v1.2

Regression Analysis

How to develop and assess a CER

"All models are wrong, but some are useful." -George Box
"In mathematics, context obscures structure. In data analysis, context provides meaning." -George Cobb
"Mathematical theorems are true; statistical methods are sometimes effective when used with skill." -David Moore

v1.2

Regression Overview

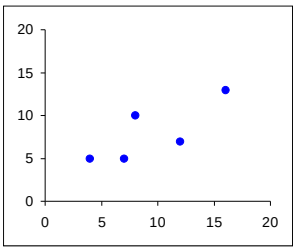
- Key Ideas $Y = a + bX + \varepsilon$ $\hat{Y} = \hat{a} + \hat{b}X$
  - Correlation
  - Best fit / minimum error
  - Homoscedasticity!
  - Statistical significance
  - Quantification of uncertainty
- Practical Applications
  - CER Development
  - Learning Curves
- Analytical Constructs
  - OLS Regression
  - Analysis of Variance (ANOVA)
  - Confidence Intervals
  - Linear algebra
- Related Topics
  - Parametrics 3
  - Distributions
    - Normal, Chi, t, F 10
  - Hypothesis testing
  - Risk Analysis 9

Regression Within The Cost Estimating Framework

v1.2

Past

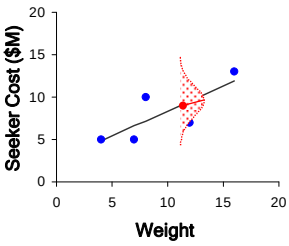
Understanding your historical data



Historical costs for similar systems

Present

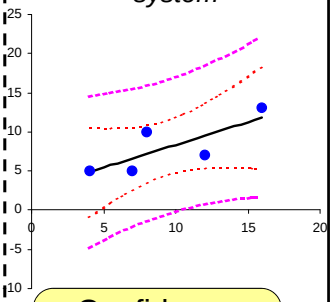
Developing estimating tools



Cost Estimating Relationships

Future

Estimating the new system



Confidence and Prediction Intervals

Bivariate Data Analysis 6

v1.2

- One independent variable and one dependent variable (i.e., y is a function of x)
- Visual display of information — What does it look like?
  - Scatter plot, residual plots
- Measures of central tendency — What's your best guess?
  - $\hat{Y}$ from regression equation
- Measures of variability — How much remains un-explained?
  - Standard Error of the Estimate (SEE), R^2
- Measures of uncertainty — How precise are you?
  - Confidence and Prediction Intervals
- Statistical Tests — How can you be sure?
  - F test, t test, ANOVA

v1.2

Regression Models

- **Factors** (y = bx)
  - Estimates made using factors are essentially a simplified form of regression where the intercept is assumed to be zero
  - Should generally be avoided and replaced with the linear regression methods covered in this module

Covered in this module

- **OLS - Linear Regression** (y = a + bx)
  - The simplest regression model
    - Seeks to minimize the sum of squared errors
  - Assumes one independent and one dependent variable
  - Can also estimate **non-linear models** (power, logarithmic, exponential) by transforming equations into linear forms
- **OLS - Multivariate Linear Regression** ($y = a + b_1x_1 + \ldots + b_kx_k$)
  - Seeks to minimize sum of squared errors
  - Assumes one dependent variable with multiple independent variables
  - Adds the danger that the model will exhibit multicollinearity

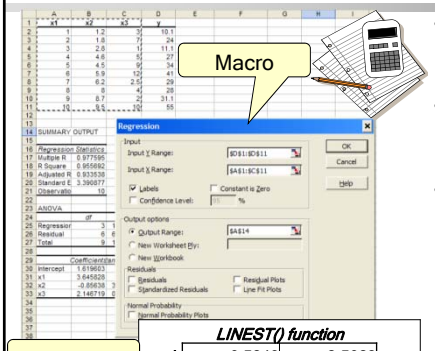
Touched on in this module

- **OLS Generalizations** (e.g., MLE) **and Non-OLS Models** (WLS, MUPE, ZMPE)
  - May have objectives other than minimizing sum squared errors, such as minimizing percent (multiplicative) errors
  - May be necessary when OLS regression techniques yield residuals that violate the heteroscedasticity assumption

'To b or Not to b' The y-intercept in Cost Estimation, R. L. Coleman, J. R. Summerville, P. J. Braxton, B. L. Cullis, E. R. Druker, SCEA, 2007.

v1.2

Regression Analysis in Excel



LINEST() function

b	0.5849	2.5023	a
SEb	0.2636	2.7110	SEa
R2	0.6214	2.4612	SEE
F	4.9243	3	d.f.
SSR	29.8280	18.1720	SSE

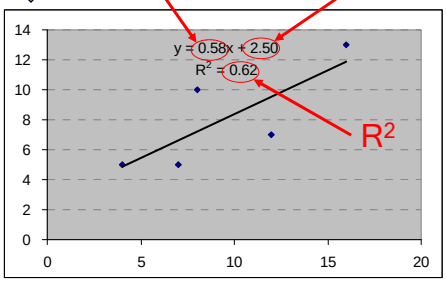
Warning: Results of macros do not update if your data changes!

- Since OLS is such a popular form of regression, Excel has the built-in capability to perform the analysis via **LINEST** function and macro
- Data must first be formatted correctly in Excel
  - Place the $x_1 \ldots x_k$ data in adjacent columns (*not* rows!)
  - Place the y data in a column
- The **LINEST** function (recommended) returns various regression statistics dynamically and is discussed in later slides
  - Select the Y and X ranges
  - Set "constant" equal to TRUE (or 1) to calculate the intercept normally, Set the constant equal to FALSE (or 0) to force the intercept to zero
  - Set "stats" equal to TRUE (or 1) to return regression statistics
- To enable the macro, the Data Analysis ToolPak must first be added
- Then regression analysis is easy to perform
  - Select the X and Y ranges
    - Put a checkmark in the "Labels" box iff the labels are included in the selected range
  - Select an Output Range for the results

v1.2

Trendlines in Excel

6



- Right-click the Data Series
  - Add Trendline…
  - Type = Linear
  - Options
    - Display equation on chart
    - Display R-squared value on chart

Warning: Not to be confused with Trend Analysis (i.e., Time Series)

- Can also choose other functional forms

Tip: Always plot your data and fit a Trendline first. This will help guide and check your Regression Analysis

v1.2

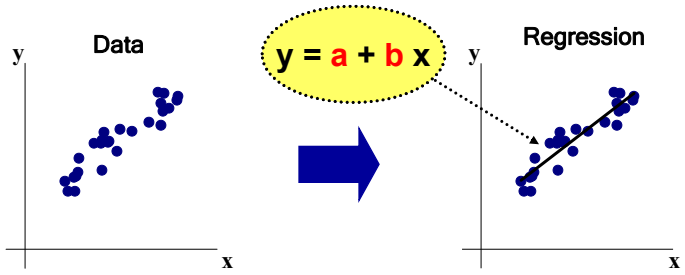
Regression Outline

- Core Knowledge
  - Definition of Regression
  - Preliminary Concepts
  - Linear Regression
    - Finding the Equation
    - Goodness of Fit
    - Confidence Intervals
  - Non-linear Models
  - Multivariate Models
  - Selecting the Best Model
- Summary
- Resources
- Related and Advanced Topics

Definition of Regression

v1.2

- Regression Analysis is used to describe a *statistical* relationship between variables
  - Specifically, it is the process of estimating the "best fit" parameters of a specified function that relates a dependent variable to one or more independent variables (including implicit uncertainty)



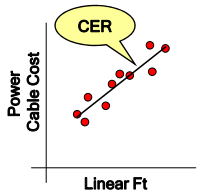
Regression Analysis in Cost Estimating

v1.2

- If the dependent variable is a cost, the regression equation is often referred to as a *Cost Estimating Relationship* or *CER*
  - The independent variable in a CER is often called a *cost driver*

Examples of cost drivers:

Cost	Cost Driver (single)
Aircraft Design	# of Drawings
Software	Lines of Code
Power Cable	Linear Feet

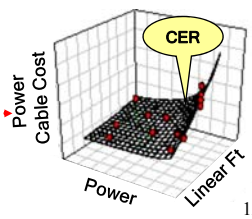


3

  - A CER may have multiple cost drivers:

Example with multiple cost drivers:

Cost	Cost Driver (multiple)
Power Cable	Linear Feet Power



v1.2

Preliminary Concepts

- Correlation
- Causation
- Types of Models

v1.2

Preliminary Concepts – Introduction

- Correlation
  - Are X and Y related?
- Causation
  - Does X "drive" Y?
- Model Type
  - What kind of function does the relationship resemble?

We will examine each in more detail…

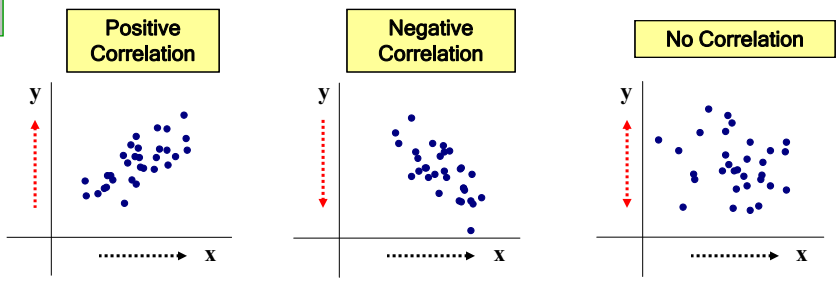
v1.2

Preliminary Concepts – Correlation

6

- Correlation is a measure of the linear relationship between two or more variables
- Variables have either positive, negative, or no correlation between them

1



"As X increases, Y increases"

"As X increases, Y decreases"

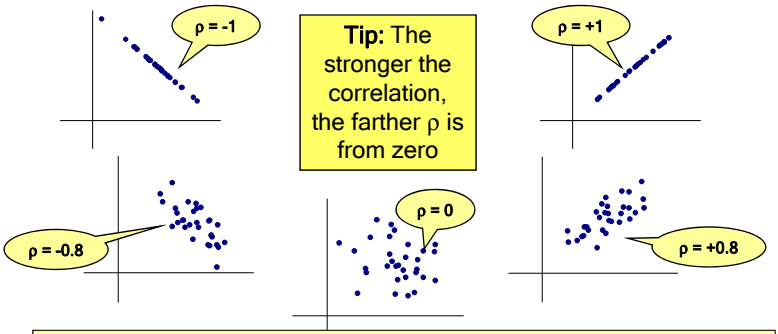
"As X increases, Y has no relation"

v1.2

Preliminary Concepts – Correlation

6

- Correlation is quantified by a *correlation coefficient* (ρ)
  - ρ can range from -1 to +1



Tip: The stronger the correlation, the farther ρ is from zero

The presence of correlation is what leads you to regression analysis. Scatter plotting is the best initial step to detect it. Explicit estimation of ρ will be addressed later…

v1.2

Preliminary Concepts – Causation

2

18

- IMPORTANT: Correlation does not imply causation!
  - e.g., Cost of damage by a fire vs. # of firemen that fought it
- On the other hand correlation is a necessary condition for causation
- There is no statistical method to verify a cause-and-effect relationship
- For CER development, cost drivers should be chosen based on
  1. Presence of correlation AND
  2. Sound engineering principles for the relationship being investigated
     - Do you believe X really "drives" the cost of Y?

3

6

v1.2

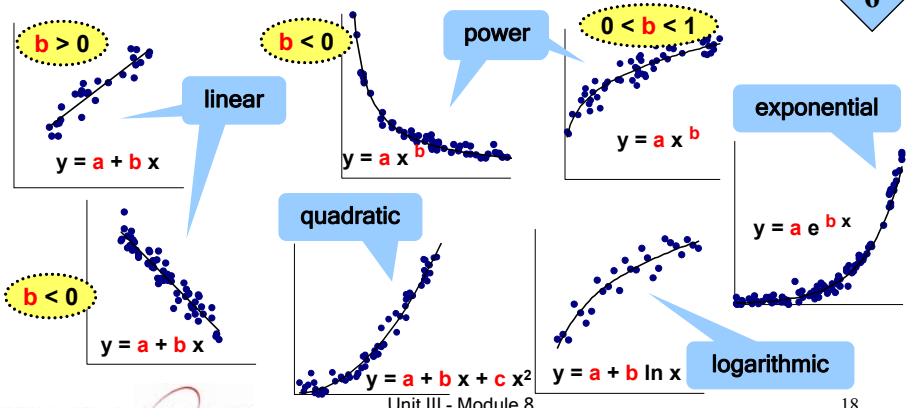
Preliminary Concepts – Types of Models

19

- A mathematical function must be specified before regression analysis is performed
  - The specified function is called a regression model
  - Many types of models may be considered:

3

6

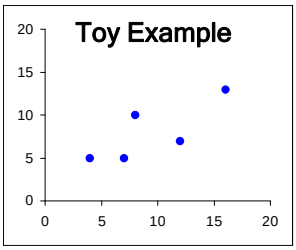


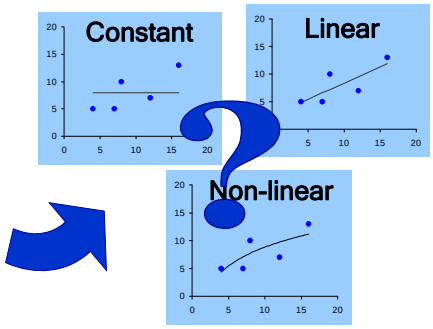
v1.2

Determining a Regression Model

- A scatter plot should always be performed first to determine what kind of model should be tested, if any at all:
  - Specifying the wrong function for a model can lead to an incorrect interpretation of the results

6





Tip: A linear model is usually the best starting point

v1.2

Linear Regression Analysis

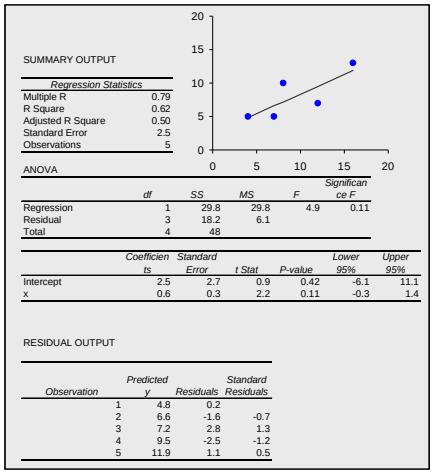
- Finding the Equation
- Goodness of Fit
- Confidence Intervals

Introduction – Toy Problem

v1.2

- We will demonstrate the elements of regression analysis using a toy problem
- We can find the results quickly by running a regression analysis in Excel
  - But what do the numbers mean?
  - Where do they come from?
  - How do we interpret the results?

We will answer these questions while providing connections to the output shown here



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

RESIDUAL OUTPUT

Observation	Predicted y	Residuals	Standard Residuals
1	4.8	0.2	
2	6.6	-1.6	-0.7
3	7.2	2.8	1.3
4	9.5	-2.5	-1.2
5	11.9	1.1	0.5

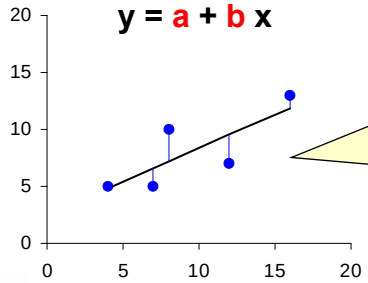
Key Concept of Regression

v1.2

3

- The regression procedure uses the *"method of least squares"* to find the "best fit" parameters of a specified function
  - We will focus on Ordinary Least Squares (OLS) Regression[1]
  - The idea is to minimize the **sum of squared deviations** (called "errors" or "residuals") between the Y data and the regression equation

Called the "Sum of Squared Errors" or "SSE"



In other words… Find the equation that minimizes the sum of the squared distances of the blue lines

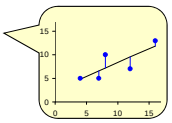
Let's look at the math…

1 OLS Regression is most commonly used and is the foundation to the understanding of all regression techniques. To learn about variations and more advanced techniques, see Resources and Advanced Topics.

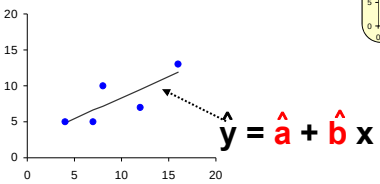
v1.2

Finding the Regression Equation

- Problem: Find $\hat{a}$ and $\hat{b}$ such that the SSE is minimized…



Data	
X	Y
4	5
7	5
8	10
12	7
16	13



The SSE is minimized when the slope is equal to the **correlation coefficient, r** multiplied by the **quotient of standard deviations, s_Y/s_X** *and* the line passes through **the means of X and Y**[1]

8

This formulation reduces to the following set of equations

$$\hat{b} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$\hat{a} = \bar{Y} - \hat{b}\,\bar{X}$$

8

Where:

$\bar{X}$ and $\bar{Y}$ are the means of the 2 variables and

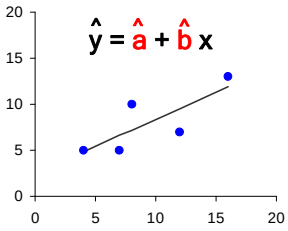
X and Y are the raw values of the 2 variables and

1 The geometry of this formulation is provided under Related and Advanced Topics

v1.2

Finding the Regression Equation: Example

4



Data		Calculations	
X	Y	XY	X^2
4	5	20	16
7	5	35	49
8	10	80	64
12	7	84	144
16	13	208	256

n = 5

$\bar{X}$ = avg of X data = 9

$\bar{Y}$ = avg of Y data = 8

ΣXY = 427

ΣX^2 = 529

$$\hat{b} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} \qquad \hat{a} = \bar{Y} - \hat{b}\,\bar{X}$$

Example Calculation

$$\hat{b} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$\hat{b} = \frac{427 - 5*9*8}{529 - 5*9^2}$$

$$\hat{b} = 0.6$$

$$\hat{a} = \bar{Y} - \hat{b}\,\bar{X}$$

$$\hat{a} = 8 - 0.6*9$$

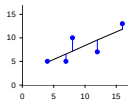
$$\hat{a} = 2.5$$

$$\hat{Y} = 2.5 + 0.6\,X$$

v1.2

Statistical Error and Residual Analysis

- In addition to the equation we have just found, we must describe the statistical error – the "fuzz" or "noise" – in the data



- This is done by adding an "error term" (ε) to the basic regression equation to model the residuals:

$$y = a + b\,x + \varepsilon$$

- There are two key assumptions in OLS regression regarding the error term:

1. It is independent with X
2. It is normally distributed with a mean of zero and *constant variance for all X*

These assumptions need to be checked, as the entire analysis hinges on their validity

A "residual plot" is useful in determining whether these assumptions apply to the data

This is called **homoscedasticity**!

- The error about the "true underlying" line generates a data cloud whose best-fit line is slightly different (and unknowable)

The error term is what makes Regression more than just Curve Fitting

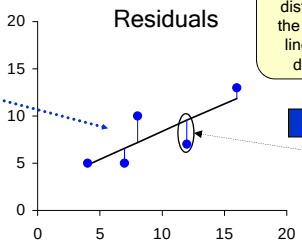
v1.2

Residual Analysis: Example

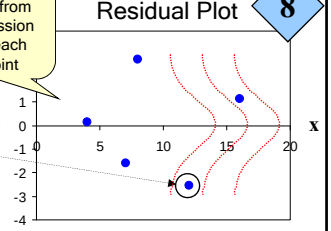
Create a Residual Plot to verify that the OLS assumptions hold:

5

$y = 2.5 + 0.6\,x + \varepsilon$



Plot the distance from the regression line for each data point



8

May still be useful given logical relationship, near significance, and explained variation

Questions to ask:

No 1. Does the residual plot show independence?

No 2. Are the points symmetric about the the x-axis with constant variability for all x?

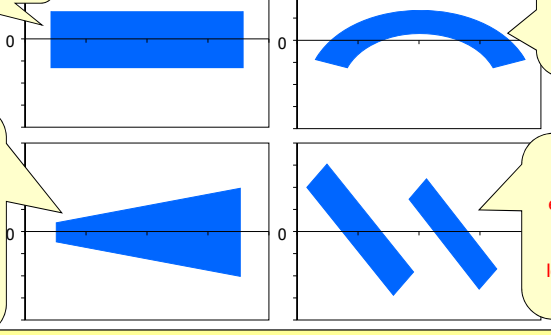
The OLS assumptions are not reasonable. This tells us: The error term ε cannot be modeled with constant variance. The MLE generalization is called for here, but we will proceed with bivariate OLS as if we had passed the White test…

v1.2

Example Residual Patterns

Tip: A residual plot is the primary way of indicating whether a non-linear model (and which one) might be appropriate

Good residual pattern:
- Independent with x
- Constant variation



Residuals **not independent of x**: A curvilinear model is probably more appropriate in this case

Residuals **do not have constant variation**: Weighted Least Squares, Multiplicative Error, or MLE approach should be examined

Residuals **not independent of x**: e.g., in learning curve analysis, this pattern might indicate loss of learning or injection of new work

7

Usually the residual plot provides enough visual insight to determine whether or not linear OLS regression is appropriate. If the picture is inconclusive, statistical tests exist to help determine if the OLS assumptions hold[1].

1 See references to learn about Weighted Least Squares Regression and statistical tests for residuals

v1.2

Excel Demo - Parameters and Residuals

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

RESIDUAL OUTPUT

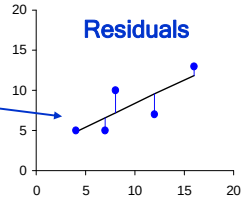
Observation	Predicted y	Residuals	Standard Residuals
1	4.8	0.2	0.1
2	6.6	-1.6	-0.7
3	7.2	2.8	1.3
4	9.5	-2.5	-1.2
5	11.9	1.1	0.5

Equation Parameters

$$\hat{Y} = \hat{a} + \hat{b}\,X = 2.5 + 0.6\,X$$

LINEST() function		
b	0.5849	2.5023a
SEb	0.2636	2.7110SEa
R2	0.6214	2.4612SEE
F	4.9243	3d.f.
SSR	29.8280	18.1720SSE

Residuals



Application

v1.2

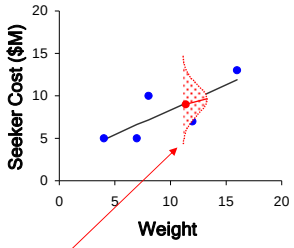
- Suppose our toy problem defines X and Y as follows:
  - Y is the cost of a seeker in a missile
  - X is the weight of the seeker
- Assuming the regression is a good fit[1], the following is true:
  - The cost estimate for a new seeker with weight of *11* is

    2.5 + 0.6*(*11*) = *$9.1M*
  - The result is the estimated mean of the distribution of all possible costs, which is assumed to be a Normal distribution[2]

1 Goodness of Fit is described in the next section
2 Standard deviation of the cost distribution will be discussed later

6

Note: A **seeker** is a component of the missile that is used to detect a target



Tip: The Mean is usually calculated in the cost model. The distribution is usually accounted for in the risk analysis.

9

v1.2

Goodness of Fit

- Analysis of Variance (ANOVA)
- Uncertainty
- Statistical Significance
- Explained Variation

Goodness of Fit – Introduction

v1.2

- The regression procedure produces quantitative measures (statistics) that allow us to determine *goodness of fit*
- We will examine 4 main topics in this section:
  1. Analysis of Variance (ANOVA)
     - Conducted as a basis to calculate most of the statistics
  2. Uncertainty of:
     - The coefficients estimated for the regression equation
     - Estimates made using the regression equation
  3. Statistical significance of:
     - The parameters of the regression equation (t stats)
     - The regression model as a whole (F stat)
  4. Variation explained by the regression (R^2)

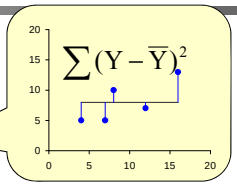
Analysis of Variance (ANOVA)

v1.2

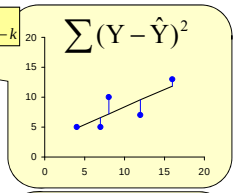
Measures of Variation

7

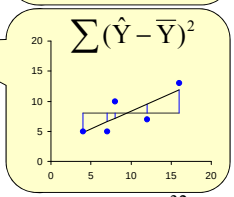
1. **Total Sum of Squares (SST):** $\sim \chi^2_{n-1}$
   The sum of the squared deviations **between the data and the average**



2. **Residual or Error Sum of Squares (SSE):** $\sim \chi^2_{(n-1)-k}$
   The sum of the squared deviations **between the data and the regression line**
   "The unexplained variation"



3. **Regression Sum of Squares (SSR):** $\sim \chi^2_k$
   The sum of the squared deviations **between the regression line and the average**
   "The explained variation"



8

SST = SSE + SSR
"total" = "unexplained" + "explained"

v1.2

Analysis of Variance (ANOVA)

Mean Measures of Variation

- **Mean Squared Error** (or Residual) (MSE):

$$MSE = \frac{SSE}{(n-1) - k}$$

- **Mean of Squares of the Regression** (MSR):

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

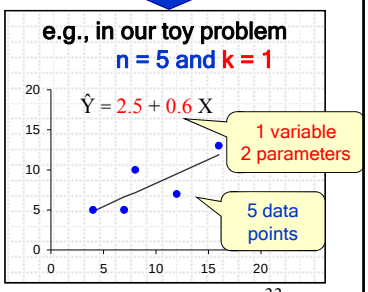
10

The denominator for each of the above is called the *degrees of freedom*, or *df*, associated with each type of variation

Warning: Some sources use *k* (or *p*) to denote the number of *parameters*, including the intercept

n = # data points

k = # independent variables



v1.2

Excel Demo: ANOVA

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

RESIDUAL OUTPUT

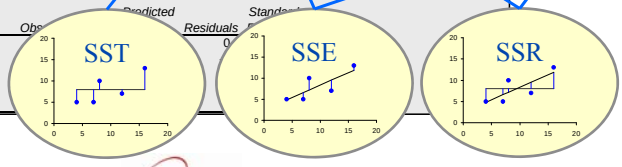
MSR = SSR/df

MSE = SSE/df

LINEST() function

b	0.5849	2.5023 a
SEb	0.2636	2.7110 SEa
R2	0.6214	2.4612 SEE
F	4.9243	3 d.f.
SSR	29.8280	18.1720 SSE
MSR	29.8280	6.0573 MSE

Note: MSR and MSE are not calculated as part of the LINEST function. They have to be calculated using the formulas at the top of the slide.



v1.2

Uncertainty

- It is important to consider the uncertainty that surrounds an estimate
- There are two important types of uncertainty that are quantified in regression analysis
  1. Standard errors of estimated regression coefficients (SE_β)
  2. Standard error of the regression equation (SEE)
- These measures play a key role in:
  - Determining statistical significance
  - Calculating confidence intervals

These topics will be discussed in detail later…

v1.2

Uncertainty of Coefficients

- Each estimated coefficient has an associated standard error

The standard error values are affected by the MSE as well as variability within the x data points

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

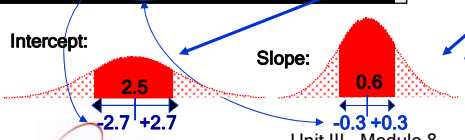
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.2	1.4

As a result of previous assumptions, coefficients are *normally* distributed

$\hat{Y} = 2.5 + 0.6\,X$

LINEST() function

b	0.5849	2.5023a
SEb	0.2636	2.7110SEa
R2	0.6211	2.4612SEE
F	4.9243	3d.f.
SSR	29.8280	18.1720SSE
MSR	29.8280	6.0573MSE



The true coefficients then have a *t* distribution about the estimated values

v1.2

Uncertainty of the Estimate

- The estimated regression equation has a standard error of the estimate, or SEE

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

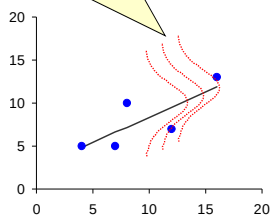
$$SEE = \sqrt{6.1} = 2.5$$

$$SEE = \sqrt{MSE}$$

LINEST() function

b	0.5849	2.5023 a
SEb	0.2636	2.7110 SEa
R2	0.6214	2.4612 SEE
F	4.9243	3 d.f.
SSR	29.8280	18.1720 SSE
MSR	29.8280	6.0573 MSE

The SEE is the estimated standard deviation of the Normal distribution that models the residuals



v1.2

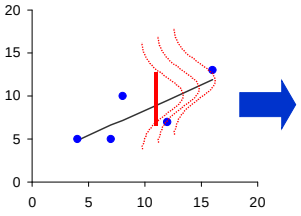
Coefficient of Variation

- The *coefficient of variation*, or *CV* is the ratio of the SEE to the mean of the dependent variable

$$CV = \frac{SEE}{\bar{Y}}$$

Tip: CV of less than 15% is desirable

The mean is the only value not found on the regression output



Example:
From our sample problem, SEE = 2.5 and $\bar{Y}$ = 8, so

CV = 2.5 / 8 = 31%

Note: Using SEE instead of S_y reduces CV, which can be seen in the Advanced Topics section

The CV expresses the standard error of the estimate as a percent (of the mean)

v1.2

Statistical Significance in Regression

- *Statistical significance* means the probability is "acceptably low" that a stated hypothesis is true.
  - The "acceptable" probability is referred to as a *significance level*, or α
  - In regression analysis, the standard hypothesis that is checked is that one or more variables has a coefficient with an actual value of zero
- The statistical tests of interest in Regression Analysis involve the x coefficient(s) in the Regression Equation

10

8

Typical significance levels are α =0.01 and 0.05

Tip: To say b = 0 implies there is no relationship between x and y

$$y = a + bx$$

Are the statistics good enough to convince us that **b** is not zero?

i.e., "Is Probability that the hypothesis "**b** = 0" is true < α ?"

v1.2

The t statistic

- For a regression coefficient, the determination of statistical significance is based on a *t test*
  - The test depends on the ratio of the coefficient's estimated value to its standard error, called a *t statistic*

10

17

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

$$y = a + bx$$

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

We cannot conclude there is a relationship between x and y

Example Setup:

Set α = 0.05

Hypothesis:

$$H_0 : b = 0$$
$$H_a : b \neq 0$$

Test Statistic:

$$t = \frac{\text{Estimated Coefficient}}{\text{Standard Error}} = \frac{0.6}{0.3} = 2.2$$

t stat

p-value: 0.11

Decision: We reject H_0 iff the p-value is less than the chosen significance level (0.05)

Since 0.11 > 0.05 we cannot reject H_o Therefore, this coefficient is not statistically significant

v1.2

t Summary and F Foreshadowing

- The t statistics tell us whether each independent variable is a good predictor
- The F statistic tells us whether the regression as a whole is a good model
  - To be discussed more fully in the multivariate regression section

9

8

Note: In a regression with one independent variable, the F test and t test will yield the same p-value

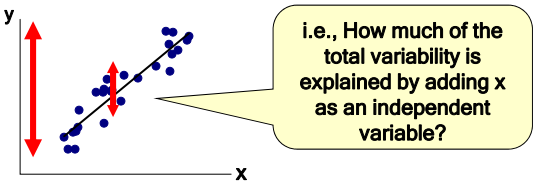
- Our first toy problem was *not* found to be statistically significant
  - This is not surprising due to the small number of data points
  - Could also be due to wrong functional form
- May still be useful given logical relationship, near significance, and explained variation; we will continue to use it for demonstration purposes

8

v1.2

Coefficient of Determination

- The *coefficient of determination, R^2*, is another measure of goodness of fit for a regression
- This statistic represents the fraction of the variability in the dependent variable that is accounted for by the regression



- Higher values of R^2 are better

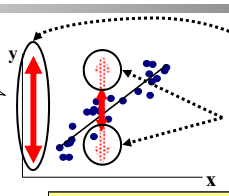
Tip: The value of R^2 is a useful indicator but it is not as important as statistical significance as a criterion for judging the acceptability of the model

Calculating R^2

v1.2

10

How much of the total variability is explained by adding x as an independent variable?

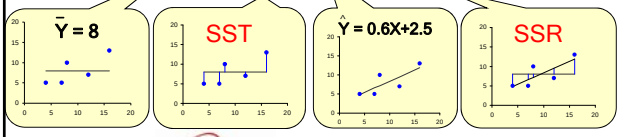


Remember...
"total variation" = SST
"explained variation" = SSR

8

$$R^2 = \frac{explained\ variation}{total\ variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Data		Calculations			
X	Y	$\bar{Y}$	$(Y - \bar{Y})^2$	$\hat{Y}$	$(\hat{Y} - \bar{Y})^2$
4	5	8	9	5	10
7	5	8	9	7	2
8	10	8	4	7	1
12	7	8	1	10	2
16	13	8	25	12	15
			Σ=48		Σ=30



Toy Problem Calculation

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$= 1 - \frac{18.2}{48.0} = 0.62$$

Excel Demo: R^2 and Correlation

v1.2

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.5	2.7	0.9	0.42	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

RESIDUAL OUTPUT

Observation	*Predicted y*	*Residuals*	*Standard Residuals*
1	4.8	0.2	
2	6.6	-1.6	-0.7
3	7.2	2.8	1.3
4	9.5	-2.5	-1.2
5	11.9	1.1	0.5

Correlation : $r = \sqrt{R^2}$

LINEST() function

b	0.5849	2.5023 a
SEb	0.2636	2.7110 SEa
R2	0.6214	2.4612 SEE
F	4.9243	3 d.f.
SSR	29.8280	18.1720 SSE
MSR	29.8280	6.0573 MSE

SST=SSR+SSE

$$R^2 = 1 - \frac{SSE}{SST}$$

Note: r is the sample statistic for correlation ρ. The appropriate sign for r can be determined from the x coefficient.

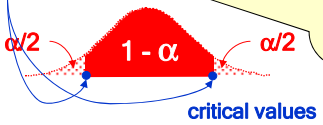
Confidence Intervals

Introduction to Confidence Intervals

- A confidence interval (CI) suggests to us that we are (1-α)*100% confident that the true value of the random variable is contained within the calculated range*
- Confidence intervals are calculated for:
  - Regression Equation Parameters
  - The Regression Equation at a given value of x
- The calculation combines the Student's t distribution with the associated standard error
  - The general formula is:

$$(\text{mean est}) \pm t_{\alpha/2,\ df} \times \text{stderr}$$

10 "t" = **Student's t distribution with (n-1)-k degrees of freedom**. Critical values can be looked up on a table or calculated in Excel



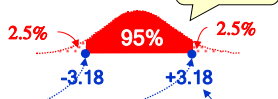
* Note this statement provides a general sense of what a confidence interval does for us in concise language for ease of understanding. The specific statistical interpretation is that if many independent samples are taken where the levels of the predictor variable are the same as in the data set, and a (1-α)*100% confidence interval is constructed for each sample, then (1-α)*100% of the intervals will contain the true value of the parameter (or value of the dependent variable at a given value of x, depending on which interval is being calculated).

Confidence Intervals
of Regression Coefficients

v1.2

- We will use our example problem to demonstrate how to calculate a 95% Confidence Interval for **b**:

t distr with 3 df



Example:

$\hat{Y} = 2.5 + 0.6\,X$

$\alpha = 0.05$

$df = 5 - 1 - 1 = 3$

So we can say with 95% confidence that the true value for b is contained in this range

Confidence Interval:

$$= \hat{b} \pm t_{\alpha/2,\,df} \times \text{stdev}$$
$$= 0.6 \pm t_{.025,3} \times 0.3$$
$$= 0.6 \pm 3.18 \times 0.3$$
$$= (-0.3, 1.4)$$

In Excel: "=T.INV.2T(0.05,3)"

Excel provides a 95% confidence interval for each parameter

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79
R Square	0.62
Adjusted R Square	0.50
Standard Error	2.5
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.8	29.8	4.9	0.11
Residual	3	18.2	6.1		
Total	4	48			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.5	2.7	0.9	0.41	-6.1	11.1
x	0.6	0.3	2.2	0.11	-0.3	1.4

Tip: This result is directly related to the t test for the x-coefficient. It is always true that if the confidence interval includes zero, then the coefficient is *not* statistically significant

CI for b	VALUE	FORMULA
Point Estimate	0.58	From LINEST(.)
Confidence Level	95%	User input
Half-Width	0.84	TINV(1-.95,3)*0.26
Lower Bound	-0.25	PE - HW
Upper Bound	1.42	PE + HW or LB + 2*HW
Includes Zero?	Yes	=IF(AND(LB<0,UB>0),"Yes","No")
Significant?	No	IF(B33="Yes","No","Yes")

Confidence Intervals
of Estimates

v1.2

- To calculate a confidence interval for the mean value of Y at a given value of X we use the following formula:

11

8

$$\hat{Y} \pm t_{\alpha/2,\,df} \times SEE\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}} \quad \leftrightarrow \quad \sqrt{\left(\frac{SEE}{\sqrt{n}}\right)^2 + \left(se_b \cdot (X - \bar{X})\right)^2}$$

Standard error of ŷ for a fixed X

"bounce"

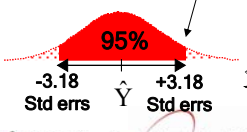
"wiggle"

Example:

The t-dist critical value for $\alpha = 0.05$, $df = 3$ is 3.18

This means +/- 3.18 std errors forms a 95% confidence interval for the mean in our toy problem

=DEVSQ(X)



Problem: Find a 95% confidence interval for ŷ at X = 11

$$= [a + b(X)] \pm 3.18 \times 2.5\sqrt{\frac{1}{5} + \frac{(X - 9.4)^2}{529 - 5*9.4^2}}$$

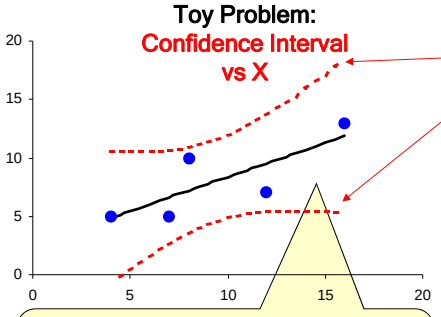
$$= [2.5 + 0.6(11)] \pm 3.18 \times 2.5\sqrt{\frac{1}{5} + \frac{(11 - 9.4)^2}{529 - 5*9.4^2}}$$

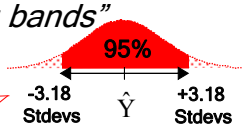
$$= 9 \pm 3.8 \rightarrow = (5.2,\ 12.8)$$

v1.2

Confidence Bands

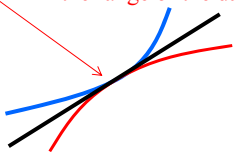
- When we plot the confidence intervals for all X, we get a pair of hyperbolic curves called the "*Confidence Bands" or "Working-Hotelling bands"*





Note that the interval becomes larger as X moves away from the mean

Warning: CI only accounts for uncertainty in model parameters. If the model form is wrong, more severe departures could occur outside the range of the data



v1.2

Prediction Intervals

16

9

- It is more common that we would want a confidence interval for a new observation of Y at a given value of X
  - This is referred to as a *Prediction Interval* and has the following formula:

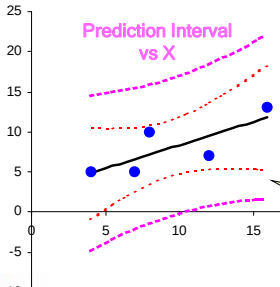
$$\hat{Y} \pm t_{\alpha/2,\,df} \times SEE \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}} \quad \Leftrightarrow \quad \sqrt{(SEE)^2 + \left(\frac{SEE}{\sqrt{n}}\right)^2 + \left(se_b \cdot (X - \bar{X})\right)^2}$$

Standard error of **Y** for a fixed X

"fuzz" or "noise"

"bounce"

"wiggle"



Tip: The Prediction Interval provides an interval for y as opposed to ŷ and should be used for risk analysis

The t distribution critical values remain the same as previously calculated

Note that prediction intervals are wider than confidence intervals

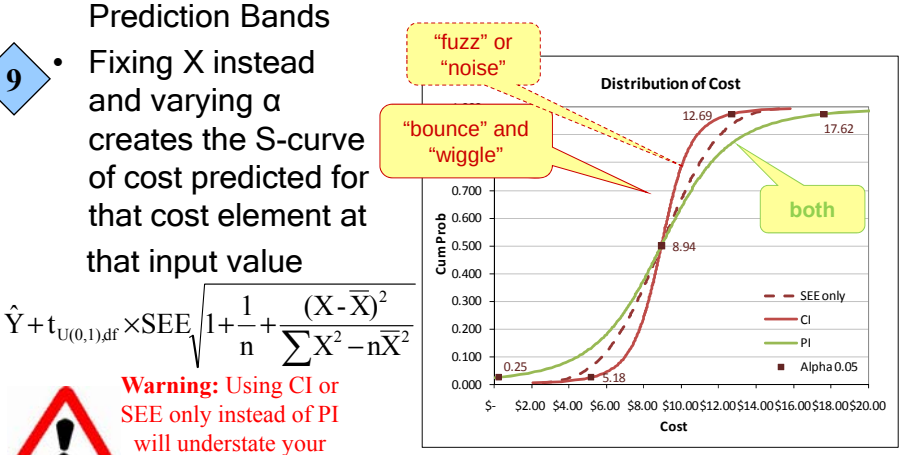
v1.2

From PI to S-Curve

- The PI formula for fixed α and varying X creates the Prediction Bands
- Fixing X instead and varying α creates the S-curve of cost predicted for that cost element at that input value

9

$$\hat{Y} + t_{U(0,1),df} \times SEE \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$



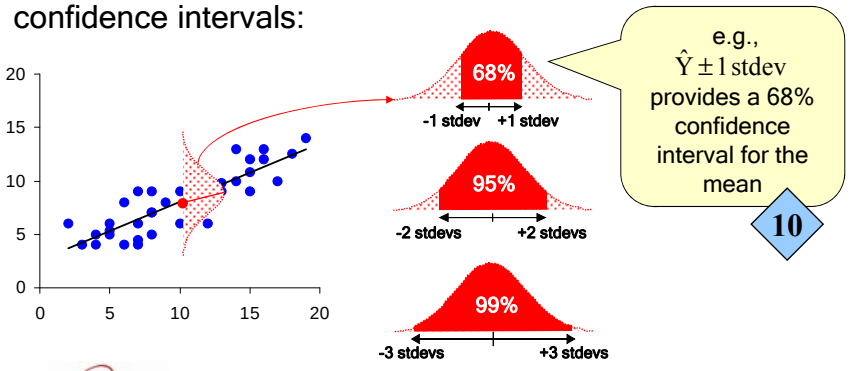
Warning: Using CI or SEE only instead of PI will understate your uncertainty!

NEW!

v1.2

Confidence Intervals when n is large

- When n is large (at least 30), the t distribution approaches a Normal distribution
- This means critical values can be approximated by the Normal, which produces the following standard confidence intervals:



e.g., $\hat{Y} \pm 1\,\text{stdev}$ provides a 68% confidence interval for the mean

10