v1.2

Multivariate and Non-linear Regression Models

© 2002-2013 ICEAA All rights reserved.

v1.2

Non-linear Models – Introduction

- To model non-linear relationships with OLS regression, the data must first be transformed in a way that makes the relationship linear
- All the steps for linear regression may then be performed on the transformed data
- The most common forms of non-linear models are:
  - Logarithmic
  - Exponential
  - Power

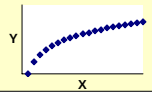
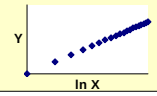
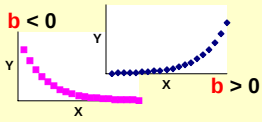
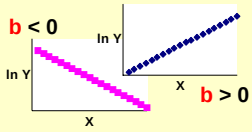
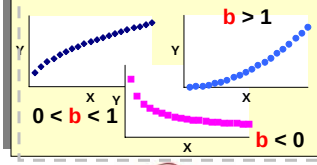
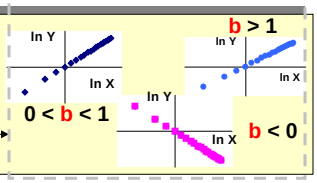
© 2002-2013 ICEAA All rights reserved.

Linear Transformations

v1.2

12

20

Unit Space	Model	Log Space
	**Logarithmic** $y = a + b \ln x$	
	**Exponential** $y = a e^{b x}$ $\ln y = \ln a + b x$	
	**Power** $y = a x^{b}$ $\ln y = \ln a + b \ln x$	

© 2002-2013 ICEAA All rights reserved.

Example: Exponential Model

v1.2

- We will use the same data as before, but apply an exponential model to it
  - Recall that the data failed the White test for homoscedasticity (p = 0.047) 6
  - In practice, a *Power model* (linear when take logs of both sides) might be called for here, but this is shown in Module 7 (Learning Curves), so Exponential (linear when take log of y) is demonstrated 7
- The next step is to conduct linear regression analysis on the data in semi-log space
- After the analysis is complete, we will transform the parameters of the linear equation back to unit space

Tip: Exponential is rare in practice

$a = e^{\ln a}$

$\ln y = \ln a + b x \Leftrightarrow y = a e^{b x}$

$b = b$

© 2002-2013 ICEAA All rights reserved.

Example: Exponential Model

v1.2

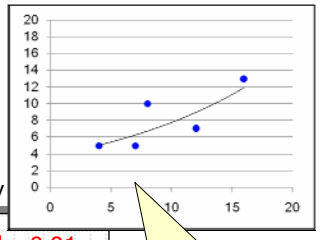
13

- First we run the regression:

Exponential Model	X	Int
Coefficients	0.07	1.34
Standard Errors	0.03	0.33
R^2, SEE	0.62	0.30
F, DF	4.81	3
SSR, SSE	0.44	0.28
MSR, MSE	0.44	0.09
Y-bar, CV	0.669	45.3%
T stats	2.19	4.00
P values	0.1160	0.0280

Tip: LOGEST() produces same output, however LOGEST coefficients are the exponentials of the LINEST coefficients.

The regression is nearly statistically significant at α = 0.10 with semi-log space R^2 = 0.62



Unit-space data showing exponential trend.

- Then we check the residual plot:
  - The residual plot is ambiguous; we expand the White test…
  - …for a formal determination on homoscedasticity

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

$y = a\,e^{b\,x}$

- Finally, we find the parameters for the unit space equation:

$a = e^{\ln a} = e^{1.34} = 3.81$

$b = 0.07$

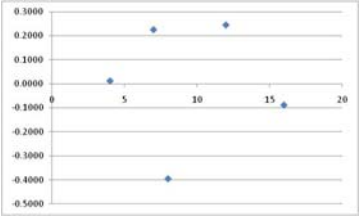
$\hat{Y} = 3.81e^{0.07X}$

© 2002-2013 ICEAA All rights reserved.

Example: Exponential Model (Expanded White Test)

v1.2

White Test	Unit Model	Expoential Model
Auxiliary R^2:	0.79	0.62
White Stat:	3.93	3.12
DF:	1	1
p-value:	0.047	0.077
Conclusion:	Heteroskedastic	Heteroskedastic



White Test Conclusions

- Homoscedasticity *still rejected* at α = 0.10 (now not rejected at α = 0.05)
- In practice, could use MLE or Power Model (or the 5% significance level), but we will proceed as if OLS assumptions were validated

Next Steps

- Calculate *unit-space goodness of fit statistics* for apples-to-apples model comparisons

NEW!

© 2002-2013 ICEAA All rights reserved.

v1.2

Unit-Space Goodness of Fit Comparison

Statistic	Linear	Exponential	How to Calculate/Comments
Fit Space R^2	0.62	0.62	From LINEST(.)
Unit Space R^2	0.62	0.64	1-SSE/SST = 1-SUMSQ(ε)/DEVSQ(y) in unit space
Unit Space Adj R^2	0.50	0.52	1-(((1-R^2)*(n-1))/df)
Fit Space SEE	2.46	0.30	From LINEST(.)
Unit Space SEE	2.46	2.40	SQRT(SSE/DF) in unit space
Fit Space CV	31%	45%	SEE/y-bar in fit space
Unit Space CV	31%	30%	SEE/y-bar in unit space

- These differences are not overwhelming, but the routine serves as a reference for comparison of more complicated, multivariate models across types

Warning: It is unusual for a power or exponential model to have better unit space than fit space statistics; generally the unit space conversion causes these stats to *worsen*

© 2002-2013 ICEAA All rights reserved.

v1.2

Minding the Intercept

- One common mistake when performing OLS regression is the omission of a y-intercept in power and exponential models when one exists
  - This has the effect of causing higher than necessary error in the regression
  - Fortunately, it can easily be detected by examining the relation of the trendline to the data
  - It can also be corrected by adding (or subtracting) a constant value to (from) the y-values in the data and examining the change in the trendline ("the simple way") or by using Solver or other packages ("the elegant way")
- The example on the next page assumes the data follows a power curve with a non-zero y-intercept

6

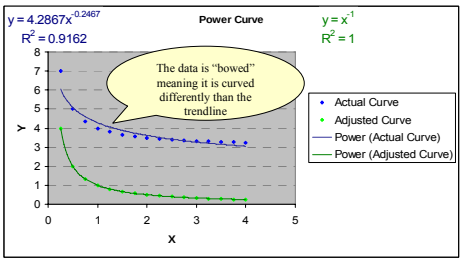
"'To b or Not to b' The y-intercept in Cost Estimation, R. L. Coleman, J. R. Summerville, P. J. Braxton, B. L. Cullis, E. R. Druker, SCEA, 2007.

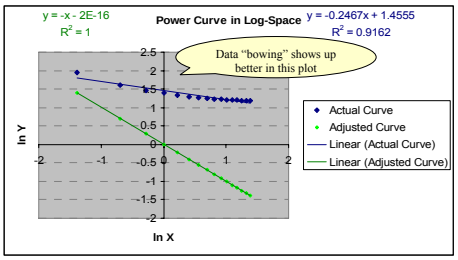
© 2002-2013 ICEAA All rights reserved.

Minding the Intercept - Example

v1.2

- The plot to the top-right shows the actual data as well as the data that has been adjusted to take the intercept into account
- The data was adjusted by subtracting a constant from all y-values until the optimal R^2 was achieved
  - This constant is the best guess for the y-intercept
  - A "bowing" of the data in relation to the trendline is the symptom that led to the belief in an intercept
- The plot to the bottom-right shows this same graph in log-space
- Adding the intercept greatly increases the R^2 of the regression
  - Without intercept: .9162
  - With intercept: 1





© 2002-2013 ICEAA All rights reserved.

Minding the Intercept - Example

v1.2

- The ANOVA statistics for the two regressions are shown to the right
- Notice the decrease in standard error and increase in R^2

Regression Statistics	
Multiple R	0.957159023
R Square	0.916153395
Adjusted R Square	0.910164352
Standard Error	0.060953166
Observations	16

ANOVA

	df	SS	MS	F	Significance F
Regression	1	0.568333545	0.568334	152.9716	6.33998E-09
Residual	14	0.052014039	0.003715		
Total	15	0.620347583			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	1.455511028	0.018553102	78.45109	6.5E-20	1.415718583
ln x	-0.246655725	0.019942786	-12.36817	6.34E-09	-0.289428747

Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Square	1
Standard Error	3.68462E-16
Observations	16

ANOVA

	df	SS	MS	F	Significance F
Regression	1	9.341591866	9.341592	6.88E+31	3.024E-216
Residual	14	1.9007E-30	1.36E-31		
Total	15	9.341591866			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	0	1.12154E-16	0	1	-2.40545E-16
ln x	-1	1.20554E-16	-8.3E+15	3E-216	-1

Warning: These "perfect" results are from a toy problem using "cooked" data

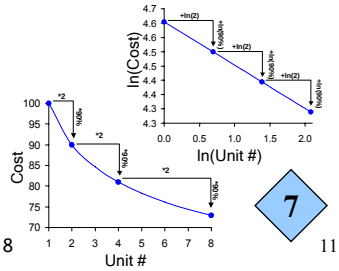
© 2002-2013 ICEAA All rights reserved.

v1.2

Non-linear Model Summary

- The same process performed on the exponential example applies to other non-linear model types
  - The only difference lies in which piece of the data set gets transformed
    - i.e. Logarithmic ⇔ take the log of the x data
      Exponential ⇔ take the log of the y data
      Power ⇔ take the log of both x and y
    - Other functions can be used to transform data (e.g., $\sqrt{x}$, sin x, etc.) but logarithms are the most common

Tip: Power models are used to analyze learning curves - they are probably the most common use of non-linear regression in cost analysis



7

© 2002-2013 ICEAA All rights reserved.

v1.2

Multivariate Regression

AKA Multiple Regression

- Basics
- ANOVA Revisited
- Adjusted R^2
- t and F Summary

© 2002-2013 ICEAA All rights reserved.

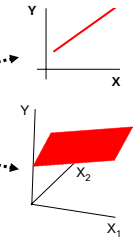
v1.2

Multivariate Regression

- If there is more than one independent variable in linear regression we call it *multivariate regression*
- The general equation is as follows:

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

  - So far, we have seen that for one independent variable, the equation forms a line in 2-dimensions
  - For two independent variables, the equation forms a plane in 3-dimensions
  - For three or more variables, we are working in higher dimensions which are difficult to display visually in Excel.



- The math is more complicated, but the results can be easily obtained from a regression tool or simple formula (LINEST()) as found in Excel

© 2002-2013 ICEAA All rights reserved.

v1.2

Multivariate Regression

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

- In general the underlying math is similar to the simple model, but matrices are used to represent the coefficients and variables
  - Understanding the math requires background in Linear Algebra
  - Demonstration is beyond the scope of the module, but can be obtained from the references
- Some key points to remember for multivariate regression include:
  - Perform residual analysis between each X variable and Y
  - Avoid multicollinearity, i.e., the situation in which high correlation among (2 or more) X variables inflates standard errors and therefore biases significance tests
  - Use the "Goodness of Fit" metrics and significance tests to guide you toward a good model

© 2002-2013 ICEAA All rights reserved.

v1.2

Identifying a Multivariate Regression

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

- In general, theory and sound reasoning should be used to determine which variables to include in a multivariate model
  - Choose variables that are correlated with the dependent variable and can be justified; i.e. show correlation and causation
  - It is hard to 'prove' that a model is correctly identified, but with correlation statistics and well developed reasoning, a model can be shown to be robust
- If a relevant variable is omitted, it may cause b estimates to be biased and will increase SSE ("omitted variable bias")

© 2002-2013 ICEAA All rights reserved.

v1.2

Coefficients in Multivariate Regression

$$y = a + \hat{b}_1x_1 + \hat{b}_2x_2 + \ldots + \hat{b}_kx_k$$

- The Excel output gives the predicted coefficients

Example Multivariate Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.844800428
R Square	0.713687763
Adjusted R Square	0.635602607
Standard Error	0.177820122
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	3	0.867007454	0.289002485	9.13986477	0.002528225
Residual	11	0.347819953	0.031619996		
Total	14	1.214827407			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.372062796	0.163187242	8.40790478	9.0585E-06	1.012890097	1.731235495
X Variable 1	0.250514367	0.22424172	1.117182172	0.28773191	-0.243038331	0.744067064
X Variable 2	-1.126402129	0.219022126	-5.142869129	0.000321795	-1.608466577	-0.644337681
X Variable 3	-0.159991407	0.173317791	-0.923110112	0.37575435	-0.541461295	0.221478482

Note: LINEST() outputs numbers in gray box. Analyst adds labels and other calculations.

Equation Parameters

$$\hat{Y} = 1.4 + 0.3\ X_1 - 1.1\ X_2 + -0.2\ X_3$$

	b3	b2	b1	a
Coefficients	-0.160	-1.126	0.251	1.372
Standard Errors	0.173	0.219	0.224	0.163
R^2, SE(y)	71.4%	0.178		
F, DF	9.140	11		
SSR, SSE	0.867	0.348		
MSR, MSE	0.289	0.032		
ADJ R^2, SEE, CV	63.6%	0.178	16.1%	
T stats	-0.92	-5.14	1.12	8.41
P values	0.3758	0.0003	0.2877	0.0000
Significance F	0.0025			

© 2002-2013 ICEAA All rights reserved.

Analysis of Variance (ANOVA)

v1.2

- **Mean Measures of Variation**

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

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

- Mean of Squares of the Regression (MSR):

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

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

15 data points

where:

n = # data points

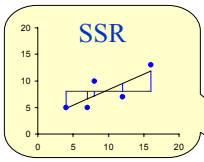
k = # equation variables

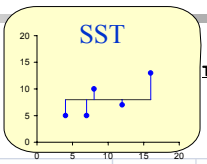
3 Variables

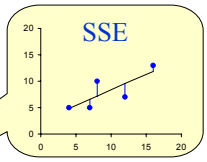
© 2002-2013 ICEAA All rights reserved.

Excel Demo: ANOVA

v1.2







Total Sum of Squares (SST): The sum of the squared deviations between the data and the average

Regression Sum of Squares (SSR): The sum of the squared deviations between the regression line and the average

"The explained variation"

Residual or Error Sum of Squares (SSE): The sum of the squared deviations between the data and the regression line

"The unexplained variation"

	b3	b2	b1	a
efficients	-0.160	-1.126	0.251	1.3
rd Errors	0.173	0.219	0.224	.03
R^2, SE	71.4%	0.178		
F, DF	9.140	11		
SSR, SSE	0.867	0.348		
MSR, MSE	0.289	0.032		
ADJ R^2, SEE, CV	63.6%	178	16.1	
T stats	-0.92	-5 4	1.12	
P values	0.3758	0.00	0.2877	0.0000
Significance F	0.0025			

8

Note: df_{resid} is provided. Df_{reg} must be caclulated using $df_{reg} = n - df_{resid}$

$MSR = SSR/df_{resid}$

$$\sim \frac{\chi_k^2}{k}$$

$F = MSR / MSE$

$$\sim F_{k,(n-1)-k}$$

$MSE = SSE/df_{reg}$

$$\sim \frac{\chi^2_{(n-1)-k}}{(n-1)-k}$$

© 2002-2013 ICEAA All rights reserved.

v1.2

Adjusted R^2

- Adjusted R^2, or R^2_a, adjusts for degrees of freedom
  - Can be used to compare coefficients of determination between models with different numbers of variables including in the same model when a variable is considered for elimination due to lack of significance
  - Can be used as justification for including near-significant variables in models if those variable improve the model's performance

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

Warning: negative values of R^2_a may occur when fitting non-OLS trends to data

$$R_a^2 = 1 - \left[\left(\frac{SSE}{SST}\right)\left(\frac{n-1}{(n-1)-k}\right)\right] = 1 - \left[\left(1-R^2\right)\left(\frac{n-1}{(n-1)-k}\right)\right]$$

% unexplained

penalty (> 1)

Tip: SSR+SSE=SST is true only in OLS. In general, we have R^2 = 1-SSE/SST but *not* R^2 = SSR/SST. Note also that R^2_a can be negative.

ICEAA CEBoK © 2002-2013 ICEAA All rights reserved.

v1.2

t statistics in Multivariate Regression

15

- In multivariate regression, a t test is conducted for each coefficient
- The results provide insight as to which variables add the most value to the prediction of cost
  - Adding additional variables will always decrease SSE and increase (unadjusted) R^2
  - An insignificant t statistic makes a variable *a candidate* to be eliminated from the regression (can compare nested vs. full model using SEE, CV, adjusted R^2 and F statistics) [1]
  - A variable whose p value is greater than 0.5 should *almost certainly* be eliminated because it is more likely than not that its (nonzero) coefficient was observed by chance

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$$

	b3	b2	b1	a
Coefficients	-0.160	-1.126	0.251	1.372
Standard Errors	0.173	0.219	0.224	0.163
R^2, SE(y)	71.4%	0.178		
F, DF	9.140	11		
SSR, SSE	0.867	0.348		
MSR, MSE	0.289	0.032		
ADJ R^2, SEE, CV	63.6%	0.178	16.1%	
T stats	-0.92	-5.14	1.12	8.41
P values	0.3758	0.0003	0.2877	0.0000
Significance F	0.0025			

Note: Correlation between the independent variables may affect results. High correlation among independent variables is often associated with ***multicollinearity*** and should be avoided. A correlation matrix is a good first step to check for multicollinearity. The example is expanded later as an Advanced Topic.

	x1	x2	x3
x1	--	0.037	0.298
x2	0.037	--	-0.030
x3	0.298	-0.030	--

The p-values suggest that x_2 is highly significant (as is the intercept, which is generally retained regardless of significance results). The remaining variables are *candidates* for elimination.

1. There are several methods such as stepwise regression for determining the best subset of independent variables. See the references for more details

ICEAA CEBoK © 2002-2013 ICEAA All rights reserved.

Calculation of t statistic

v1.2

- As before, the t statistic for each variable may be calculated as ratio of the estimated coefficient to the corresponding standard error:

$$t = \frac{\hat{b}_i}{se_{b_i}}$$

- t is also the square root of the partial F statistic, F*

$$t = \sqrt{F^*}$$

$$F^* = \frac{SSR_{FullModel} - SSR_{ReducedModel}}{\frac{SSE_{FullModel}}{df}} = \frac{SS(b_i \mid a, b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_k)}{\frac{SSE_{FullModel}}{df}}$$

Partial sum of squares - captures the value of adding the variable in question

© 2002-2013 ICEAA All rights reserved.

The F statistic

v1.2

- An *F test* is used to determine whether the coefficients of ***all*** the independent variables are zero
  - Depends on the ratio of the MSR to the MSE, called an *F statistic*

$y = a + b_1x_1 + b_2x_2 + b_3x_3$

Are all coefficients = 0?

	b3	b2	b1	a
Coefficients	-0.160	-1.126	0.251	1.372
Standard Errors	0.173	0.219	0.224	0.163
R^2, SE(y)	71.4%	0.178		
F, DF	9.140	11		
SSR, SSE	0.867	0.348		
MSR, MSE	0.289	0.032		
ADJ R^2, SEE, CV	63.6%	0.178	16.1%	
T stats	-0.92	-5.14	1.12	8.41
P values	0.3758	0.0003	0.2877	0.0000
Significance F	0.0025			

Example Setup:

Set $\alpha = 0.05$

Hypothesis:

$H_0 : b_1 = b_2 = b_3 = 0$

H_a : at least one $b_i \neq 0$

Test Statistic:

$$F = \frac{MSR}{MSE} = \frac{0.289}{0.032} = 9.140$$

F stat

P-value: 0.0025

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

Since 0.0025 < 0.05
We reject H_0
This regression as a whole *is* statistically significant

We conclude the regression is a good model as a whole. Note, the results from the t test should still be addressed.

© 2002-2013 ICEAA All rights reserved.

Paring Down the Multivariate Regression Model

v1.2

- You may have a model in which some of the coefficients are significant, and some not
  - Note: If only the y-intercept is significant, then it is not really a linear model, it devolves down to a simple average
- If the F statistic is significant, but only some of the t statistics are, then you may be able to achieve a better model by removing the non-significant variables
  - Re-run the model with the least significant variable excluded
  - Compare SEE, CV, F stat, and R^2_a for the two models
  - Continue the above step until all coefficients are significant
  - Compare goodness of fit and significance statistics across all models you've seen. Using these, and sound engineering judgment, select the final model.
    - Only the y-intercept may be non-significant … in practice, it is used "as is" even if it is not significant. This is important because without the y-intercept, OLS estimators are *not* Best Linear Unbiased Estimators (BLUE).

Tip: Given logical relationship, near significance, and explained variation, it may be beneficial to keep non-significant variables in a model. Such variables should only be retained if they improve d.f.- adjusted metrics

© 2002-2013 ICEAA All rights reserved.

t and F Summary

v1.2

- The t statistics tell us if each independent variable is a good predictor
- The F statistic tells us if the regression as a whole is a good model

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

- In our example, the model was found to be significant (large F), but two of the three variables were not (small t)

Tip: If possible, test the resulting model on an independent data set.

© 2002-2013 ICEAA All rights reserved.

v1.2

Selecting the Best Model

© 2002-2013 ICEAA All rights reserved.

v1.2

Choosing a Model

- We have seen what the linear model is, and explored it in depth
- We have looked briefly at how to generalize the approach to non-linear models
- You may, at this point, have several significant models from regressions
  - One or more linear models, with one or more significant variables
  - One or more non-linear models
- Now we will learn how to choose the “best model”

© 2002-2013 ICEAA All rights reserved.

Steps for Selecting the "Best Model"

v1.2

- You should already have rejected all non-significant models first
  - If the F statistic is not significant
- You should already have stripped out all non-useful variables and made the model "minimal"
  - Variables that do not incrementally contribute to goodness of fit, overall model significance, (adjusted) variation explained, etc. were already removed
- Select "within type" based on (adjusted) R^2
  - When comparing multivariate regression models, select based on adjusted R^2, which compensates for the number of independent variables
- Select "across type" based on SSE (SEE for multivariate models)

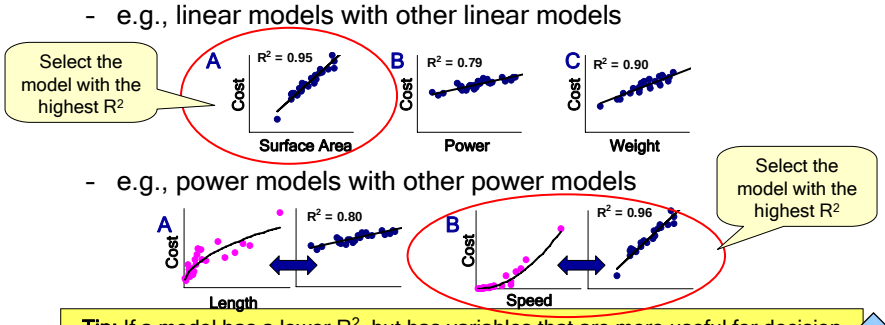
We will examine each in more detail…

© 2002-2013 ICEAA All rights reserved.

Selecting "Within Type"

v1.2

- Start with only significant, "minimal" models
- In choosing among "models of a similar form", R^2 is the criterion
- "Models of a similar form" means that you will compare
  - e.g., linear models with other linear models



  - e.g., power models with other power models

Tip: If a model has a lower R^2, but has variables that are more useful for decision makers, retain these, and consider using them for CAIV trades and the like

16

© 2002-2013 ICEAA All rights reserved.

Selecting "Across Type"

v1.2

14

- Start with only significant, "minimal" models
- In choosing among "models of a different form", the SSE in unit space is the criterion (SEE if degrees of freedom change; CV if dependent variables changes)
- "Models of a different form" means that you will compare:
  - e.g., linear models with non-linear models
  - e.g., power models with logarithmic models
- We must compute the SSE by:
  - Computing $\hat{Y}$ ***in unit space*** for each data point
  - Subtracting each $\hat{Y}$ from its corresponding actual Y value
  - Sum the squared values, this is the SSE
- An example follows…

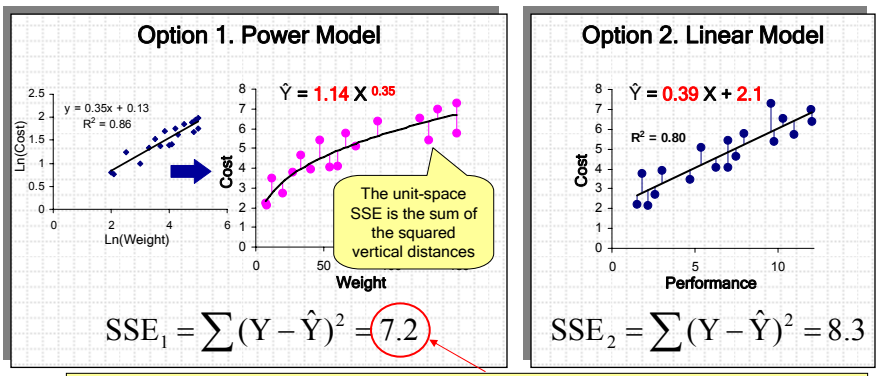
Warning: We cannot use R^2 to compare models of different forms because the R^2 from the regression is computed on the transformed data, and thus is distorted by the transformation

© 2002-2013 ICEAA All rights reserved.

Selecting "Across Type" Example

v1.2

- Suppose we want to choose between the following models for a method of estimating cost:



We choose the power model because it has the lower unit-space SSE (SEE if the two had different number of vars.)

© 2002-2013 ICEAA All rights reserved.

v1.2

Comparing Nested Models Example

- Suppose we want to choose between the following models for a method of estimating cost:

	b3	b2	b1	a
Coefficients	-0.160	-1.126	0.251	1.372
Standard Errors	0.173	0.219	0.224	0.163
R^2, SE(y)	71.4%	0.178		
F, DF	9.140	11		
SSR, SSE	0.867	0.348		
MSR, MSE	0.289	0.032		
ADJ R^2, SEE, CV	63.6%	0.178	16.1%	
T stats	-0.92	-5.14	1.12	8.41
P values	0.3758	0.0003	0.2877	0.0000
Significance F	0.0025			

		b2	b1	a
Coefficients		-1.118	0.189	1.313
Standard Errors		0.217	0.213	0.149
R^2, SE(y)		69.2%	0.177	
F, DF		13.45	12	
SSR, SSE		0.840	0.375	
MSR, MSE		0.280	0.031	
ADJ R^2, SEE, CV		64.0%	0.177	15.9%
T stats		-5.14	0.89	8.80
P values		0.0003	0.3942	0.0000
Significance F		0.0009		

- Reduced model eliminates the least significant variable (b_3). We can see by removing the least significant variable R^2, SEE, CV, significance F and adjusted R^2 all improve when b_3 is removed. A (possible) next step would be to also eliminate b_1 and compare again.

© 2002-2013 ICEAA All rights reserved.
NEW!

v1.2

Regression Summary

- Regression analysis is a powerful tool in cost analysis, particularly for developing CERs
- Two of the most important results of OLS Regression are:
  - Statistical significance
  - Uncertainty
- This module has covered:
  - The basic math behind the analysis
  - How to interpret the results from a regression tool such as Excel
  - How to apply the results and choose among models
- Many other regression techniques extend beyond the scope of this module, but can be found in the resources provided

© 2002-2013 ICEAA All rights reserved.

v1.2

Resources - Textbooks

- *An Introduction to Mathematical Statistics and It's Applications*, 3rd ed., Richard J. Larsen and Morris L. Marx, Prentice Hall, 2000
- *Applied Linear Regression Models*, Neter et al., Irwin Inc., 1996
- *Introductory Econometrics with Applications*, R. Ramanathan, Dryden Press, 1997
- *Applied Regression Analysis,* N.R. Draper and H. Smith, Wiley, 1998
- *Regression Analysis by Example,* S. Chatterjee, A. Hadi, and B. Price, Wiley, 1999
- *Regression With Graphics*, L. Hamilton, Brooks/Cole Publishing, 1992
- *Econometric Models and Economic Forecasts*, R. Pindyck and D. Rubinfeld, McGraw-Hill (College Division), 1997
- *Using Econometrics - A Practical Guide*, A. H. Studenmand, Addison-Wesley, 2000
- *A Guide to Econometrics*, P. Kennedy, MIT Press, 1998

© 2002-2013 ICEAA All rights reserved.

v1.2

Resources - Papers

- "The Multicollinearity Problem: Coping with the Persistent Beast" Kevin Cincotta, David Lee, LMI, February 2007
- "Modern Techniques of Multiplicative-Error Regression," Steve Book, SCEA, 2006
- "The Minimum-Unbiased-Percentage Error (MUPE) Method In CER Development" Shu-Ping Hu, DoDCAS, 2001
- "Why ZMPE When You Can MUPE?" Dr. Shu-Ping Hu, Alfred Smith, SCEA/ISPA, 2007
- "Testing for the Significance of Cost Drivers Using Bootstrap Sampling," Daniel I. Feldman, SCEA/ISPA, 2010
- "New Research in General Error Regression Model (GERM) Significance Testing," Kevin Cincotta, SCEA/ISPA, 2010
- "The Business Case for Bootstrapping: When You're Stuck with Incomplete Data, Here's How You Make it Work!" Brett Gelso, Glenn Grossman, Eric Druker

© 2002-2013 ICEAA All rights reserved.

v1.2

Related and Advanced Topics

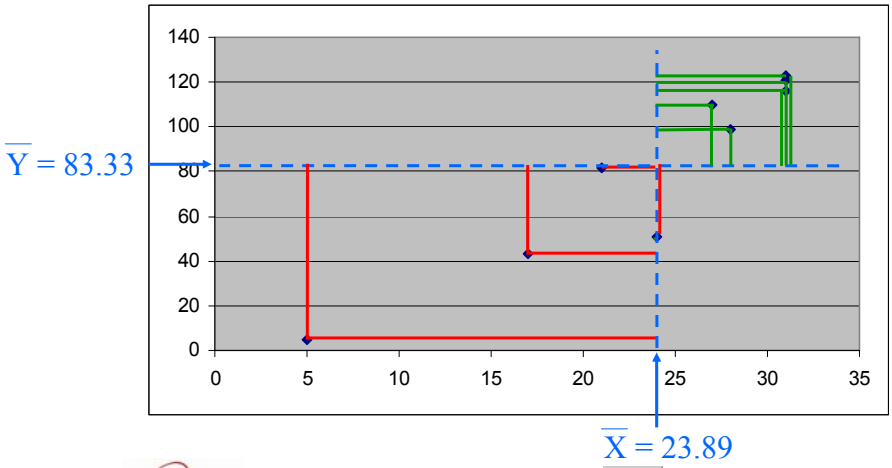
- Geometric Interpretations
- Derivation of Formulae
- White Test
- ANOVA Redux
- The Bivariate Normal Distribution and the Geometry of Regression
- Correction Factors
- Multicollinearity
- Non-OLS Models
- Maximum Likelihood Estimation

© 2002-2013 ICEAA All rights reserved.

v1.2

Geometric Interpretations

- Means = "center of gravity"

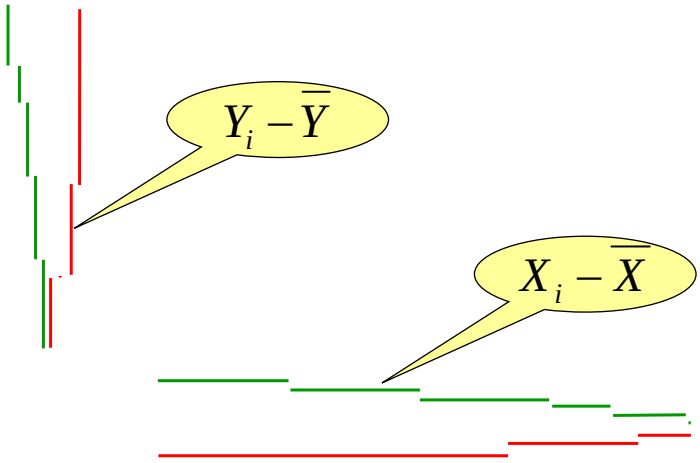


© 2002-2013 ICEAA All rights reserved.

v1.2

Geometric Interpretations

- Deviations from mean sum to zero

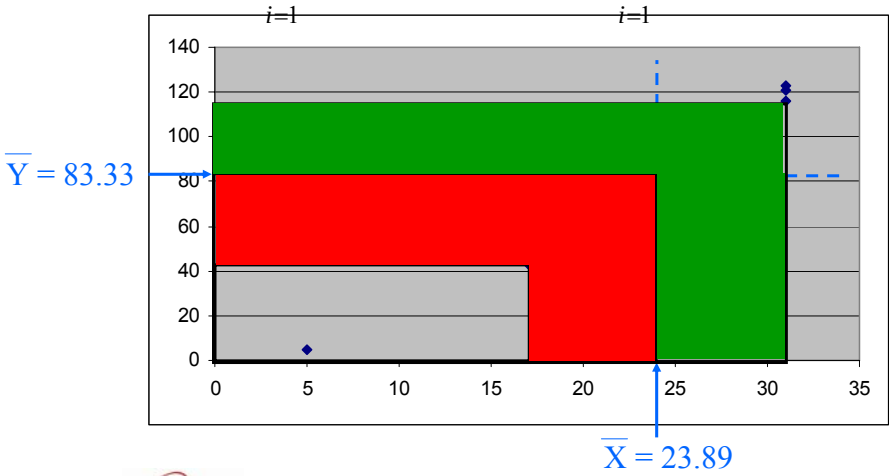


© 2002-2013 ICEAA All rights reserved.

v1.2

Geometric Interpretations

- Lemma: $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$

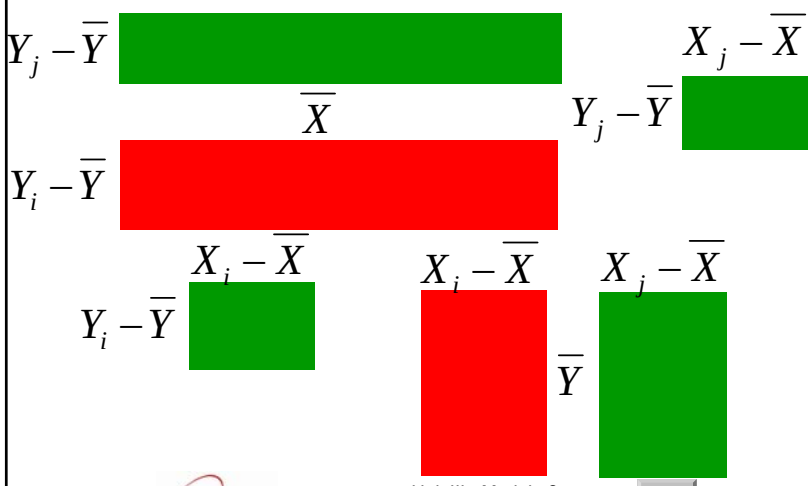


© 2002-2013 ICEAA All rights reserved.

v1.2

Geometric Interpretations

- Deviations from mean sum to zero



© 2002-2013 ICEAA All rights reserved.

v1.2

Deriving the Equations

$$SSE(a,b) = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2 = \sum_{i=1}^{n} \left(a + bX_i - Y_i\right)^2$$

$$\frac{\partial SSE}{\partial a} = 2\sum_{i=1}^{n} \left(a + bX_i - Y_i\right) = 0 \Rightarrow \sum_{i=1}^{n} Y_i = an + b\sum_{i=1}^{n} X_i$$

$$\frac{\partial SSE}{\partial b} = 2\sum_{i=1}^{n} X_i \left(a + bX_i - Y_i\right) = 0$$

$$a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^{\,2} - \sum_{i=1}^{n} X_i Y_i = 0$$

© 2002-2013 ICEAA All rights reserved.

v1.2

Deriving the Equations (cont'd.)

$$\overline{Y} = a + b\overline{X} \Rightarrow a = \overline{Y} - b\overline{X}$$

$$\left(\overline{Y} - b\overline{X}\right)\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^{\,2} - \sum_{i=1}^{n} X_i Y_i = 0$$

$$b\left(\sum_{i=1}^{n} X_i^{\,2} - \overline{X}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i$$

"easy to calculate"

"easy to remember"

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

8

© 2002-2013 ICEAA All rights reserved.

v1.2

White Test

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

© 2002-2013 ICEAA All rights reserved.

White Test Applied to Toy Problem

v1.2

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!

© 2002-2013 ICEAA All rights reserved.

Sums of Squares Shortcuts

v1.2

- These formulae are more computationally efficient:
  - Total Sum of Squares (SST): $\sum_{i=1}^{n} Y_i^2 - \overline{Y}\sum_{i=1}^{n} Y_i$
  - Residual or Error Sum of Squares (SSE): $\sum_{i=1}^{n} Y_i^2 - a\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i Y_i$
  - Regression Sum of Squares (SSR): $b\left(\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i\right)$
- Can you verify the identity using these?

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

© 2002-2013 ICEAA All rights reserved.

R^2 and Reducing CV

v1.2

- We said that one of the goals of running CERs is to reduce CV, and that R^2 is the percent explained variation
  - But how are the two related?
  - We can show that the reduction of CV is a function of R^2

for large n

$$CV_{old} = \frac{s_Y}{\bar{Y}} = \frac{1}{\bar{Y}}\sqrt{\frac{SST}{n-1}} \qquad CV_{new} = \frac{SEE}{\bar{Y}} = \frac{1}{\bar{Y}}\sqrt{\frac{SSE}{n-k}} \qquad R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$CV_{redux} = \frac{CV_{new}}{CV_{old}} = \sqrt{\frac{n-1}{n-k}}\sqrt{\frac{SSE}{SST}} = \sqrt{\frac{n-1}{n-k}}\sqrt{1-R^2} \approx \sqrt{1-R^2}$$

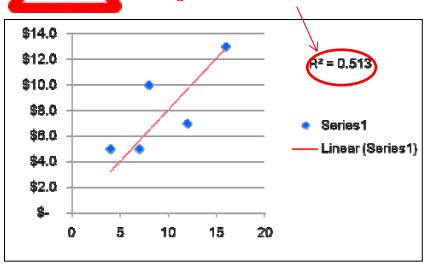
$$CV_{redux} = \frac{SEE}{s_Y} = \frac{2.46}{3.36} = \sqrt{\frac{5-1}{5-2}}\sqrt{1-0.62} = 0.71 = \frac{30.8\%}{43.3\%}$$

© 2002-2013 ICEAA All rights reserved.

Zero-Intercept R^2

v1.2

Warning: When the regression line is forced through the origin (0), R^2 and R^2_a in the trendline can be different than in the LINEST or macro output



LINEST() function

b	0.8072	0.0000	a
SEb	0.1050	#N/A	SEa
R2	0.9366	2.4152	SEE
F	59.0874	4	d.f.
SSR	344.6673	23.3327	SSE

$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$ is preferred to $R^2 = \frac{\text{SSR}}{\text{SST}}$, because

$R^2 = \frac{\text{SSR}}{\text{SST}}$ can only be used when $SSR + SSE = SST$

Only the case for OLS regression in fit space

© 2002-2013 ICEAA All rights reserved.

v1.2

Leverage

- Leverage is a measure of how far an observation is from the average values of all the independent variables in the equation

C.I. Formula

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

= Leverage

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

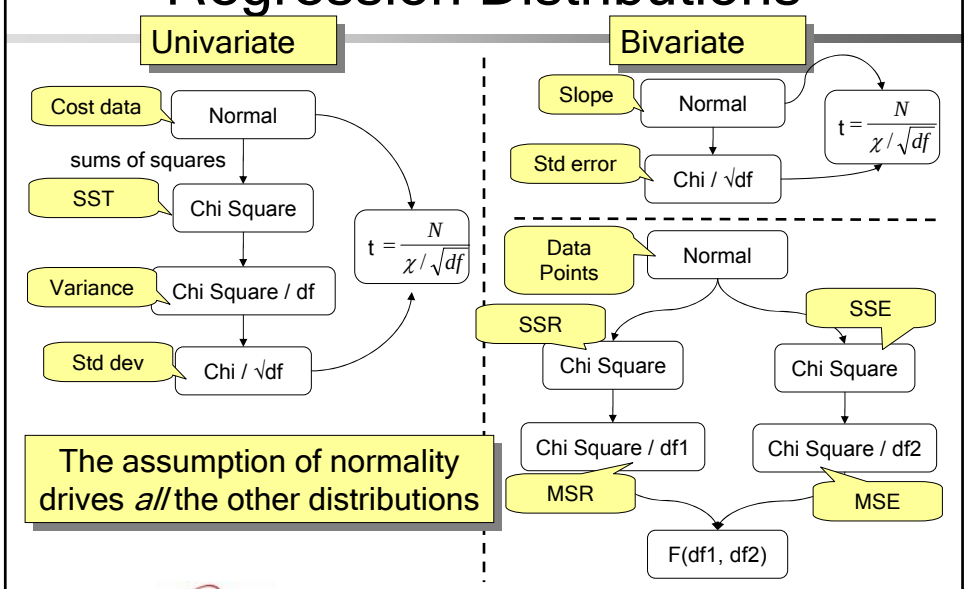
P.I. Formula

An observation is considered a potential outlier with respect to X if its leverage value is greater than 2*(p/n), where p is the number of parameters and n is the number of observations.

© 2002-2013 ICEAA All rights reserved.

v1.2

Regression Distributions



The assumption of normality drives *all* the other distributions

© 2002-2013 ICEAA All rights reserved.

The Geometry of Regression

v1.2

- The following charts show the geometry of regression by building up a picture
  - The picture provides a mental image that aids in understanding the regression equation
  - This visual framework has potential applications in risk analysis

9

- The below facts enable us to derive the picture
  - For any two jointly distributed variables, there is a regression line
    - The slope is:

$$b = \rho*(\sigma_y / \sigma_x)$$

    - The y intercept is:

$$a = \mu_y - \rho(\sigma_y / \sigma_x) * \mu_x$$

  - If the variables are joint bivariate normal, then ρ is the correlation coefficient

10

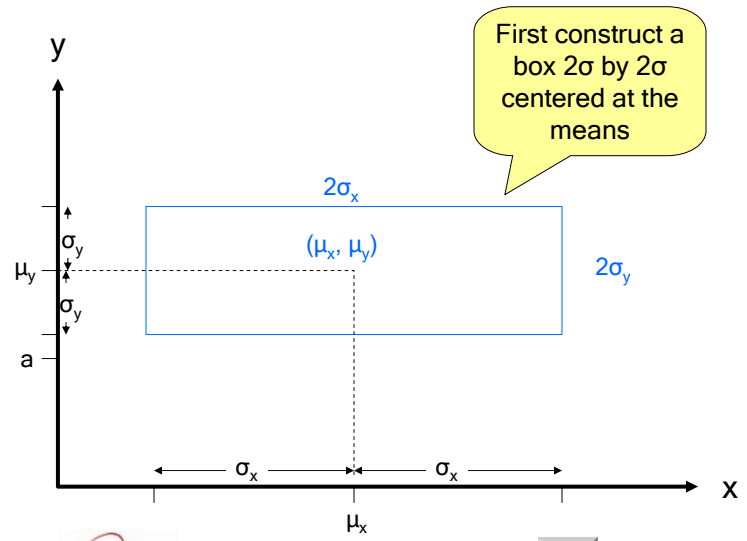
Let's look at the graph…

© 2002-2013 ICEAA All rights reserved.

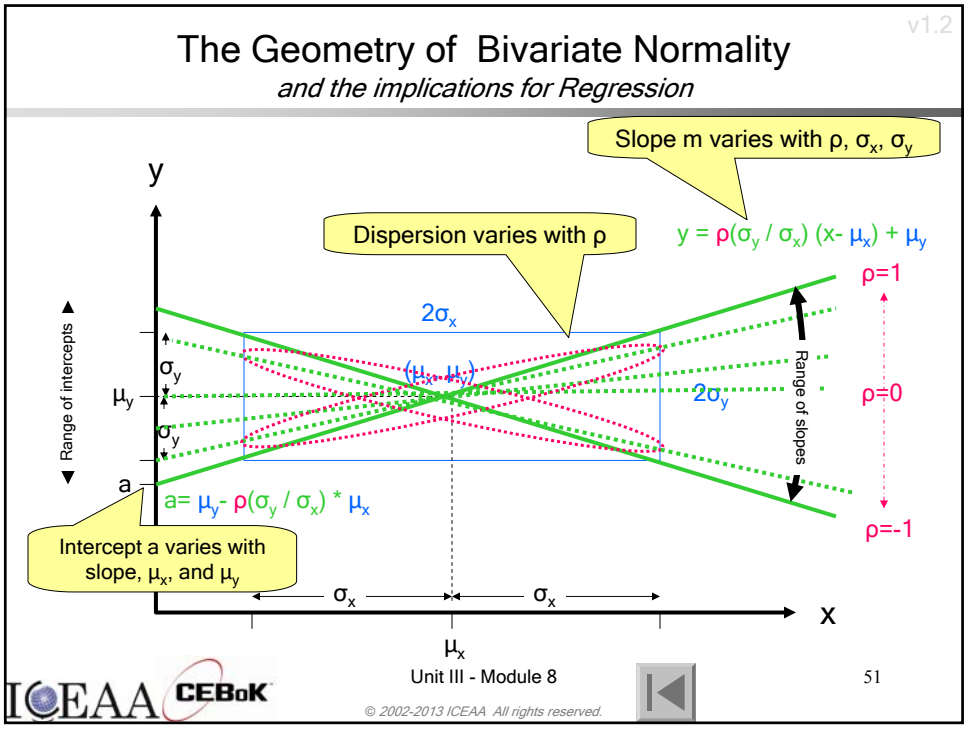
The Geometry of Bivariate Normality

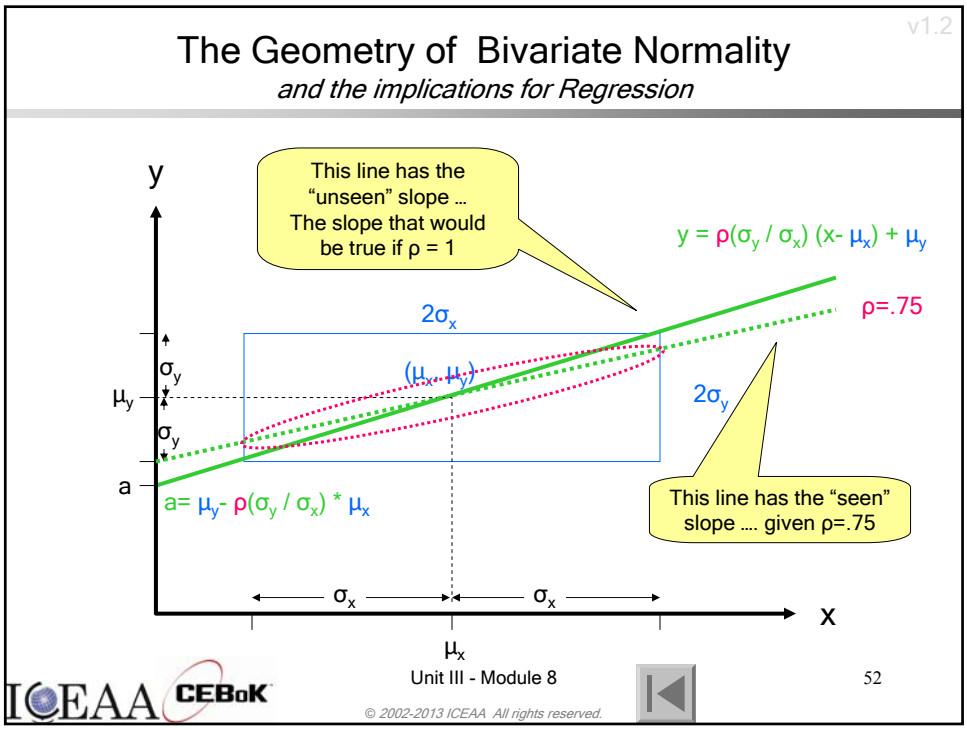
and the implications for Regression

v1.2



© 2002-2013 ICEAA All rights reserved.


v1.2
The Geometry of Bivariate Normality
and the implications for Regression
Slope m varies with ρ, σx, σy
Dispersion varies with ρ
y = ρ(σy / σx) (x- μx) + μy
ρ=1
ρ=0
ρ=-1
Range of intercepts
Range of slopes
2σx
2σy
(μx, μy)
σy
μy
a
a= μy- ρ(σy / σx) * μx
Intercept a varies with slope, μx, and μy
σx
μx
y
x
Unit III - Module 8
51
© 2002-2013 ICEAA All rights reserved.


v1.2
The Geometry of Bivariate Normality
and the implications for Regression
This line has the "unseen" slope … The slope that would be true if ρ = 1
y = ρ(σy / σx) (x- μx) + μy
ρ=.75
2σx
2σy
(μx, μy)
σy
μy
a
a= μy- ρ(σy / σx) * μx
This line has the "seen" slope …. given ρ=.75
σx
μx
y
x
Unit III - Module 8
52
© 2002-2013 ICEAA All rights reserved.

v1.2

Correction Factors

- When converting CERs developed in log-linear space to not log-linear space, the CER will predict closer to the median than the mean

$$\text{Goldberger Factor}(GF) \cong e^{\left((1-r_0)\frac{s^2}{2}\right)}$$

s = standard error of the estimate

r_0 = leverage value in log space if $\mathbf{x}_o$ is a vector of independent variables in the data matrix

$$\text{PING Factor}(PF) \cong e^{\left(\left(1-\frac{p}{n}\right)\frac{s^2}{2}\right)}$$

P = # of estimated coefficients

n = sample size

s = standard error of the estimate

9

© 2002-2013 ICEAA All rights reserved.

v1.2

Multicollinearity

- Multicollinearity occurs when there is a strong linear relationship among two or more independent variables
  - The model form of this linear relationship must match the model form of the regression in order for multicollinearity to occur
- Some symptoms that multicollinearity may be occurring are:
  - Large changes in the values of the regression coefficients when another variable is added or deleted
  - Regression coefficients having an opposite sign from what intuition predicts
  - Two (independent) variables thought to be similar have large (in absolute value) but "opposite" signs
  - Variables expected to be significant are not
  - A high overall R^2 with several non-significant independent variables
  - High Variance Inflation Factors (VIFs) or Variance Amplification Factors (VAFs)
- The existence of Multicollinearity has a couple of adverse effects on the results of OLS:
  - Biases coefficients and inflates their standard errors
    - This in turn biases t-tests and p-values; also makes them imprecise
  - Makes it difficult to understand the effect each independent variable has on predicting the outcome
    - **It is important to note that multicollinearity does not affect the reliability of the model predictions; it simply biases individual coefficient values and their estimated significance.**

NEW!

© 2002-2013 ICEAA All rights reserved.

v1.2

Multicollinearity

- One of the most robust methods to address multicollinearity is to find the VIF of each independent variable
- The VIF of an independent variable is defined as $\frac{1}{(1-R_{U-\beta}{}^2)}$ where $R_{U-B}{}^2$ is the coefficient of determination when the dependent variable β is regressed against all other dependent variables
  - The VIF of β is the multiplicative factor (>=1) by which the variance of β is increased due to correlation among the regressors
  - All information needed to compute the VIF for each variable is found as part of Excel's Linest() function
- In general, a VIF of over 10 indicates a severe enough problem to take a second look. If no VIF exceeds 4, you may reasonably conclude that there is no issue with multicollinearity.
- If multicollinearity is clouding the results of a regression model, consider removing the variable with the largest VIF and re-running the model, understanding that the variable to be removed may be the intercept!
- Continue until no VIF exceeds 4 (ideally) or 10 (if desperate)

Note: VIFs may be calculated in this manner in standard OLS regression. For zero-intercept regression, $R^2_{U-\beta}$ can't be used because it assumes a constant term. However, the VIF can be also be calculated as $SE_{\beta j}{}^2/SE_{\beta j}{}^2{}_{native}$ where

$$\mathbf{SE_{\beta j}{}^2{}_{native} = SEE^2/[(n-1)Var(X_j)]}$$

NEW!

© 2002-2013 ICEAA All rights reserved.

v1.2

Multicollinearity - Example

Before removing multicollinearity

Regression Statistics	
Multiple R	0.977595073
R Square	0.955692126
Adjusted R Square	0.933538189
Standard Error	3.390877227
Observations	10

Notice neither independent variable is significant

ANOVA

	df	SS	MS	F	Significance F
Regression	3	1488.03271	496.0109	43.138704	0.000187089
Residual	6	68.9882902	11.498048		
Total	9	1557.021			

	Coefficients	Standard Error	t Stat	P-value	VIF*
Intercept	1.619602821	2.704279863	0.5989036	0.571141	
x1	3.645827501	3.059240194	1.1917428	0.2783581	67.15169364
x2	-0.856380194	3.226126284	-0.2654515	0.7995466	67.07078039
x3	2.146718658	0.306108425	7.0129356	0.0004193	1.036809323

- The VIFs for x1 and x2 are high, indicating that multicollinearity is present
  - This is further verified by scatter plotting them together (see below)
- By rerunning the regression with each of the variables removed, the best regression is found

After removing multicollinearity

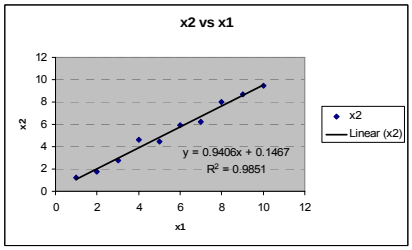
Regression Statistics	
Multiple R	0.977328896
R Square	0.955171771
Adjusted R	0.942363705
Standard E	3.157722838
Observatior	10

ANOVA

	df	SS	MS	F	Significance F
Regression	2	1487.222505	743.6113	74.5758	1.90735E-05
Residual	7	69.79849467	9.971214		
Total	9	1557.021			

	Coefficients	Standard Error	t Stat	P-value	VIF*
Intercept	1.489038566	2.47633097	0.601308	0.56659	
x1	2.840039751	0.353958217	8.023658	8.94E-05	1.036596958
x3	2.147881586	0.285031425	7.535596	0.000133	1.036596958

*VIF's were computed separately, not as part of Excel ANOVA results



© 2002-2013 ICEAA All rights reserved.

v1.2

Multicollinearity – Ridge Regression

- Ridge regression is one of several ways of regularizing a regression model so that all the independent variables may remain in the analysis
  - Should mainly be used when there is extremely high correlation between the independent variables
- In ridge regression, a ridge variable is added to the SSE expression, such that solutions with inflated SE_β values are no longer optimal
- Advantages of Ridge Regression
  - Reduces the standard errors of the estimated coefficients
  - No independent variable is removed from the analysis
- Disadvantages of Ridge Regression
  - Model estimates will be biased!
  - Coefficients lose some of their interpretability
  - Set the ridge too high, and estimates are biased beyond recognition (recall: multicollinearity does *not* bias overall model estimates). Set it too low, and the multicollinearity problem is not remedied.

Warning: Ridge regression is trial-and-error intensive!

© 2002-2013 ICEAA All rights reserved.

v1.2

A *Very* Brief Overview of Two Non-OLS Regression Techniques

Weighted Least Squares Regression
Multiplicative Error Regression

*These slides are meant as a top-level overview of these techniques, not an instruction guide. For more detailed information, seek the resources provided in the Resources section.

© 2002-2013 ICEAA All rights reserved.

Non-OLS Regression Techniques
Weighted Least Squares

v1.2

- Weighted Least Squares regression is similar to Ordinary Least Squares in that it still works by minimizing the sum squared error
  - The difference is instead of treating the errors associated with the data points equally, certain points are weighted

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 \quad \text{vs.} \quad \sum_{i=1}^{n} w_i (y_i - f(x_i))^2$$

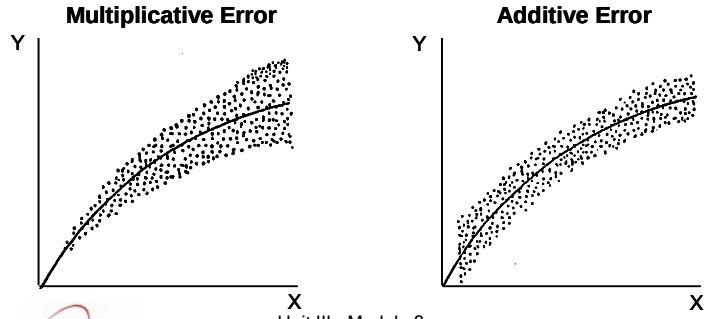
  - One common method is setting w_i to $\frac{1}{\sigma^2{}_i}$
    - Thus giving higher weight to points with lower variance in measurement of the x's. When (and only when) this weighting convention is used, WLS estimators are BLUE.
- WLS regression is useful in many cases
  - To compensate for a violation of the homoscedasticity assumption of OLS (funnel-shaped residual plots)
  - When certain data points are believed to be more correct or applicable than other data points

© 2002-2013 ICEAA All rights reserved.

Non-OLS Regression Techniques
Multiplicative Error Regression

v1.2

- OLS seeks to minimize the additive error of the regression
  - $Y = ax^b + \varepsilon$
- However, non-linear functions may exhibit multiplicative error instead
  - $Y = (ax^b) \cdot \varepsilon$
- When this is the case, multiplicative error techniques must be used
  - Examples include MUPE and ZMPE
  - Papers on these techniques are listed in a special section on the resources slide
- When prediction intervals around OLS-transformed regressions are produced they demonstrate a multiplicative error pattern as well



© 2002-2013 ICEAA All rights reserved.

Non-OLS Regression Techniques: Maximum Likelihood Estimation

v1.2

- If you believe the function to be linear but still have issues with heteroscedasticity, maximum likelihood estimation (MLE) may be in order
- This *generalization of OLS* accounts for non-constant error term variances
- MLE solutions reduce to OLS solutions when variance is held constant, so MLE estimators are OLS estimators when OLS assumptions hold
- MLE estimators are asymptotically BLUE for large data sets; even with heteroskedasticity, as long as other OLS assumptions hold (e.g. zero-mean, normal i.i.d. error term)
- If s is constant, log likelihood objective function = $\Sigma\{\ln(1/\sigma) - \varepsilon^2/2\sigma^2\}$ is *maximized* when SSE is minimized
- Must specify σ^2 as a function of x
- Other remedies for heteroscedasticity include generalized least squares (GLS, a generalization of WLS) and transformation to log space

ICEAA CEBoK

Unit III - Module 8

© 2002-2013 ICEAA All rights reserved.

NEW!

61