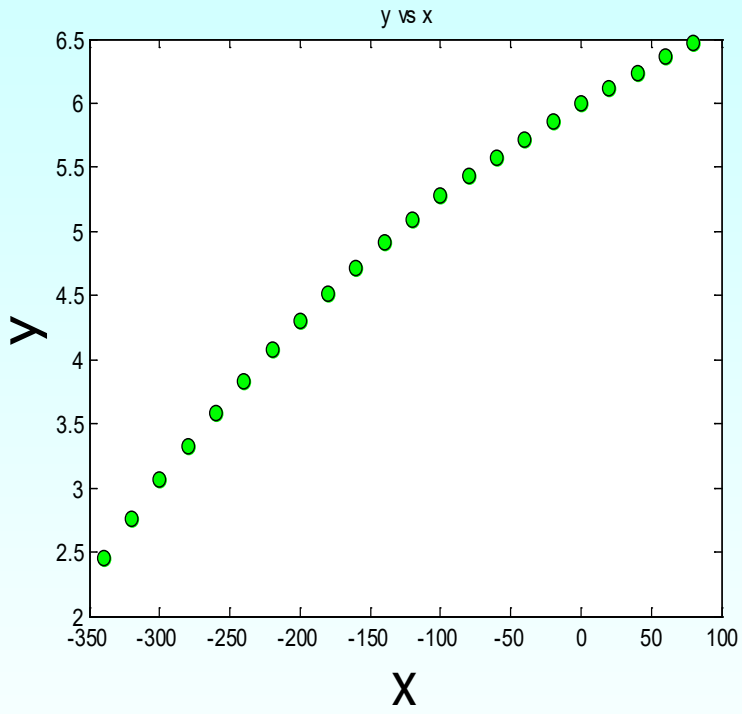
Adequacy of Linear Regression Models

http://numericalmethods.eng.usf.edu

Transforming Numerical Methods Education for STEM Undergraduates

Data



Therm exp coeff vs temperature

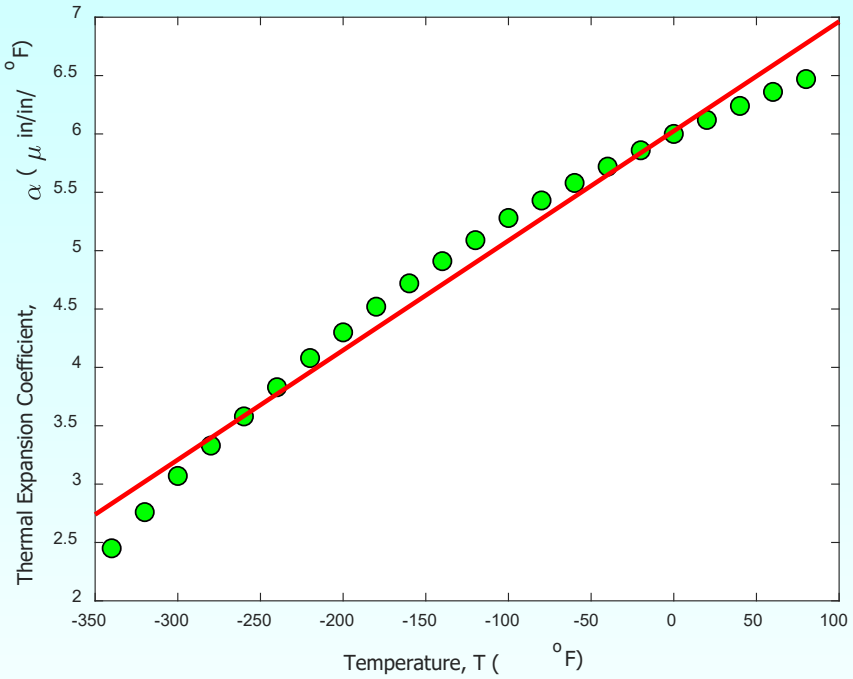
T	α
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.2
-60	5.58
-80	5.43
-100	5.28
-120	5.09

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 ^{o}F

α is in μin/in/ ^{o}F

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**

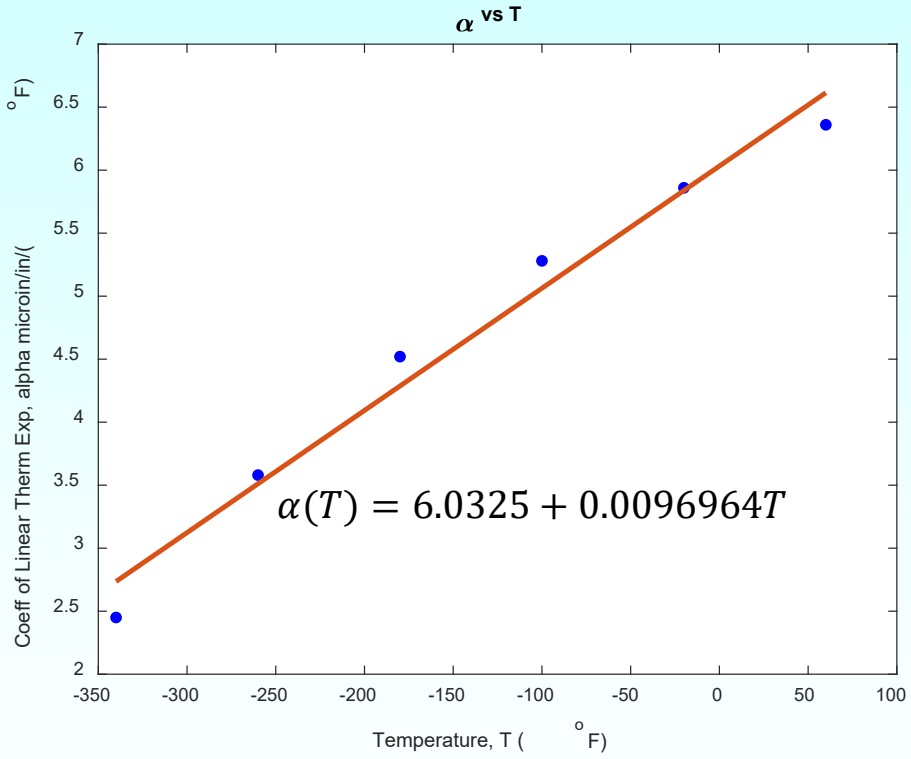
Four checks

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?

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 with ±2 standard error of estimate?

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

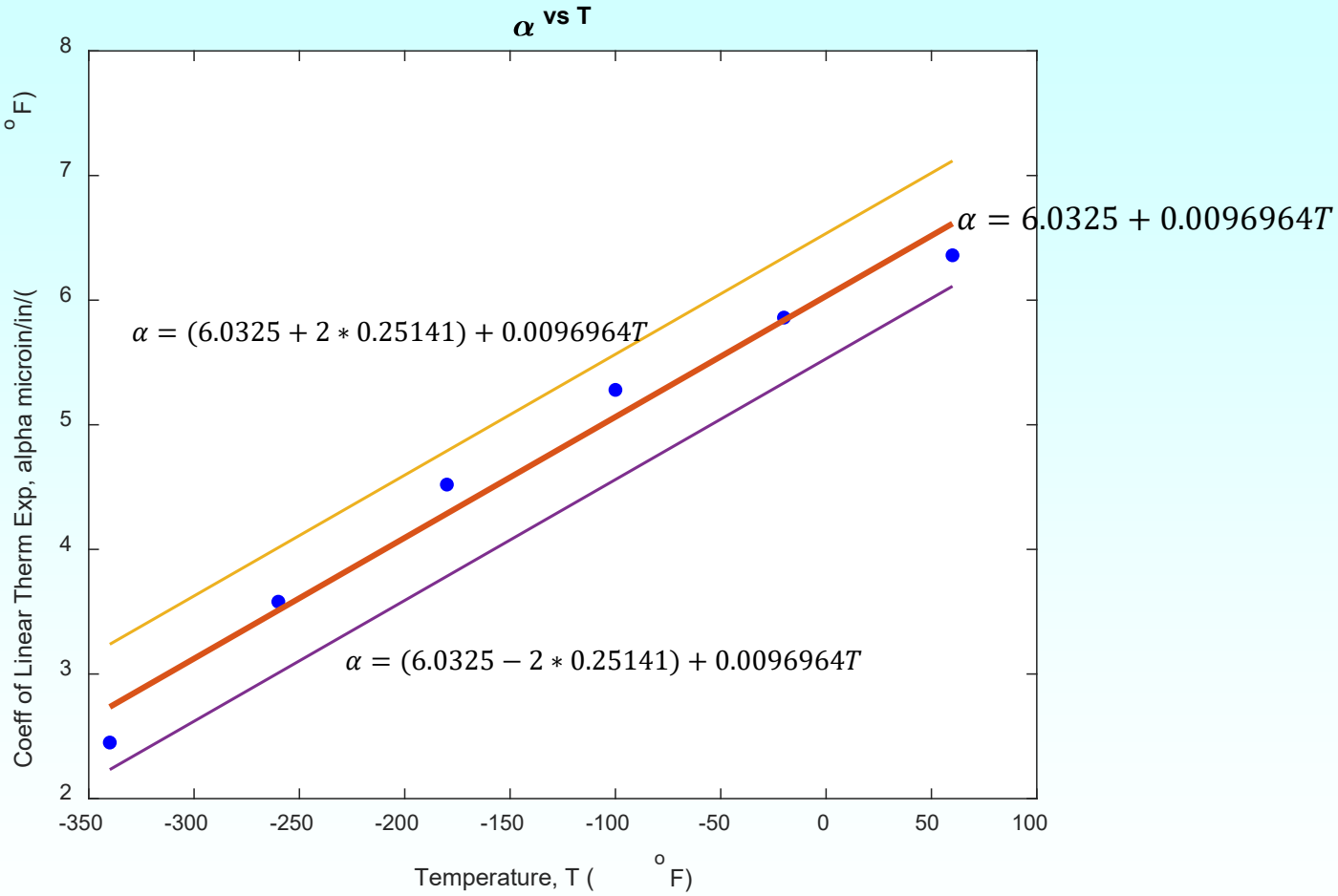
$$S_r = 0.25283$$

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

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

$$= 0.25141$$

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?

Coefficient of determination

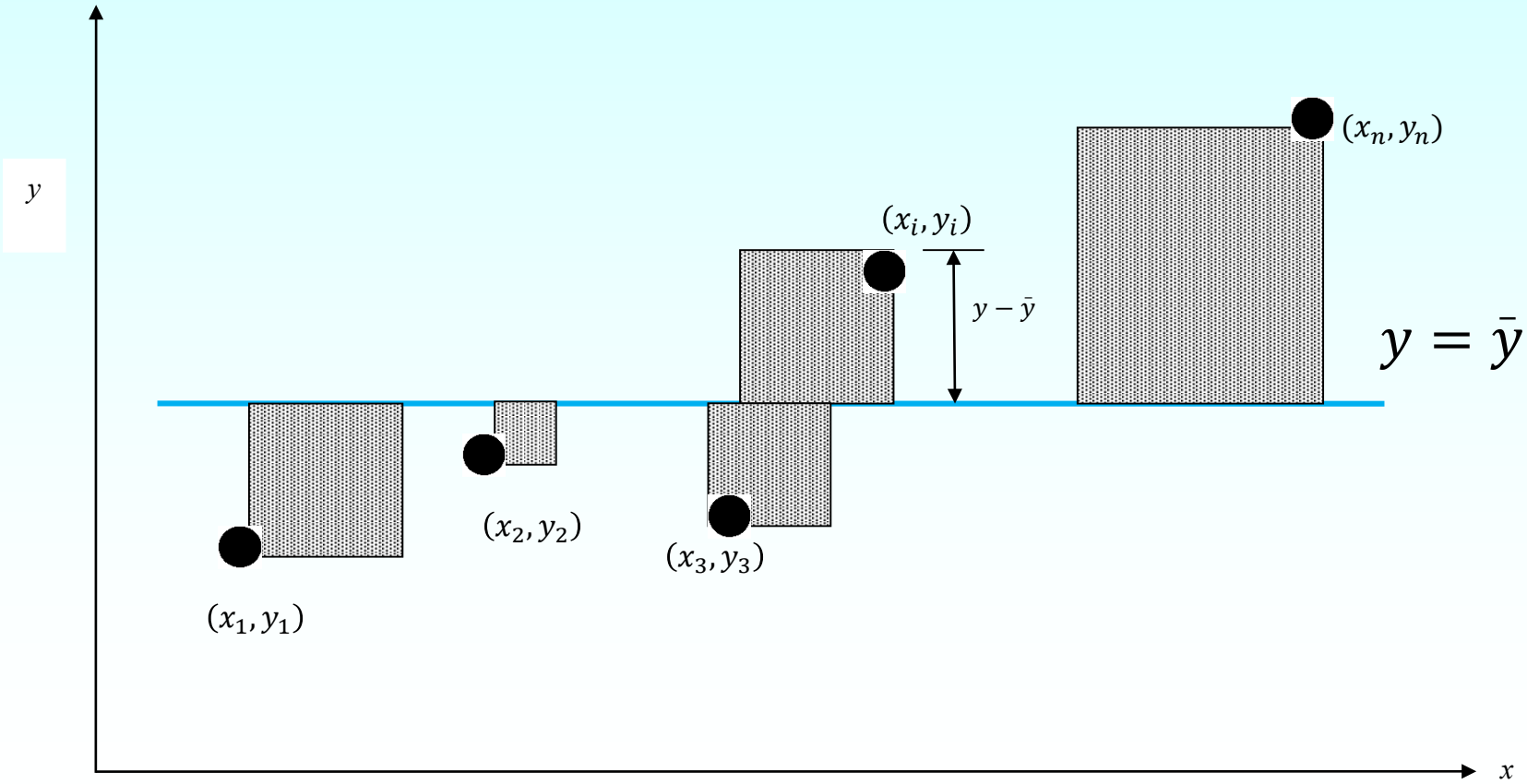
$$S_t = \sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2$$

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

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

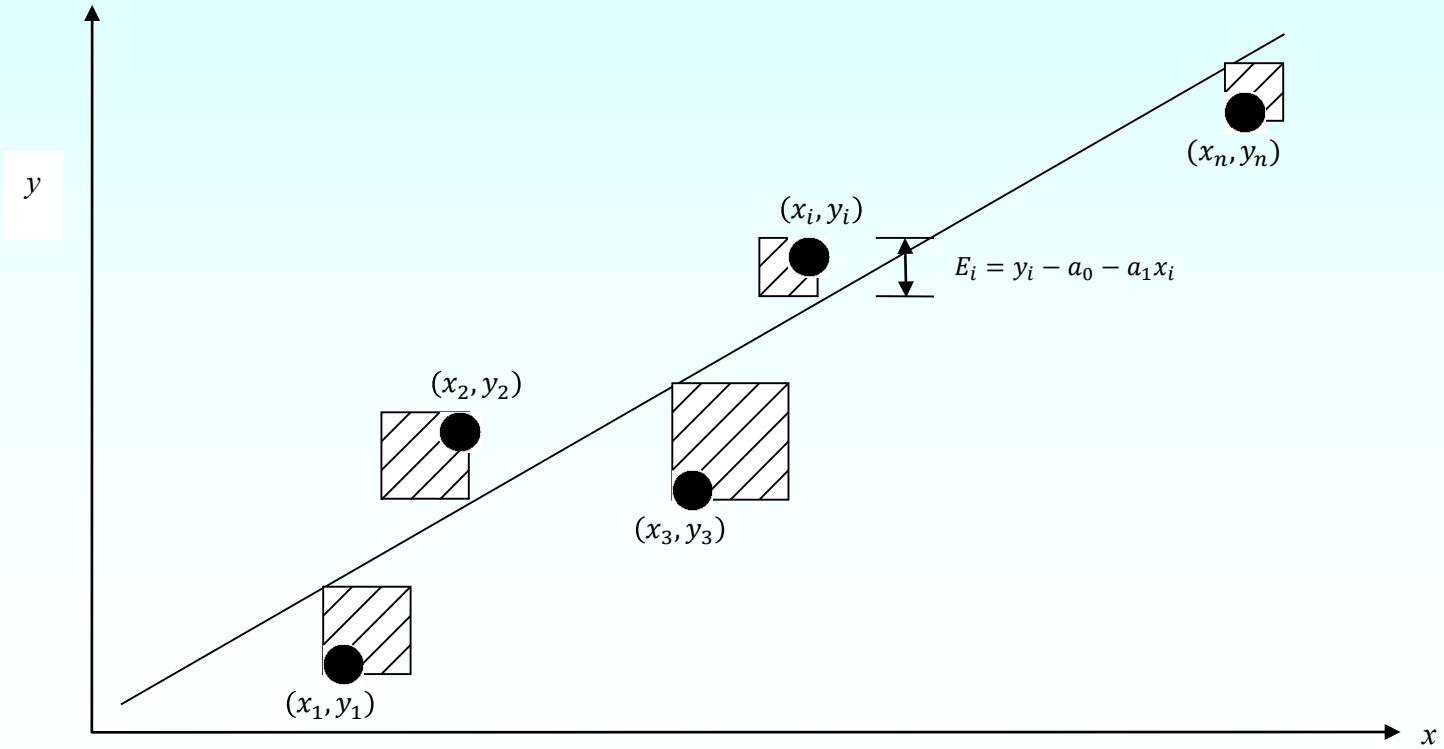
Sum of square of residuals between data and mean

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



Sum of square of residuals between observed and predicted

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



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

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

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

$$= 0.98820$$

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

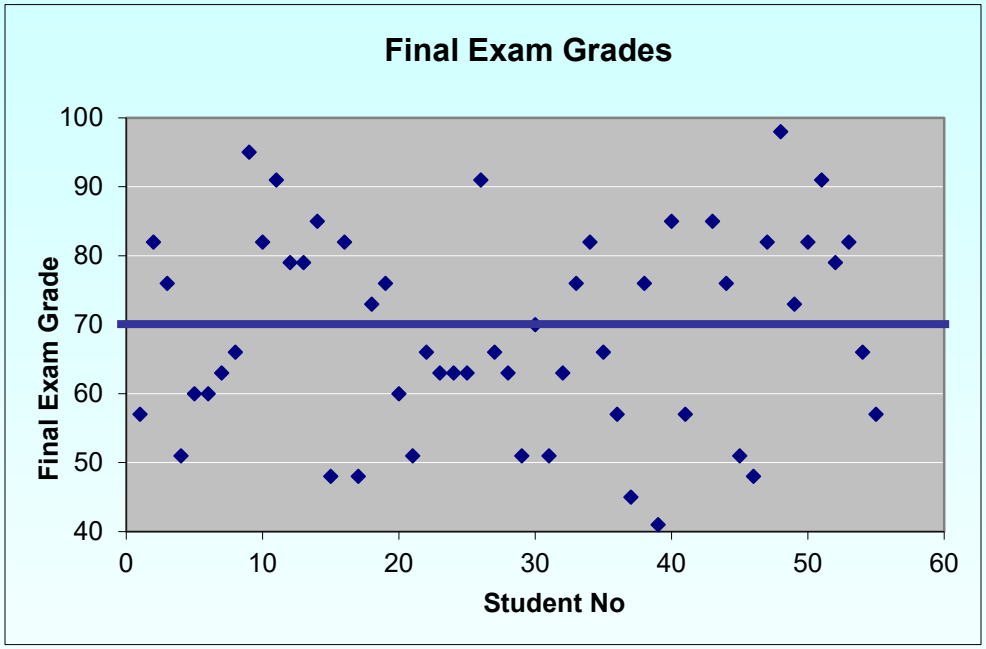
0.6 to 0.8 - Strong relationship

0.4 to 0.6 - Moderate relationship

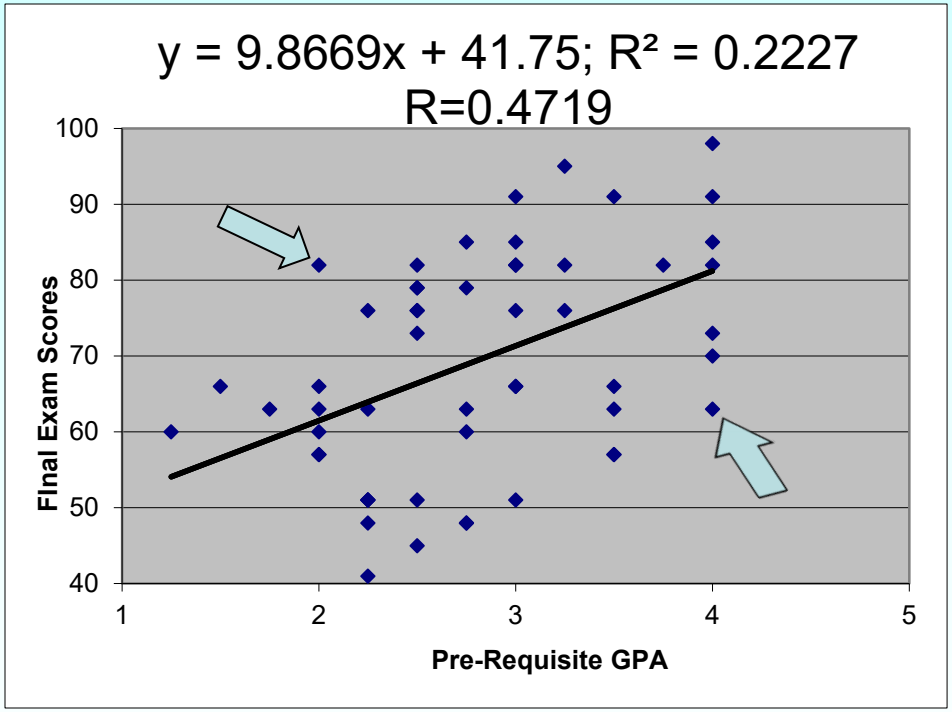
0.2 to 0.4 - Weak relationship

0.0 to 0.2 - Weak or no relationship

Final Exam Grade



Final Exam Grade vs Pre-Req GPA

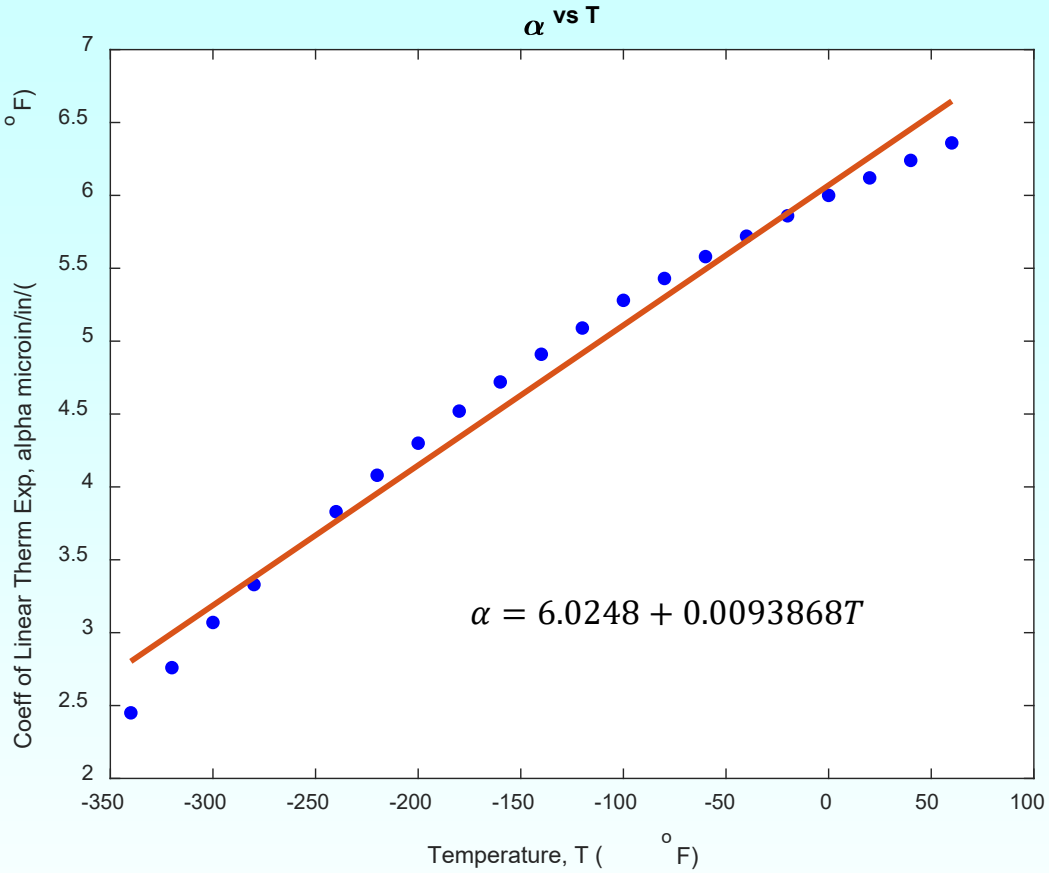


Redoing Check 1, 2 and 3 with 22 data points

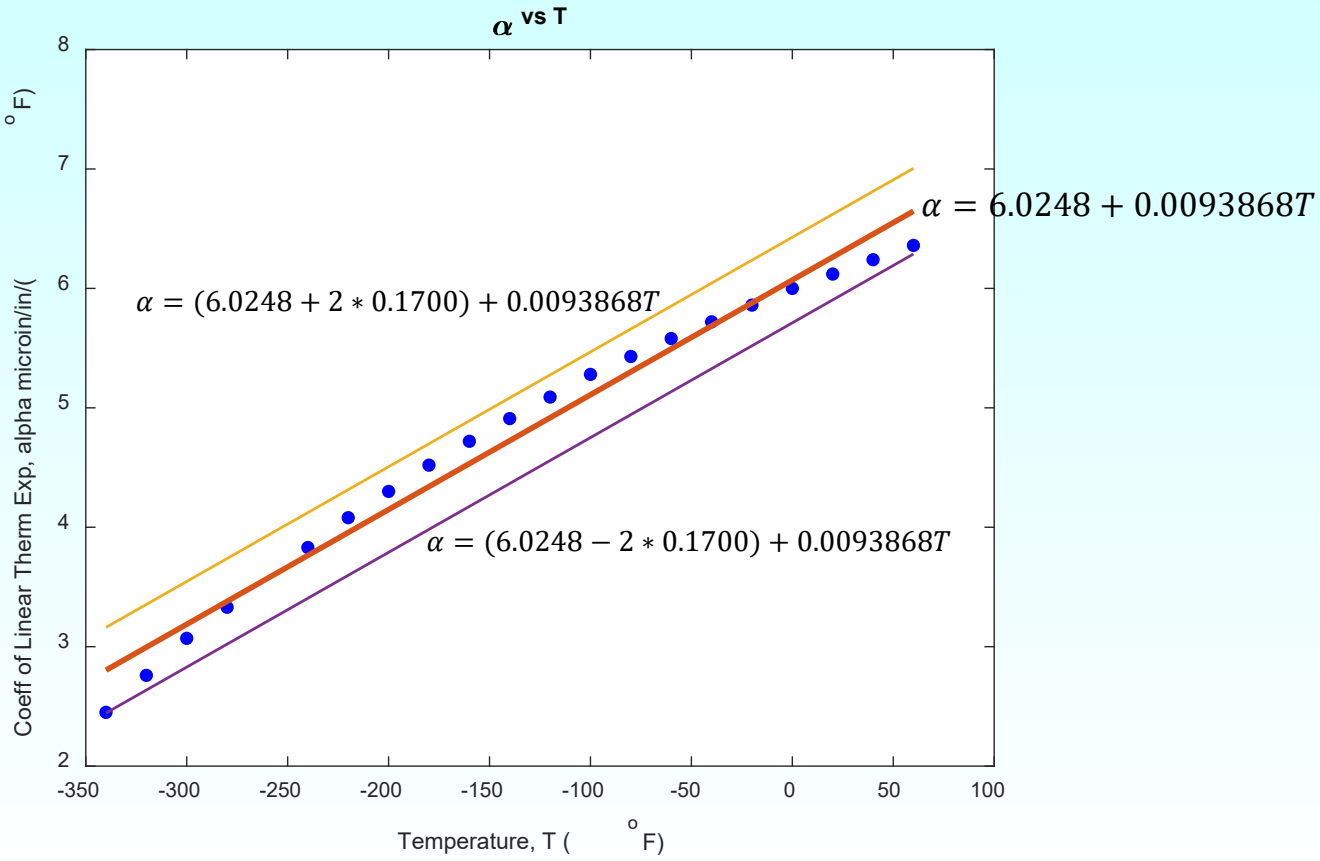
Check 1:Plot Model and Data

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: Using Standard Error of Estimate



$s_{\alpha/T}$ =0.1700

Check 3: Using Coefficient of Determination

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

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

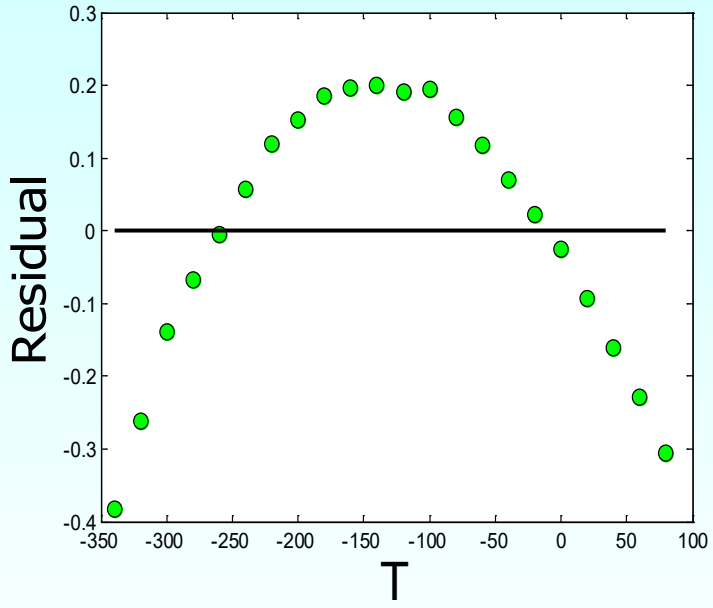
$$= 0.9791$$

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

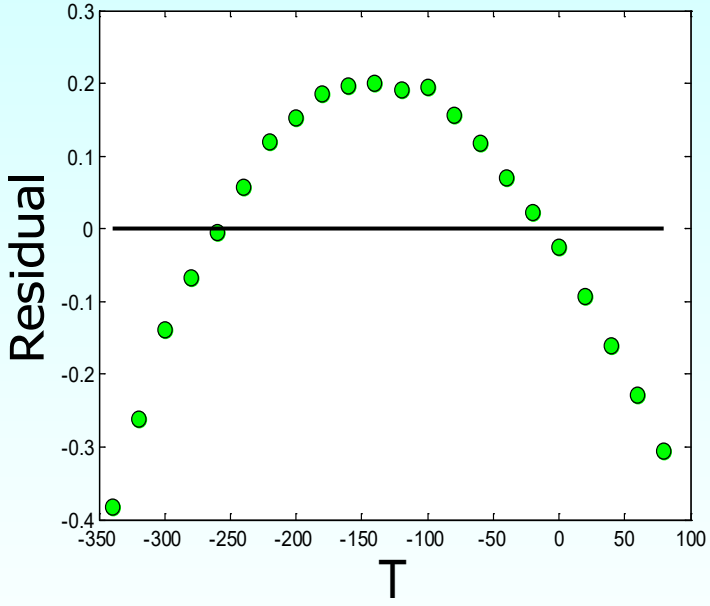
Model meets assumption of random errors

- 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.~~

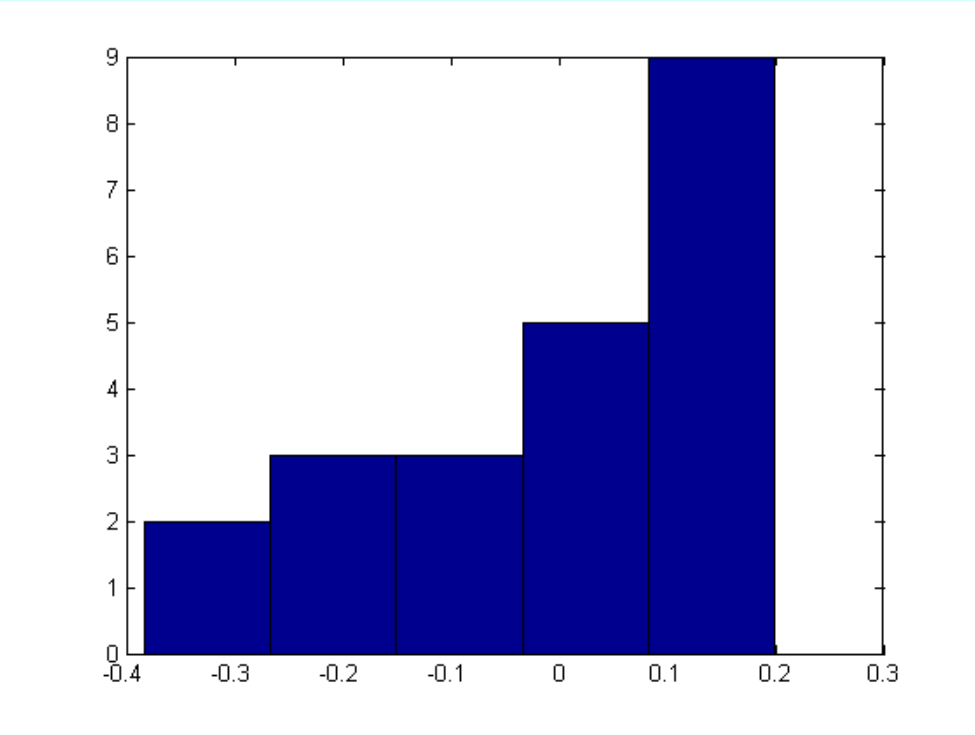
Are residuals negative and positive?



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



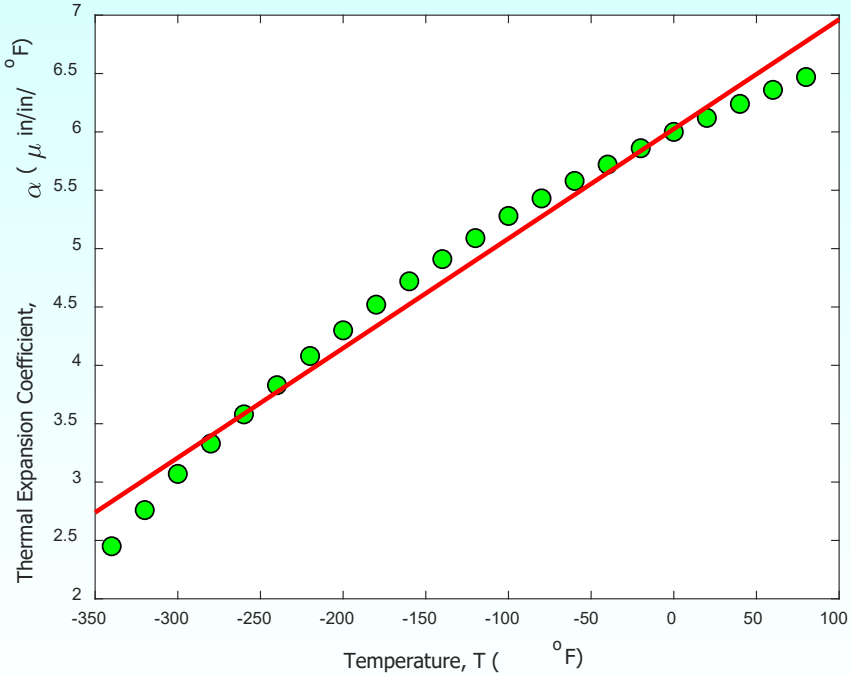
Do the residuals follow normal distribution?



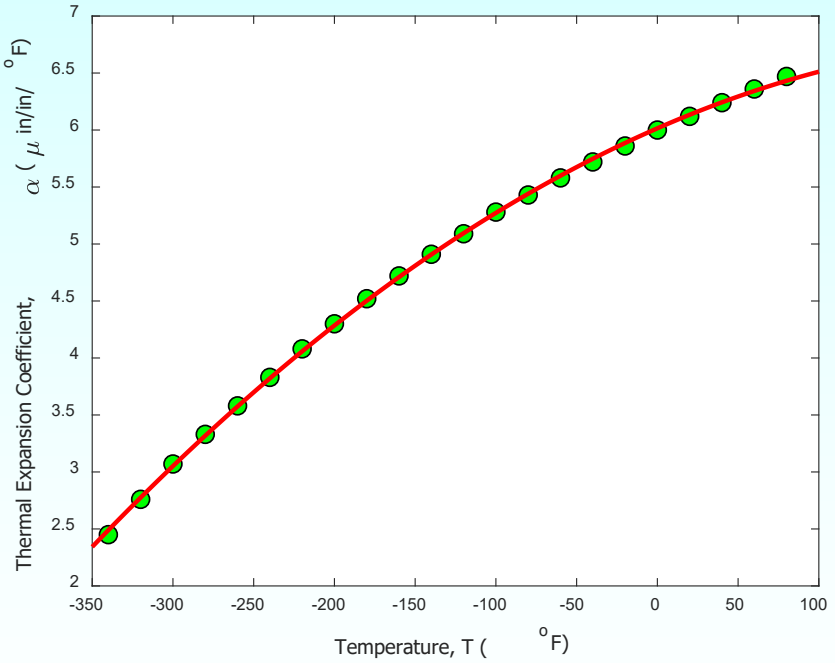
END

What polynomial model to choose if one needs to be chosen?

Which model to choose?

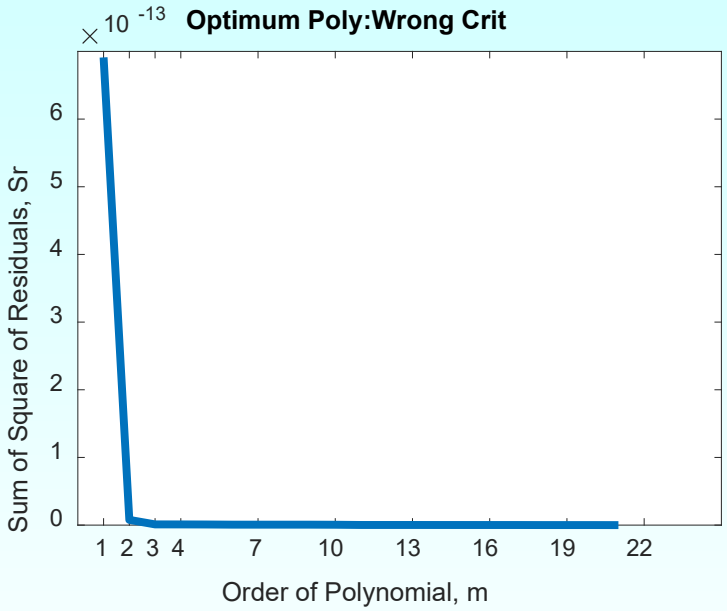


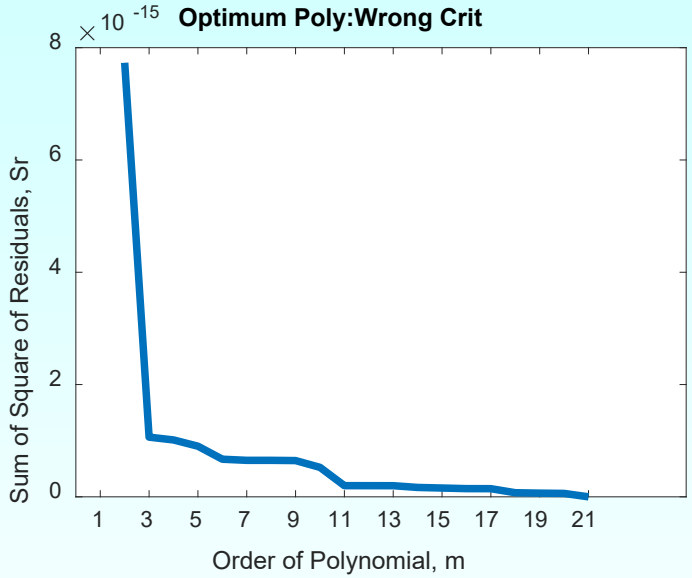
$\alpha = 0.009387T + 6.025$



$\alpha = -0.00001228T^2 + 0.006196T + 6.015$

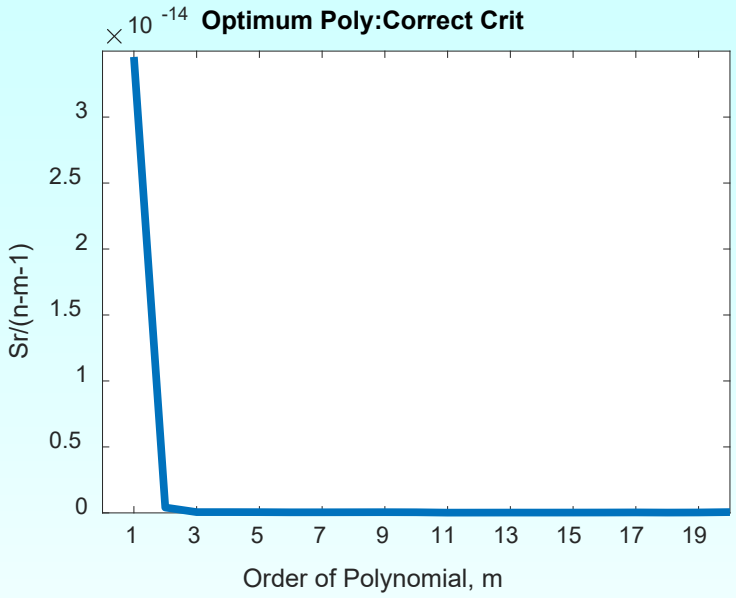
Optimum Polynomial: Wrong Criterion

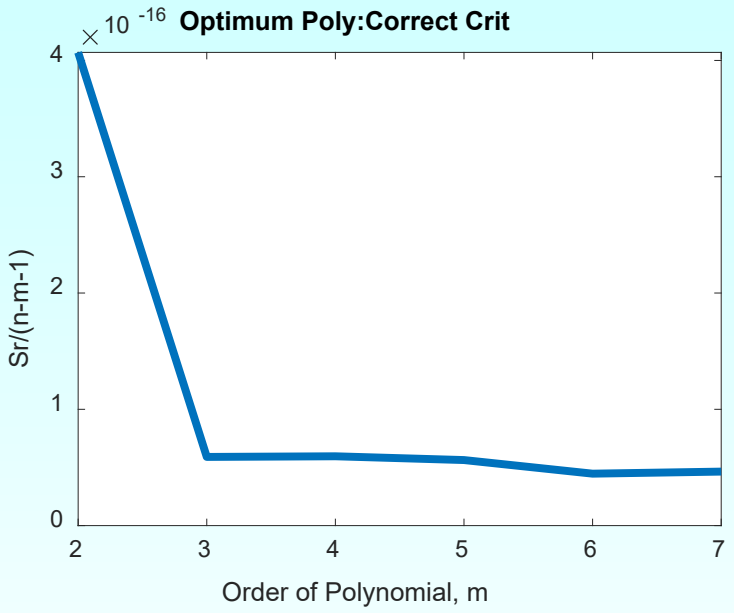




Both graphs are same
Left one starts at m=1
Right one starts at m=2

Optimum Polynomial: Correct Criterion



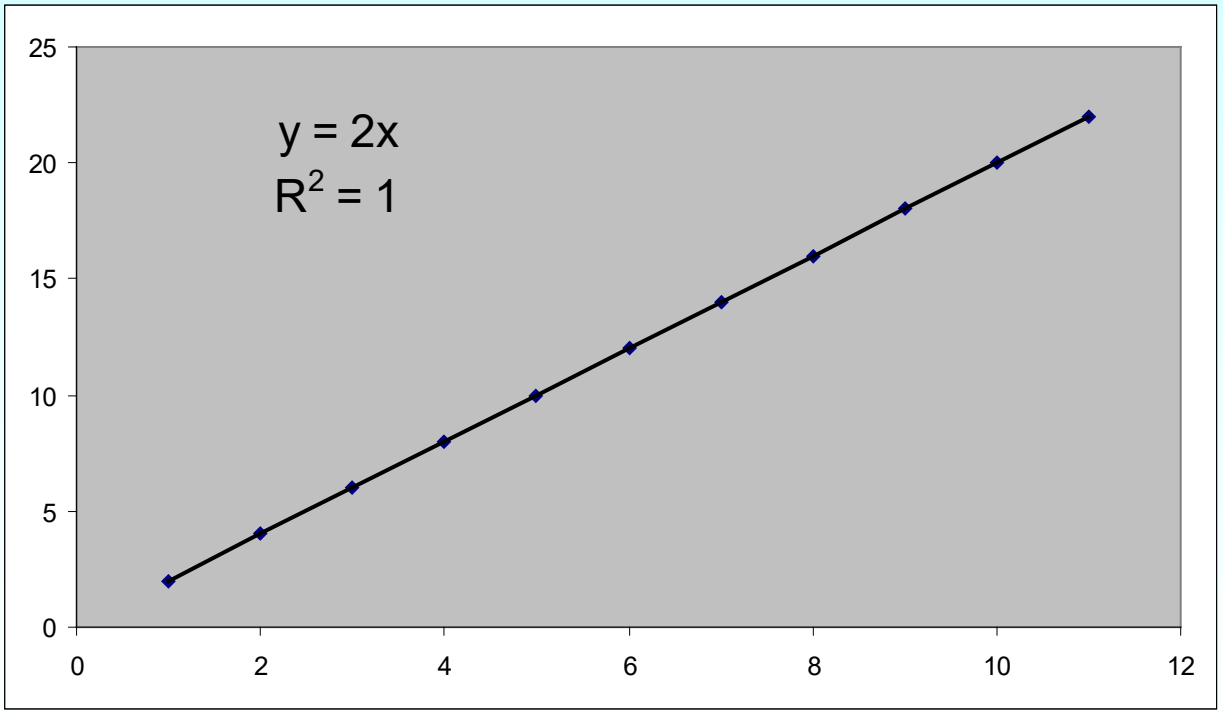


Both graphs are same
Left one starts at m=1
Right one starts at m=2

END

Effect of an Outlier

Effect of Outlier



Effect of Outlier