Data Analytics for Business
Nonlinear Transformation Models

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

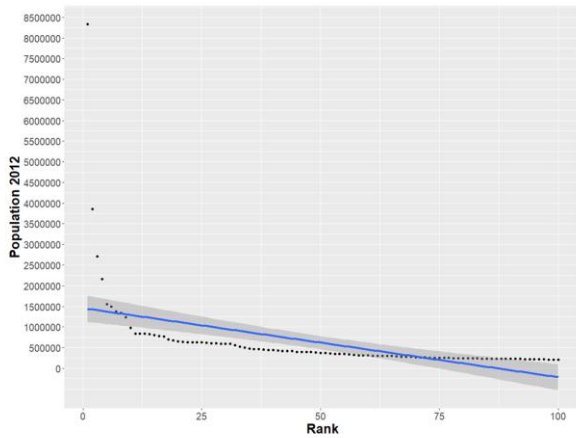
Introduction to Nonlinear Models

Lessons in this Module

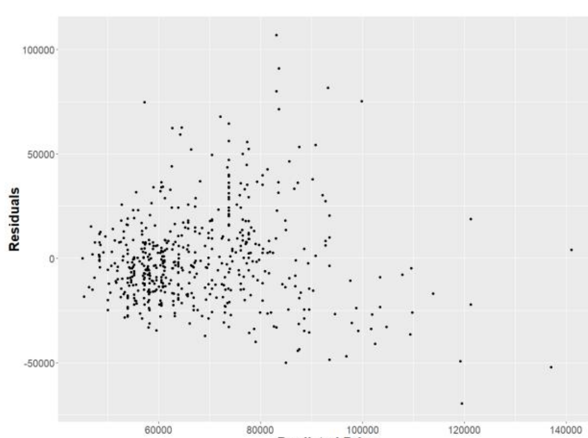
A. Introduction to Nonlinear Models
B. Linear-Log Model
C. Log-Linear Model
D. Log-Log Model
E. Polynomial Model

An Example of a Nonlinear Relationship

US City Population and Rank



Another Example (From the Housing Dataset in the Ecdat Package in R)



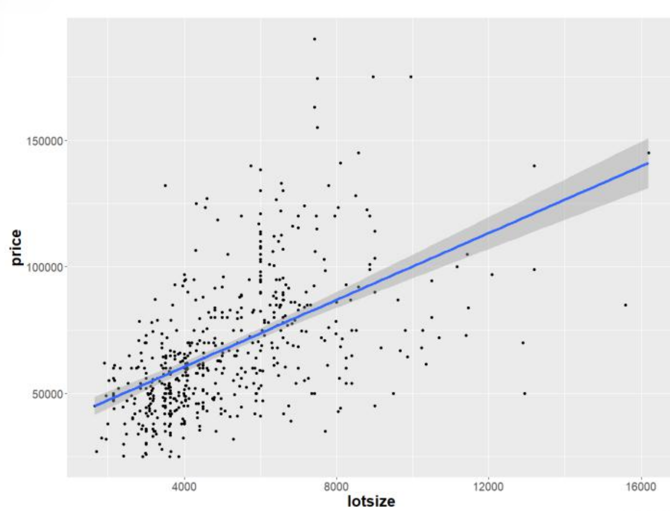
Model A: Linear-Linear Model

	Estimate	S.E.	t Value	Pr>|t|
Intercept	34140	2,491	13.7	<.0001
lotsize	6.599	0.4458	14.8	<.0001

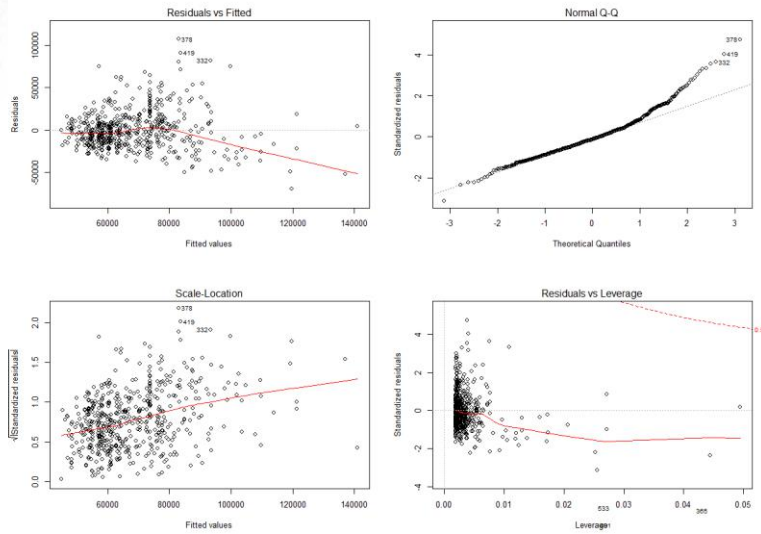
R-Squared	Adjusted R-Squared
0.2871	0.2858

- Model A: $price = b_0 + b_1 * lotsize$
- As X (lotsize) increases by 1 unit, Y (price) changes by b_1 (6.599) units, holding all other factors constant

Model A: Scatter Plot



Model A: Diagnostics Plot



Model A: $price = b_0 + b_1{*}lotsize$ Assumptions

- We need to check if the assumptions for a linear regression model hold
- The residuals vs. fitted values plot indicates that Model A has heteroscedasticity (non-constant variance); the QQ-plot suggest that there is non-linearity
- Hence, we need to start exploring non-linear models
- We focus on natural log transformations of the variables because they are easier to interpret. Note that in R, log() computes the natural log

The Various Log Transformations

	Y	log(*Y*)
X	***Model A: Level-level model*** $Y = b_0 + b_1*X$	
log(*x*)		

The Various Log Transformations

	Y	log(*Y*)
X	*Model A: Level-level model* $Y = b_0 + b_1*X$	
log(*x*)	***Model B: Linear-Log Model*** $Y = b_0 + b_1*log(X)$	

The Various Log Transformations

	Y	log(*Y*)
X	*Model A: Level-level model* $Y = b_0 + b_1 * X$	***Model C: Log-linear Model*** $\log(Y) = b_0 + b_1 * X$
log(*x*)	*Model B: Linear-Log Model* $Y = b_0 + b_1 * \log(X)$	

The Various Log Transformations

	Y	log(*Y*)
X	*Model A: Level-level model* $Y = b_0 + b_1 * X$	*Model C: Log-linear Model* $\log(Y) = b_0 + b_1 * X$
log(*x*)	*Model B: Linear-Log Model* $Y = b_0 + b_1 * \log(X)$	***Model D: Log-Log*** $\log(Y) = b_0 + b_1 * \log(X)$

Note:
- The log() function in R computes the natural logarithm
- If a variable *x* has some values = 0, then use log(*x*+1) transformation

Data Analytics for Business

Nonlinear Transformation Models

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Linear-Log Model

Model B: Linear-Log Model Independent Variable Transformed

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-250,728	20,184	-12.42	<.0001
Ln_lotsize	37660	2381	15.81	<.0001

R-Squared	Adjusted R-Squared
0.315	0.3137

- Model B: $price = b_0 + b_1 * log(lotsize)$
- Create a new variable Ln_lotsize which is the natural log of lotsize
- Run Model B using the Housing dataset
- How would you interpret the coefficient of Ln_lotsize?

Model B: $price = b_0 + b_1*log(lotsize)$

What does the coefficient $b_1 = 37660$ imply?

A. When lotsize increases by 1 sq foot, price increases (on average) by $37,660.
B. When lotsize decreases by 1 sq foot, price increases (on average) by $37,660.
C. When lotsize decreases by log(1), price increases (on average) by $376.60.
D. When lotsize increases by 1%, price increases (on average) by $376.60.

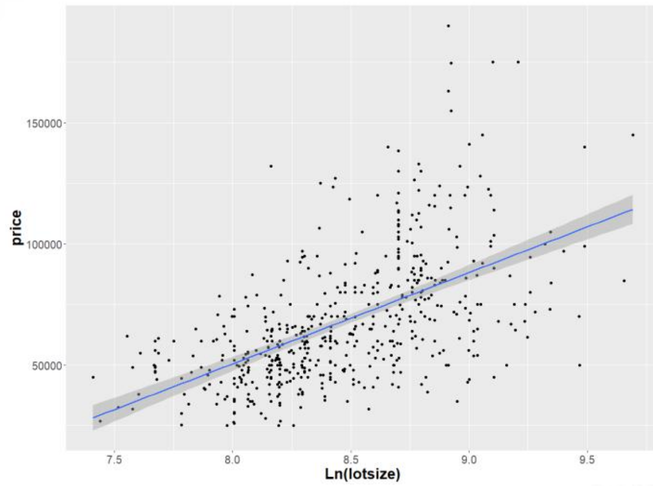
What is the correct answer?

D. *A one percent increase in the independent variable increases (or decreases) the dependent variable by (coefficient/100) units*.

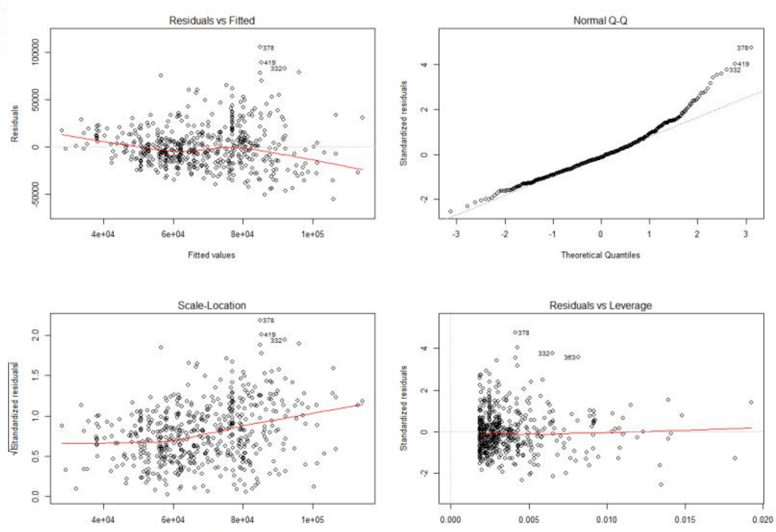
Interpreting a Linear-Log Model

- Linear-log model (and most other models) needs to be interpreted carefully
- It does not make much practical sense to increase the “log(lotprice)” by one unit
- But increasing X by 1 percent is almost equivalent to increasing (natural) $log(X)$ by 0.01 units
- Hence, a 1 percent increase in X increases (natural) $log(X)$ by .01 and, therefore, changes the Y variable by $.01*b_1$

Model B: Scatter Plot

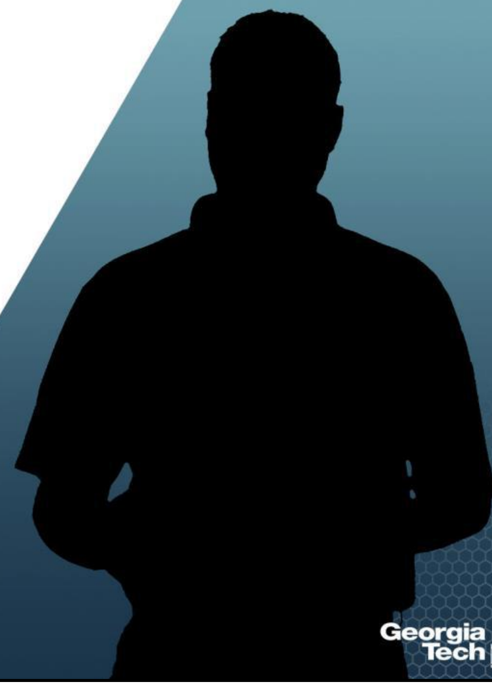


Model B: Diagnostics Plots



Data Analytics for Business

Nonlinear Transformation Models



Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Log-Linear Model

Model C: Log-Linear Model Dependent Variable Transformed

	Estimate	S.E.	t Value	Pr>|t|
Intercept	10.58	.03451	306.51	<.0001
lotsize	.00009315	.000006177	15.08	<.0001

R-Squared	Adjusted R-Squared
0.2947	0.2935

- Model C: $log(price) = b_0 + b_1*lotsize$
- Create a new variable Ln_price which is the natural log of price
- How would you interpret the coefficient of lotsize?

Model C: *log(price)* = b_0 + b_1*lotsize*

What does the coefficient b_1 = .00009315 imply in a log-linear model?

A. When lotsize increases by .01 sq. ft., price increases by .009315% (on average)
B. When lotsize increases by 1 sq. ft., price increases by .009315% (on average)
C. When lotsize increases by 1 sq. ft., price decreases by .009315% (on average)
D. When lotsize decreases by 1 sq. ft., price increases by .009315% (on average)

What is the correct answer?

B. ***The dependent variable changes by 100*(coefficient) percent for a one unit increase in the independent variable while all other variables in the model are held constant***

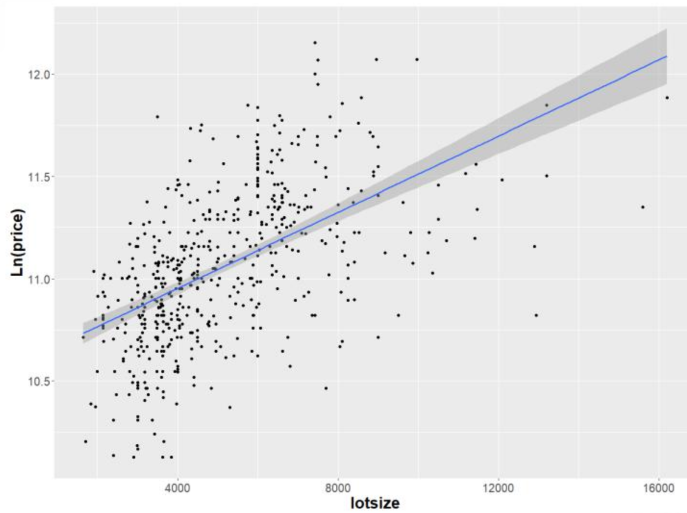
Georgia Tech

Interpreting the Log-linear Model *log(price)* = b_0 + b_1*lotsize*

- Increasing *x* by one unit will increase (natural) *log(y)* by b_1 units
- With x = lotsize, the model *log(price)* = b_0 + b_1*x is the same as $y = e^{(b_0+b_1 x)}$. Hence, *dy/dx* = $b_1 y$, or *dy/y* = b_1*dx*
- Multiplying both sides by 100, we get 100*dy/y* = 100*b_1*dx*
- Note that (100*dy/y) is the percentage change in Y
- If *dx* = 1, then this one unit change in *x* leads to a 100*b_1 percentage change in Y
- Note: this interpretation works when b_0 + b_1*x is very small
  The accurate percentage change in Y = (e^{b_1} - 1)*100 for a one unit change in X

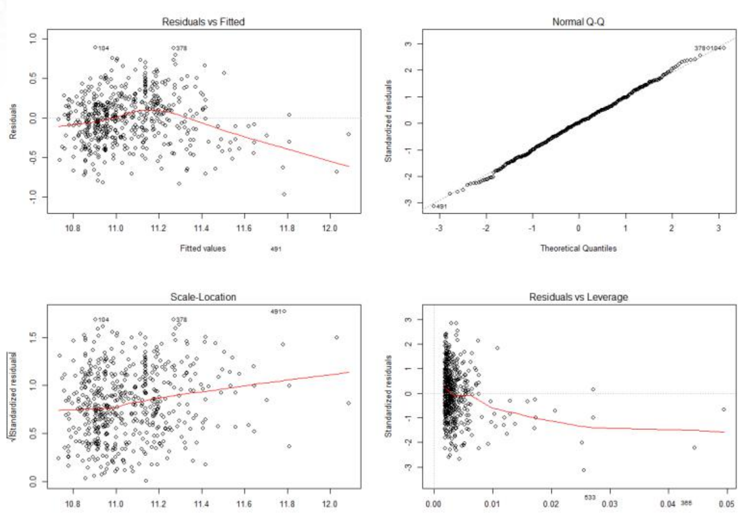
Georgia Tech

Model C: Scatter Plot



Georgia Tech

Model C: Diagnostic Plots

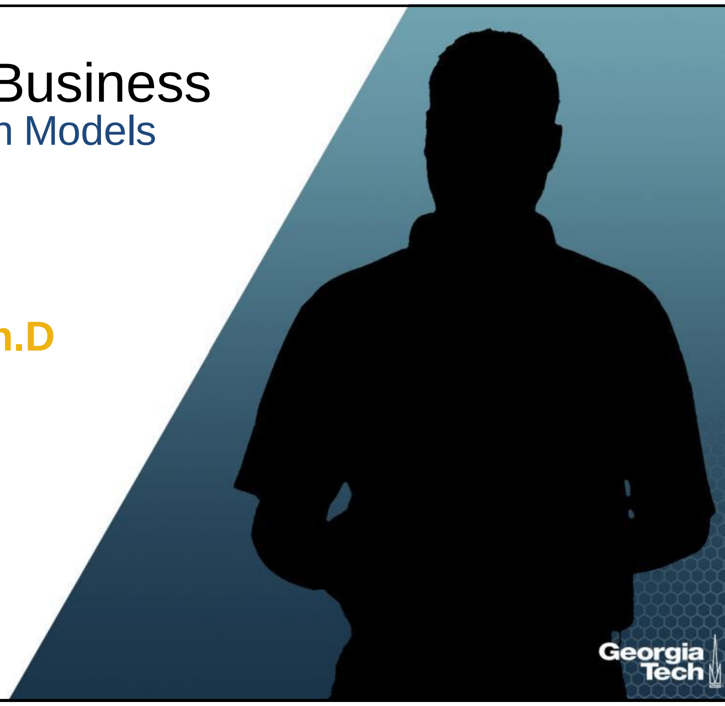


Georgia Tech

Data Analytics for Business

Nonlinear Transformation Models

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business



Log-Log Model

Model D: Log-Log Model

	Estimate	S.E.	t Value	Pr>|t|
Intercept	6.46853	0.27674	23.37	<.001
lotsize	0.54218	0.03265	16.61	<.001

R-Squared	Adjusted R-Squared
0.3364	0.3352

- Model D: *log(price)* = $b_0 + b_1$**log(lotsize)*
- The dependent and the independent variables are log transformed
- How do you interpret the coefficient of lotsize?

Model D: Log-Log Model
*log(price) = b_0 + b_1*log(lotsize)*

b_1 = 0.54218 implies that

A. When lotsize increases by 1%, price increases (on average) by 0.54218%

B. When lotsize decreases by 1%, price increases (on average) by 0.54218%

C. When lotsize increases by 1 sq. ft., price increases (on average) by 0.54218%

D. When lotsize decreases by 1 sq. ft., price increases (on average) by 0.54218%

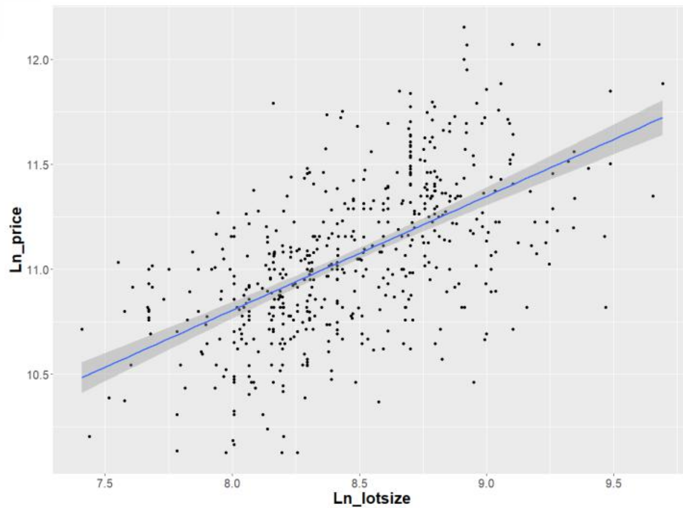
What is the correct answer?

Answer -- A. ***The dependent variable changes by b_1% percent for a one percent increase in the independent variable while all other variables in the model are held constant***

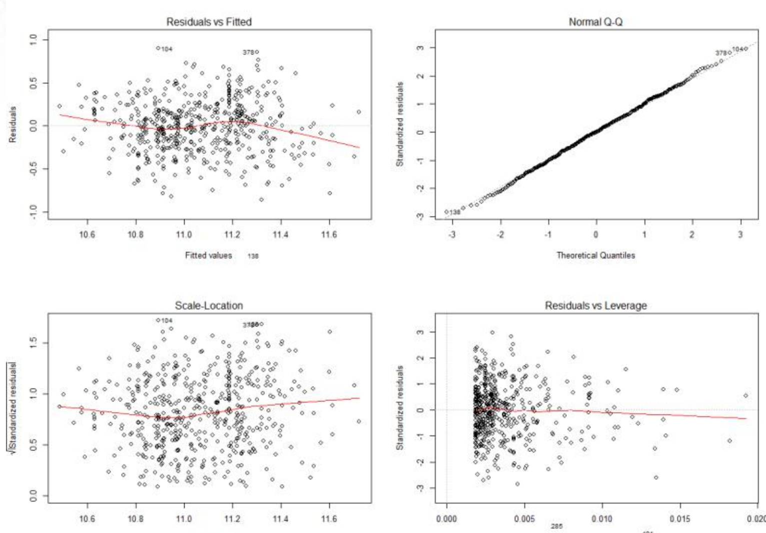
Interpreting the Log-Log Model

- Increasing (natural) log(*X*) by 0.01 leads to increasing (natural) log(*Y*) by b_1 *0.01 units
- Increasing (natural) log(*X*) by 0.01 is almost equivalent to increasing *X* by 1 percent, which implies changing *Y* by b_1 percent
- In a regression setting, we'd interpret elasticity as the percent change in *Y* (the dependent variable), when *X* (the independent variable) increases by one percent
- Hence b_1 captures elasticity

Model D: Scatter Plot



Model D: Diagnostics Plot



Comparing the Models

	R-Squared	Adjusted R-Squared
Model A: Level-Level	0.2871	0.2858
Model B: Linear-Log	0.3150	0.3137
Model C: Log-Linear	0.2947	0.2935
Model D: Log-Log	0.3364	0.3352

Reasons for (log) Transforming Data

- To achieve a (more) linear relationship
- To make a distribution more normal
- To make the variance more constant
- To get a better fit in the model – i.e., increase R-Squared

Log Transformations Cheat Sheet

	Y	log(Y)
X	**Model A: Level-Level model** $Y = b_0 + b_1*X$ As X increases by 1 unit, Y changes by b_1 units, holding all other factors constant.	**Model C: Log-Linear Model** $log(Y) = b_0 + b_1*X$ As X increases by 1 unit, Y increases by (b_1*100)%, holding all other factors constant.
log(x)	**Model B: Linear-Log Model** $Y = b_0 + b_1*log(X)$ As X increases by 1%, Y increases by (b_1/100) units, holding all other factors constant.	**Model D: Log-Log** $log(Y) = b_0 + b_1*log(X)$ As X increases by 1%, Y changes by b_1%, holding all other factors constant.

Georgia Tech

Data Analytics for Business

Nonlinear Transformation Models

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

Polynomial (Quadratic) Model

Georgia Tech

Model E: Polynomial (Quadratic) Model

	Estimate	S.E.	t Value	Pr>|t|
Intercept	11,340	4,892	2.317	.0209*
lotsize	14.81	1.589	9.317	<.0001
lot_square	-6.238e-04	1.162e-04	-5.370	<.0001

R-Squared	Adjusted R-Squared
0.323	0.3205

- Model E: $price = b_0 + b_1{*}lotsize + b_2{*}lotsize^2$
- Create a new variable lot_square = $lotsize^2$
- Fit the model, $price = b_0 + b_1{*}lotsize + b_2{*}lotsize^2$

Model E: $price = b_0 + b_1{*}lotsize + b_2{*}lotsize^2$

The coefficient $b_1 = 14.81$ implies that

A. When lotsize increases by 1 sq. ft., price will increase by $14.81

B. When lotsize decreases by 1 sq. ft., price will increase by $14.81

C. When lotsize increases by 1%, price will increase by $14.81

D. None of the above

What is the correct answer?

D: None of the above

Model E: Interpreting a Polynomial (Quadratic) Model

price = b_0 + b_1*lotsize + b_2*lotsize^2

- Coefficients b_1 and b_2 cannot be interpreted individually because when *lotsize* is increased by 1 unit, it is not possible (or meaningful) to hold *lotsize*^{2} constant
- A quadratic model does not allow for <u>an isolated interpretation of coefficients</u> since $\frac{d(price)}{d(lotsize)} = b_1 + 2b_2 * lotsize$
- This means that the slope (impact of 1 unit increase in x) is not a constant. It changes at every point of the quadratic curve (if you plot y vs. x)

Georgia Tech

Recap of this Module

A. Introduction to Nonlinear Models
B. Linear-Log Model
C. Log-Linear Model
D. Log-Log Model
E. Polynomial Model

Georgia Tech