Introduction to Statistics
Part 2

January 18, 2018

GSA

D2D

DATA TO DECISIONS

Course Outline

- Intro to Stats part 1 review
- Significance
- Correlation
- Regression
- Cluster analysis http://www.listendata.com/2016/01/cluster-analysis-with-r.html

Intro to Statistics Part 1 Review

- What is statistics
- Major definitions
- Exploratory analysis
- Descriptive statistics
- Inferential statistics
- Probability and statistics

Statistical Significance

- Be careful about results of your statistic analysis
- Statistics isn’t an exact science, think of it as finely tuned guesswork
- The statistical significance let’s you assess how good your “guess” is
  - Usually is denoted as α (“alpha”)
  - Alpha level is largely arbitrary
  - Depends on an industry
  - If analysis satisfies industry-accepted alpha level, your “guess” is good

GSA

D2D
DATA TO DECISIONS

Probability in Statistics

- Statistical analysis makes certain claims about the data
- Statistics uses probability math to determine significance of the results
- Significance is level of probability considered to be "good enough"
- In probability speak the claim about the data is *Hypothesis*
- P-Value (short for probability):
  - ➢ A measure of the strength of evidence against the hypothesis
  - ➢ The smaller the p-value, the greater is the evidence against the hypothesis

What is Probability

- A branch of math calculating likelihood of a given event
- Expressed as a number between 1 and 0
  - 1 for certainty
  - 0 for impossibility
- In a simple case of two discrete events (e.g. coin toss) the probability is
  - P(a) = N(a) / [N(a) + N(b)
  - N is number of events
  - Important: events (a) and (b) are independent, either (a) or (b)
  - For large number of events we use mass probability function
- In cases with continuous events/variables we use probability density function

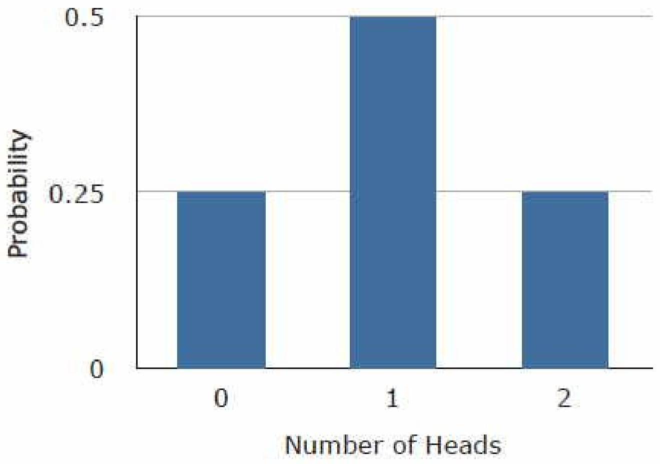
Two-coin Toss Example

Binomial Distribution

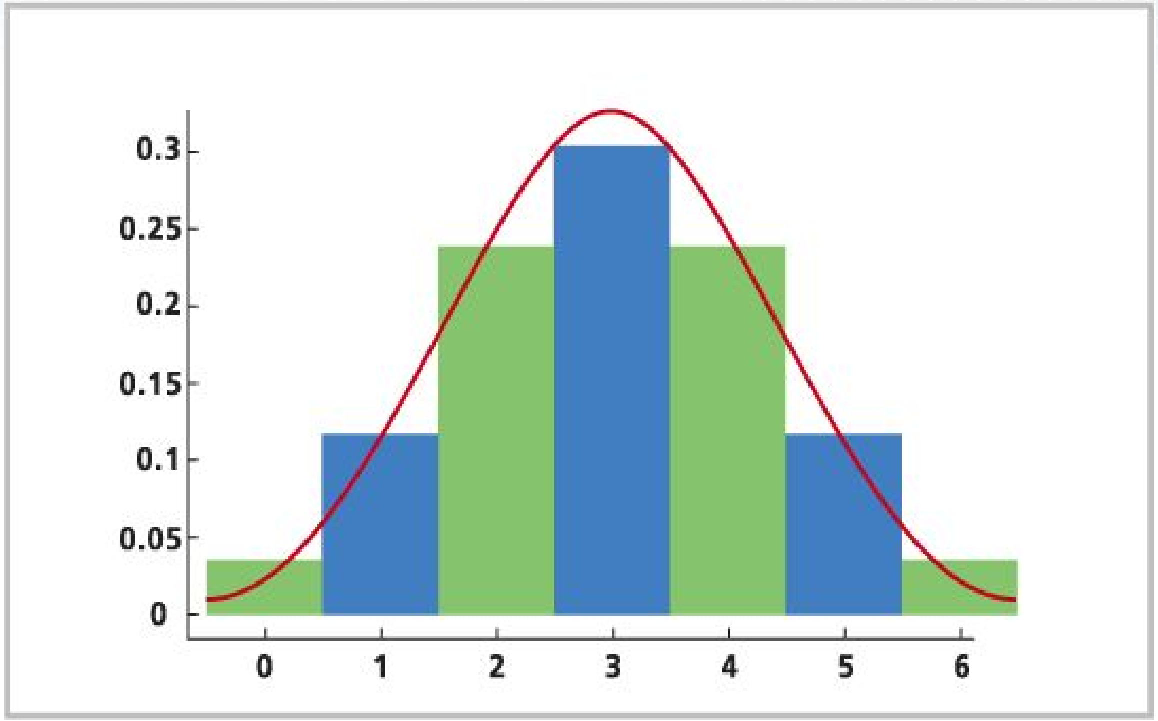
Outcome	First Flip	Second Flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails

Number of Heads	Probability
0	1/4
1	1/2
2	1/4

Note: if you replace Probability with Frequency, you will get histogram of a two-coin test



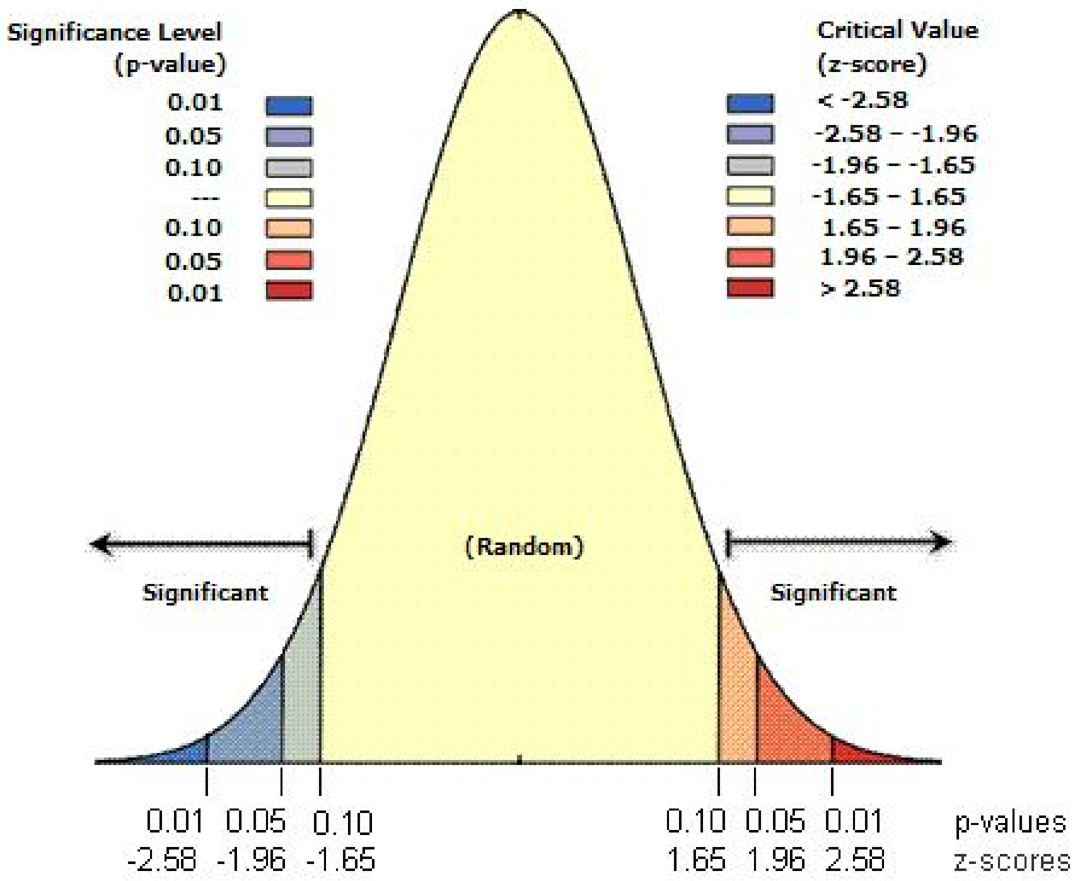
Binomial and Normal Distributions



For large number of events (i.e. from discrete to continuous) binomial becomes normal

Binomial Distribution is Mass Probability Function
Normal Distribution is Probability Density Function

What is P-value



Hypothesis Testing

- The claim on trial is called *Null Hypothesis*
- Null (“non existing”) Hypothesis H_0, e.g.:
  - ➢ no significant difference between two data samples
- To determine the significance the P-value is compared to an α level
  - ➢ Rough guideline:

    P-value < 0.01 - very strong evidence against H_0

    0.01 < P-value < 0.05 – strong evidence against H_0

    0.05 < P-value < 0.1 – weak evidence against H_0

    P-value > 0.1 – little or no evidence against H_0

Hypothesis Testing Example

- Null Hypothesis: leaves of plants grown in a shadow are different from leaves of same plants in the sun
- Leaves in a shadow, length
  - c(0,1,3,4,6,11,11,6,5,2,1)
- Leaves in the sun, length
  - c(3,3,7,11,9,7,5,3,1,0,0)
- Evaluate with descriptive statistics

```
> sun = c(0,1,3,4,6,11,11,6,5,2,1)
> shadow = c(3,3,7,11,9,7,5,3,1,0,0)
> mean(sun)
[1] 4.545455
> mean(shadow)
[1] 4.454545
> sd(sun)
[1] 3.777926
> sd(shadow)
[1] 3.670521
```

Hypothesis Testing w/ T-test

P-value >> 0.1, the null hypothesis is rejected

- **T.test**

```
> sun = c(0,1,3,4,6,11,11,6,5,2,1)
> shadow = c(3,3,7,11,9,7,5,3,1,0,0)
>  t.test(sun,shadow)
 Welch Two Sample t-test
data:  sun and shadow
t = 0.057241, df = 19.983, p-value = 0.9549
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.222152  3.403970
sample estimates:
mean of x mean of y
 4.545455  4.454545
```

Hypothesis Testing Based on Descriptive Statistics

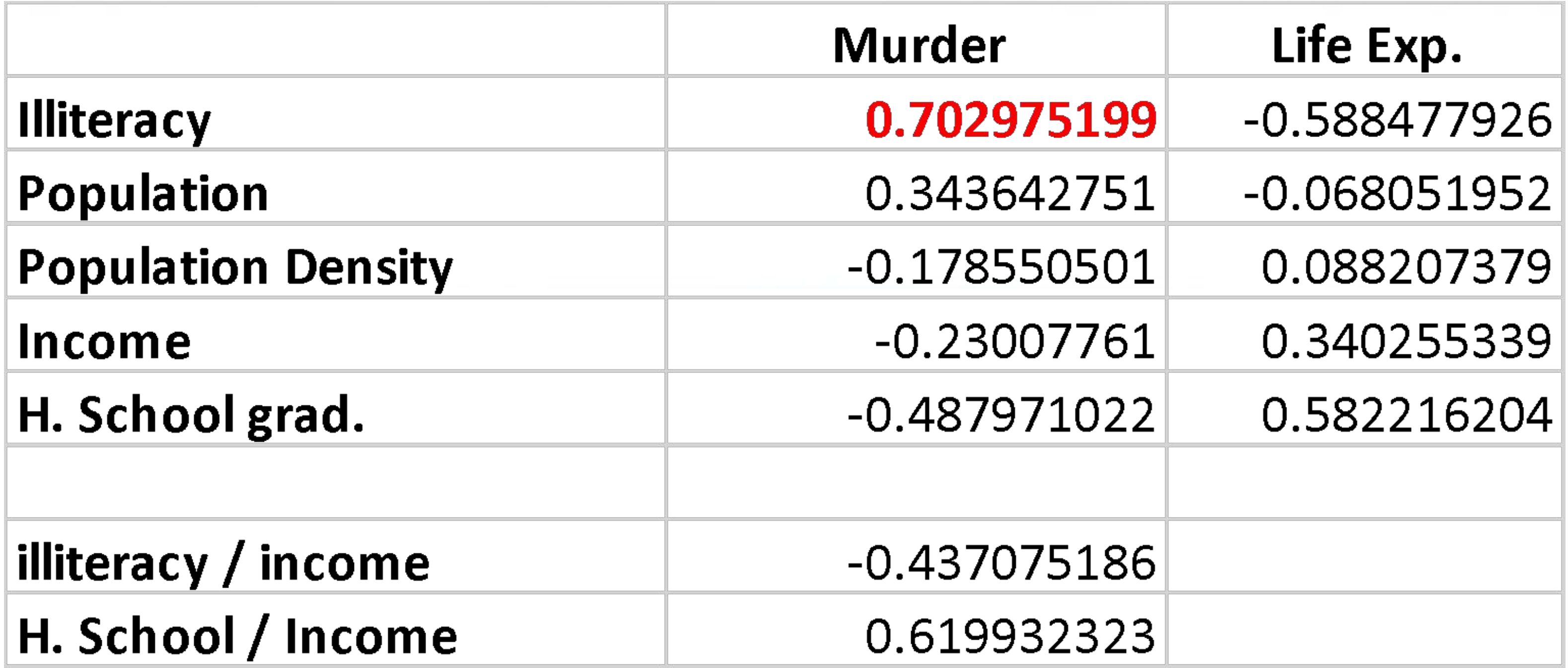
- **T.test**
  - One sample t.test: $t = (m-\mu)/S/\sqrt{n}$

    where m is mean, μ is theoretical mean, S is sum of squares, and n is number of samples
  - Independent two-sample t.test: $t = (m_A - m_B)/\sqrt{(S^2/n_A + S^2/n_B)}$
  - Paired t.test: similar to one sample where one of the samples serves as theoretical
- **Z.test**
  - One sample z.test: $z = (x - \mu) / \sigma$, where σ is standard deviation
- **T Vs. Z:** use Z if you know standard deviation
- **T- and Z- tests compare means ("normalized" by variances)**
- **Use ANOVA to compare variances**

Hypothesis Testing Based on Inferential Statistics

- Prime inferential method is Correlation
- T/Z tests Vs. Correlation
  - ➢ Descriptive: can compare measurements in same units
  - ➢ Inferential: can compare measurements in same and in different units
  - ➢ Cannot use correlation when the data is not “ordered”, the samples are not “tied” to each other, e.g. you can correlate width and length of leaves of the same plant sample

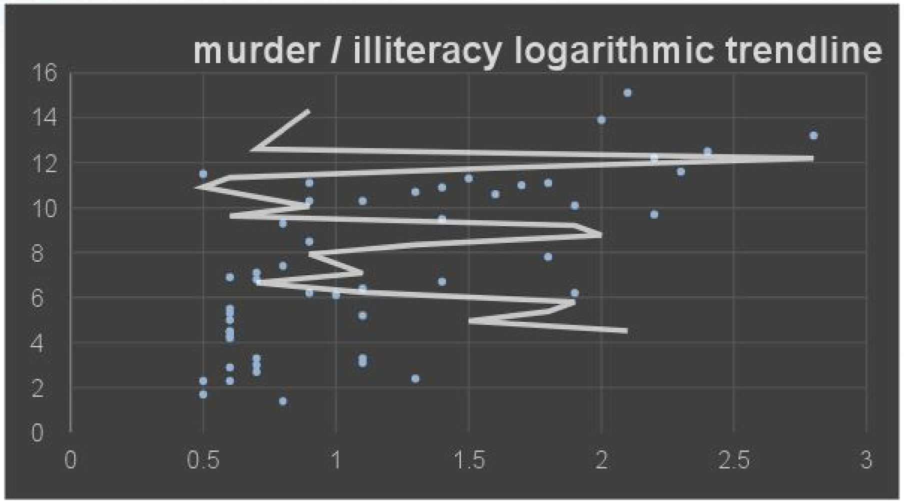
Correlation Analysis Example StatesDataForR.xlxs

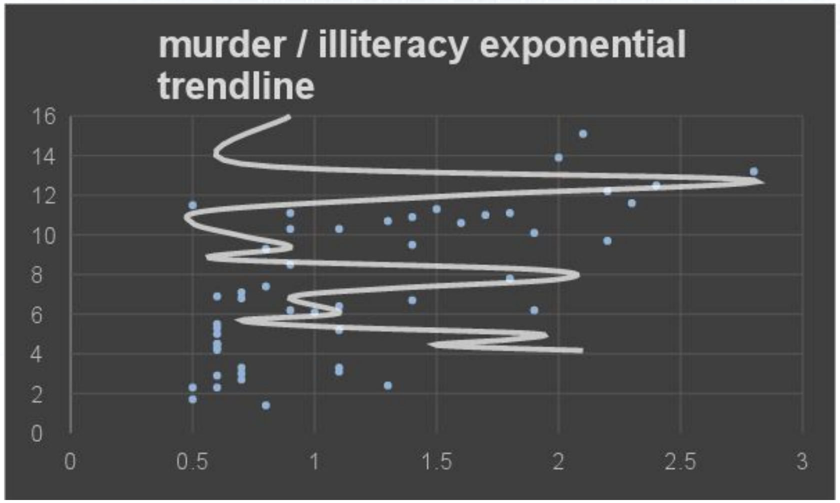


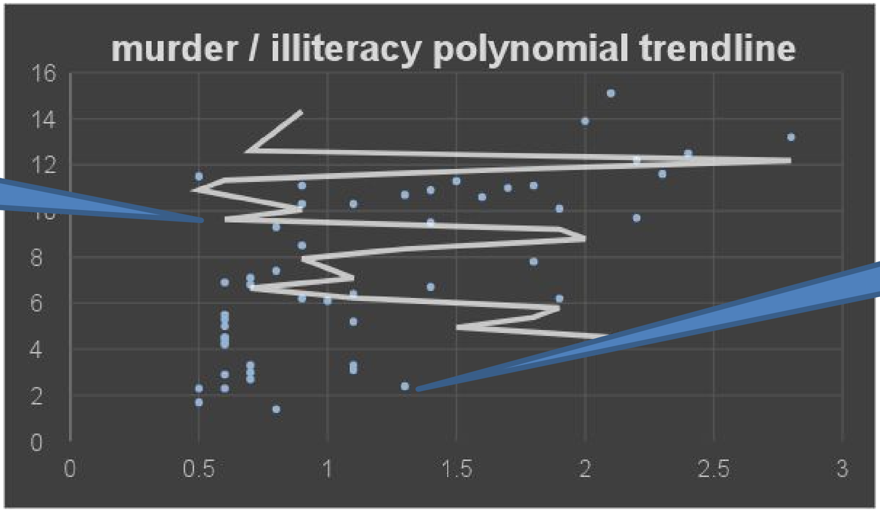
	Murder	Life Exp.
Illiteracy	**0.702975199**	-0.588477926
Population	0.343642751	-0.068051952
Population Density	-0.178550501	0.088207379
Income	-0.23007761	0.340255339
H. School grad.	-0.487971022	0.582216204
illiteracy / income	-0.437075186	
H. School / Income	0.619932323	

Regression Analysis Example

Scatter Plot / StatesDataForR.xlxs





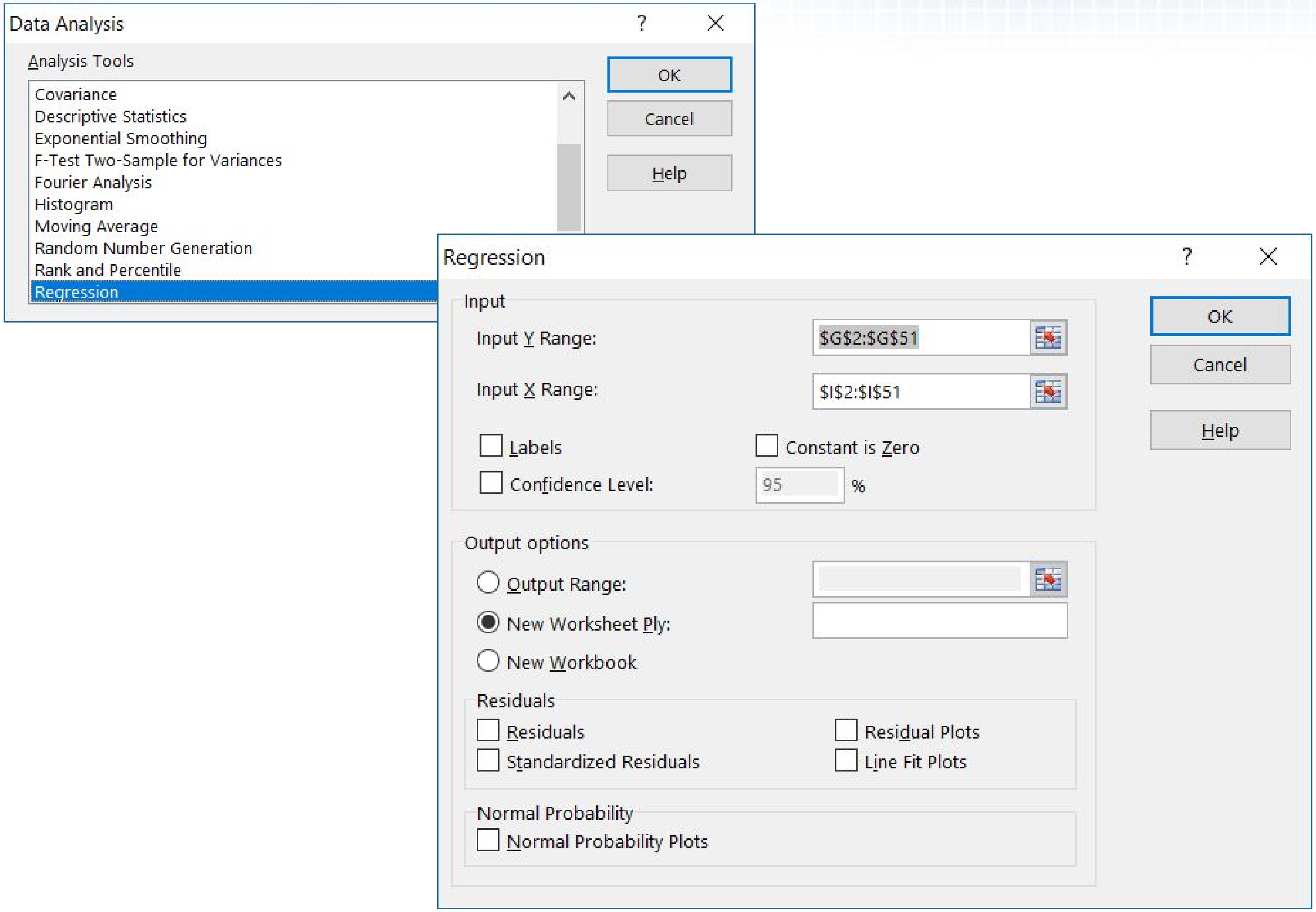


Nevada / Las Vegas

Mississippi

Regression Analysis Example

Data Analysis / Regression / StatesDataForR.xlxs



Input

Input Y Range: $G$2:$G$51

Input X Range: $I$2:$I$51

☐ Labels　☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals　☐ Residual Plots

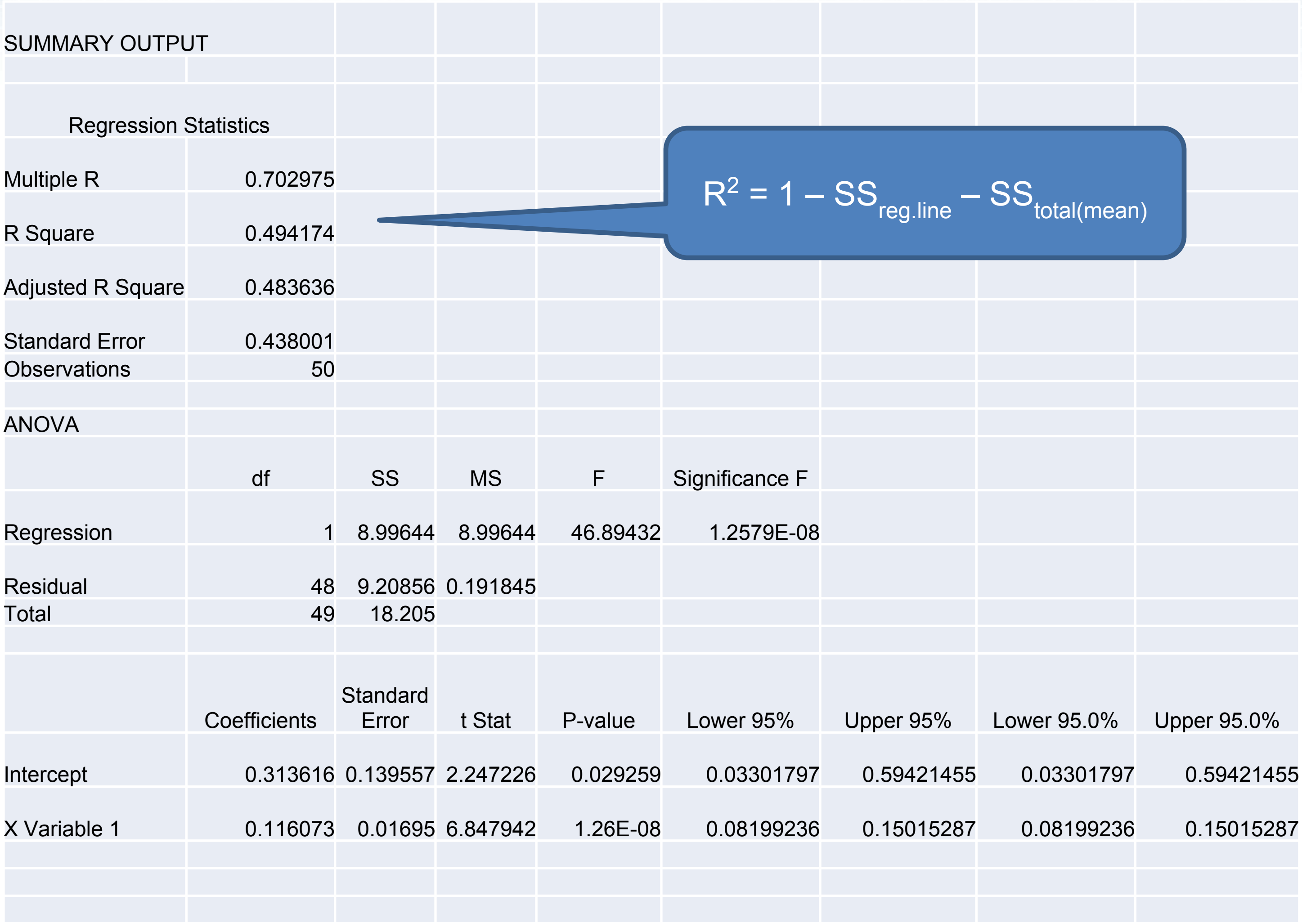
☐ Standardized Residuals　☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Regression Analysis Example (Cont.)

Data Analysis / Regression / StatesDataForR.xlxs



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702975
R Square	0.494174
Adjusted R Square	0.483636
Standard Error	0.438001
Observations	50

ANOVA

	df	SS	MS	F	Significance F
Regression	1	8.99644	8.99644	46.89432	1.2579E-08
Residual	48	9.20856	0.191845		
Total	49	18.205			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.313616	0.139557	2.247226	0.029259	0.03301797	0.59421455	0.03301797	0.59421455
X Variable 1	0.116073	0.01695	6.847942	1.26E-08	0.08199236	0.15015287	0.08199236	0.15015287

$R^2 = 1 – SS_{reg.line} – SS_{total(mean)}$

Useful Links

- http://www.statisticshowto.com/statistics-basics/
- https://www.socialresearchmethods.net/kb/index.php
- A semester-long course
- http://online.stanford.edu/course/probability-and-statistics-self-paced
- R tutorials
- http://www.statmethods.net/index.html
- http://www.cengage.com/resource_uploads/downloads/1305115341_450336.pdf

Q & A