2015 ICEAA Professional Development & Training Workshop
June 09-12, 2015 • San Diego, California

Advanced Probability and Statistics

Use and Application of Probability
and Statistics Concepts

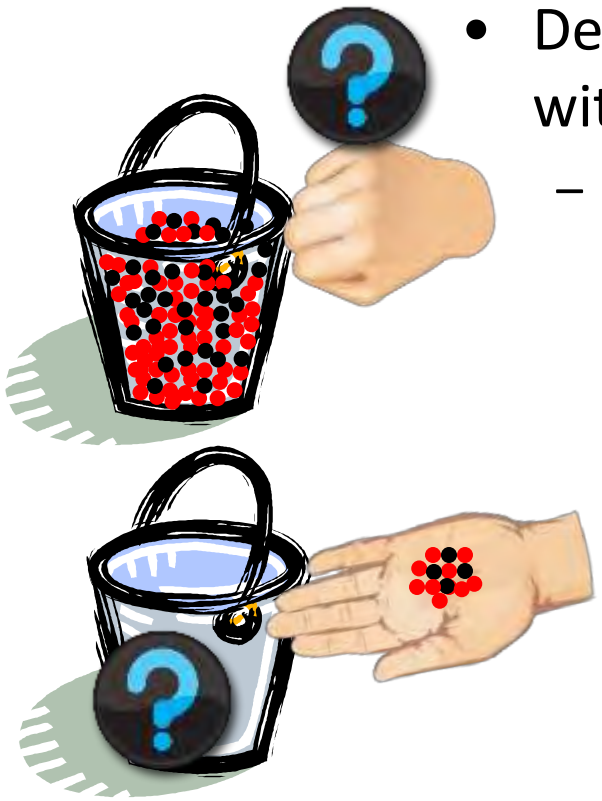
Agenda

- Review of Basic Probability/Stats Concepts
- Application / Use of Probability Distributions
- Advanced Topics
  - Ordinary Least Squares (OLS)
  - Outlier Analysis
  - Tests for Veracity of OLS Assumptions
  - Maximum Likelihood Estimation (MLE)
  - Method of Moments (MoM)

Review Topics

- Measures of Central Tendency
  - Mean, median, and mode
- Measures of Dispersion
  - Variance, standard deviation, and coefficient of variation
- Probability Density Function
- Cumulative Distribution Function
- Random Variables
  - Discrete and continuous

Review—Definitions



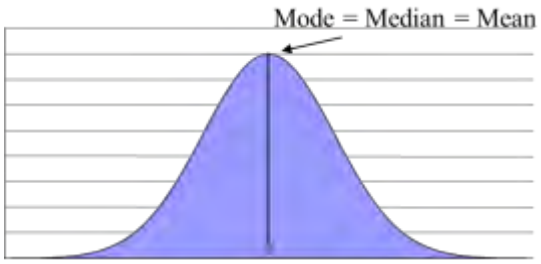
- Probability and Statistics are two sides of the same coin
- Definitions in terms of the classic "Pail filled with marbles of varying colors":
  - Probability: Given the color density of marbles in the pail, what is the likelihood you have red marbles in your hand if you've drawn a random sample?

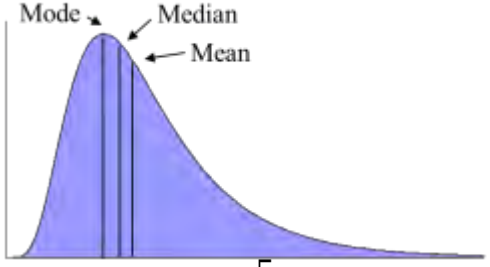
Statistics: Given the number of red and black marbles you randomly drew, what is the color density of the marbles in the pail?

Hat top: Glen B. Alleman at http://herdingcats.typepad.com/my_weblog/2013/03/probabilistic-cost-and-schedule-processes.html

Review - Measures of Central Tendency

- These statistics describe the "middle region" of the distribution
  - Mean
    - The weighted average, expected value, of the distribution
    - Includes all values or observations; affected by outliers (extreme low or high values of data in the sample or population)
  - Median
    - The "middle" of the distribution; value exactly divides the distribution into equal halves
    - Does not includes all values or observations, only ranks; not affected by outliers (extreme low or high values of data in the sample or population)
  - Mode
    - The value in the distribution that occurs most frequently
    - Includes all values or observations





Review- Measures of Dispersion

- Variance
  - Measure of how far a set of numbers is spread out
  - Describes how far the numbers lie from the mean
  - Mean squared deviations of a value from the mean

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

- Standard Deviation
  - Square root of the variance
  - Measures amount of variation of values around the mean

$$s = \sqrt{s^2}$$

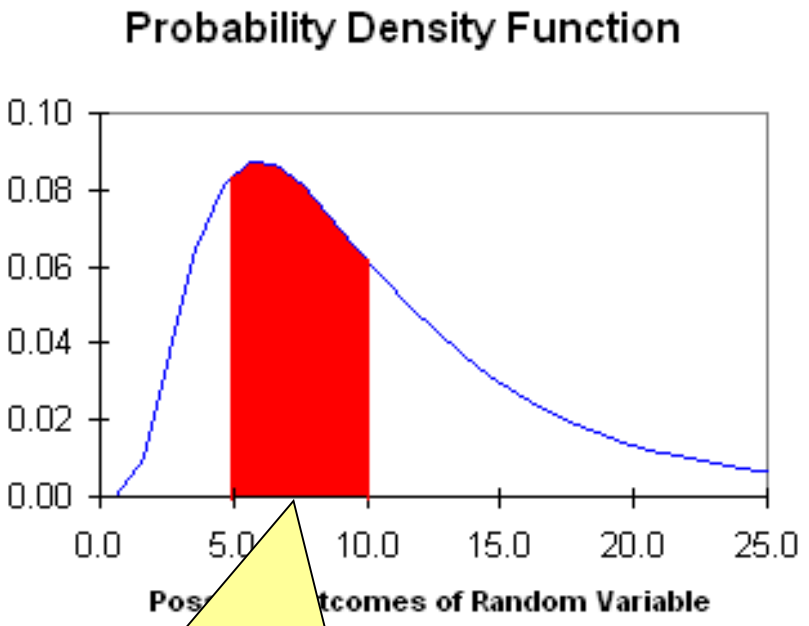
- Coefficient of Variation (CV)
  - Normalized measure of dispersion
  - Expresses the standard deviation as a percent of the mean
  - Large CVs indicate that mean is a poor representation of the distribution

$$CV = \frac{s}{\bar{X}}(100)$$

Review – PDF

- Probability Density Function (PDF)
  - Total area under curve is 1
    - Probability that random variable takes on some value in the range is 1
  - Probability that a specific random variable will take a value between *a* and *b* is the area under the curve between *a* and *b*

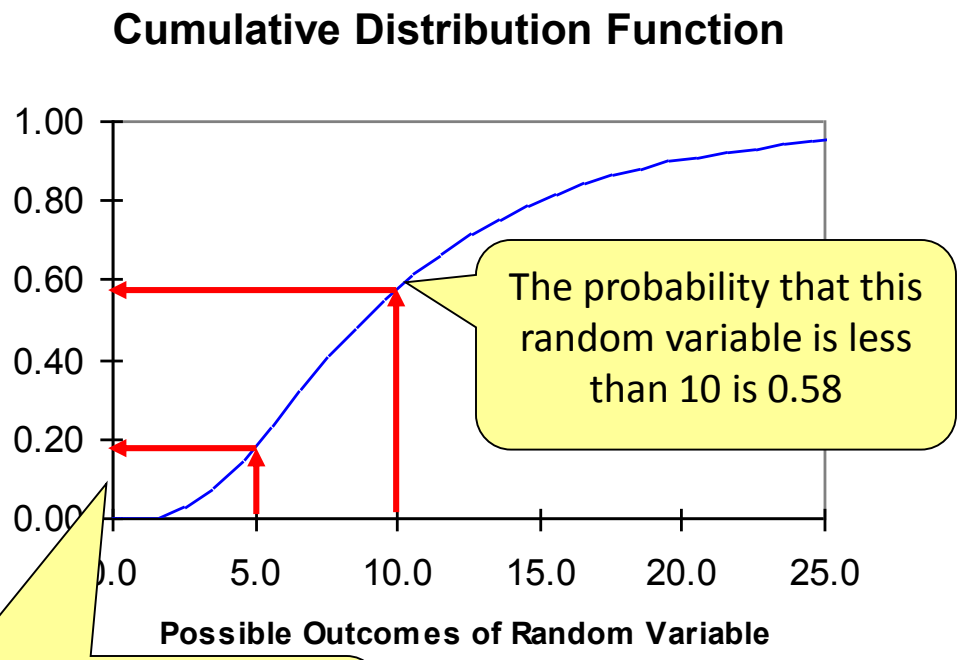
$$P(a \le X \le b) = \int_a^b p(x)dx$$



Probability that random variable is between 5 and 10 is equal to shaded area (40%)

Review– CDF

- Cumulative Distribution Function (CDF)
  - Computes cumulative probabilities across the range of *X,* from left to right
  - This curve shows probability that the random variable is *less* than x (a particular value of X)
  - The CDF reaches/ approaches 1
    - Probability that the random variable is less than the maximum value (may be infinite) is 1 (100%)



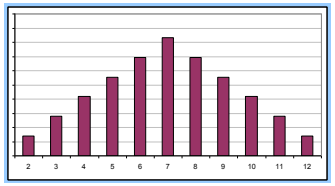
$$P(a \le X \le b) = F(b) - F(a)$$

$$= \int_{-\infty}^{b} p(x)dx - \int_{-\infty}^{a} p(x)dx = \int_{a}^{b} p(x)dx$$

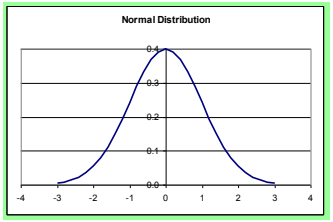
PDF shows the *shape* of distribution
CDF shows the *percentiles*

Random Variables

- A random variable takes on values that represent outcomes in the sample space
  - It cannot be fully controlled or exactly predicted
- Discrete vs. Continuous
  - A set is discrete if it consists of a finite or countably infinite number of values



    - e.g., number of circuits, number of test failures
    - e.g., {1,2,3, …} – the random variable can only have a positive integer (natural number) value
  - A set is continuous if it consists of an interval of real numbers (finite or infinite length)

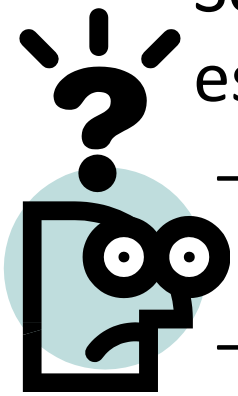


    - e.g., time, height, weight
    - e.g., [-3,3] – the random variable can take on any value in this interval

Probability Distributions

Application / Use

Using Probability Distributions



- So how are probability distributions useful in cost estimating?
  - The need for an estimate implies use of probability and statistics
  - Need to know underlying statistical behaviors to speak about confidence intervals, not single point (deterministic) values
    - We forecast to complete at or below $260M with 75% confidence
    - Our Estimate to Complete for the remaining work is $267,000 with a 85% confidence.
  - With probability and statistics foundation, credible estimates driven by underlying statistical and resulting probabilistic (stochastic) processes, are achievable

Sources of Estimate Uncertainty

- Uncertainty is indefiniteness about outcome of a situation
- Is assessed in cost models to estimate the likelihood that a specific funding level will be exceeded
- Uncertainties come from lack of knowledge about the future; historical data inaccuracies; unclear requirements; misuse, misrepresentation, or misinterpretation of estimating data; misapplied assumptions or estimating methods; or intentionally ignoring probabilistic nature of work

Generating Random Variables

- Models involving random occurrences require the specification of a probability distribution for each random variable
- To aid in the selection, a number of named distributions have been identified (i.e., Normal, Uniform, Bernoulli, etc.)
- The named distribution is specified by the mathematical statement of the probability density function (PDFs)
- Each has one or more parameters that determine the shape and location of the distribution *(μ, σ, a, b,* etc.)
- Cumulative distributions may be expressed as mathematical functions; or, for cases when integration is impossible, extensive tables are available for evaluation of the CDF
- Special software can be used, but many times sufficient functionality exists in MS Excel
  - Examples follow

Inverse Transform Technique

- Representing uncertainty in cost estimating frameworks is frequently achieved through the use of random variables
- Inverse Transform technique is commonly applied in generation of random variables
  - CDF maps a value (x) to probability that random variables takes on a value less an or equal to x
  - Inverse of CDF maps a probability (0 -1) to a value from specified distribution
  - Generating uniform (0,1) random number allows calculation of value from any distribution with an invertible CDF
  - In Excel, RAND() function generates uniform random number*
- The generated random variables from identified probability distributions are then used in a Monte Carlo method to develop estimates

*Or tables available at: http://www.nist.gov/pml/wmd/pubs/upload/AppenB-HB133-05-Z.pdf

Generating RVs in MS Excel (1 of 2)

Distribution	Formula
Normal	With mean, μ, and standard deviation, σ $= NORMINV(RAND(), \mu, \sigma)$ With mean, 0, and standard deviation, 1 $= \mu + NORMSINV(RAND()) * \sigma$
Lognormal	With mean, μ, and standard deviation, σ $= LOGINV(RAND(), \mu, \sigma)$
Triangular	With minimum, *a*, maximum, *b*, and most likely, *c* $r = RAND()$ $= IF(r \le \frac{(c-a)}{(b-a)}, a + \sqrt{r * (b-a) * (c-a)}, b - \sqrt{(1-r) * (b-a) * (b-c)}$

Generating RVs in MS Excel (2 of 2)

Distribution	Formula
Uniform	With lower bound, *a*, and upper bound, *b* Discrete: $= RANDBETWEEN(a,b)$ Continuous: $= a + (b-a) * RAND()$
Bernoulli	With *p* as the probability of X=1 and *1-p* as the probability of X=0 $= IF(RAND() < p,1,0)$
Empirical	Build a table structure with: possible values (*x*), frequency (likelihood) of each possible value, and cumulative likelihood (*r*). Generate a random variable (R), look-up in the table to find the value of cumulative likelihood so that: $r_{i\leq}R \leq r_{i+1}$ Random variable then will be the corresponding value of *x* in the identified interval (use LOOKUP function)

Use of Probability Distributions Summary

- MS Excel provides a straightforward method to generate probabilistic estimates for simple problems
- Other methods / tools exists
- Only representative methodology has been shown here given the usual availability of MS Excel to most practitioners
- References for other tools and for generating random variables for other distributions are easily locatable through Internet searches (or statistics textbooks!)

Related and Advanced Topics

- Ordinary Least Squares (OLS)
- Outlier Analysis
- Tests for Veracity of OLS Assumptions
- Maximum Likelihood Estimation (MLE)
- Method of Moments (MoM)

Introduction

- Recall, a sample is a collection of observations
- A statistic is any function of the sample (e.g., the sample mean and sample variance)
  - Typically used to estimate an unknown (population) parameter of the *distribution* of the observations
  - Such a statistic is called an *estimator* of the population parameter
  - For example, the sample variance (S^2) is often used as an *estimator* of the population variance, but it is not the only estimator for population variance that exists:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

  - Similarly, there are multiple methods for calculating estimators of the parameter(s) of a distribution

Related and Advanced Topics

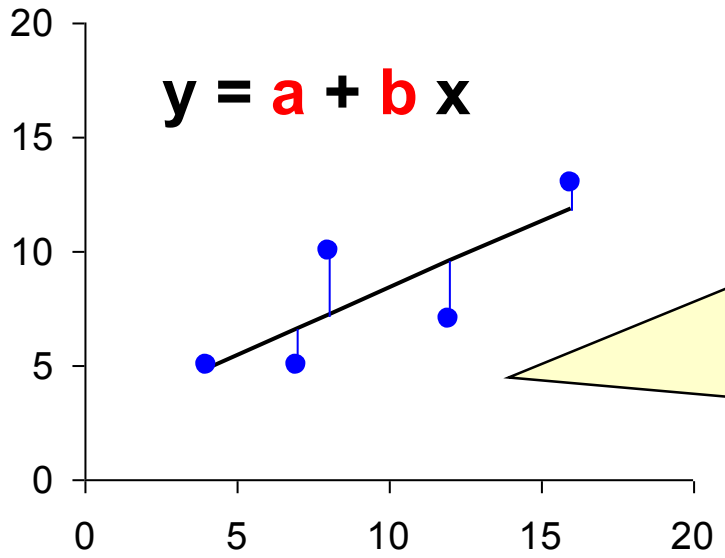
- **Ordinary Least Squares (OLS)**
- Outlier Analysis
- Tests for Veracity of OLS Assumptions
- Maximum Likelihood Estimation (MLE)
- Method of Moments (MoM)

Ordinary Least Squares (OLS)

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"



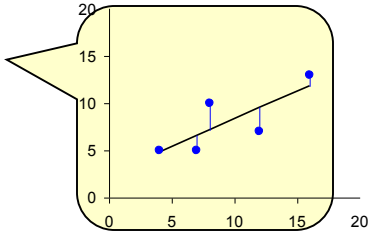
In other words… Find the equation that minimizes the sum of the squared distances of the blue lines

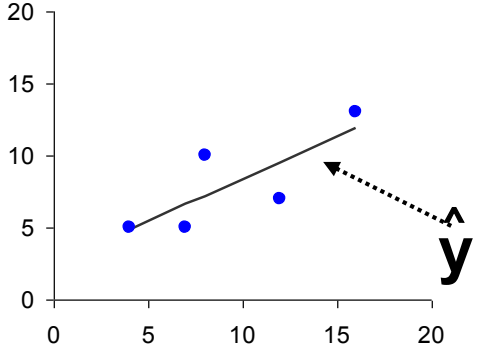
8

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 in Module 08

Finding the Regression Equation

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





$$\hat{y} = \hat{a} + \hat{b}\,x$$

- 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:

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

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

Related and Advanced Topics

- Ordinary Least Squares (OLS)
- **Outlier Analysis**
- Tests for Veracity of OLS Assumptions
- Maximum Likelihood Estimation (MLE)
- Method of Moments (MoM)

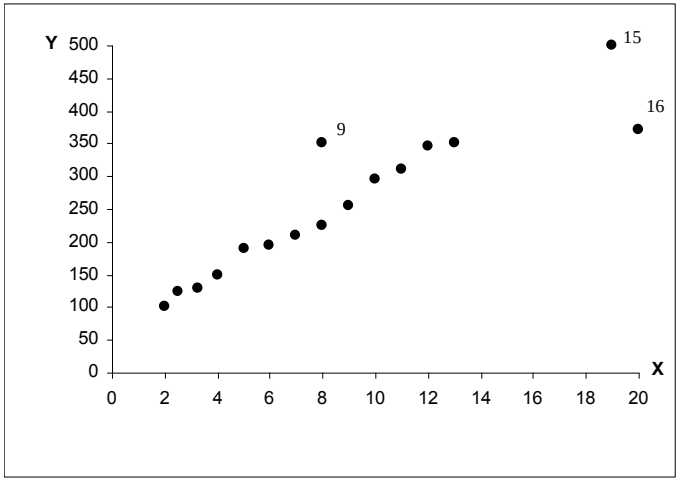
Outlier Analysis

- **The underlying principle of outlier analysis is to:**
  - detect whether a small minority of data observations (e.g. 3 or less) have an unusual amount of influence on the regression line, and
  - apply techniques to mitigate this "unusual" amount of influence
- **Determining what is deemed "an outlier" does require some judgment on the part of the analyst.**
  - For example, there is no true consensus in the cost community on "outlier" thresholds for (X, Y) values. Some analysts prefer 2 standard deviations from the mean, others prefer 3 standard deviations.
  - We deal with similar challenges with other statistical measures such as lowest acceptable t-stat, R-threshold for determining multicollinearity, "most preferred" confidence level, etc.

Note: Examples in this outlier analysis section assume data that's normally distributed. The last slide of this section summarizes other methods that tend to also work reasonably well for data not normally distributed.

Outlier Analysis

- **Outliers can have a significant effect on the regression coefficients**
- **Which points on the graph would you predict to be influential observations?**
- **How do we tell? 3 ways**
  - Outliers w/ respect to X
  - Outliers w/ respect to Y
  - Outliers w/ respect to Y_x
- **A best practice is to test each data point in all 3 ways**
  - Test each as a possible outlier w/respect to X, Y and/or Y_x



Assume for our example, that we are evaluating a CER where $Y_x = b_1 + b_2 X$

Outliers with Respect to X: # Std. Devs

- **All data should be from the same population**
  - Assumes data is normally distributed
- **Analyze observations**
  - Based on the values of X for each data point, are there any data points that look very different from the rest?
- **How to identify potential outliers with respect to X**
  - Calculate mean and standard deviation of X_i values in the dataset
  - Divide the difference between each X_i and $\overline{X}$ by the S_x

$$\text{\# Std Deviations} = \frac{(x_i - \overline{x})}{s_x}$$

s_x ← Standard Deviation of X data

  - Identify observations that fall more than 2 standard deviations from the mean (or 3 standard deviations from the mean, if preferred)

Outliers with Respect to X: # Std. Devs

- Calculated mean and standard deviation of X_i values in the dataset
- Divided the difference between each X_i and $\overline{X}$ by the s_x ----->

ID	X_i	$\overline{X}$=Mean	$x_i-\overline{x}$	$(x_i-\overline{x})^2$	$(x_i-\overline{x})/s_x$
1	2	8.73	-6.73	45.2929	-1.256
2	2.5	8.73	-6.23	38.8129	-1.163
3	3.2	8.73	-5.53	30.5809	-1.032
4	4	8.73	-4.73	22.3729	-0.883
5	5	8.73	-3.73	13.9129	-0.696
6	6	8.73	-2.73	7.4529	-0.509
7	7	8.73	-1.73	2.9929	-0.323
8	8	8.73	-0.73	0.5329	-0.136
9	8	8.73	-0.73	0.5329	-0.136
10	9	8.73	0.27	0.0729	0.050
11	10	8.73	1.27	1.6129	0.237
12	11	8.73	2.27	5.1529	0.424
13	12	8.73	3.27	10.6929	0.610
14	13	8.73	4.27	18.2329	0.797
15	19	8.73	10.27	105.4729	1.917
16	20	8.73	11.27	127.0129	2.103
Σ =	139.7			430.7344	
n =	16			28.72	
mean =	8.73125		std dev =	5.36	

Mean of X_i 's = 8.73

Std Dev of X_i 's = 5.36

Standard Deviations from $\overline{X}$ = $\frac{(X_i - 8.73)}{5.36}$

X_{15} = 19 and X_{16} = 20 are about 2 standard deviations from $\overline{X}$

Therefore, 2 of the 16 observations are potential outliers.

Outliers with Respect to X: **Leverage**

- **Leverage Value is one indicator on the degree of influence a given X_i may have on regression coefficients**
- **Looks at how much influence an observation could have on the coefficients of the regression equation**
  - Leverage values sum up to p (# of parameters)
  - Average leverage value = p/n (n = # of observations)
  - An observation is considered a potential outlier with respect to X if its leverage value is greater than 2(p/n) to 3(p/n)

$$\text{Leverage} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}$$

Outliers with Respect to X: Leverage

p = 2 & n = 16, (p/n) = 2/16 = 0.125. 2(p/n) = **0.25**. 3(p/n) = **0.375**., $\bar{X}$ = 8.73 & SD of X_i 's = 5.36

ID	X	$(x_i-\bar{x})^2$	LV*	2 (p/n)**	2.5 (p/n)	3 (p/n)
1	2	45.29	0.168	0.25	0.3125	0.375
2	2.5	38.81	0.153	0.25	0.3125	0.375
3	3.2	30.58	0.133	0.25	0.3125	0.375
4	4	22.37	0.114	0.25	0.3125	0.375
5	5	13.91	0.095	0.25	0.3125	0.375
6	6	7.45	0.080	0.25	0.3125	0.375
7	7	2.99	0.069	0.25	0.3125	0.375
8	8	0.53	0.064	0.25	0.3125	0.375
9	8	0.53	0.064	0.25	0.3125	0.375
10	9	0.07	0.063	0.25	0.3125	0.375
11	10	1.61	0.066	0.25	0.3125	0.375
12	11	5.15	0.074	0.25	0.3125	0.375
13	12	10.69	0.087	0.25	0.3125	0.375
14	13	18.23	0.105	0.25	0.3125	0.375
15	19	105.47	0.307	0.25	0.3125	0.375
16	20	127.01	**0.357**	0.25	0.3125	0.375
Σ =	139.7	430.73				
n =	16	28.72	**= variance**			
mean =	8.73125	5.36	**= std deviation**			
p =	2					

Recall from Slide 30 that we have a CER with 2 population parameters, Therefore, p = 2.

X_{15} = 19 and X_{16} = 20 have leverage values exceeding 2(p/n).

Therefore, these 2 observations are potential outliers.

* Example for X_1 : $LV = \frac{1}{n} + \frac{(Xi-\bar{X})^2}{\sum(Xi-\bar{X})^2} = \frac{1}{16} + \frac{45.29}{430.73} = .063 + .105 = .168$

Outliers with Respect to Y and Y_x

- **Outliers with Respect to Y: Use same method for Outliers with Respect to X**
  - Refer to the 2 methods shown in Slides 31 – 34, but instead apply to Y_i
- **Outliers with respect to Y_x : These represent observations that the model doesn't predict well**
  - The further the observation is from the regression line, the larger the estimating error
  - Approaches in evaluating size of residual
  - Compare with the standard error of the estimate (SE, SEE, s_{yx}): *Standardized Residual*
  - Individual variance on the residual: *Studentized Residual*

Outliers with Respect to Y_x

- **Observations that are not predicted well by the regression equation**
  - Calculate predicted cost and standard error of the dataset
  - Calculate difference between each Y_i and Y_x and divide by the standard error of Y_x (denoted as S_{Yx})

$$\text{\# of Standard Errors} = \frac{(Y_i - Y_x)}{S_{Yx}}$$

S_{Yx} ← Standard Error of Y_x Data

  - Identify observations that fall more than 2 standard errors from the calculated Y_x (or 3 standard errors from Y_x, if preferred)

Outliers with Respect to Y and Y_x

Evaluating "flagged" Obs. #9 and #16 by calculating **Standard Deviations** and **Standard Errors**

Evaluate actual Y's *

Standard Deviations from $\bar{Y}$ = $\frac{(Y_i - \bar{Y})}{S_Y}$

St Devs = $\frac{(Y_i - 245.9)}{108.25}$

St Devs for Y_9 = $\frac{(345 - 245.9)}{108.25}$

St Devs for Y_9 = 0.915

St Devs for Y_{16} = $\frac{(350 - 245.9)}{108.25}$

St Devs for Y_{16} = 0.961

Obs	Given X	Given Actual Y Y	Estimated or Calculated Y Y_x	$(Y_x - Y)$ Error e_i Residual	$(Y_x - Y)^2$ Square Error e_i^2 Residual^2
1	2	100	122.4	-22.39	501
2	2.5	125	131.6	-6.56	43
3	3.2	130	144.4	-14.41	208
4	4	140	159.1	-19.10	365
5	5	180	177.5	2.55	7
6	6	185	195.8	-10.81	117
7	7	200	214.2	-14.16	201
8	8	205	232.5	-27.52	757
9	8	345	232.5	112.48	12,652
10	9	240	250.9	-10.87	118
11	10	280	269.2	10.77	116
12	11	290	287.6	2.42	6
13	12	330	305.9	24.07	579
14	13	335	324.3	10.71	115
15	19	500	434.4	65.58	4,301
16	20	350	452.8	-102.77	10,562
	Mean of Y =	245.9	245.9	SSE =	30,646
	SE of Y =	108.25	Mean of Y-hat	SE of Y_x =	46.79

Evaluate calculated Y_x's *

Standard Deviations from Y_x = $\frac{(Y_i - Y_x)}{S_{Y_x}}$

St Errors for Y_9 = $\frac{(345 - 232.5)}{46.79}$

St Errors for Y_9 = 2.404

St Errors for Y_{16} = $\frac{(350 - 452.8)}{46.79}$

St Errors for Y_{16} = -2.196

Using the ± 2 std dev rule, neither observation is an outlier with respect to Y

Using the ± 2 std dev rule, both observations ARE outliers with respect to Y_x

* Note 1: S_Y = SQRT($\Sigma(Y_i - \bar{Y})^2 / (n-1)$) = SQRT ((175,761 / (16 -1)) = 108.25
* Note 2: S_{Y_x} = SQRT($\Sigma (Y_i - Y_x)^2 / (n-p)$) = SQRT ((30,646 / (16 - 2)) = 46.79

Outliers with Respect Y_x

Highlighting steps to calculate **Leverage (LV)** and **Studentized Residual (e_i^*)** for Obs. #16

		Evaluation of Y_x					
Obs	Estimated or Calculated Y Y_x	$(Y_x - Y)$ *Error* e_i Residual	$(Y_x - Y)^2$ *Square Error* e_i^2 Residual^2	The Square of Calculated Ys vs Mean of Calculated Ys	Sum of LV = p = *2.00* Leverage LV	Sq root of unbiasd estim of variance $s\{e_i\}$	$e_i^* = e_i / s\{e_i\}$ Internally Studentized Residual
1	122.4	-22.39	501	15,264	0.1677	42.7	-0.52
2	131.6	-6.56	43	13,082	0.1527	43.1	-0.15
3	144.4	-14.41	208	10,308	0.1335	43.6	-0.33
4	159.1	-19.10	365	7,541	0.1145	44.0	-0.43
5	177.5	2.55	7	4,691	0.0948	44.5	0.06
6	195.8	-10.81	117	2,513	0.0798	44.9	-0.24
7	214.2	-14.16	201	1,010	0.0695	45.1	-0.31
8	232.5	-27.52	757	180	0.0637	45.3	-0.61
9	232.5	112.48	12,652	180	0.0637	45.3	**2.48** **R**
10	250.9	-10.87	118	24	0.0627	45.3	-0.24
11	269.2	10.77	116	543	0.0662	45.2	0.24
12	287.6	2.42	6	1,734	0.0744	45.0	0.05
13	305.9	24.07	579	3,599	0.0873	44.7	0.54
14	324.3	10.71	115	6,139	0.1048	44.3	0.24
15	434.4	65.58	4,301	35,525	0.3073	38.9	1.68
16	452.8	-102.77	10,562	42,779	0.3573	37.5	**-2.74** **R, D**
	245.9	*SSE =*	*30,646*	***145,112***	**MSE** = SSE =	30,646	= ***2,189***
	Mean of Y-hat	*SE of Y_x =*	*46.79*	*Sum*	n- k -1	14	

R: an observation w/an unusual Dependent variable D: an observation w/an unusual Cook's D-statistics value

$$LV = \frac{1}{n} + \frac{(Y_x - \bar{Y}_x)^2}{\sum(Y_x - \bar{Y}_x)^2}$$

$$LV = \frac{1}{16} + \frac{(452.8 - 245.9)^2}{145{,}112}$$

$$LV = 0.0625 + 0.2948$$

$$LV = 0.3573$$

$$s^2\{e_i\} = MSE\,(1 - LV)$$

$$s^2\{e_i\} = 2{,}189\,(1 - 0.3573)$$

$$s^2\{e_i\} = 1{,}406.9$$

$$s\{e_i\} = 37.5$$

$$e_i^* = e_i / s\{e_i\}$$

$$e_i^* = -102.77 / 37.5$$

$$\mathbf{e_i^* = -2.74}$$

Internally Studentized Residual for observation #16

As already noted in slide 34, obs. #16 has a Leverage > 2(p/n). Driven by its high e_i, obs. #16 has an e_i^* > 2 std dev (unusual Y_x)

Outliers with Respect to Y_X

Observations Influencing the Regression Coefficients

An observation is considered **influential** by having:

1. a moderate leverage value and a large residual,
2. a large leverage value and a moderate residual, or
3. a large leverage value and a large residual.

Cook's Distance (Cook's D) is a statistic that is commonly used to determine if an observation is influential.

- The distance an observation would be from a regression equation built with this observation omitted from the dataset.

$$D_i = \left[\frac{e_i^2}{p(MSE)} \right] \left[\frac{\text{Leverage}}{(1 - \text{Leverage})^2} \right]$$

$p = \#$ of population parameters in the equation

MSE = MSE from the equation with all the observations

If Cook's D > 50th percentile of the F distribution for (p, n-p) degrees of freedom, then the observation is considered influential.

Outliers with Respect to Y_X

Observations Influencing the Regression Coefficients

Calculation of Cook's D statistic for observation #16:

$$D_i = \left[\frac{e_i^2}{p(MSE)}\right] \left[\frac{\text{Leverage}}{(1 - \text{Leverage})^2}\right]$$

p = # of population parameters in the equation = 2

MSE = MSE from the equation with all the observations = 2,189

$$D_i = \left[\frac{-102.77^2}{2(2{,}189)}\right] \left[\frac{0.3573}{(1-0.3573)}\right] = 2.412 \text{ x } 0.556 = \mathbf{1.341}$$

Lookup the **50th percentile** of the F distribution for (p, n-p) degrees of freedom:

- F distribution (2, 16-2) degrees of freedom = **F distribution (2, 14) = 0.729**
  - Excel's F.INV function provided this reference value for F (a =0.50, numerator = 2, denominator =14)

Therefore, evaluating observation #16: Cook's D > $F_{(0.50, 2, 14)}$

1.341 > 0.729

Cook's D indicates that observation #16 is influential (aka "an unusual value")

What if You Find an Outlier?

Part 1: Evaluate Outlier with respect to X or Y

A. Investigation

- Do you have the right value for the observation?
- Has the observation been normalized correctly?
- Is the observation part of the population?
- How different is the outlier?
- Were there any unusual events that impacted the value of the observation?

B. Actions based upon results of Investigation

- Correct data entry errors
- Improve normalization process
- Remove data point if not part of population
- Determine if unusual program events make a difference

Part 2: Outlier with respect to Y_x (note: do this after completing part 1)

A. Investigation

- Did you choose the correct functional form?
- Are there any omitted cost drivers?
- Was the same criteria applied to all outliers?

B. Actions based upon results of Investigation

- Add another cost driver and/or choose another functional form
- Dampen or lessen Y_x influence by transforming X or Y data
- Create and compare two regression equations:
  - One with and one without the outlier(s)

Other Outlier Detection Methods

- Median and Median Absolute Deviation Method (MAD)
  - For this outlier detection method, the median of the residuals is calculated. Then, the difference is calculated between each historical value and this median. These differences are expressed as their absolute values, and a new median is calculated and multiplied by an empirically derived constant to yield the MAD.
  - If a value is a certain number of MAD away from the median of the residuals, that value is classified as an outlier. The default threshold is 3 MAD.
    - This method is generally more effective than the mean and standard deviation method for detecting outliers, but it can be too aggressive in classifying values that are not really extremely different. Also, if more than 50% of the data points have the same value, MAD is computed to be 0, so any value different from the residual median is classified as an outlier.
- Median and Interquartile Deviation Method (IQD)
  - For this outlier detection method, the median of the residuals is calculated, along with the 25th percentile and the 75th percentile. The difference between the 25th and 75th percentile is the IQD. Then, the difference is calculated between each historical value and the residual median. If the historical value is a certain number of MAD away from the median of the residuals, that value is classified as an outlier.
  - The default threshold is 2.22, which is equivalent to 3 standard deviations or MADs.
    - This method is somewhat susceptible to influence from extreme outliers, but less so than the mean and standard deviation method. Box plots are based on this approach. The median and interquartile deviation method can be used for both symmetric and asymmetric data.

Related and Advanced Topics

- Ordinary Least Squares (OLS)
- Outlier Analysis
- **Tests for Veracity of OLS Assumptions**
- Maximum Likelihood Estimation (MLE)
- Method of Moments (MoM)

Tests for Veracity of OLS Assumptions

- There are two key assumptions about the error term when performing OLS:
  - Error term normality
  - Homoscedasticity

- In the case of multivariate regressions ($y = a + b_1x_1 + \ldots + b_kx_k$), there is an additional assumption:
  - Independence of regressors

If any of these assumptions do not hold, neither do the results of OLS. Therefore, we must always test the veracity of OLS assumptions.

Tests for Veracity of OLS Assumptions

- **Independence of regressors**
- Error term normality
- Homoscedasticity

Tests for Veracity of OLS Assumptions: Independence of Regressors

- In OLS regression, the x-variables (regressors) are assumed to be independent of each other
- When this assumption is violated, such that one or more x-variables can be expressed as (an approximately) linear combination of the other x-variables, this is called multicollinearity
- In this context, correlation among the regressors is *sufficient*, but not *necessary*, for multicollinearity to occur
- Multicollinearity does not typically cause issues with the predictive capability of a CER but it can create a wrong signs problem where even though an independent variable has a positive correlation with the dependent variable, the coefficient may be negative
- Several ways to address – can exclude variables with wrong signs, or combine highly correlated variables

1. This topic is covered in greater detail in PT04 Multivariate Regression

Tests for Veracity of OLS Assumptions

- Independence of regressors
- **Error term normality**
- Homoscedasticity

Tests for Veracity of OLS Assumptions: Error Term Normality

- In OLS regression, the additive error term (ε) is assumed to be normally distributed
- This assumption by applying a test for normality to the residuals
- Myriad tests for normality are available. The three most commonly used in cost analysis are [1]:
  - Kolmorogov-Smirnov (K-S) test
  - Anderson-Darling test
  - Chi square test
- Less commonly used normality tests include:
  - D'Agostino's K-Squared test
  - Jarque-Bera test
  - Cramer-von Mises criterion
  - Shapiro-Wilk test
  - Various *adaptations* of the K-S test

The following slides focus on the K-S test as an example.

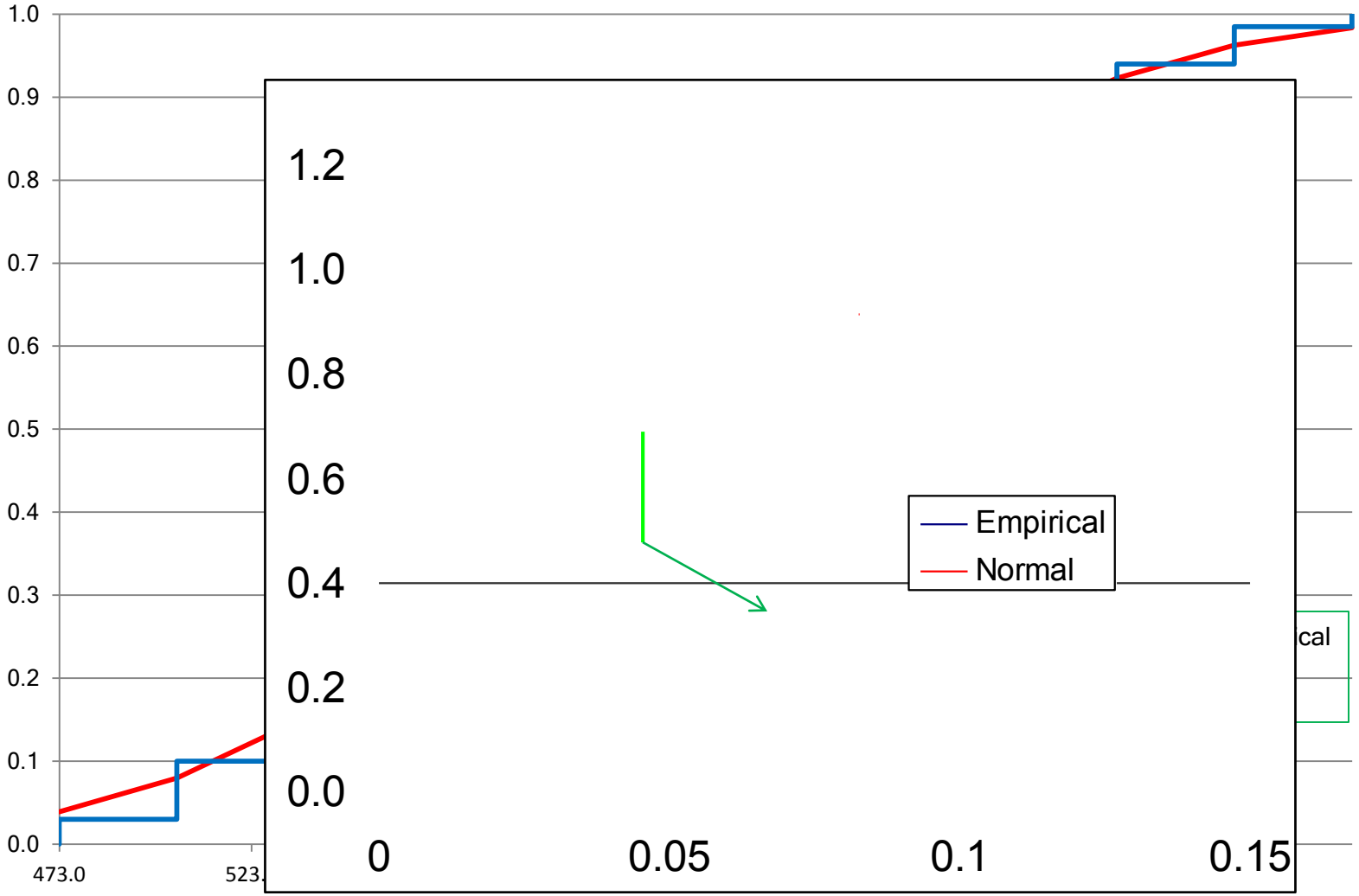
1. Many of these tests can be used to test whether data follow any distribution, not just the normal.

K-S Test: Step by Step Instructions

1. Rank each residual, from lowest to highest. Assign "1" to the minimum residual, "2" to the next lowest, and n to the highest (where n = number of data points).
2. Create an empirical cumulative distribution function (ECDF) of the residuals. This function assigns to each residual the probability that any residual does not exceed that value. The ECDF value of the i^{th} residual is $\mathbf{R_i/n}$, where R_i is the rank of the i^{th} residual.
3. Create a theoretical cumulative distribution function (TCDF) of the residuals. This is a normal distribution with $\mu = 0$ and σ = standard error of the estimate (SEE) from the regression. The TCDF value for the ith residual, ε_i, is[1] **NORMDIST(ε_i,0, SEE, 1)**
4. Calculate the maximum absolute distance between the ECDF and the TCDF ($K\text{-}S_{max}$)
5. Compare $K\text{-}S_{max}$ to a table of critical values to determine whether the normality hypothesis is rejected

1. This example uses Microsoft Excel syntax. Its respective arguments are the value at which the distribution is to be evaluated, the hypothesized mean, the hypothesized standard deviation, and a binary variable indicating whether the distribution is to be evaluated cumulatively.

K-S Test Illustrated



K-S Test: Rejection Criteria and Possible Remedies

- Reject normality if K-S_{max} exceeds critical values as defined in an authoritative source [1]
- For example, at α = 0.1 and n=15, the critical value is **0.244**
- Because K-S_{max} = **0.134**, we do not reject the (null) hypothesis of normality and conclude that the error term is normally distributed
- Possible remedies for when normality assumption is violated:
  - Alternative combination(s) of independent variables
  - Alternative functional forms (especially log space)
  - Alternative regression techniques (e.g. GERM)
  - Repeat test using one or more other techniques (e.g. Anderson-Darling, Chi Square)
  - Proceed anyway, noting the K-S result as a caveat (significance tests are invalid, but estimates are still unbiased)

1. For example, http://www.scribd.com/doc/54045029/79/Table-of-critical-values-for-Kolmogorov-Smirnov-test

Tests for Veracity of OLS Assumptions

- Independence of regressors
- Error term normality
- **Homoscedasticity**

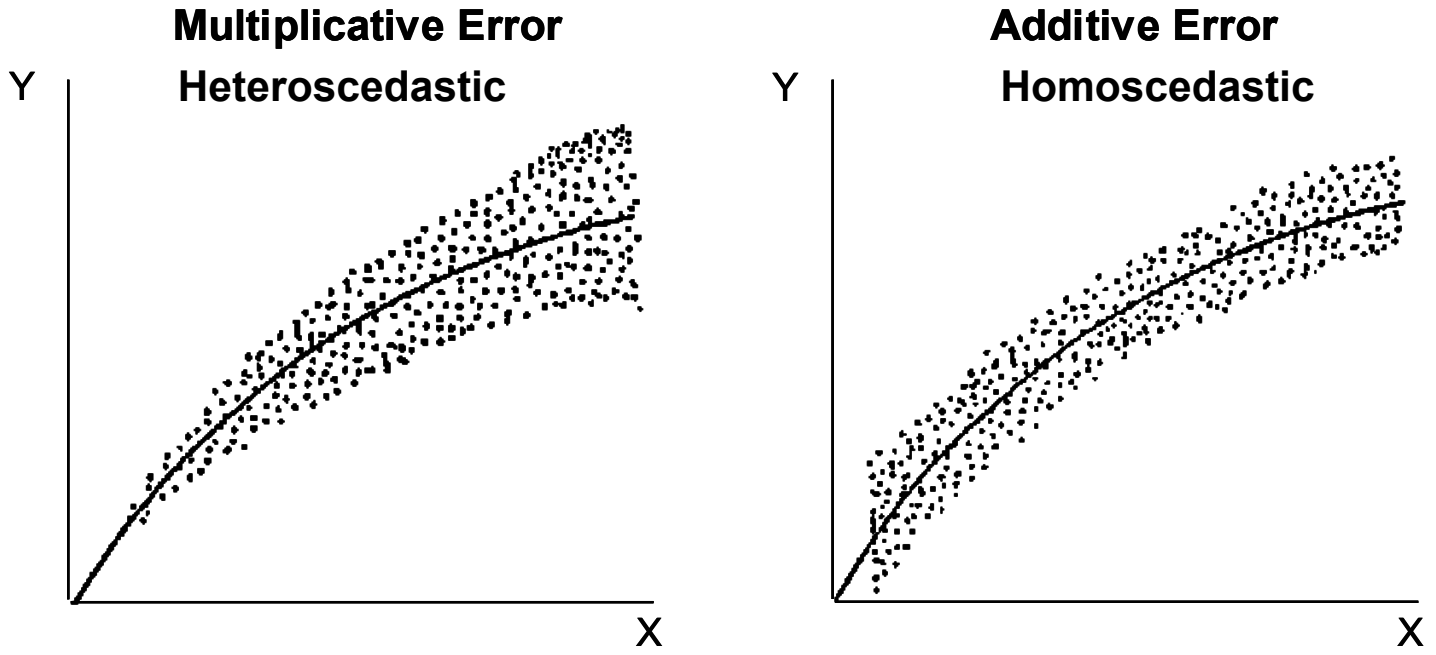
Tests for Veracity of OLS Assumptions: Homoscedasticity

- In OLS regression, e is assumed to have a constant variance. This is called homoscedasticity
- When this assumption is violated (i.e., the error term has non-constant variance), this is called heteroscedasticity
- Heteroscedasticity can be easily detected using the industry standard White Test

1. This topic is covered in greater detail in PT04 Multivariate Regression

Example of Homoscedasticity

- The case of multiplicative error terms results in heteroscedasticity in linear space

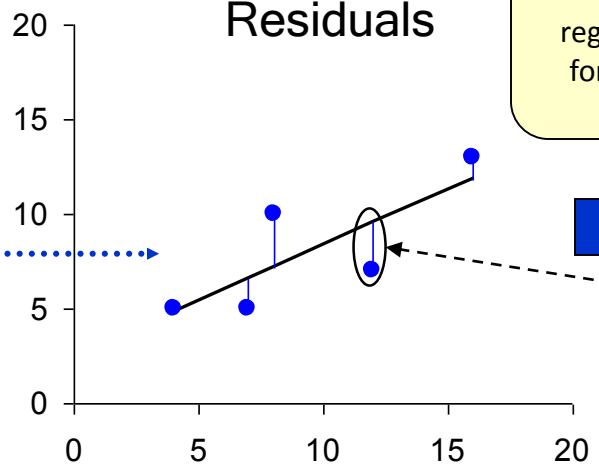


Edited Excerpt from CEBoK Module 8

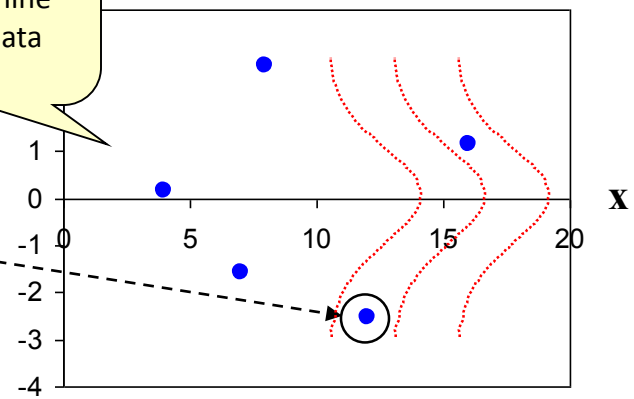
Create a Residual Plot to verify that the OLS assumptions hold:

$y = 2.5 + 0.6x + \varepsilon$

5



Plot the distance from the regression line for each data point



Questions to ask:

1. Does the residual plot show independence? yes(?)
2. Are the points symmetric about the the x-axis with constant variability for all x? yes (?)

The OLS assumptions are reasonable (?) This tells us:

- **That our assumption of linearity is reasonable**
- **The error term e can be modeled as Normal with mean of zero and perhaps constant variance**

ICEAA

White Test: Step-by-Step Instructions

1. Perform regression as usual to generate squared errors (ε^2)
2. Regress ε^2 on each regressor, squared regressor, pairwise crossproduct, and an intercept
   1. For 1 x: Regress on x^2, x
   2. For 2 x's: Regress on x_1, x_1^2, x^2, x_2^2, and x_1x_2
   3. For 3 x's: Regress on x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, and x_2x_3
   4. For *k* x's: C(k+2,2) = (k+2)(k+1)/2 regressors
3. Calculate the R^2 from the *auxiliary* regression
4. White statistic = nR^2 follows a chi square [1] distribution with (m-1) degrees of freedom, where m = number of estimated parameters (not including intercept) from *auxiliary* regression
5. Reject the null hypothesis of homoscedasticity and conclude that OLS cannot be used if p-value is less than a specified critical value (say, 0.10)

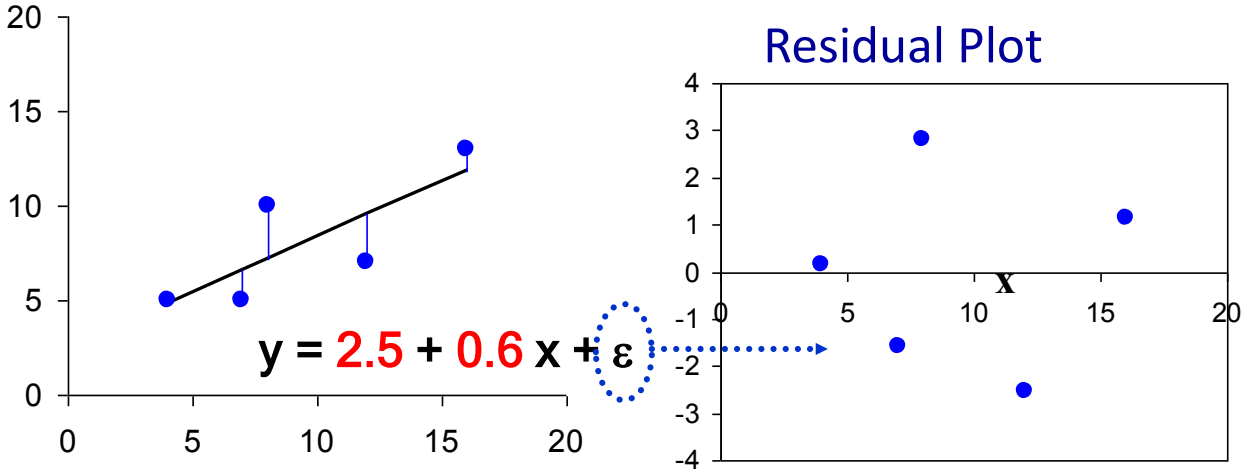
1. Recall that the OLS error term is normally distributed, and that the product of two normally distributed random variables can be written as a chi square random variable.

The White Test Applied to CEBoK Toy Problem

Data		Calculations				
X	Y	XY	X^2	ε^2	x	x^2
4	5	20	16	0.03	4	16
7	5	35	49	2.55	7	49
8	10	80	64	7.95	8	64
12	7	84	144	6.35	12	144
16	13	208	256	1.30	16	256

White Test	
Auxiliary R^2:	0.79
White Stat:	3.93
DF:	1
p-value:	0.047
Conclusion:	UH-OH!

CEBoK Excerpt Revisited



Questions to ask:

1. Does the residual plot show independence? **No!**
2. Are the points symmetric about the the x-axis with constant variability for all x? **No!**

The OLS assumptions are not reasonable (!) This tells us:

- **The error term e cannot be modeled with constant variance. The MLE generalization is called for here.**

Heteroscedasticity: Possible Remedies

- In order of preference:
  - Perform Maximum Likelihood Estimation (MLE)-based regression, a generalization of OLS that allows for heteroscedastic error terms
    - Investigate nonlinear functional forms
  - Investigate alternative regressors

Related and Advanced Topics

- Ordinary Least Squares (OLS)
- Outlier Analysis
- Tests for Veracity of OLS Assumptions
- **Maximum Likelihood Estimation (MLE)**
- Method of Moments (MoM)

Maximum Likelihood Estimation (MLE)

- Method of estimating the parameters of a distribution (when applied to a data set and given a distribution)
- The maximum likelihood estimator(s) of unknown distribution parameter(s) is the value of the parameter(s) that make the observed values (of the data set) most likely
- In other words, parameters that maximize the **likelihood function (**where the x's are the observed values):

$$f_{x_1},\ldots,f_{x_n}(x_1,\ldots,x_n;\theta) = f_x(x_1;\theta)\,f_x(x_2;\theta)\bullet\bullet\bullet\, f_x(x_n;\theta)$$

$$p_{x_1},\ldots,p_{x_n}(x_1,\ldots,x_n;\theta) = \prod_{j=1}^{n} p_x(x_j;\theta)$$

- For example, given a Bernoulli distribution with two observations of x=1 and one of x=0, the likelihood function would be:

$$p_{x_1}, p_{x_2}, p_{x_3}(x_1,x_2,x_3;p) = p + p + (1-p)$$

- In practice, it is often easier to maximize the logarithm of the likelihood function, the **log-likelihood function**

MLE Excel Implementation Overview

- In practice, the lognormal has been found to model error distributions well
- The log-likelihood function for the lognormal distribution is

$$l(\beta,\theta) = -\frac{n}{2}\ln\theta - \sum_{i=1}^{n}\ln y_i - \frac{1}{2\theta}\sum_{i=1}^{n}\left(\ln y_i - \ln f(x_i,\beta)\right)^2$$

- To re-case this in terms with which we are familiar, maximizing the above is the same as minimizing its negative, which is

$$l(\beta,\theta) = \frac{n}{2}\ln\theta + \sum_{i=1}^{n}\ln y_i + \frac{1}{2\theta}\sum_{i=1}^{n}\left(\ln y_i - \ln f(x_i,\beta)\right)^2$$

- We minimize with respect to the parameter θ
- Taking derivatives we find that we have to minimize

$$L^*(\beta) = \sum_{i=1}^{n}\left(\ln y_i - \ln f(x_i,\beta)\right)^2$$

- Notice the similarity to least-squares regression

Log-Transformed Least Squares

- What we have derived is a generalization of log-transformed ordinary least squares in the context of maximum likelihood
- In Log-Transformed Ordinary Least Squares, apply a logarithmic transform to both the actual and the estimated costs
- For the power equation $Y=aX^b$ this transforms the equation from a nonlinear equation to a linear one:

$$\ln Y = \ln\left(aX^b\right) = \ln a + b \ln X$$

- The parameters can be easily calculated in Excel
- Must remember to transform the *a* parameter.
- The Maximum Likelihood Median Estimator is More General.
  - Any equation form may be used, but unless the log transformed equation is linear, may need computer to solve (e.g., Excel's Solver Capability).
  - Nothing in the MLE derivation forces any particular functional form.

Related and Advanced Topics

- Ordinary Least Squares (OLS)
- Outlier Analysis
- Tests for Veracity of OLS Assumptions
- Maximum Likelihood Estimation (MLE)
- **Method of Moments (MoM)**

Method of Moments (MoM)

- Method of estimating the parameters of a statistical model (when applied to a data set and given a statistical model)
- Equates sample **moments** with unobservable population moments
  - Raw Moments- Centered about zero
  - Central Moments- Centered about the mean

	Raw	Central
1st Moment	$E(X) = \mu$	$E[(X-E[X])] = 0$
2nd Moment	$E(X^2)$	$E[(X-E[X])^2] = \sigma^2$
3rd Moment	$E(X^3)$	$E[(X-E[X])^3]$
nth Moment	$E(X^n)$	$E[(X-E[X])^n]$

MoM Excel Implementation Overview

Method of Moments Steps:

1. Choose the number of moments to use to be equal to the number of unknown parameters, k.

   -e.g., if we want to calculate both population parameters of the Beta distribution, we must calculate two moments

2. Write each population moment as a function of the unknown parameter
3. Solve for the unknown parameter(s)

MoM Excel Implementation- Detailed Steps

1. Given the same normally distributed dataset as in previous example, calculate the sample mean and standard deviation. These are the estimates of the distribution parameters
2. Calculate the theoretical value for each lower bound Calculate the theoretical (cumulative) value for each lower bound as in MLE example, except the mean and standard deviation are equated to sample mean and sample standard deviation
3. Determine the theoretical values (to arrive at your expected theoretical values) by subtracting the value of each cumulative value by the previous lower bound's cumulative value
4. Unlike in MLE, these values will not change as we already have arrived at our MoM estimates of the distribution parameters

MoM Excel Implementation- Example

- Calculate the sample mean and sample standard deviation of the dataset (using AVERAGE and STDEV functions in excel, respectively):

Sample Mean	0.4576
Sample Standard Deviation	1.1533

- Calculate the cumulative theoretical and theoretical values of each lower bound using the sample mean and sample standard deviation as the parameters of the normal distribution:

A	B	C	D	E	I	J
Lower Bound	**Upper Bound**	**Mid-Point**	**CUMFREQ**	**Observed**	**CUM Expected$_{(MoM)}$**	**Expected $_{(MOM)}$**
-2.2	-2	-2.1	10	4	21.2013	7.9818
-2	-1.8	-1.9	16	6	**33.0924**	**11.8912**
-1.8	-1.6	-1.7	26	10	50.2841	17.1917
-1.6	-1.4	-1.5	45	19	74.4045	24.1204
-1.4	-1.2	-1.3	69	24	107.2458	32.8413

=$Q$5*NORMDIST(A2,$R$1,$R$2,1)
Where: Q5= Number of Observations
R1= Sample Mean
R2= Sample Standard Deviation

=I2-I1

Conclusion

- It is important to understand probability distributions and the assumptions that underlie CER development and risk analysis
- Be sure to test your OLS assumptions first
- Use Maximum Likelihood Estimators where appropriate

Backup Slides

MS Excel Example for Applying Random Variables

Simple Example

- We'll prepare a probabilistic cost estimate for a testing project with the following parameters

Cost Element		Distribution	Parameters
	Labor	Triangular	4,9,22
	Test Cycles	Emipirical	See table
	Travel	Normal	10,2
	PM/SE	Uniform	2,7

Determine probabilistic total cost (Cost of each test cycle is 10)

# Cyles	Frequency	Cumulative
15	0.3	0
25	0.45	0.3
35	0.15	0.75
45	0.1	0.9

Construct in Excel

	E	F	G	H
2				=RAND()
3				
4		Distribution	Parameters	
5	Labor	Triangular	4,9,22	=IF(H2<=(22-4)/(9-4),4+SQRT(H2*(9-4)*(22-4)),9-SQRT((1-H2)*(9-4)*(9-22)))
6	Test Cycles @ 10/test	Emipirical	See table	=LOOKUP(RAND(),$G$12:$G$15,$E$12:$E$15)*10
7	Travel	Normal	10,2	=NORMINV(RAND(),10,2)
8	PM/SE	Uniform	2,7	=RANDBETWEEN(2,7)
9	Total			=SUM(H5:H8)
10				
11	# Cyles	Frequency	Cumulative	
12	15	0.3	0	
13	25	0.45	0.3	
14	35	0.15	0.75	
15	45	0.1	0.9	
16				

- Duplicate the calculations across a large number of columns (i.e., 1 column for each iteration)
- Thousands / tens of thousands can be quickly replicated
- Note: Best practice is to use references to cells or named ranges; values are entered in equations show only for clarity in presentation here

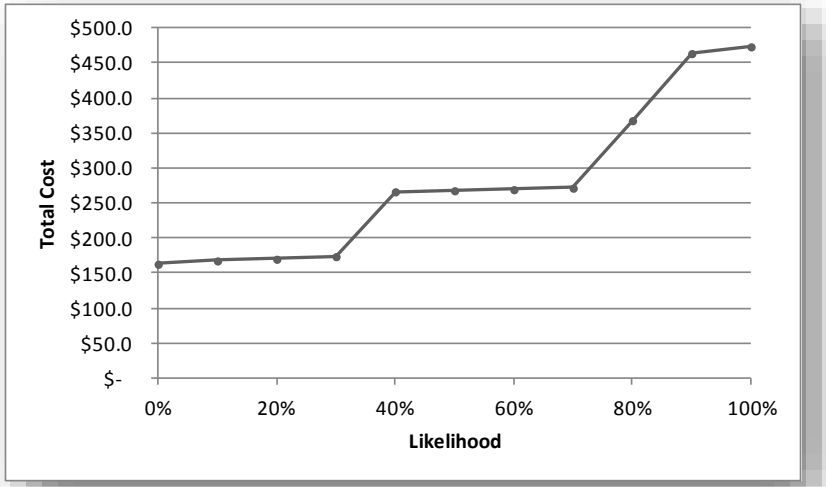
Results Summary

	Total	Labor	Testing	Travel	PM/SE
Range	$ 310.7	$ 8.4	$ 300.0	$ 11.9	$ 5.0
Min	$ 162.7	$ 5.0	$ 150.0	$ 5.2	$ 2.0
Max	$ 473.5	$ 13.5	$ 450.0	$ 17.2	$ 7.0
Mean	$ 273.3	$ 10.4	$ 254.8	$ 10.0	$ 4.5
Median	$ 267.8	$ 10.8	$ 250.0	$ 9.8	$ 5.0
50%	$ 267.8	$ 10.8	$ 250.0	$ 9.8	$ 5.0
70%	$ 271.5	$ 12.0	$ 250.0	$ 11.0	$ 6.0
90%	$ 463.8	$ 13.0	$ 450.0	$ 12.5	$ 7.0
95%	$ 468.4	$ 13.3	$ 450.0	$ 13.3	$ 7.0
99%	$ 471.5	$ 13.4	$ 450.0	$ 14.1	$ 7.0

- We forecast total cost at or below $267.8M with 50% confidence
- We forecast total cost at or below $271.5M with 70% confidence
- We forecast total cost at or below $463.8M with 90% confidence

Results Summary

- Shows decision makers what the likelihood of different funding alternatives will imply
- Useful in choosing a defensible level of contingency reserves



- No specific confidence level is considered a best practice
- Program cost estimates should be budgeted to at least the 50 % confidence level
- Budgeting to a higher level (e.g.,70 % - 80 %, or the mean) is common practice

Results Formulas

	L	M	N	O	P
1	Total				
2	162.736021717812				
3	162.785459278975				Total
4	162.912712466124			Range	=P6-P5
5	163.257865899794			Min	=MIN($L$2:$L$501)
6	163.398134516749			Max	=MAX($L$2:$L$501)
7	163.501443791051			Mean	=AVERAGE($L$2:$L$501)
8	163.803436255014			Median	=MEDIAN($L$2:$L$501)
9	163.848303996661			0.5	=PERCENTILE($L$2:$L$501,0.5)
10	163.851092779599			0.7	=PERCENTILE($L$2:$L$501,0.7)
11	163.972005952186			0.9	=PERCENTILE($L$2:$L$501,0.9)
12	164.015823866487			0.95	=PERCENTILE($L$2:$L$501,0.95)
13	164.330704975783			0.99	=PERCENTILE($L$2:$L$501,0.99)
14	164.350767613553				

Built-In Analysis Options

- Alternatively, overall results can be calculated using options from the Data Analysis tab on the Data ribbon