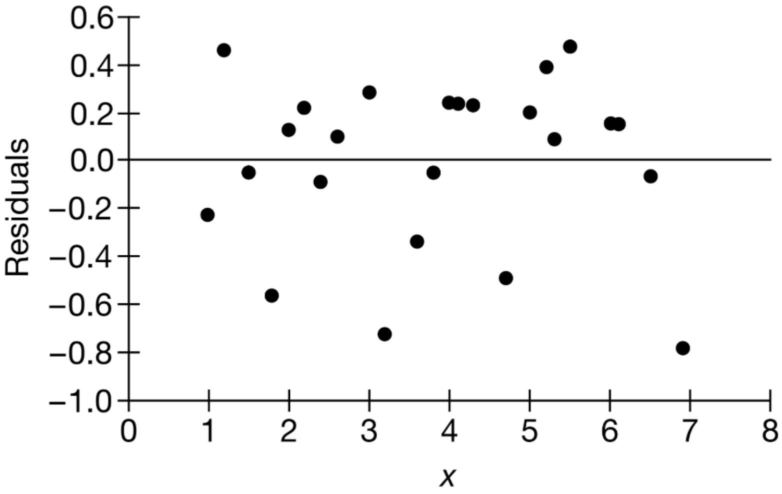
Spring 2021

Section 4 (DSSA)

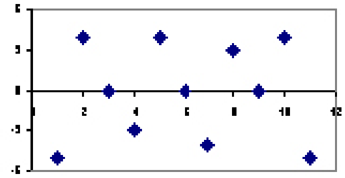
Regression – Part 2

First: Regression Model Analysis

- A plot of the errors *“A Residual Plot”* may highlight problems with the model
  - Residual (error) = Actual Y value − Predicted Y value
  - A *residual plot* is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.

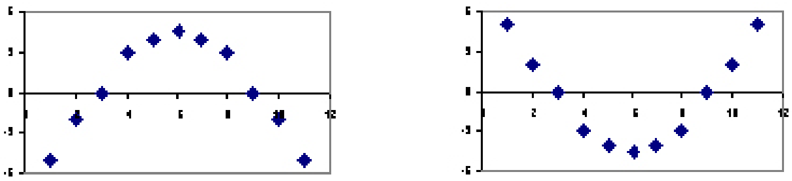


- There are four assumptions associated with a linear regression model:
  - Linearity:
    1. If the points in *a residual plot* are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.



A random pattern,

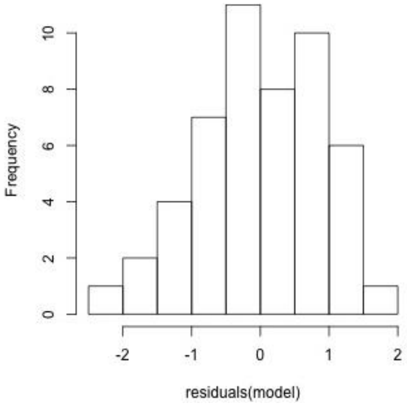
indicating a good fit for a linear model



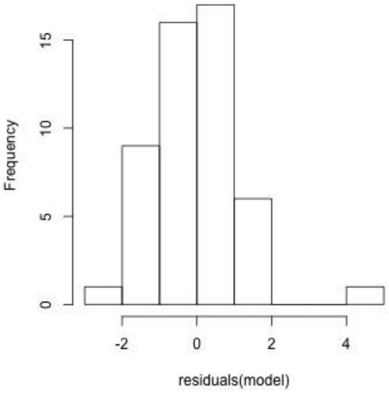
Non-random plot patterns,

suggesting a better fit for a nonlinear model

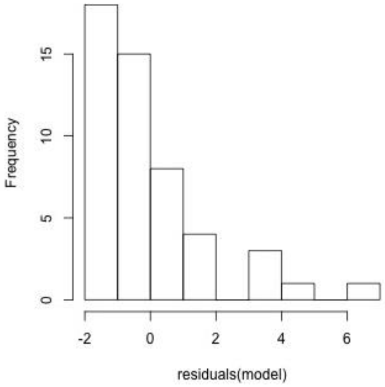
- Normality: The errors are normally distributed, with a mean of zero.
    1. If the *histogram* of standardized residuals is symmetric bell-shaped (evenly distributed around zero indicates), then the normality assumption is likely to be true.
    2. If the *histogram* indicates that random error is not normally distributed, it suggests that the model's underlying assumptions may have been violated.



The Residuals are
normally distributed

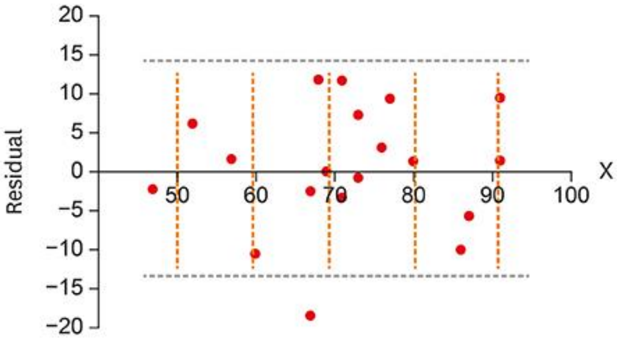


Residuals are normally distributed
But, there is one extreme outlier

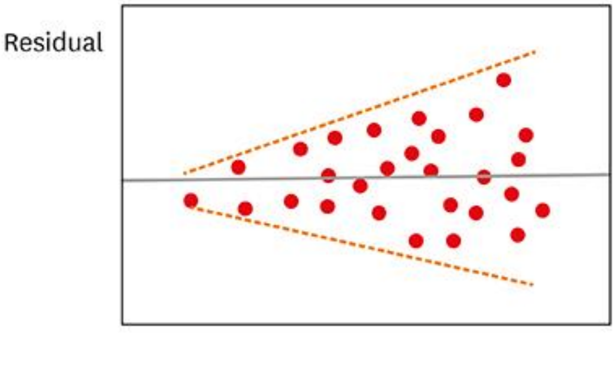


Residuals are
not normally
distributed (skewed)

- Homoscedasticity: The residuals have a constant variance regardless of the value of X
    1. Data are homoscedastic if the *residuals plot* is the same width for all values of the independent variable.
    2. Heteroscedasticity is when the variability in the response is changing as the predicted value increases.

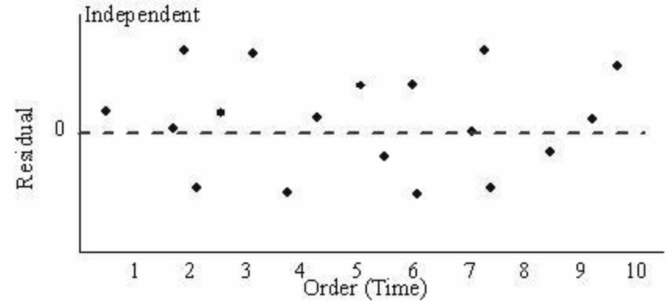


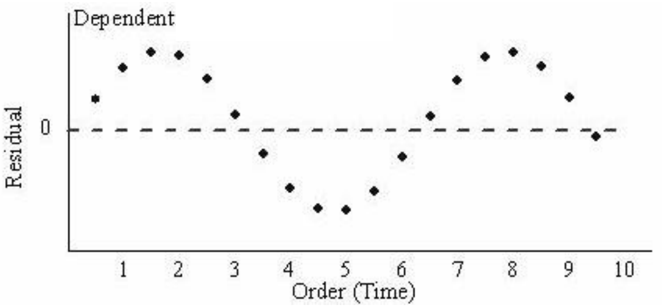
Residual plot showing homoscedasticity



Residual plot showing heteroscedasticity

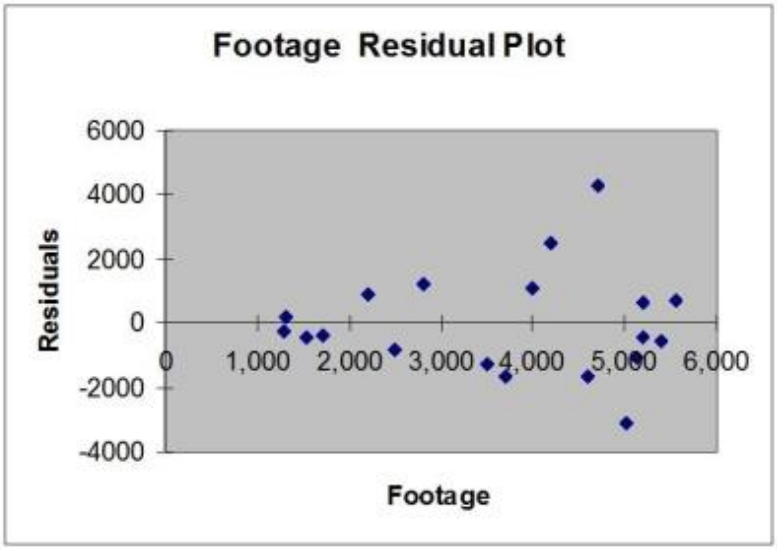
- Independence: the distribution of errors is random and not influenced by or correlated to the errors in prior observations.





Example 1

Based on the residual plot below, you will conclude that there might be a violation of which of the following assumptions?



A) independence of errors
B) homoscedasticity
C) linearity of the relationship
D) normality of errors

Ans. B

Second: Regression Model Significance

- The F-test of overall significance is performed to indicate whether the resulting linear regression model is a true representation of the population or not (to avoid sampling error)

$$\boldsymbol{Y}\text{^}=\boldsymbol{\beta 0}+\boldsymbol{\beta 1 X}$$

  - Hypothesis Testing:
    1. H0: population slope coefficient $\boldsymbol{\beta 1} = 0$
    2. H1: population slope coefficient $\boldsymbol{\beta 1} \neq 0$
  - If ($\boldsymbol{\beta 1} = 0$) the null hypothesis is that there is no relationship between Y and X
  - The alternate hypothesis is that there is a linear relationship ($\boldsymbol{\beta 1} \neq 0$)
  - If the null hypothesis is rejected, then the independent variable is responsible in the variation in Y

- An F-statistic is the ratio of two variances,

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{K}$$

    - MSR (mean squared regression)
    - MSE (mean squared error)
    - SSR (sum of squares due to regression/ Total Variability/ Explained Variability)
    - k (number of independent variables in the model/x)

  - Whenever the F value is large, the significance level (p-value) will be low, indicating that the model is useful!

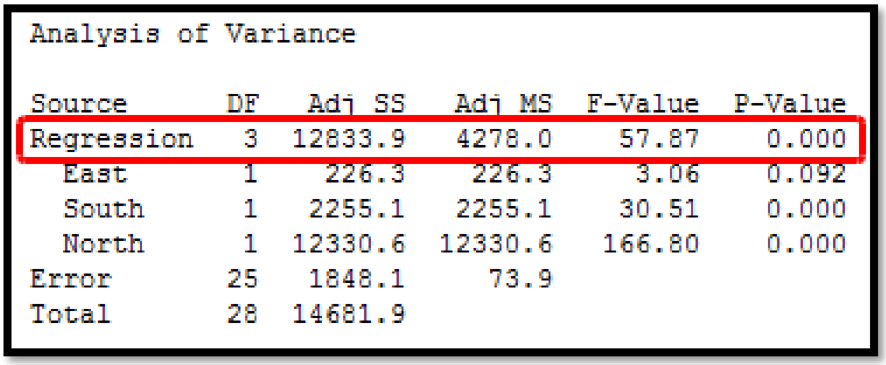
$$F_{calculated} > F_{critical}$$

$$or$$

$$P - Value < \propto$$

    - $F_{critical} = F_{\alpha,df1,df2}$ value from F distribution table
      - df1= k
      - df2= n-k-1
      - k is number of independent variables, n is the sample size
    - $P - value$ = Probability ($F_{critical} > F_{calculated}$)
    - Alpha or α, is the significance level
    - A significance level of 0.05 signifies a 5% risk of deciding that an effect exists when it does not exist (95% confidence)

- You can find the overall F-test in the ANOVA table

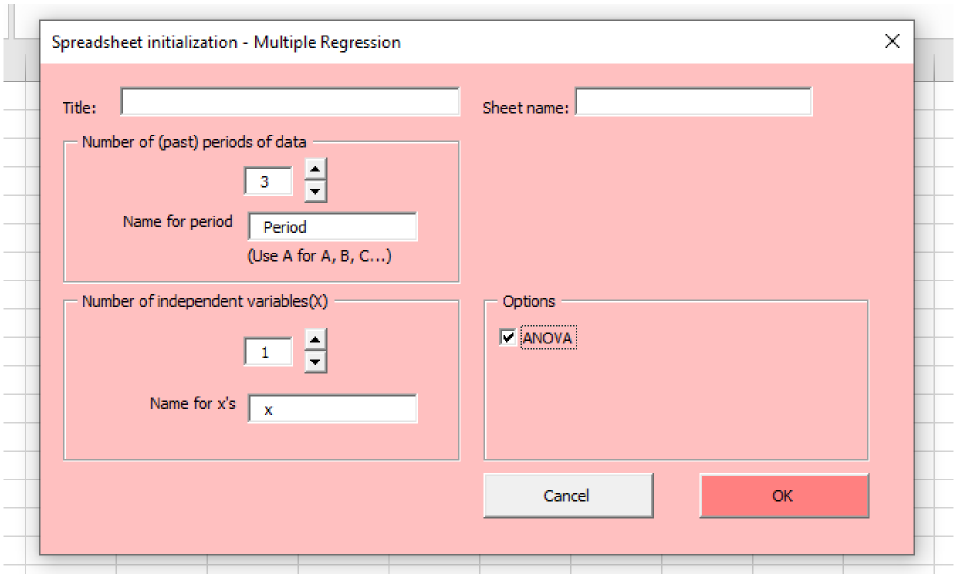


```
Analysis of Variance

Source        DF    Adj SS    Adj MS  F-Value  P-Value
Regression     3   12833.9    4278.0    57.87    0.000
  East         1     226.3     226.3     3.06    0.092
  South        1    2255.1    2255.1    30.51    0.000
  North        1   12330.6   12330.6   166.80    0.000
Error         25    1848.1      73.9
Total         28   14681.9
```

- **Utilize excel QM to build the ANOVA table**

1. Open the Excel QM.
2. Click on the “By chapter tab” and choose (chapter 4: regression models), then choose multiple regression for both simple or multiple regression examples.
3. Make sure to choose ANOVA from Options



Spreadsheet initialization - Multiple Regression

Title:

Sheet name:

Number of (past) periods of data: 3

Name for period: Period

(Use A for A, B, C...)

Number of independent variables(X): 1

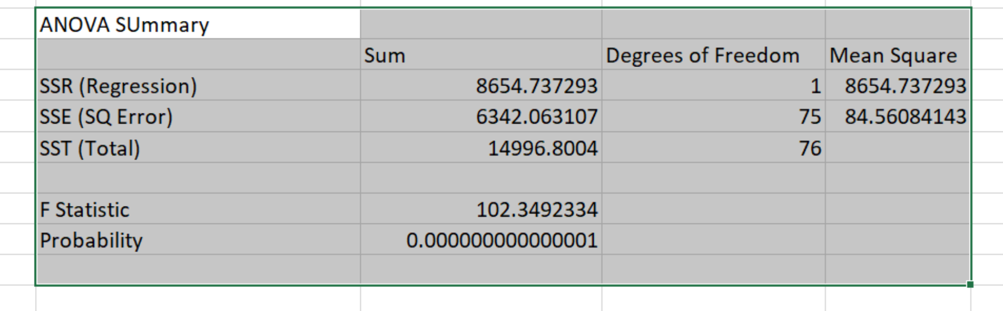
Name for x's: x

Options

☑ ANOVA

4. Enter the number of past observations and the number of independent (X) variables. You can also enter a name or title for the problem. This will initialize the size of the spreadsheet.
5. Enter the data in the shaded part under Y and X1 and the calculations will be automatically added.

6. ANOVA Table will be displayed with all its relevant data



ANOVA SUmmary			
	Sum	Degrees of Freedom	Mean Square
SSR (Regression)	8654.737293	1	8654.737293
SSE (SQ Error)	6342.063107	75	84.56084143
SST (Total)	14996.8004	76	
F Statistic	102.3492334		
Probability	0.000000000000001		

Example 2

The dataset "Cereal" contains, among other variables, the consumer reports ratings of 77 cereals available in many grocery stores and the number of grams of sugar contained in each serving. Considering "Sugars" as the explanatory variable and "Rating" as the response variable, generate the regression model and the ANOVA table. Comment on the results. (Let alpha = 0.05)

Ans.

- **Rating = 59.3 - 2.40 Sugars**

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	**1**	**8654.7**	**8654.7**	**102.35**	**0.000**
Error	**75**	**6342.1**	**84.6**		
Total	**76**	**14996.8**			

- **In the ANOVA table, the** $F_{calculated} = \frac{8654.7}{84.6} = 102.35$
- **From the F table in Appendix D in the book,** $F_{critical}\ is\ F(0.05, 1, 75) = 3.97$
- **Since** $F_{calculated} > F_{critical}$, **therefore our model is significant (there is strong evidence that $\boldsymbol{\beta}$1 is not equal to zero)**

- **Another Solution: In the ANOVA table, the** $P - value = 0$
- **Since** $P - value < \propto$, **therefore our model is significant**

- **The r² term = (0.759^2) = 0.577 indicating that 57.7% of the variability in the response variable “rating” is explained by the explanatory variable “sugar”.**