Introduction to
Data Science and Analytics

Stephan Sorger
www.StephanSorger.com

Unit 4. Excel Regression

Disclaimer:

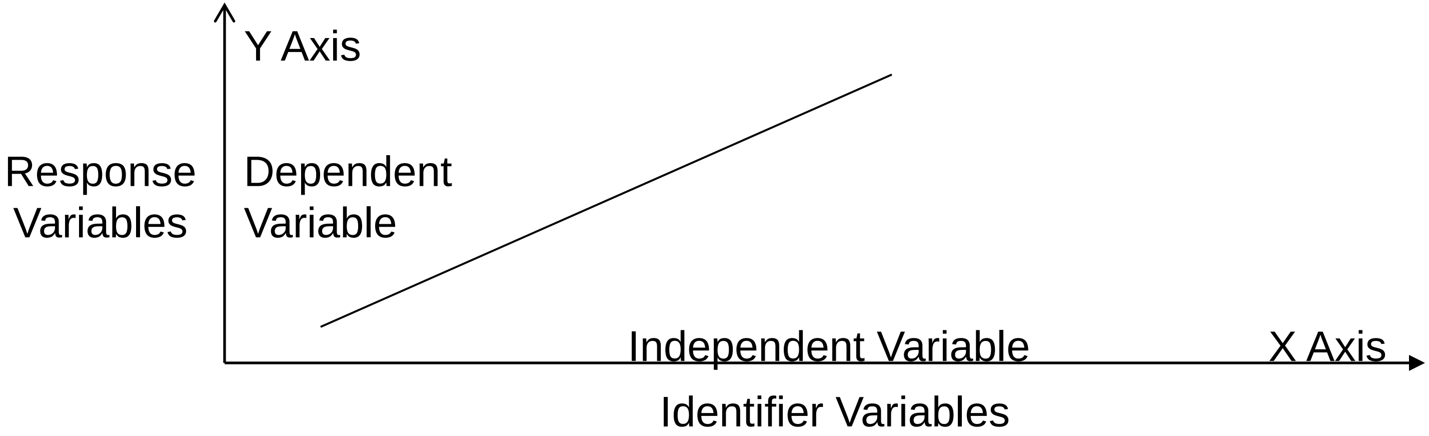
- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, "Marketing Analytics: Strategic Models and Metrics"

Outline/ Learning Objectives

Topic	Description
Background	The goal of regression analysis
Statistics	Basic statistics governing regression performance
Tests	F tests, t tests, p tests
Procedure	Executing regression analysis in Microsoft Excel
Multivariate	Executing cases with two or more independent variables

Regression Analysis

Goal is to establish the relationship between
Independent variables (the “inputs”)
and dependent variables (also called response variables)



Y Axis

Response Variables

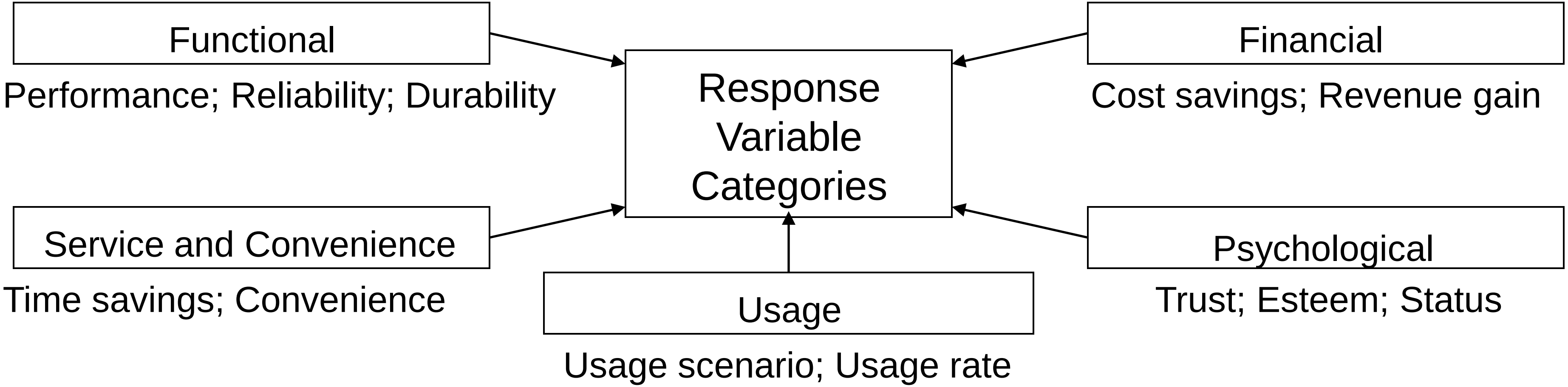
Dependent Variable

Independent Variable

X Axis

Identifier Variables

Response (Dependent) Variable Categories



Response Variable Categories

Functional

Performance; Reliability; Durability

Financial

Cost savings; Revenue gain

Service and Convenience

Time savings; Convenience

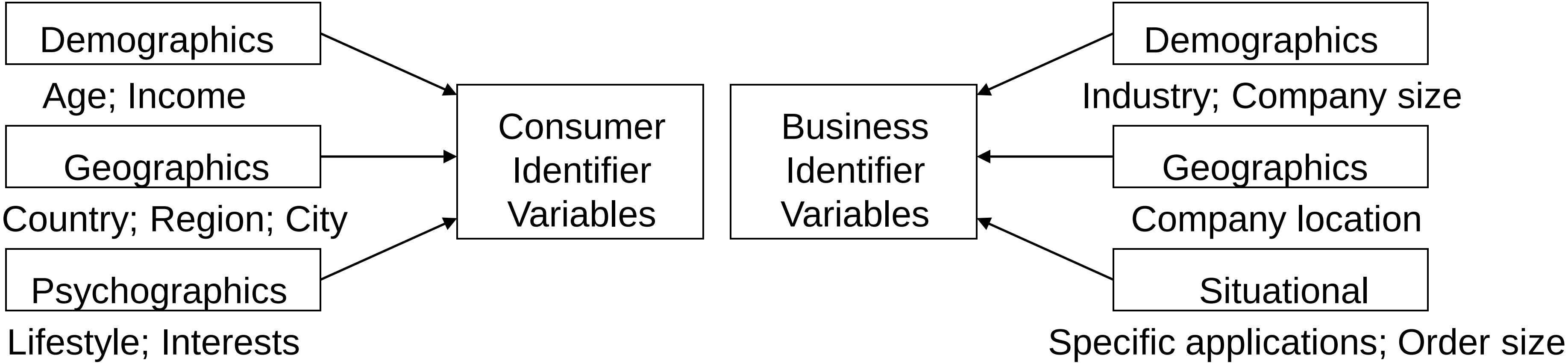
Usage

Usage scenario; Usage rate

Psychological

Trust; Esteem; Status

Independent (Input) Variables



Demographics

Age; Income

Geographics

Country; Region; City

Psychographics

Lifestyle; Interests

Consumer Identifier Variables

Business Identifier Variables

Demographics

Industry; Company size

Geographics

Company location

Situational

Specific applications; Order size

Many other independent variables possible: See next slide

Regression Example

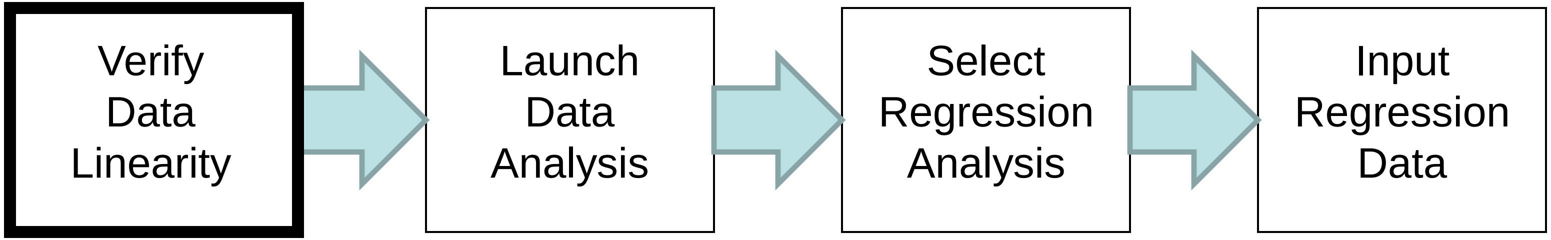
Scenario: Moving into a New Apartment
(regular apt: not mansion; not rent control)

Reponse Variable: S.F. Monthly Rent Paid

Independent Variables:
(want to predict how much people will pay)

Demographics: Age
Demographics: Income
Geographics: Location of workplace
Psychographics: Status required
Psychographics: Entertaining requirements

Regression Analysis: Process

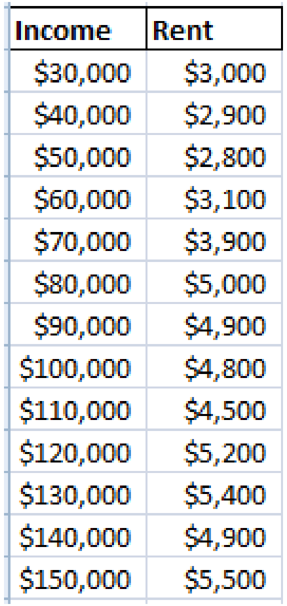


Verify Data Linearity

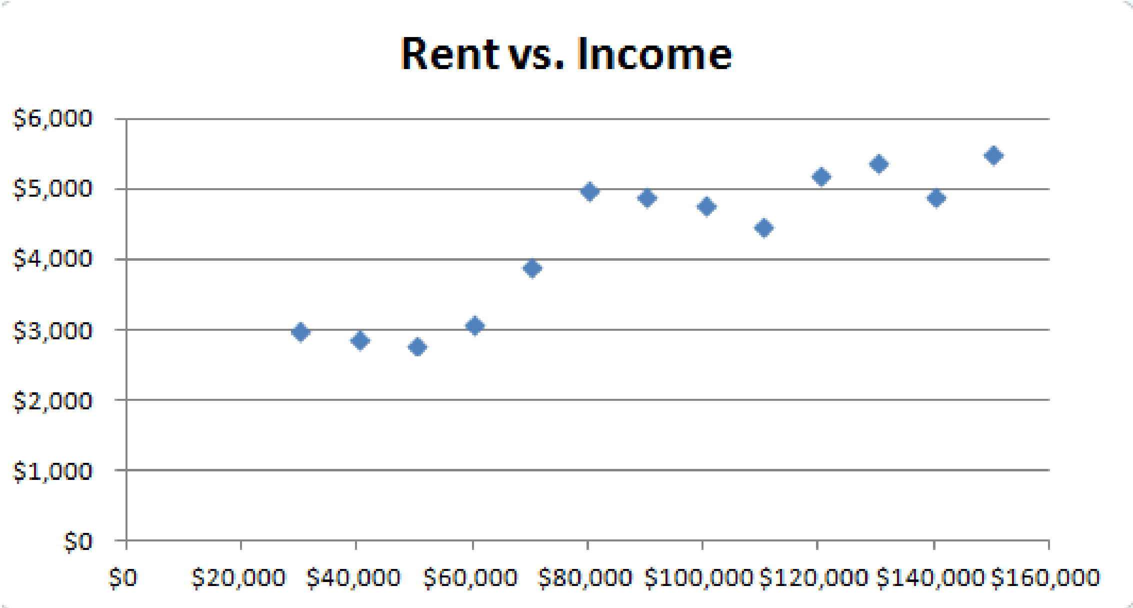
Launch Data Analysis

Select Regression Analysis

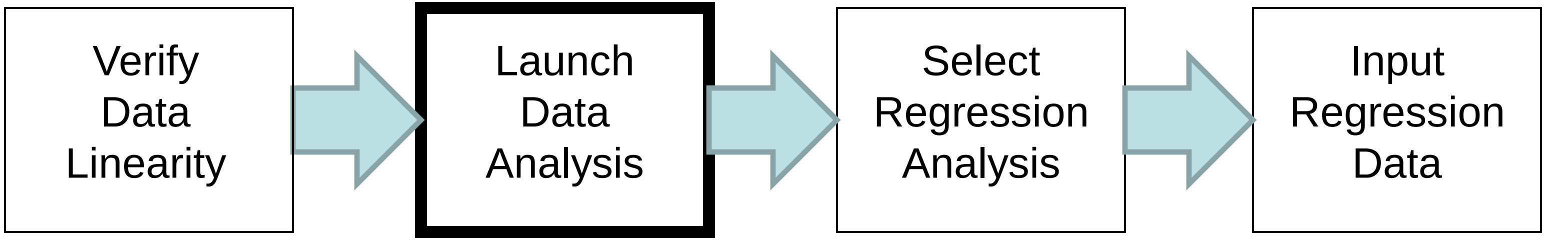
Input Regression Data



Income	Rent
$30,000	$3,000
$40,000	$2,900
$50,000	$2,800
$60,000	$3,100
$70,000	$3,900
$80,000	$5,000
$90,000	$4,900
$100,000	$4,800
$110,000	$4,500
$120,000	$5,200
$130,000	$5,400
$140,000	$4,900
$150,000	$5,500



Regression Analysis: Process

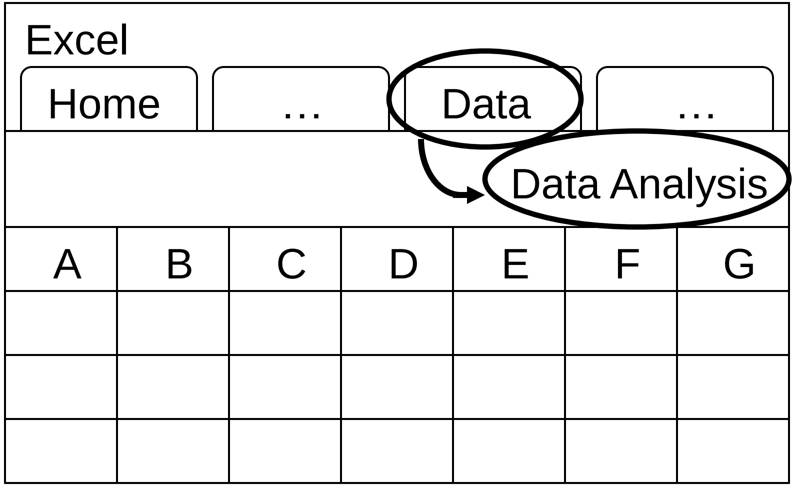


Verify Data Linearity

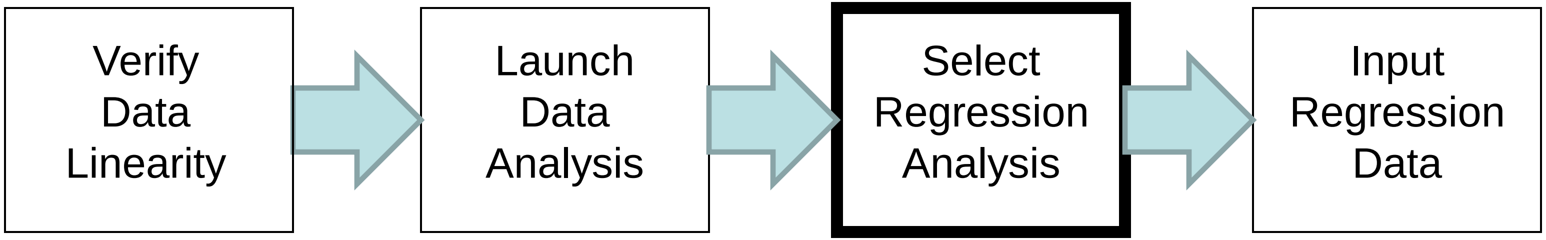
Launch Data Analysis

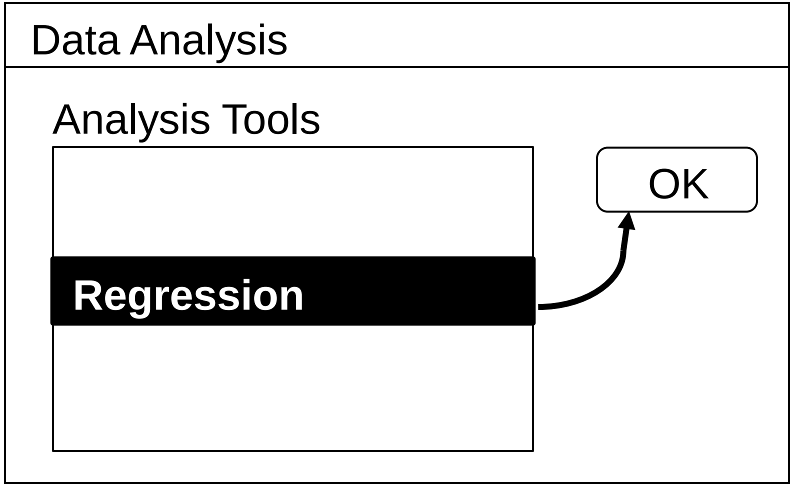
Select Regression Analysis

Input Regression Data

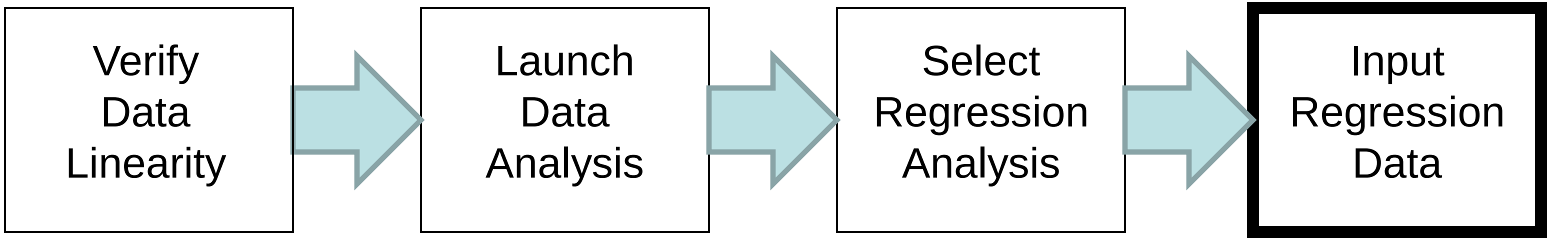


Regression Analysis: Process

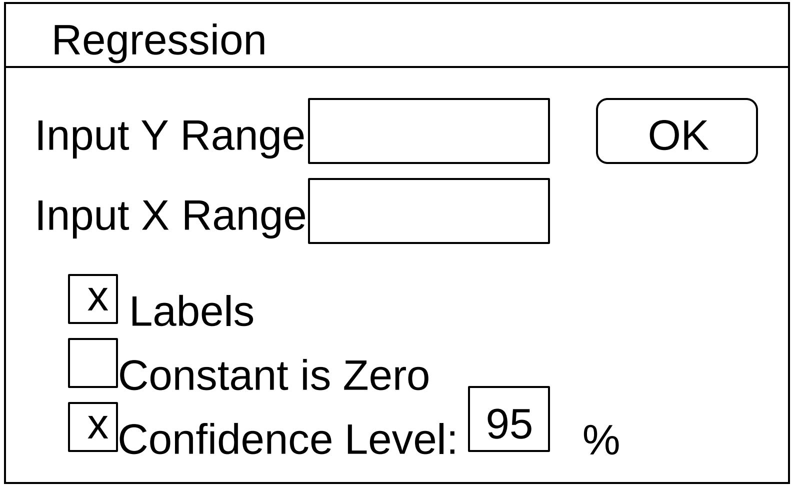




Regression Analysis: Process



Verify Data Linearity → Launch Data Analysis → Select Regression Analysis → Input Regression Data



Regression

Input Y Range

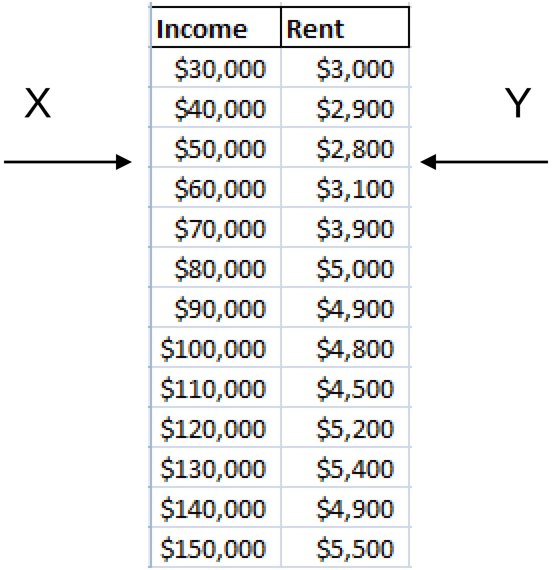
Input X Range

OK

☒ Labels

□ Constant is Zero

☒ Confidence Level: 95 %



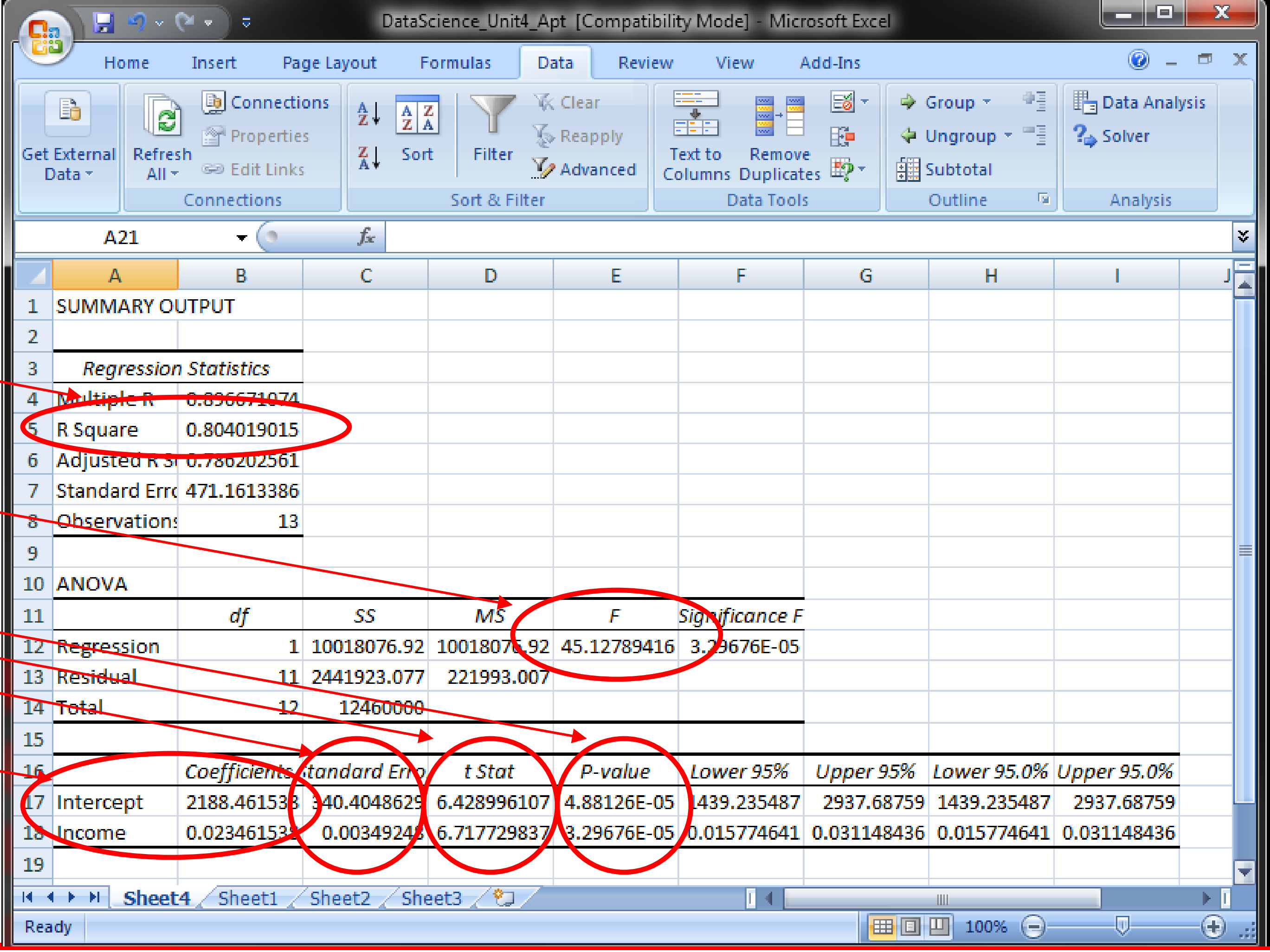
Income	Rent
$30,000	$3,000
$40,000	$2,900
$50,000	$2,800
$60,000	$3,100
$70,000	$3,900
$80,000	$5,000
$90,000	$4,900
$100,000	$4,800
$110,000	$4,500
$120,000	$5,200
$130,000	$5,400
$140,000	$4,900
$150,000	$5,500

Excel Output

R-Square

F

P value
T stat
Standard Error
Coefficients



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.896671074
R Square	0.804019015
Adjusted R S	0.786202561
Standard Erro	471.1613386
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10018076.92	10018076.92	45.12789416	3.29676E-05
Residual	11	2441923.077	221993.007		
Total	12	12460000			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2188.461538	340.4048629	6.428996107	4.88126E-05	1439.235487	2937.68759	1439.235487	2937.68759
Income	0.023461538	0.00349248	6.717729837	3.29676E-05	0.015774641	0.031148436	0.015774641	0.031148436

Regression Analysis: R-Squared

Scenario	R-Squared
No Relationship	0.0
Social Science Studies	0.3
Marketing Research	0.6
Scientific Applications	0.9
Perfect Relationship	1.0

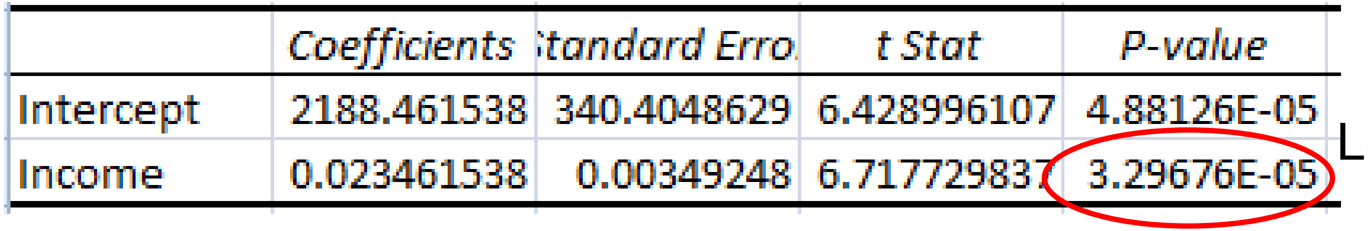
R-Squared, the Coefficient of Determination
Also known as “Goodness of Fit”,
from 0 (no fit) to 1 (perfect fit)

Hypothesis Testing: t-Stat and P-value

Statistic	Description
Standard Error	Estimate of standard deviation of the coefficient
t-Stat	Coefficient divided by the Standard Error
P-value	Probability of encountering equal t value in random data P-value should be 5% or lower

Hypothesis Testing: Test H_0 (null hypothesis)

Null hypothesis: No correlation between x and y



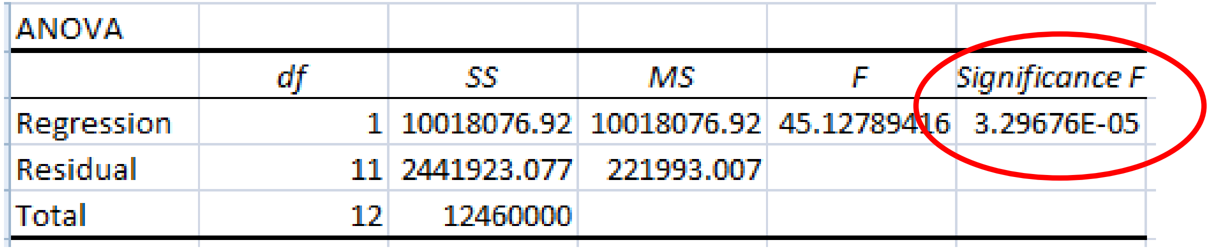
	Coefficients	*tandard Erro*	*t Stat*	*P-value*
Intercept	2188.461538	340.4048629	6.428996107	4.88126E-05
Income	0.023461538	0.00349248	6.717729837	3.29676E-05

Less than 5% → OK

Hypothesis Testing: F value

Statistic	Description
F value	Tests overall significance of the regression model
H_0	Tests null hypothesis that all regression coefficients = 0 Tests full model against a model with no variables
Significance F	Associated P value; Less than 0.05 to invalidate H_0

Hypothesis Testing: Test H_0 (null hypothesis)
Null hypothesis: No correlation between x and y



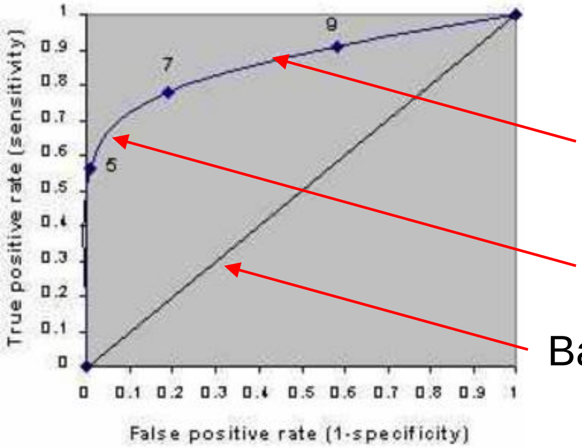
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10018076.92	10018076.92	45.12789416	3.29676E-05
Residual	11	2441923.077	221993.007		
Total	12	12460000			

Less than 5% → OK

Regression Analysis: ROC Curves

Topic	Description
ROC	Receiver Operating Characteristic Plot of true positive rate against false positive rate at different cutpoints
History	Developed during World War II by RADAR engineers
Tradeoff	Shows tradeoffs, such as sensitivity and specificity for experiments



Good→ close to top edge

Good→ close to left edge

Bad → close to diagonal

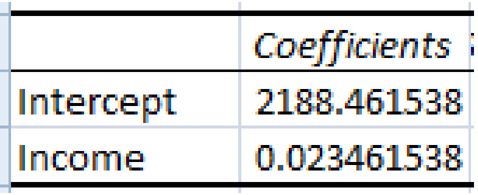
Cutpoint	Sensitivity	Specificity
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

Cutpoint	True Pos.	False Pos.
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

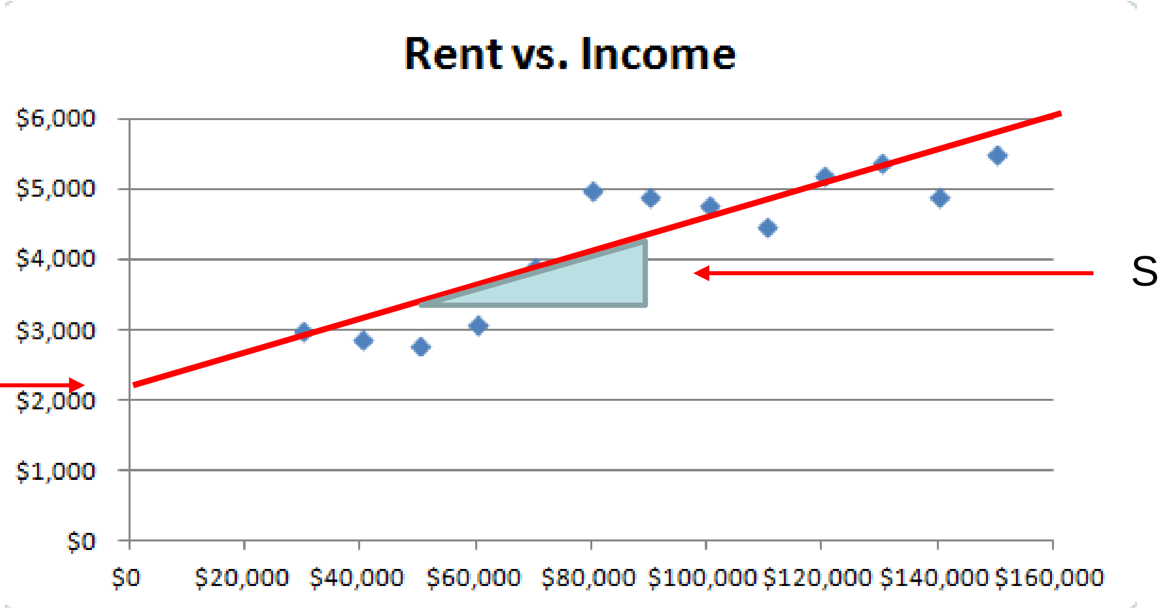
Regression Analysis: ROC Curves

Topic	Description
Tests	Test predictive performance of model; how to select cutoffs At 95% level of confidence, we test H_0 at 5% (alpha = 5%)
True positive	Correctly identified; High income people rent expensive apts.
True negative	Correctly rejected; Low income people rent cheap apartments
False positive	Null hypothesis is true; No correlation (type I error) To address type I error, reduce alpha (in our case, 5%)
False negative	Failing to reject null hypothesis which is false (type II error) We thought model doesn’t work, but it does
Tradeoff	As we decrease alpha from 5% to 1%... Type I error decreases, but Type II error increases (typical) Selecting cutoff a business decision; alpha = 5% usually good

Regression Analysis: Coefficients



	Coefficients
Intercept	2188.461538
Income	0.023461538



Slope: 0.023 / 1

Y-intercept: 2,188

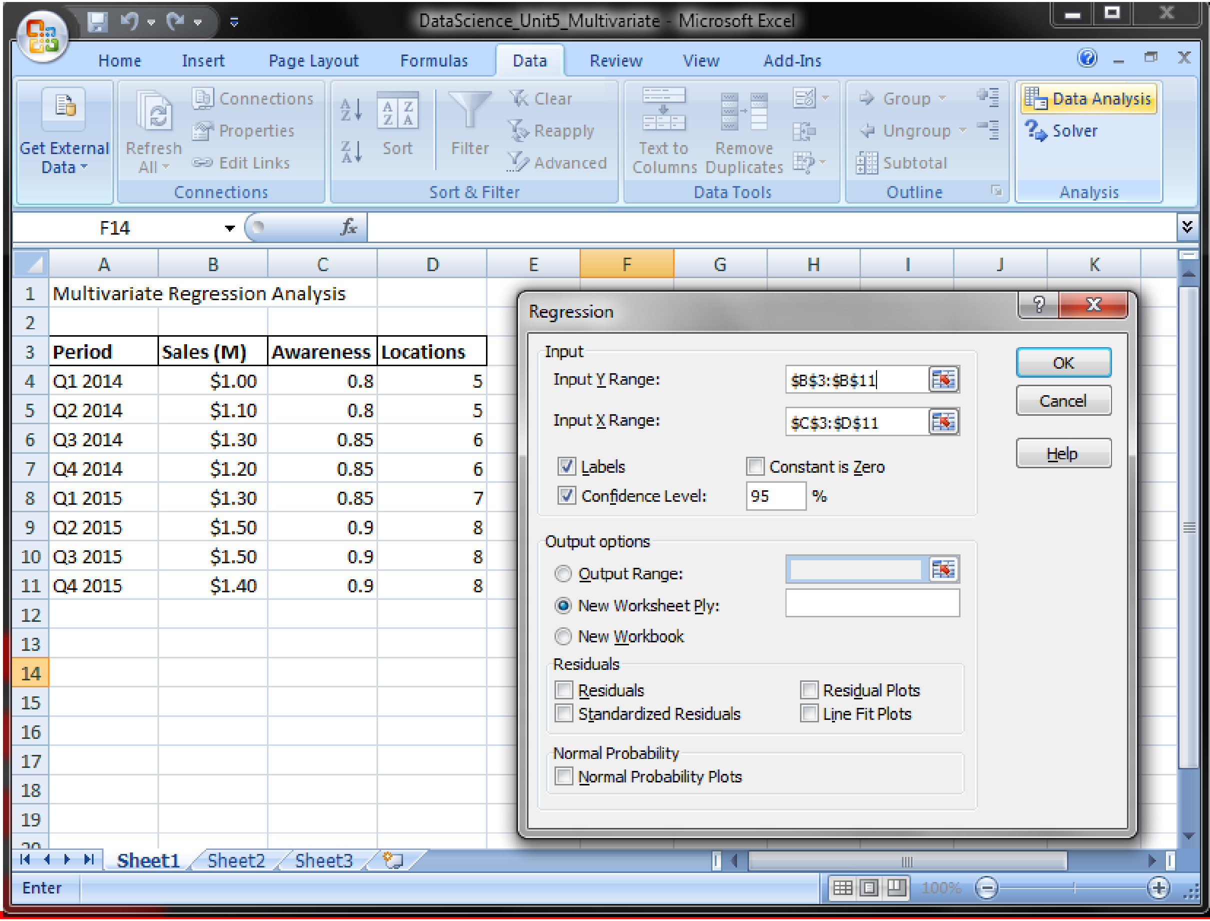
Regression Analysis: Multivariate

Period	Sales Level	Market Awareness	Number of Locations
Q1 2012	$1.0 million	80%	5
Q2 2012	$1.1 million	80%	5
Q3 2012	$1.3 million	85%	6
Q4 2012	$1.2 million	85%	6
Q1 2013	$1.3 million	85%	7
Q2 2013	$1.5 million	90%	8
Q3 2013	$1.5 million	90%	8
Q4 2013	$1.4 million	90%	8

What would happen if we opened 2 new stores, while holding awareness at 90%?

Regression Analysis: Multivariate

Y Range: Sales
X Range: Awareness & Locations



Multivariate Regression Analysis

Period	Sales (M)	Awareness	Locations
Q1 2014	$1.00	0.8	5
Q2 2014	$1.10	0.8	5
Q3 2014	$1.30	0.85	6
Q4 2014	$1.20	0.85	6
Q1 2015	$1.30	0.85	7
Q2 2015	$1.50	0.9	8
Q3 2015	$1.50	0.9	8
Q4 2015	$1.40	0.9	8

Regression

Input

Input Y Range: $B$3:$B$11

Input X Range: $C$3:$D$11

☑ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

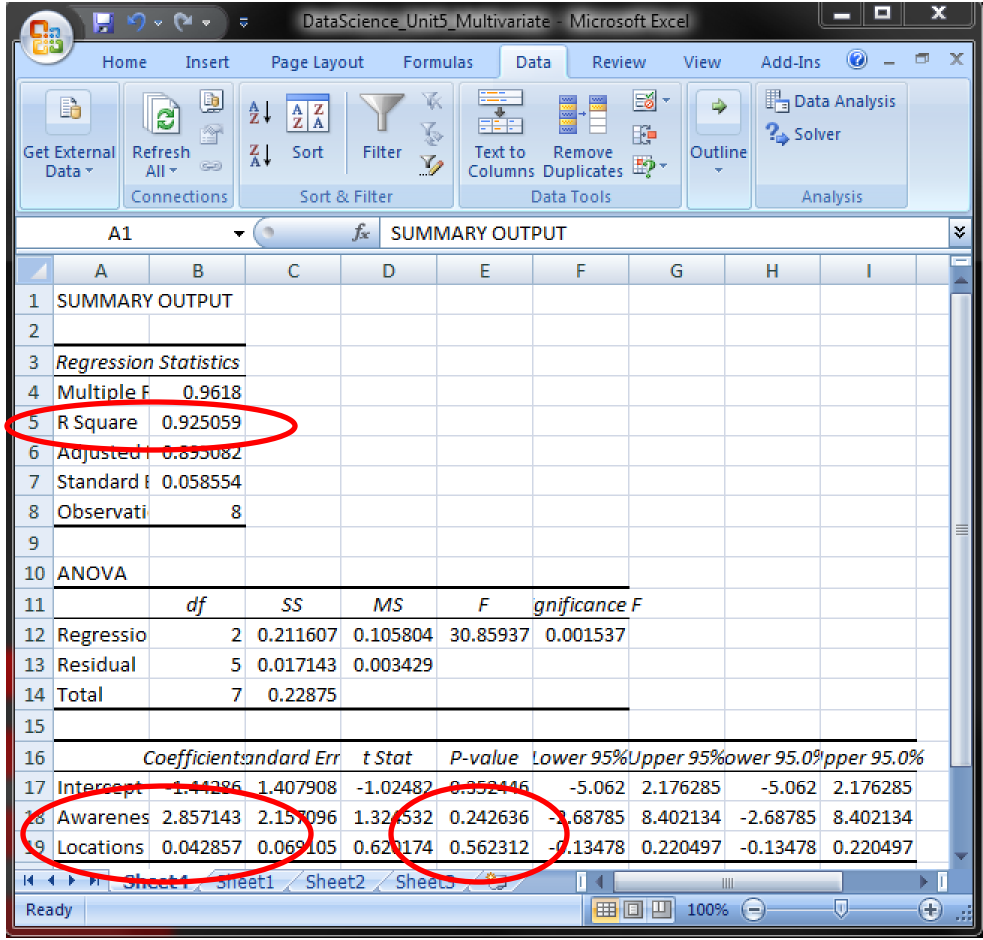
Regression Analysis: Multivariate

R – squared: 0.92

Y-intercept = -1.44286

Coefficient for Awareness: 2.857143
at a P-value of 24.2% (not very good)

Coefficient for Locations: 0.042857
at a P-value of 56.2% (poor)



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	0.9618							
5	R Square	0.925059							
6	Adjusted I	0.895082							
7	Standard E	0.058554							
8	Observati	8							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	2	0.211607	0.105804	30.85937	0.001537			
13	Residual	5	0.017143	0.003429					
14	Total	7	0.22875						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-1.44286	1.407908	-1.02482	0.352446	-5.062	2.176285	-5.062	2.176285
18	Awarenes	2.857143	2.157096	1.324532	0.242636	-2.68785	8.402134	-2.68785	8.402134
19	Locations	0.042857	0.069105	0.620174	0.562312	-0.13478	0.220497	-0.13478	0.220497

Regression Analysis: Multivariate

Output	Description	Values in Our Sales Example
R-Square	Goodness of fit of model to data	0.93
Intercept	Point where line crosses Y axis	-1.44
Coefficient 1	Coefficient for Market Awareness	2.857
Coefficient 2	Coefficient for Number of Locations	0.043

Sales = (Intercept) + (Coefficient 1) * (Market Awareness) + (Coefficient 2) * (Number of Locations)

= (- 1.44) + (2.857) * (Market Awareness) + (0.043) * (Number of Locations)

Example: Maintain brand awareness at 90%; Open two new retail stores (10 total)

= (-1.44) + (2.857) * (0.90) + (0.043) * (10) = $1.56 Million

Outline/ Learning Objectives

Topic	Description
Background	The goal of regression analysis
Statistics	Basic statistics governing regression performance
Tests	F tests, t tests, p tests
Procedure	Executing regression analysis in Microsoft Excel
Multivariate	Executing cases with two or more independent variables