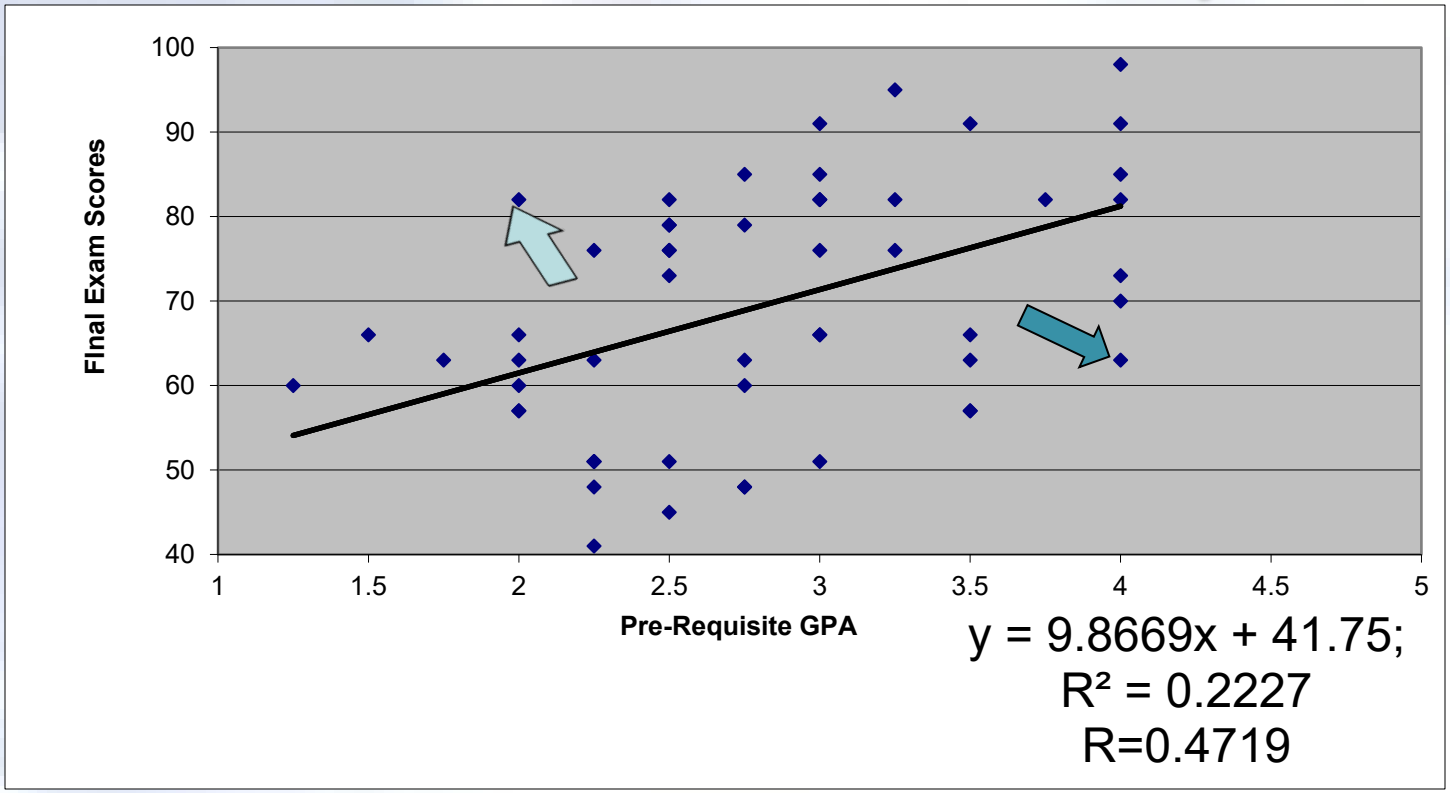


06.XX
Parting Thoughts

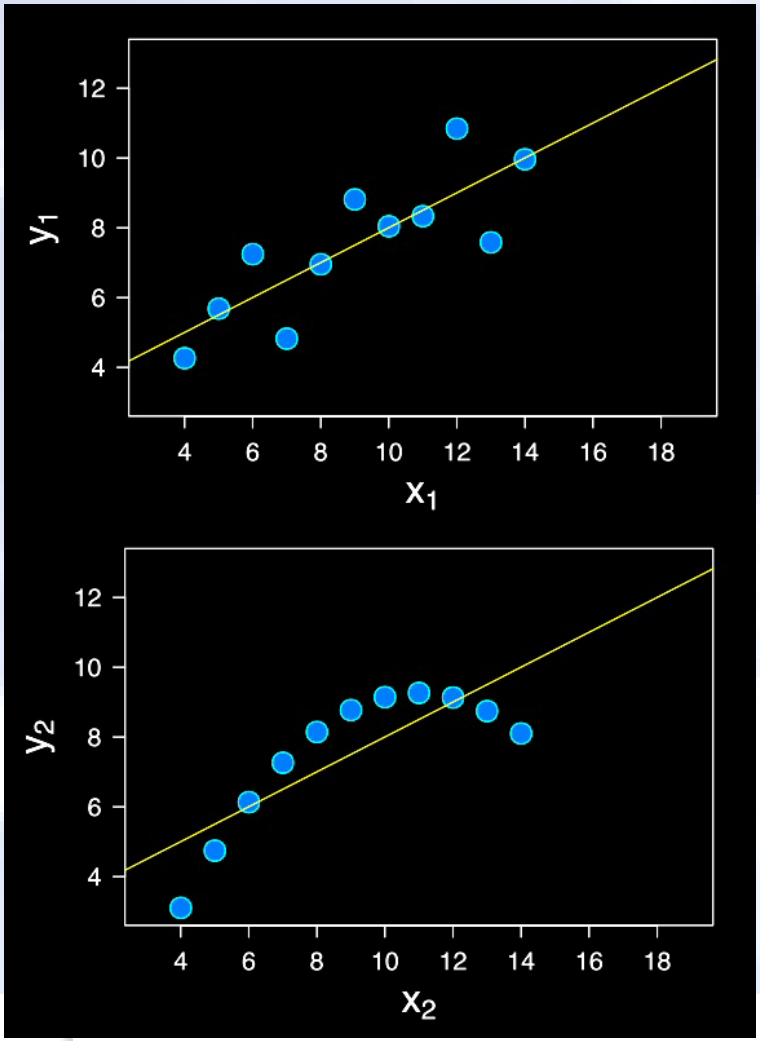
Final Exam Grade



Final Exam Grade vs Pre-Req GPA



Same but different

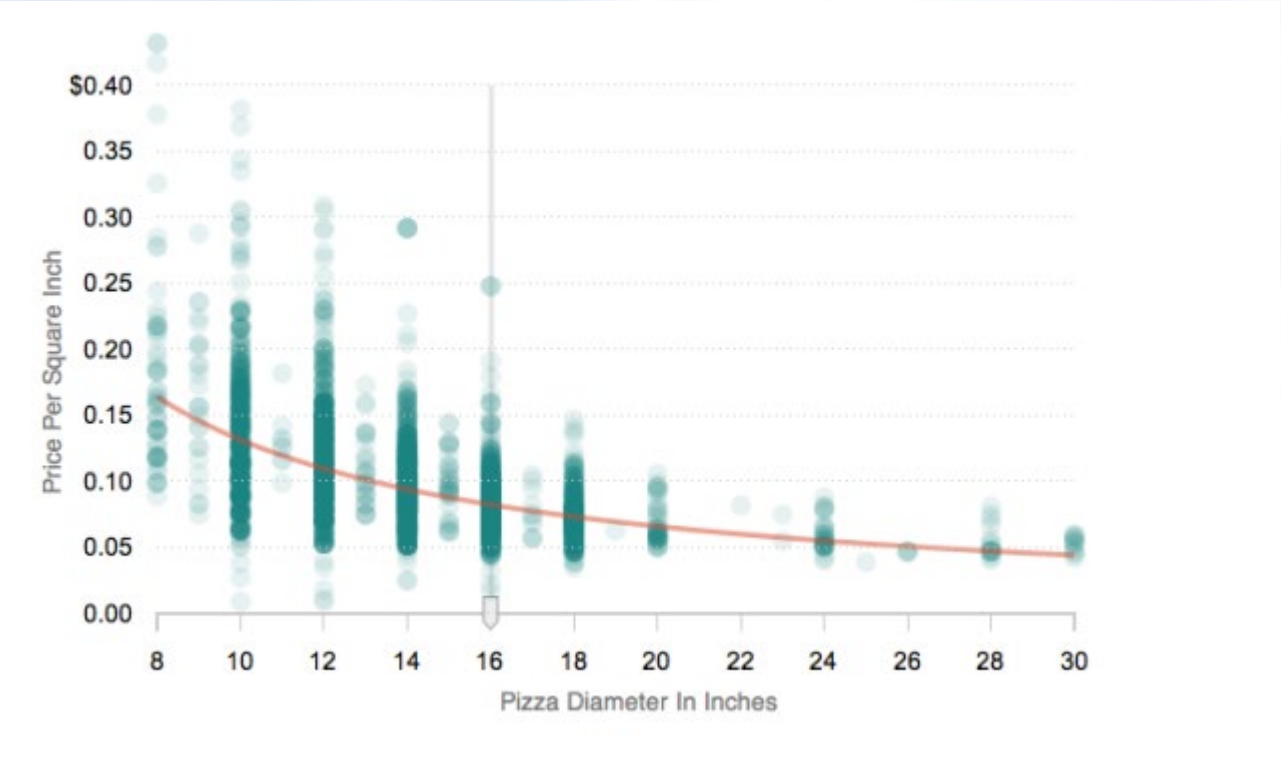


The following are the same for both lines.

- Mean of x
- Sample variance of x
- Mean of y
- Sample variance of y
- Correlation between x and y
- Linear regression line
- Coefficient of determination of the linear regression

Source: Cliff Pickover: https://twitter.com/pickover/status/1576630790309707776/photo/1

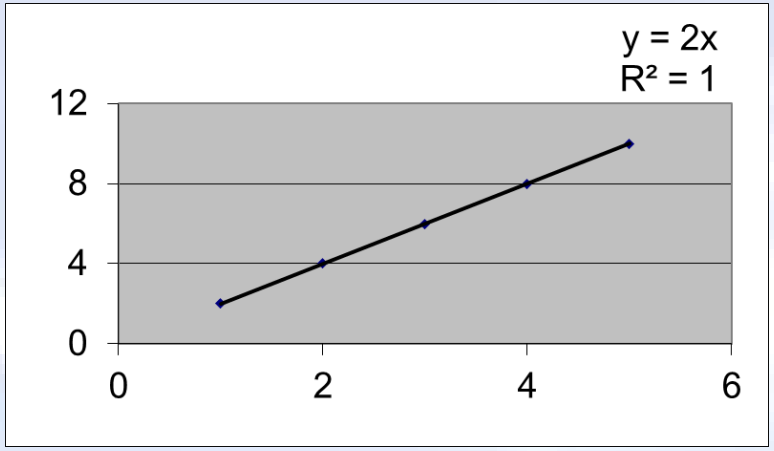
Pizza price vs Pizza Diameter



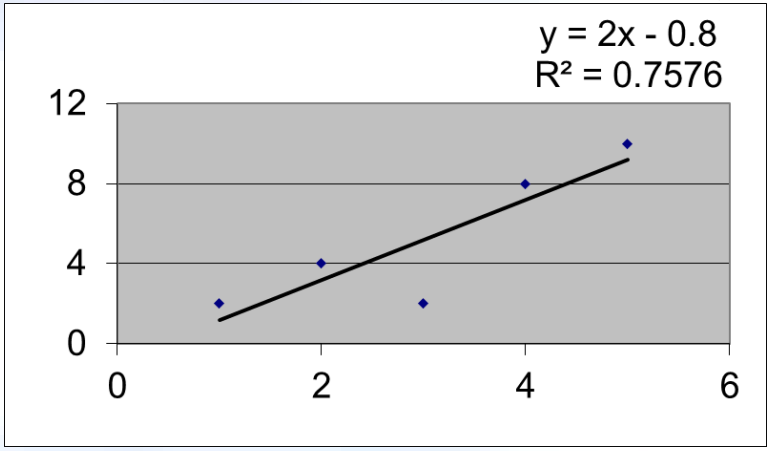
Sources: https://www.npr.org/sections/money/2014/02/26/282132576/74-476-reasons-you-should-always-get-the-bigger-pizza
https://www.themarysue.com/mpr-pizza-graph/

Effect of Outlier

1	2
2	4
3	6
4	8
5	10



1	2
2	4
3	2
4	8
5	10



Problem Assigned

Given (2,4), (2,5), (3,5) and (3,6) as data points

1) Regress to a general straight line, $y = a_0 + a_1 x$. (Answer: y=1x+2.5)

2) Find the standard error of estimate (Ans: 0.7071).

3) Find the scaled residuals (Answer: -0.7071 0.7071 -0.7071 0.7071).

Problem Assigned

Given (2,4), (2,5), (3,5) and (3,6) as data points (extension of previous problem)

1) Find the sum of the square of the differences with the mean (Ans: 2).

2) Find the sum of the square of the residuals. (Ans: 1)

3) Find the coefficient of determination (Ans: 0.5).

4) Find the correlation coefficient (Ans: 0.7071).

END