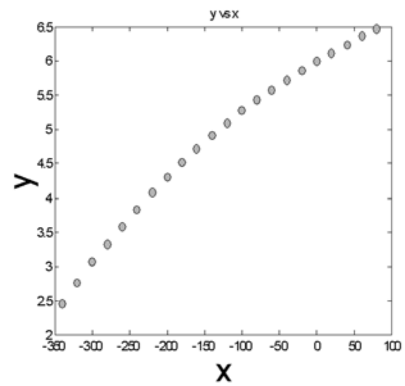
# Adequacy of Linear Regression Models

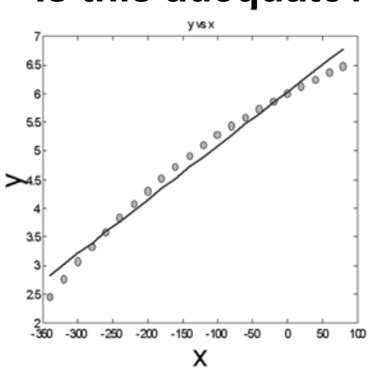
http://numericalmethods.eng.usf.edu
Transforming Numerical Methods Education for STEM Undergraduates



## Data




## 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. Plot the data and the model.
2. Find standard error of estimate.
3. Calculate the coefficient of determination.
4. Check if the model meets the assumption of random errors.

Example: Check the adequacy of the straight line model for given data

$$\alpha = a_0 + a_1 T$$

| T (F) | α (μin/in/F) |
|---|---|
| -340 | 2.45 |
| -260 | 3.58 |
| -180 | 4.52 |
| -100 | 5.28 |
| -20 | 5.86 |
| 60 | 6.36 |

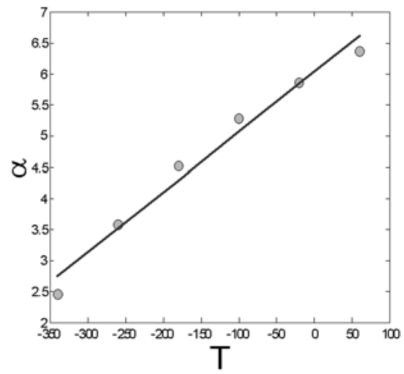
**END**

**1. Plot the data and the model**

## Data and model

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

| T (F) | α (μin/in/F) |
|---|---|
| -340 | 2.45 |
| -260 | 3.58 |
| -180 | 4.52 |
| -100 | 5.28 |
| -20 | 5.86 |
| 60 | 6.36 |




# END

**2. Find the standard error of estimate**

## Standard error of estimate

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

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

## Standard Error of Estimate

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

| $T_i$ | $\alpha_i$ | $a_0 + a_1 T_i$ | $\alpha_i - a_0 - a_1 T_i$ |
|---|---|---|---|
| **-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

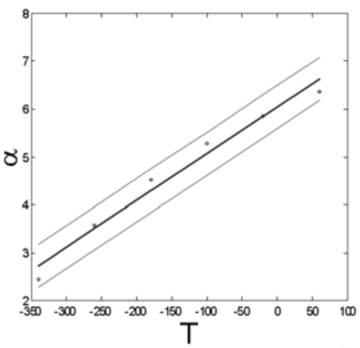
$$S_r = 0.25283$$

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

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

$$= 0.25141$$

## Standard Error of Estimate




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

## 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

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

| $T_i$ | $\alpha_i$ | 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 |

## END

## 3. Find the coefficient of determination

## Coefficient of determination

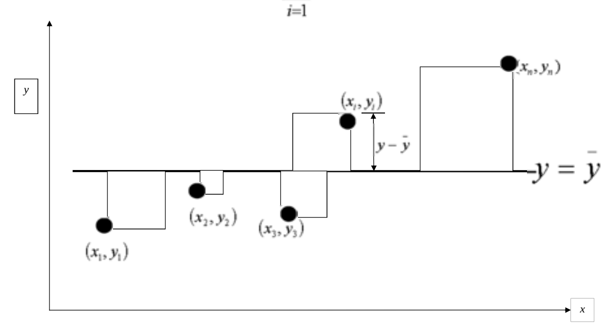
$$S_t = \sum_{i=1}^{n} \left(\alpha_i - \overline{\alpha}\right)^2$$

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

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

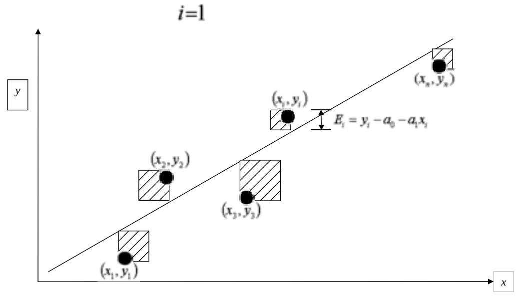
## Sum of square of residuals between data and mean

$$S_t = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$




## Sum of square of residuals between observed and predicted

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$




## Limits of Coefficient of Determination

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

$$0 \le r^2 \le 1$$

## Calculation of $S_t$

| $T_i$ | $\alpha_i$ | $\alpha_i - \overline{\alpha}$ |
|---|---|---|
| -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 |

$$\overline{\alpha} = 4.6750$$

$$S_t = 10.783$$

## Calculation of $S_r$

| $T_i$ | $\alpha_i$ | $a_0 + a_1 T_i$ | $\alpha_i - a_0 - a_1 T_i$ |
|---|---|---|---|
| -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

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

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

$$= 0.97655$$

## 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

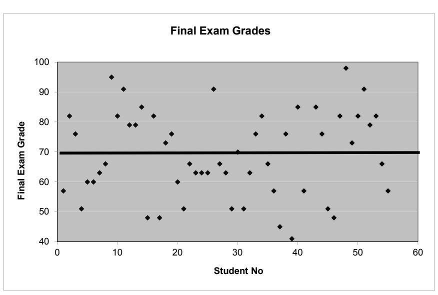
0.2 to 0.4 - Weak relationship

0.0 to 0.2 - Weak or no relationship

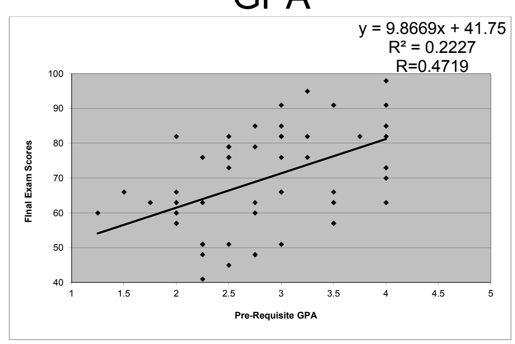
## Caution in use of $r^2$

- Increase in spread of regressor variable (x) in y vs. x increases $r^2$
- Large regression slope artificially yields high $r^2$
- Large $r^2$ does not measure appropriateness of the linear model
- Large $r^2$ does not imply regression model will predict accurately

## Final Exam Grade




## Final Exam Grade vs Pre-Req GPA



# END

# 4. Model meets 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.

## Therm exp coeff vs temperature

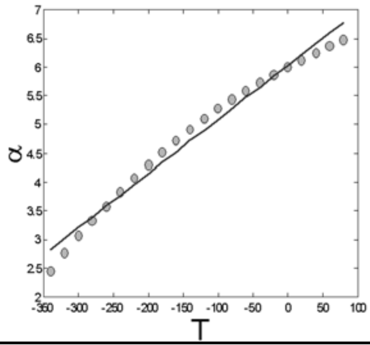
| T | α |
|---|---|
| 60 | 6.36 |
| 40 | 6.24 |
| 20 | 6.12 |
| 0 | 6.00 |
| -20 | 5.86 |
| -40 | 5.72 |
| -60 | 5.58 |
| -80 | 5.43 |

| T | α |
|---|---|
| -100 | 5.28 |
| -120 | 5.09 |
| -140 | 4.91 |
| -160 | 4.72 |
| -180 | 4.52 |
| -200 | 4.30 |
| -220 | 4.08 |
| -240 | 3.83 |

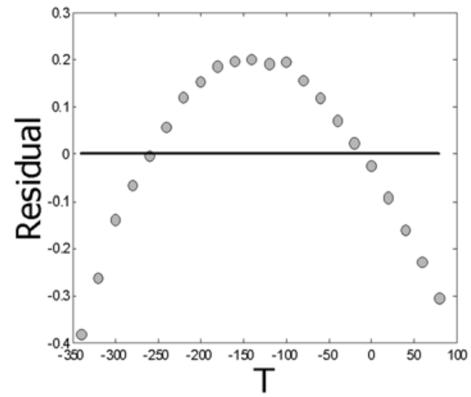
| T | α |
|---|---|
| -280 | 3.33 |
| -300 | 3.07 |
| -320 | 2.76 |
| -340 | 2.45 |

## Data and model

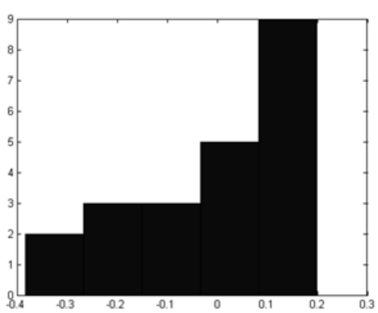
$$\alpha = 6.0248 + 0.0093868T$$




## Plot of Residuals




## Histograms of Residuals



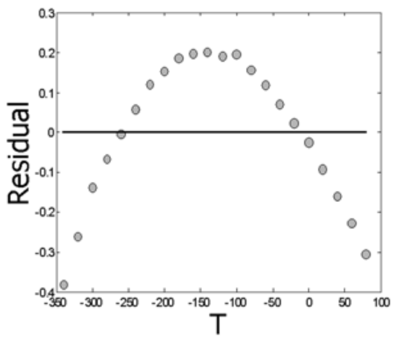
## **Check for Autocorrelation**

- Find the number of times, $q$ the sign of the residual changes for the *n* data points.
- If $(n-1)/2-\sqrt{(n-1)} \le q \le (n-1)/2+\sqrt{(n-1)}$, you most likely do not have an autocorrelation.

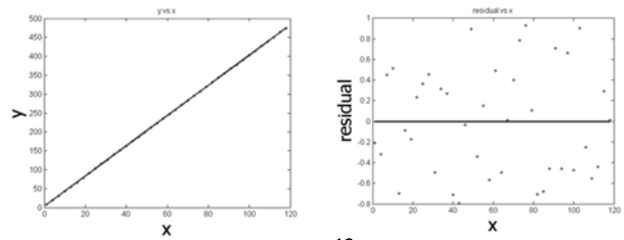
$$\frac{(22-1)}{2} - \sqrt{22-1} \le q \le \frac{22-1}{2} + \sqrt{22-1}$$

$$5.9174 \le q \le 15.083$$

## Is there autocorrelation?




## y vs x fit and residuals

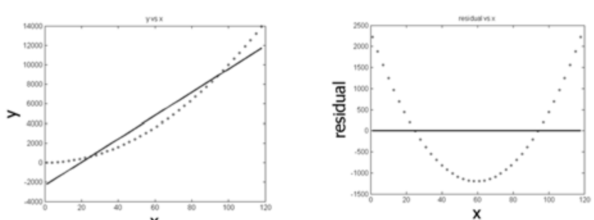



n=40

(n-1)/2-√(n-1) ≤p≤ (n-1)/2+√(n-1)

Is 13.3≤21≤ 25.7? Yes!

## y vs x fit and residuals




n=40

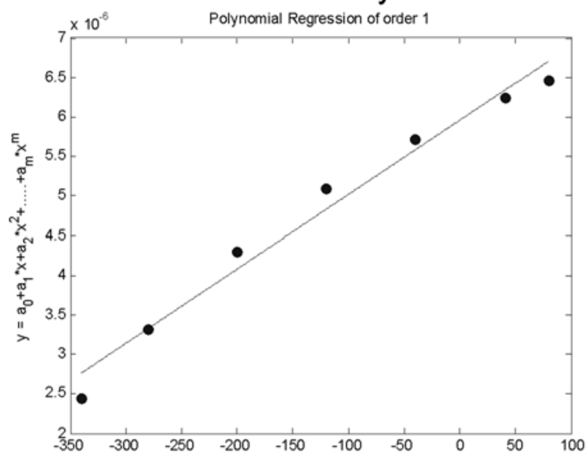
(n-1)/2-√(n-1) ≤p≤ (n-1)/2+√(n-1)

Is 13.3≤2≤ 25.7? No!

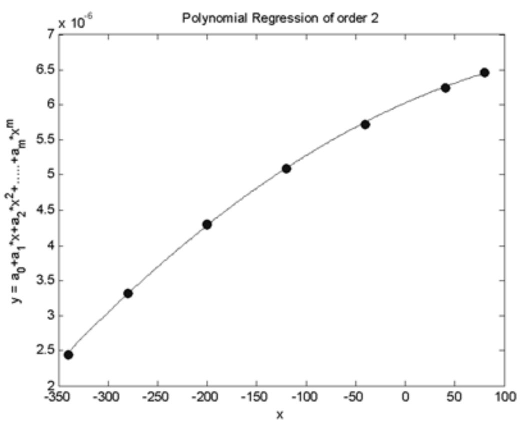
## END

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

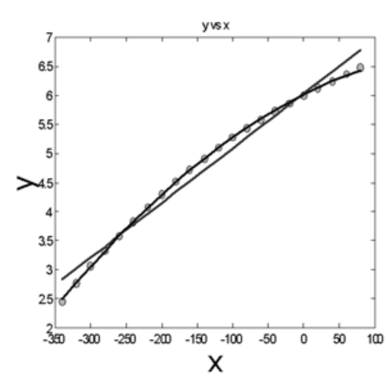
## First Order of Polynomial



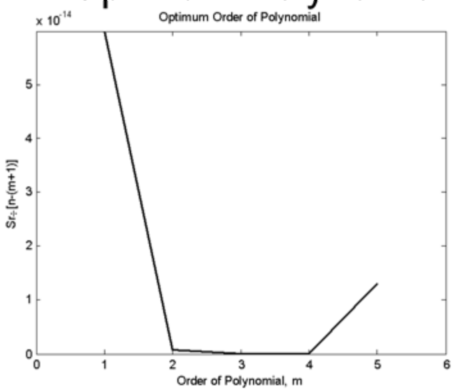

## Second Order Polynomial




## Which model to choose?



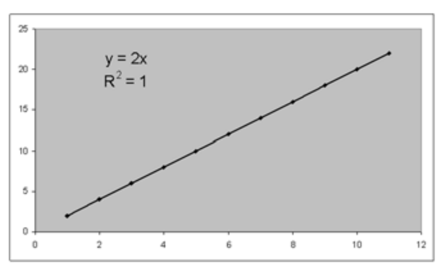
## Optimum Polynomial




**THE END**

**Effect of an Outlier**

## Effect of Outlier



## Effect of Outlier