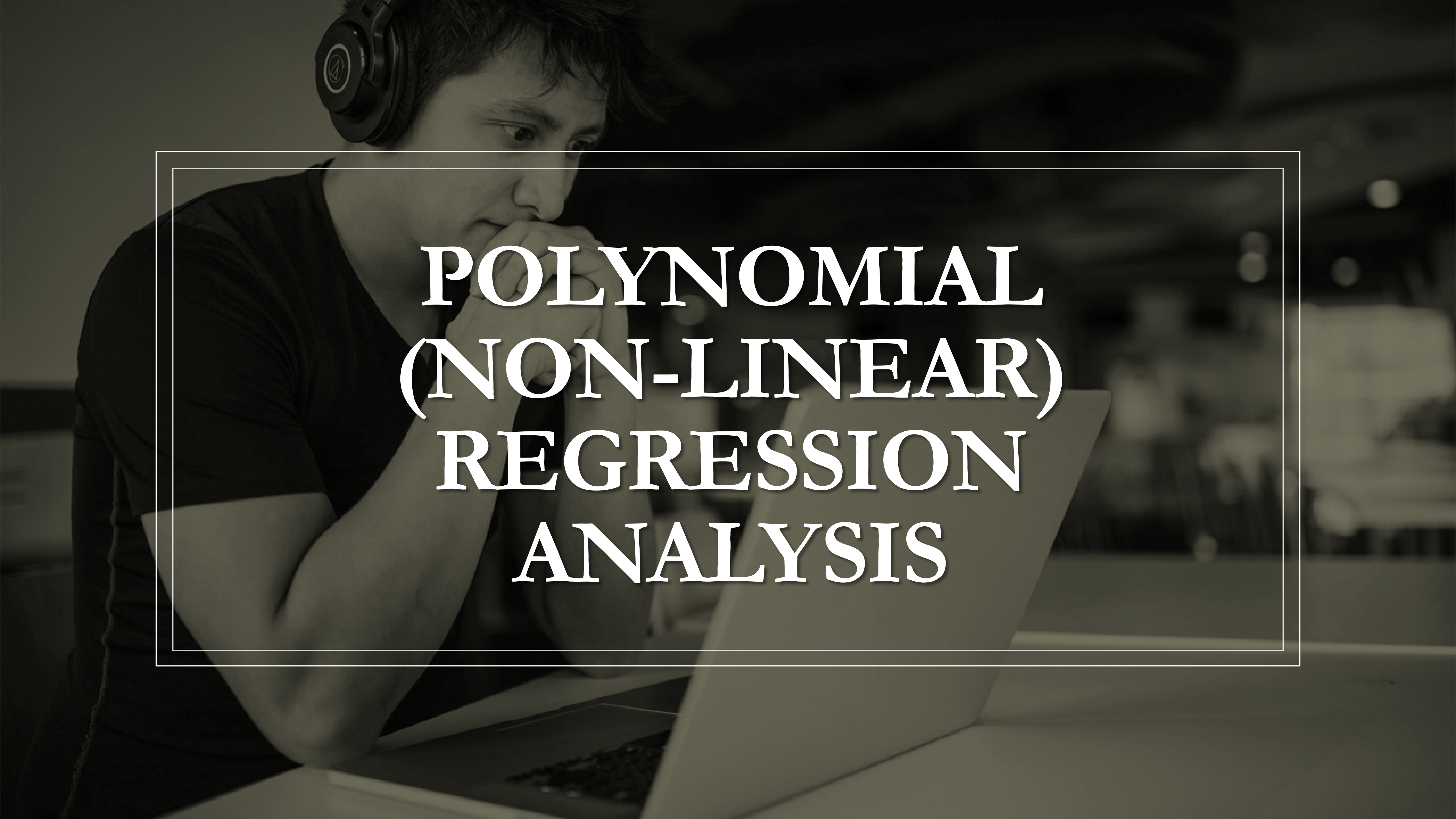
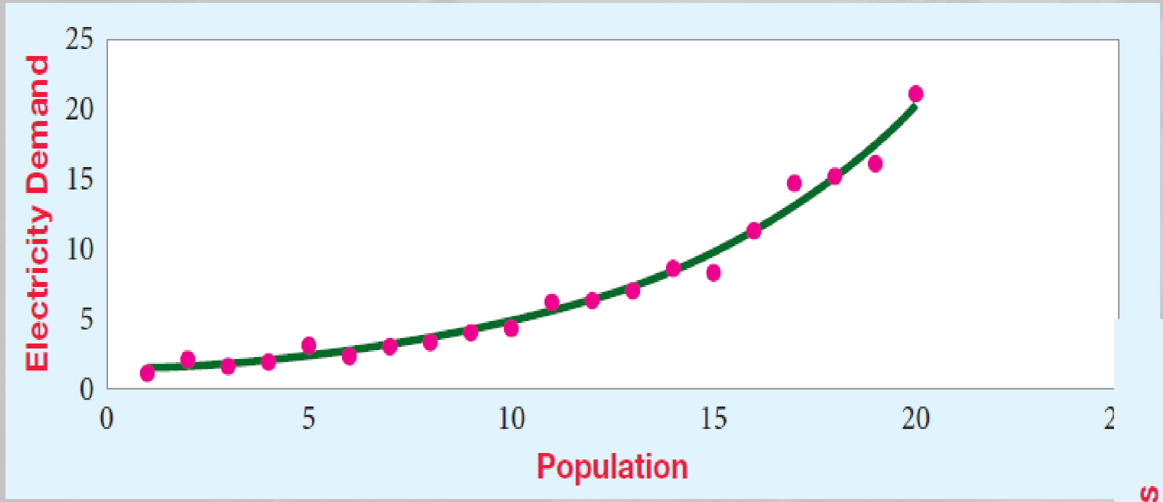
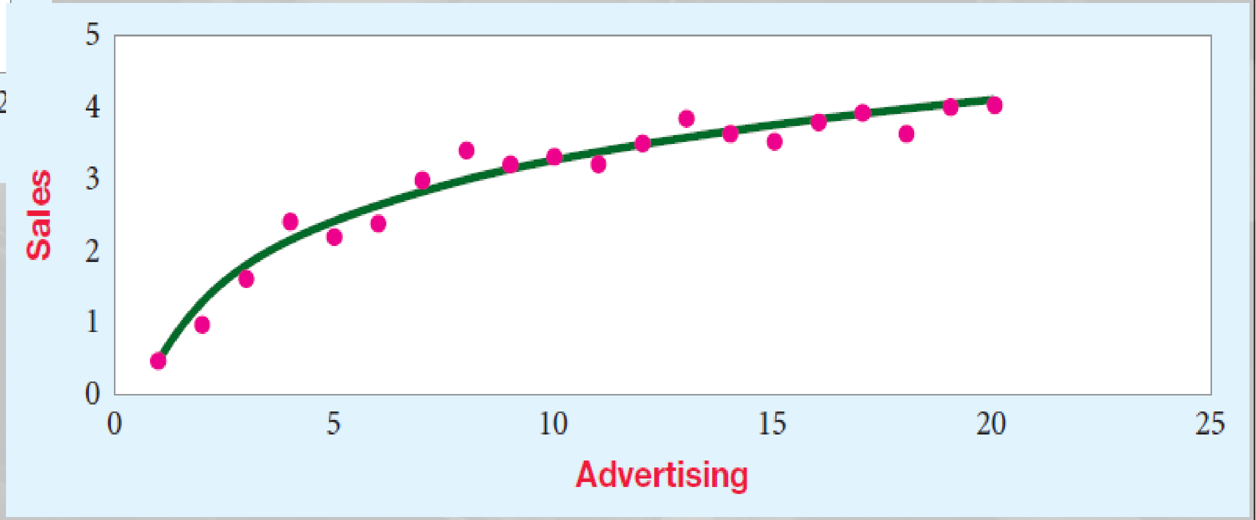


POLYNOMIAL (NON-LINEAR) REGRESSION ANALYSIS

There are also many instances in which the relationship between two variables will be curvilinear, rather than linear.





- To model such curvilinear relationships, we must incorporate terms into the multiple regression model that will create "curves" in the model we are building.

- The model which possesses the curvilinear is refer as a *polynomial model*. The general equation for a polynomial with one independent variable is given as below

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

where:

β_0 = Population regression's constant
β_j = Population's regression coefficient for variable x^j; $j = 1, 2, \ldots, p$
p = Order (or degree) of the polynomial
ε = Model error

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

where:

β_0 = Population regression's constant

β_j = Population's regression coefficient for variable x^j; $j = 1, 2, \ldots, p$

p = Order (or degree) of the polynomial

ε = Model error

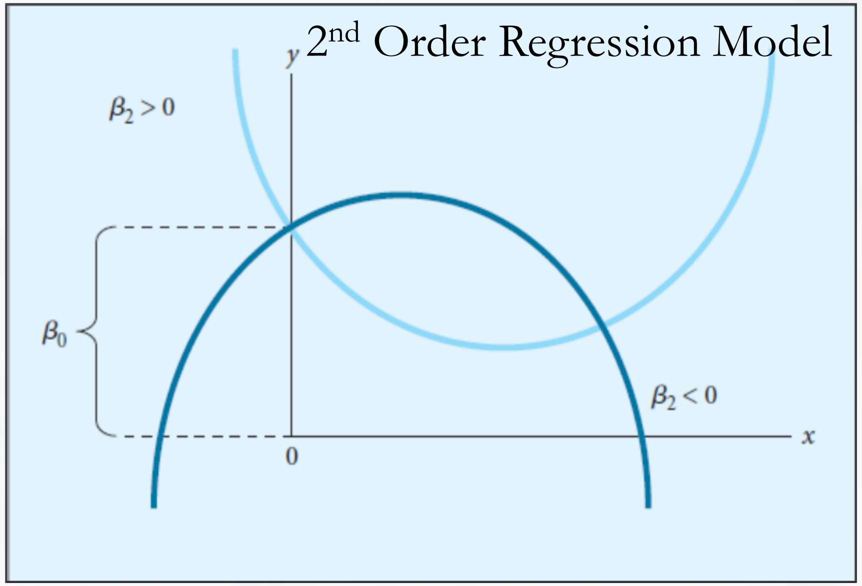
The order/degree of the model is determined by the largest exponent of the independent variable in the model.

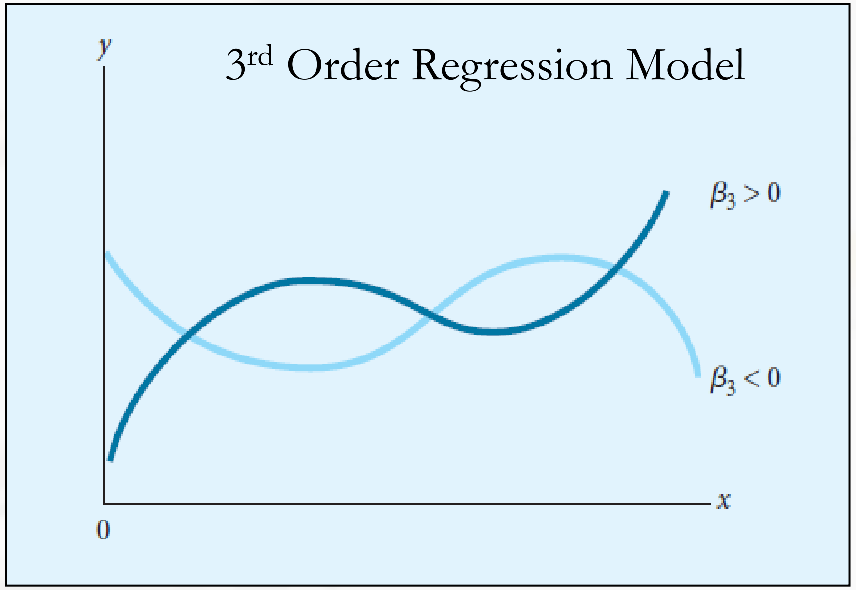
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

The model above is a second-order polynomial because the largest exponent in any term of the polynomial is 2.

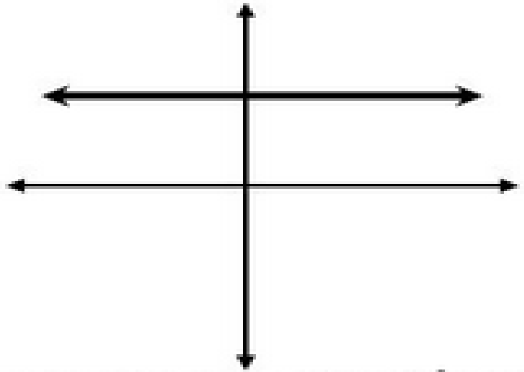
As more curves appear in the data, the order of the polynomial must be increased. Example of third-order polynomial

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

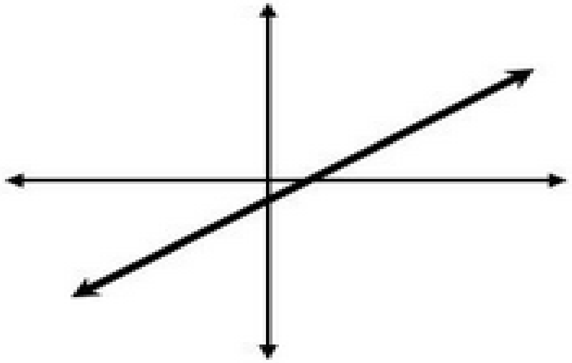

2nd Order Regression Model
y
$\beta_2 > 0$
β_0
$\beta_2 < 0$
0
x


3rd Order Regression Model
y
$\beta_3 > 0$
$\beta_3 < 0$
0
x

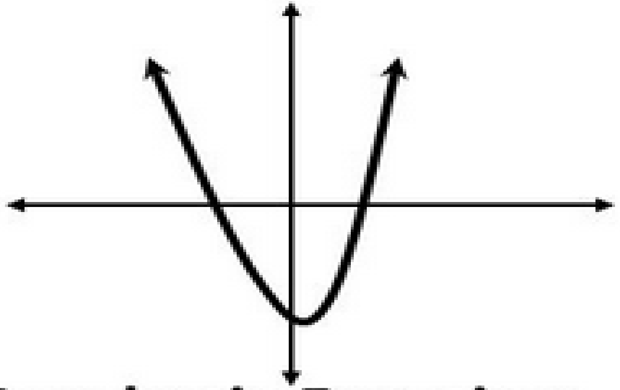
Graphs of Polynomial Functions:



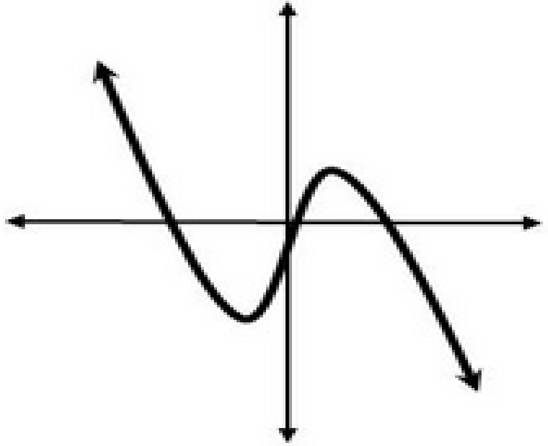
Constant Function
(degree = 0)



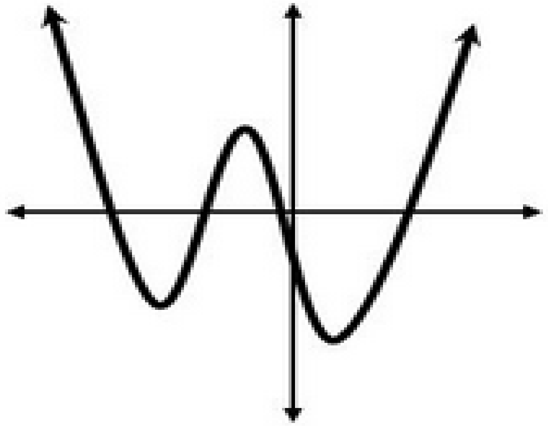
Linear Function
(degree = 1)



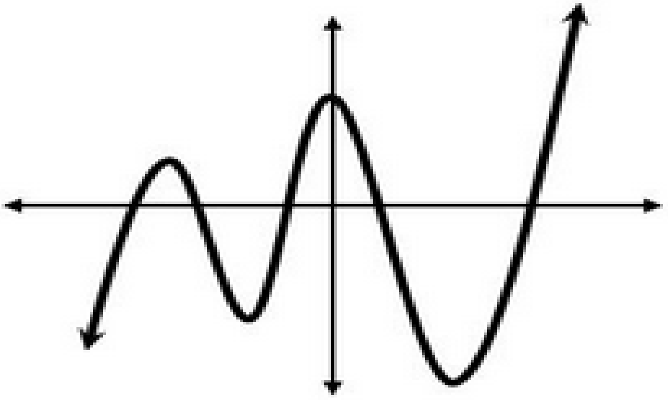
Quadratic Function
(degree = 2)



Cubic Function
(deg. = 3)



Quartic Function
(deg. = 4)



Quintic Function
(deg. = 5)

PROBLEM REQUIREMENTS

A mental health clinic wants to understand how weekly exercise duration (hours) affects patients' stress levels.

Data Set: stress_vs_exercise.xlsx

1. Fit a linear model and evaluate fit.
2. Visualize the curve to confirm the non-linear pattern.
3. Fit a polynomial model and compare R^2 and residuals.

DATA UNDERSTANDING

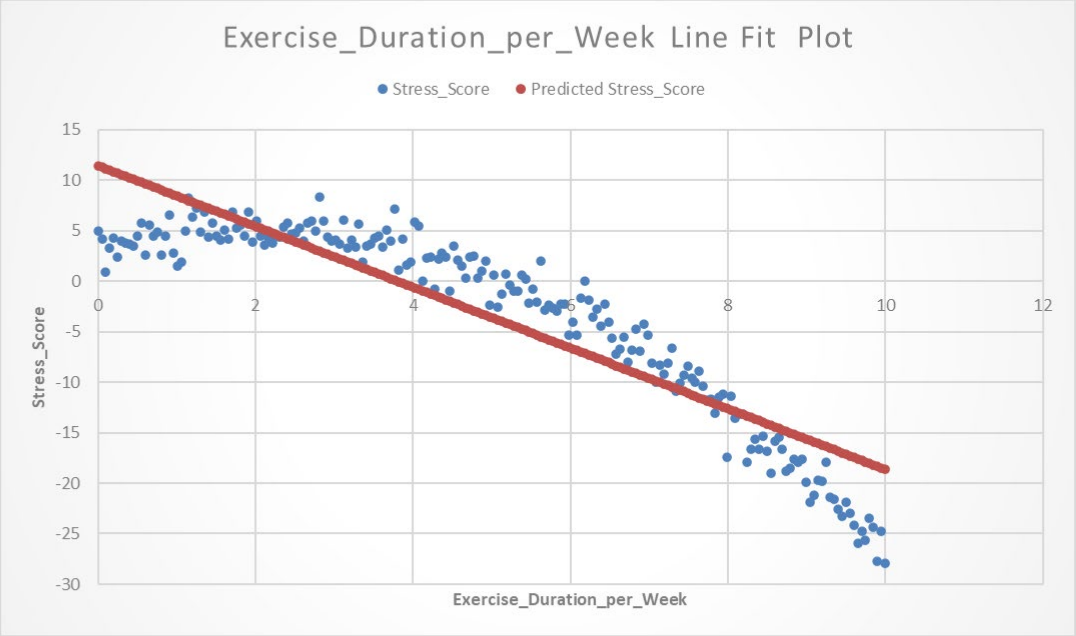
Specification:

- Variables & Correlation Analysis

Build regression model with linear regression

Diagnose Model and Improve the model

MICROSOFT EXCEL



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909302
R Square	0.82683
Adjusted R Square	0.825956
Standard Error	4.016004
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15247.47	15247.47	945.3866	2.52717E-77
Residual	198	3193.401	16.12829		
Total	199	18440.87			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11.4523	0.565826	20.23998	4.22E-50	10.33648072	12.56812
Exercise_Duration_per_Week	-3.00956	0.097881	-30.7471	2.53E-77	-3.202582673	-2.81654

REGRESSION MODEL DIAGNOSTIC

Examine Regression Prediction Model Significance

p-value = 0.000 which is <α= 0.05, therefore H0 is rejected.
Conclusion: The regression model does explain a significant proportion of the variation in sales price. Thus, the overall model is statistically significant.

Measure Model Fitness

Adjusted R^2 value, has improved to 82.6% of the variation

Examine Regression Slope Significance

P-value = 0.000 which is <α= 0.05 for ALL reg. coeff, , therefore H0 is rejected. Thus ALL variables is significance to be included in the model

Examine the Error Rate

RMSE (residual standard error) decreases = 4.02

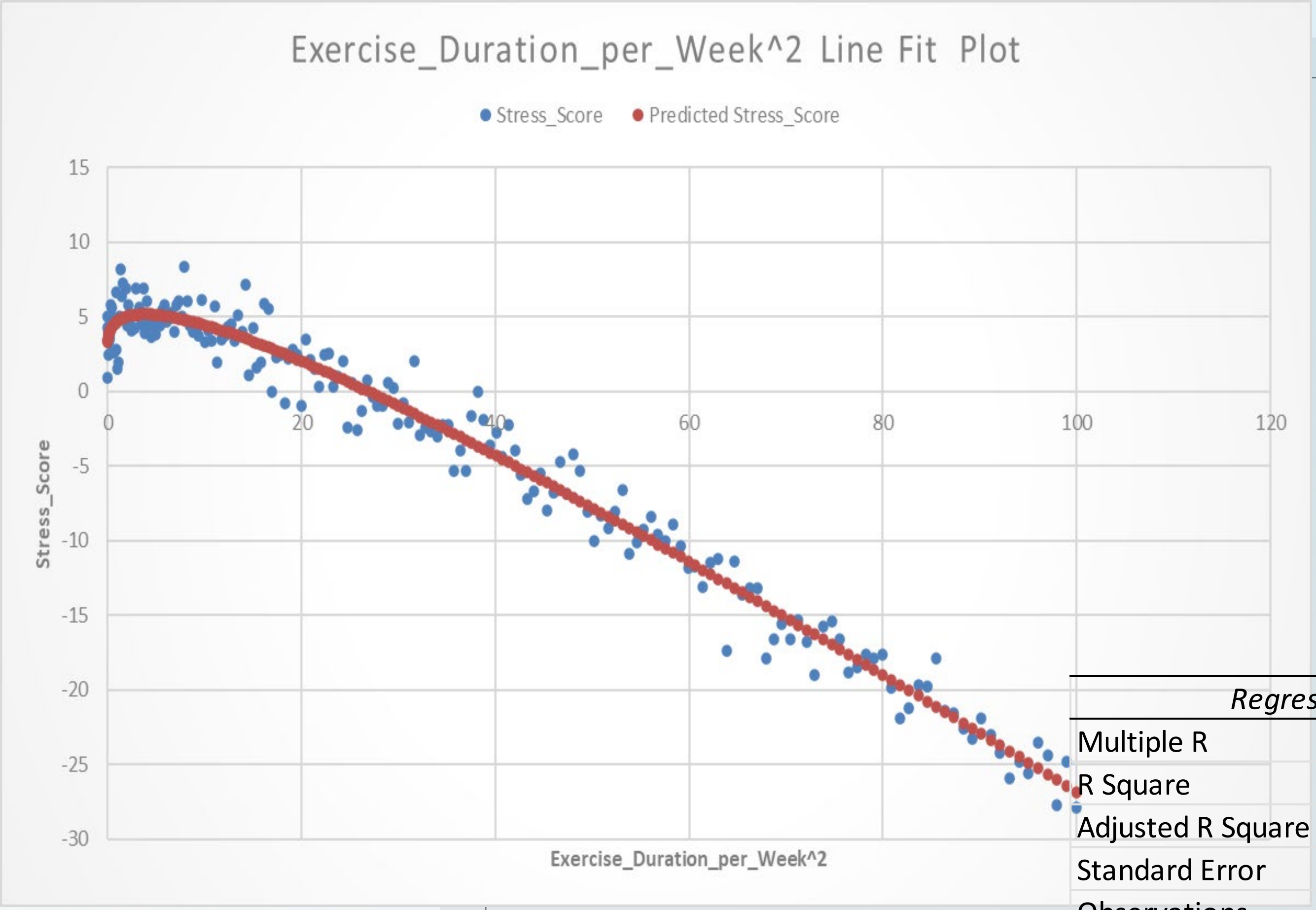
SUMMARY

Can be improved with non-linear regression. Introduce 2nd degree curve:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

IMPROVED MODEL WITH POLYNOMIAL

MICROSOFT EXCEL



Regression Statistics	
Multiple R	0.98832
R Square	0.976776
Adjusted R Square	0.97654
Standard Error	1.474437
Observations	200

ANOVA

	df	SS	MS	F	Significance F
Regression	2	18012.59524	9006.298	4142.802	1.1087E-161
Residual	197	428.2707111	2.173963		
Total	199	18440.86595			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.261746	0.309673321	10.53286	7.29E-21	2.651045932	3.872446
Exercise_Duration_per_Week	1.929592	0.143077165	13.48637	8.58E-30	1.647432065	2.211751
Exercise_Duration_per_Week^2	-0.49392	0.013849071	-35.6641	6.92E-88	-0.521226586	-0.4666

REGRESSION MODEL DIAGNOSTIC

Examine Regression Prediction Model Significance

p-value = 0.000 which is <α= 0.05, therefore H0 is rejected.
Conclusion: The regression model does explain a significant proportion of the variation in sales price. Thus, the overall model is statistically significant.

Measure Model Fitness

Adjusted R^2 value, has improved to 97.7% of the variation

Examine Regression Slope Significance

P-value = 0.000 which is <α= 0.05 for ALL reg. coeff, , therefore H0 is rejected. Thus ALL variables is significance to be included in the model

Examine the Error Rate

RMSE (residual standard error) decreases = 1.47

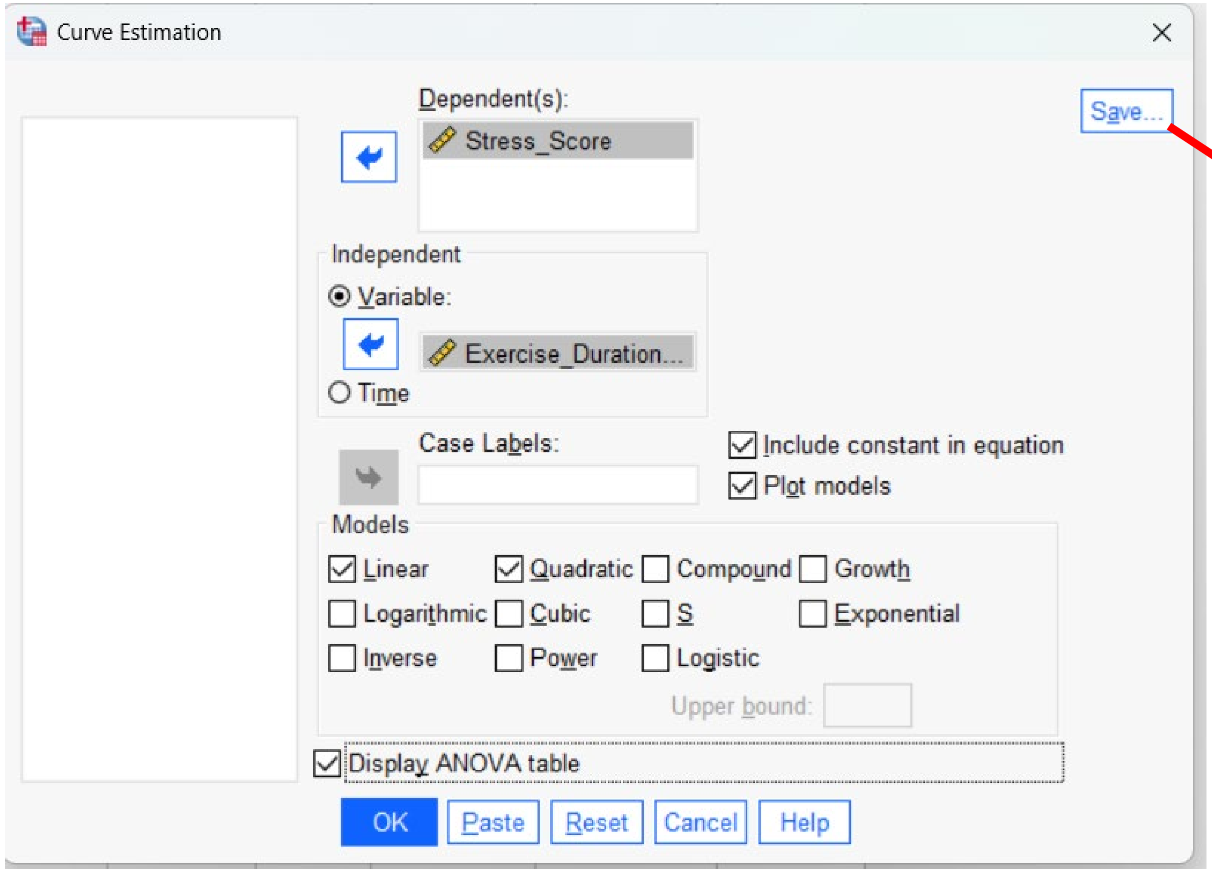
SUMMARY

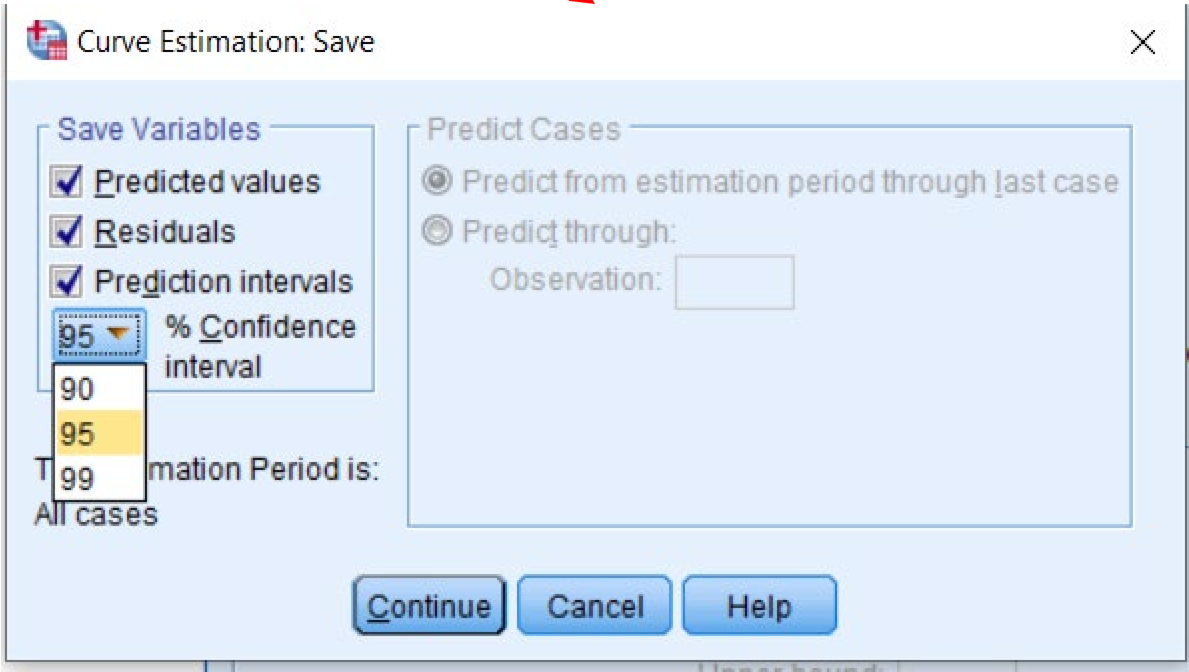
Model have improved with polynomial degree 2

IMPROVED MODEL WITH POLYNOMIAL

SPSS

Analyze -> Regression -> Curve Estimation





LINEAR OUTPUT

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.909	.827	.826	4.016

The independent variable is Exercise_Duration_per_Week.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	15247.465	1	15247.465	945.387	<.001
Residual	3193.401	198	16.128		
Total	18440.866	199			

The independent variable is Exercise_Duration_per_Week.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Exercise_Duration_per_Week	-3.010	.098	-.909	-30.747	<.001
(Constant)	11.452	.566		20.240	<.001

NON-LINEAR OUTPUT

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.988	.977	.977	1.474

The independent variable is Exercise_Duration_per_Week.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	18012.595	2	9006.298	4142.802	<.001
Residual	428.271	197	2.174		
Total	18440.866	199			

The independent variable is Exercise_Duration_per_Week.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Exercise_Duration_per_Week	1.930	.143	.583	13.486	<.001
Exercise_Duration_per_Week ** 2	-.494	.014	-1.542	-35.664	<.001
(Constant)	3.262	.310		10.533	<.001

Stress_Score
Observed
Linear
Quadratic
10.0
.0
-10.0
-20.0
-30.0
.0000
2.0000
4.0000
6.0000
8.0000
10.0000
Exercise_Duration_per_Week

EXPLORING MORE POLYNOMIAL REGRESSION

A pharmaceutical researcher aims to optimize medication dosage for maximum effectiveness. Preliminary trials indicate that both very low and very high doses reduce effectiveness.

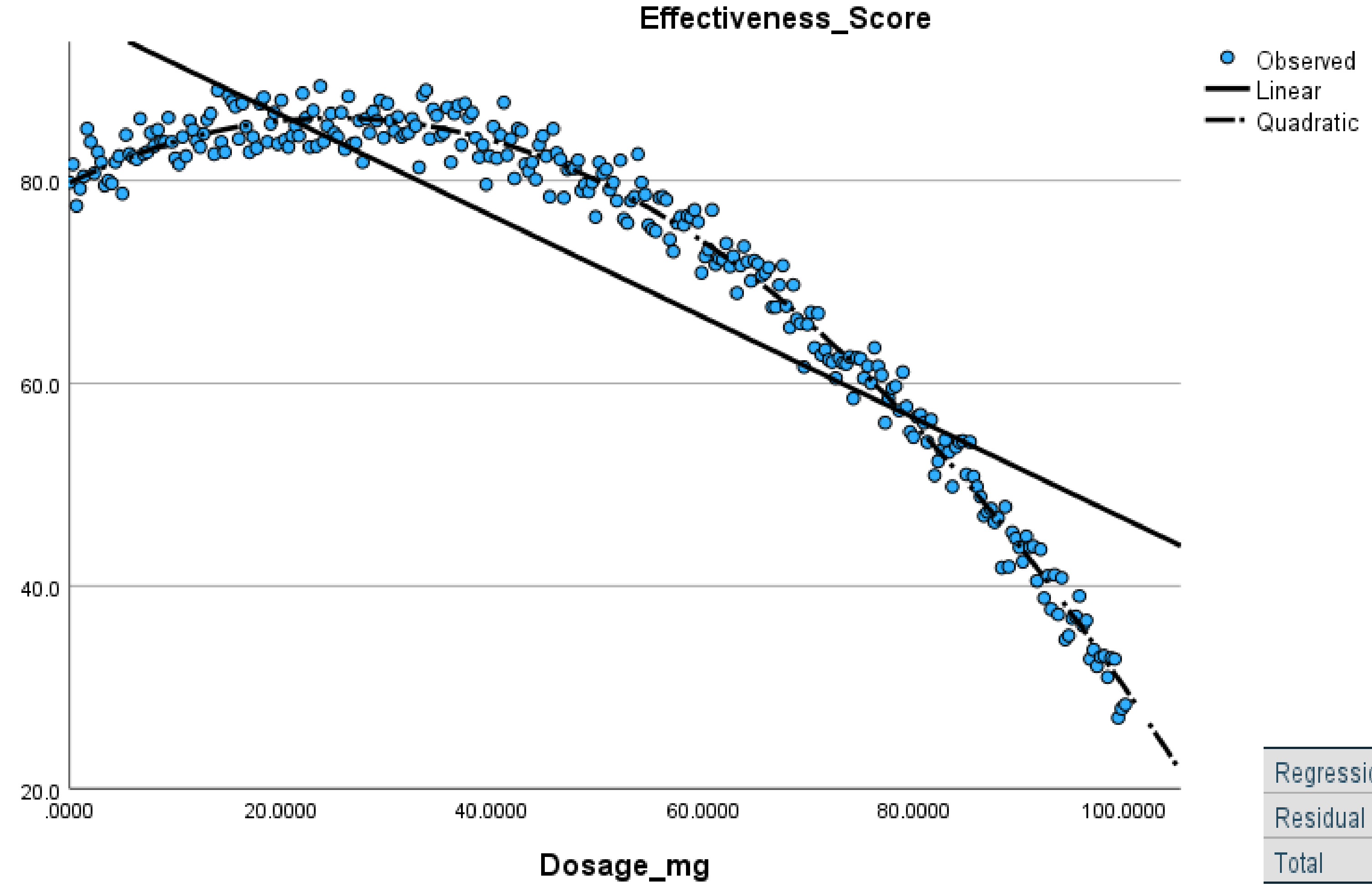
Data Set: dosage_vs_effectiveness.xlsx

Dataset Variables:

- **Dosage_mg**: Medication dose (0 to 100 mg)
- **Effectiveness_Score**: Patient response score (higher is better, scale 0–100)

What is the optimal dosage of a new medication that maximizes patient **effectiveness score**, and how does the dosage level influence overall treatment effectiveness?

ANSWER



Optimal dosage derived from the **vertex formula of a parabola** in a **quadratic regression model** before the curve declines due to overmedication or adverse effects.

$$\text{Optimal Dosage} = -\frac{\beta_1}{2 \cdot \beta_2}$$

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.993	.986	.986	1.920

The independent variable is Dosage_mg.

$$\text{Effectiveness\_Score} = \beta_0 + \beta_1 \cdot \text{Dosage\_mg} + \beta_2 \cdot \text{Dosage\_mg}^2$$

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	79735.637	2	39867.818	10817.209	<.001
Residual	1094.621	297	3.686		
Total	80830.258	299			

The independent variable is Dosage_mg.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Dosage_mg	.507	.015	.895	33.255	<.001
Dosage_mg ** 2	-.010	.000	-1.834	-68.124	<.001
(Constant)	79.756	.330		241.456	<.001

THANK YOU!