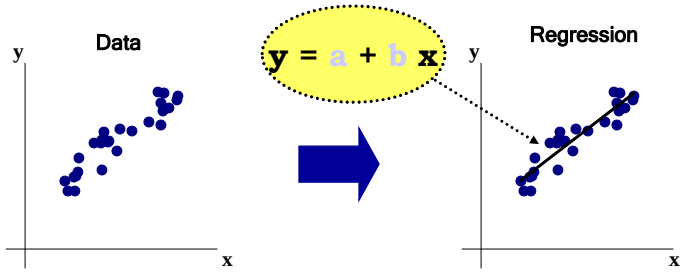
Cost Estimation

Chapter 7:
Single Variable Linear Regression

Gregory K. Mislick, LtCol, USMC (Ret)
Department of Operations Research
Naval Postgraduate School

Definition of Regression

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



6 - 2

Basic Statistics vs. Regression Example #1

- **Consider the following data set of the cost of 15 homes that are for sale in a particular town, in the column to the right:**

Home #	Price ($)
1	300,000
2	400,000
3	350,000
4	800,000
5	450,000
6	250,000
7	225,000
8	450,000
9	550,000
10	400,000
11	220,000
12	350,000
13	365,000
14	600,000
15	750,000

Basic Statistics vs. Regression Example #1 (cont)

- **From Chapter 5, we learned how to find the Descriptive Statistics from this data set. Results are shown:**
- **Questions: What if you wanted to buy a house that was specifically 1,200 square feet in size? Or 2,000 sq ft?**
- **You cannot determine the cost for any house size from this data set!**

Descriptive Statistics	
Mean	430666.6667
Standard Error	45651.02269
Median	400000
Mode	400000
Standard Deviation	176805.6506
Sample Variance	31260238095
Kurtosis	0.194942263
Skewness	0.92371377
Range	580000
Minimum	220000
Maximum	800000
Sum	6460000
Count	15

Basic Statistics vs. Regression Example #1 (cont)

- **Here is where Regression Analysis can help greatly.**
- **What was missing from the previous data set was the size of each home with its associated cost.**
- **Note the new data set to the right now showing the size of each home along with its cost.**

Home#	Price($)	Square Feet
1	300,000	1400
2	400,000	1800
3	350,000	1600
4	800,000	2200
5	450,000	1800
6	250,000	1200
7	225,000	1200
8	450,000	1900
9	550,000	2000
10	400,000	1700
11	220,000	1000
12	350,000	1450
13	365,000	1400
14	600,000	1950
15	750,000	2100

Basic Statistics vs. Regression Example #1 (cont)

- With Cost as the Dependent variable, and Square Feet as the Independent (or explanatory) variable, we calculate the following regression results:

  Cost = -$311,221.87 + $450.539 * # Square Feet

- We can now answer the Questions from Slide #4:
- ***Predictions: 1200 sq ft = $229,424.93***
- **2,000 sq ft home = *$589,856.13***

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.919362365
R Square	0.845227157
Adjusted R Square	0.833321554
Standard Error	72183.15525
Observations	15

Cost vs Square Feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.69908E+11	3.69908E+11	70.99406372	1.26178E-06
Residual	13	67735302725	5210407902		
Total	14	4.37643E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-311221.8767	90000.56578	-3.45799911	0.004242478	-505656.2781
Square Feet	450.5396012	53.47144891	8.425797513	1.26178E-06	335.021559

Basic Statistics vs. Regression Example #1: Conclusions

- **You are now able to come up with a "prediction" for any given house size due to the regression equation:**

$$\hat{Y} = -311{,}221.87 + 450.539 * X$$

or in words,

Cost = -$311,221.87 + $450.539 * # Square Feet

- **Clearly, this is much more helpful and more informative than just using descriptive statistics.**
- **Note: Regression is not ALWAYS better than the Descriptive Statistics. If the R-squared is very low, and Standard Error is high, the basic statistics may be preferable to the regression.**

6 - 7

In A Linear Regression Model

- **Cost is the dependent (or unknown) variable; generally denoted by the symbol Y.**
- **The system's physical or performance characteristics form the model's known, or independent, variables which are generally denoted by the symbol X.**
- **The linear regression model takes the following form:**

$$Y_i = b_0 + b_1 X_i + e_i$$

where b_0 (the Y intercept) and b_1 (the slope of the regression line) are the unknown regression parameters and e_i is a random error term.

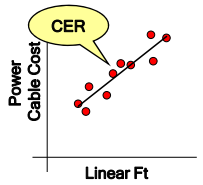
6 - 8

Regression Analysis in Cost Estimating

- **If the dependent variable is a cost, the regression equation is often referred to as a *Cost Estimating Relationship,* or *CER***
  - **The independent variable in a CER is often called a *cost driver.* A CER may have one or multiple cost drivers:**

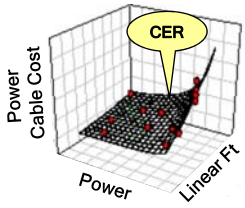
Examples of cost drivers:

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



Example with multiple (2) cost drivers:

Cost	Cost Driver (multiple)
Power Cable	Linear Feet
	Power



6 - 9

Three Primary Symbols in Regression

You will see these on the next slide and continually throughout the regression chapters:

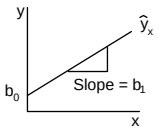
- **Y_i = any of the data points in your data set, and there are "i" of them**
- **$\hat{Y}$ = Y(hat) = the estimate or "prediction" of Y provided by the regression equation**
- **$\bar{Y}$ = Y(bar) = the mean or average of all "i" cost data points**

6 - 10

Linear Regression Model

- **We desire a model of the form:**

$$\hat{y}_x = b_0 + b_1 x$$

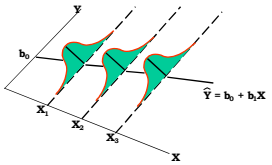


- **This model is estimated on the basis of historical data as:**

$$y_i = b_0 + b_1 x_i + e_i$$

where

$$e_i \sim N(0, \sigma_x^2), \text{ and iid}$$



- **<u>In words</u>: "Actual Cost" = "Estimated Cost" + "Error of Estimation"**

"The True Purpose of What Regression is Trying to Accomplish"

The slope and intercept (b_1 and b_0) are chosen such that the sum of the squared residuals is minimized (Least Squares Best Fit). You are trying to minimize the difference between the actual cost (Yi) and your predicted cost ($\hat{Y}$). Solving for the Error of Estimation, we get:

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y} = \text{residuals}$$

$$\sum (y_i - \hat{y})^2 = \text{minimum}$$

Least Squares Best Fit (LSBF)

- **To find the values of b_0 and b_1 that minimizes $\sum (y_i - \hat{y})^2$ one may refer to the "Normal Equations."**

$$\sum Y = nb_0 + b_1 \sum X$$
$$\sum XY = b_0 \sum X + b_1 \sum X^2$$

- **With two equations and two unknowns, we can solve for b_0 and b_1.**

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \overline{Y} - b_1 \overline{X}$$

Example #1 Revisited

- **Let's re-analyze the cost of the 15 homes for sale in the Example #1 data set.**
- **After computation, we found that the average sale price of all the homes in your data set was $430,666.67. Thus,**

  $\overline{Y}$ = $430,666.67
- ***Then* you developed an estimating relationship between home price and its size in square feet using LSBF regression:**

  $\hat{Y}$ = -$311,221.87 + $450.539 * X
- **Now you want to estimate the home price of a 2,000 square foot home:**

  $\hat{Y}$ = -311,221.87 + 450.539 * X
  = -311,221.87 + (450.539 * 2,000)
  = $589,856.13

Example #1 Conclusions

- **What do these numbers mean?**
- **$430,666.67 is the estimate of the average sale price of all homes in that data set.**
- **$589,856.13 is the estimate for a home in the data set *that has a size of 2,000 sq ft.***

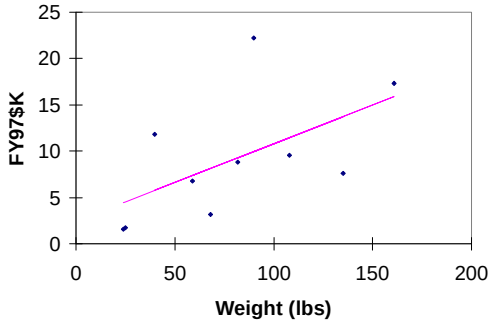
- **We use regression to try to get a better "prediction" than just using the mean.**
- ***Key Point*: If the statistics for the regression are not very good, you can always go back and use the mean. A good regression means you prefer the regression equation as an estimator, instead of using the mean.**

Another Example

- **Recall the radio data in Ch 5 used on the mean and standard deviation. Now let's look at the relationship between the average unit production costs and their associated weight:**

Historic Transmogrifier
Average Unit Production Cost

System	FY97$K	Weight (lbs)
1	22.2	90
2	17.3	161
3	11.8	40
4	9.6	108
5	8.8	82
6	7.6	135
7	6.8	59
8	3.2	68
9	1.7	25
10	1.6	24



The Regression Model

- **The first time, we'll crank it out by hand…**

Y = Average Unit Production Cost (FY97$K)

X = Weight (lbs)

$n = 10$

$$\overline{X} = \frac{\sum X_i}{n} = 79.2 \text{ lbs} \qquad \overline{Y} = \frac{\sum Y_i}{n} = \$9.06K$$

$$\sum XY = 8{,}739.4 \qquad \sum X^2 = 81{,}540$$

$$\hat{b}_1 = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2} = \frac{8{,}739.4 - (10)(79.2)(9.06)}{81{,}540 - (10)(79.2)^2} = 0.0831$$

$$\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X} = 9.06 - 0.0831(79.2) = 2.48$$

$$\therefore \hat{Y}_X = \$2.48K + (\$0.0831K)X$$

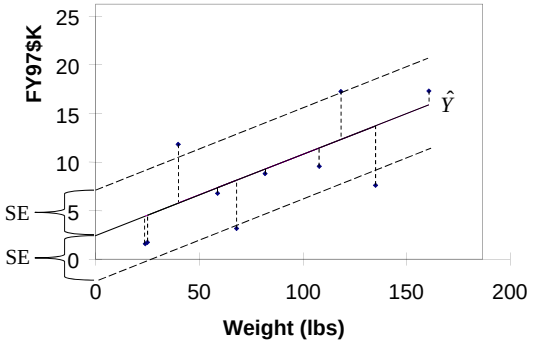
- **In words, this says that the cost of the radio equals: $2.48k + (.0831k * the weight of the radio)**

Standard Error

- **Standard Error = $S_{\hat{y}}$ = the standard deviation about the regression line. The smaller the better.**

$$SE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

n-k-1, where k is number of independent variables

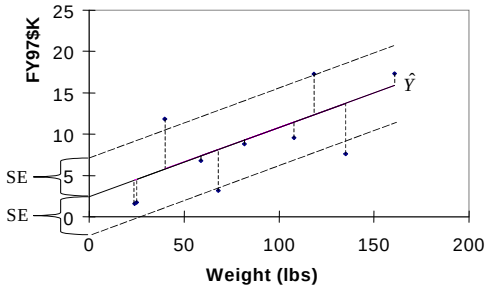


Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
22.2	10.0	12.2	149.87
17.3	15.9	1.4	2.07
11.8	5.8	6.0	35.98
9.6	11.5	-1.9	3.44
8.8	9.3	-0.5	0.24
7.6	13.7	-6.1	37.19
6.8	7.4	-0.6	0.34
3.2	8.1	-4.9	24.30
1.7	4.6	-2.9	8.15
1.6	4.5	-2.9	8.25
		Sum	269.83

Standard Error

$$SE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{269.83}{8}} = \$5.8K$$

- **For the radio data, the standard error is $5.8K.**
- **This means that *on "average" when predicting the cost of future systems we will be off by +/- $5.8K in one standard error.***

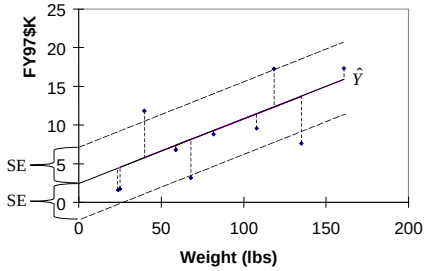


Standard Error vs Standard Deviation: Similar Concept, but.....

When working with Basic Statistics, you use *Standard Deviation*

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$
$$= \frac{172.7 + 67.9 + \cdots + 55.7}{10-1}$$
$$= \frac{399.8}{9} = 44.4 (\$K^2)$$

When performing a regression, you now use *Standard Error*



Coefficient of Variation

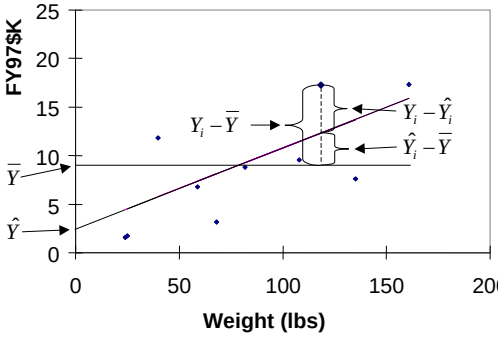
- **Coefficient of Variation (CV):**

$$CV = \frac{SE}{\overline{Y}} = \frac{\$5.8K}{\$9.06K} = 64\%$$

- **This says that *on "average", we'll be off by 64% when predicting the cost of future systems*. The smaller the better. In Ch5, CV was 73% with the same data, so using the regression is slightly better than just using univariate statistics.**

Analysis of Variance

- **Analysis of Variance (ANOVA)**



SST = Total Sum of Squares = $\sum (Y_i - \overline{Y})^2$

SSE = Sum of Squared Errors = $\sum (Y_i - \hat{Y}_i)^2$

SSR = Sum of Squares Regression = $\sum (\hat{Y}_i - \overline{Y})^2$

$SST = SSE + SSR$

SST = Total Variation

SSE = Unexplained Variation

SSR = Variation explained by Regression

	df	*SS*	*MS = SS/df*	*F = MSR/MSE*	*Significance F P(b1=b2=0)*
SSR	1	130.00	130 (MSR)	3.85	0.0852
SSE	8	269.83	33.7 (MSE)		
SST	9	399.82			

Coefficient of Determination

- **Coefficient of Determination (R^2) represents *the percentage of total variation explained by the regression model.* The larger the better (as close to 1.0 as you can)**

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2 = \frac{130}{399.8} = 1 - \frac{269.8}{399.8} = 0.3252 = 32.5\%$$

- **Since R^2 always increases when independent variables are added, Adj. R^2 helps to adjust for the number of independent variables. This is necessary when comparing regression models with an *unequal number of independent variables* (i.e., one independent variable vs. two)**

$$R^2_{adj} = 1 - \frac{SSE/(n-(k+1))}{SST/(n-1)} = 1 - \frac{269.8/8}{399.8/9} = 0.2408 = 24.1\%$$

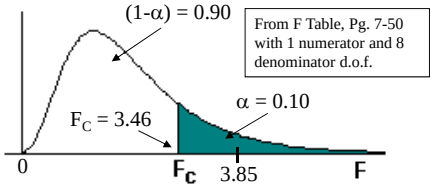
The F-Statistic

- **The F statistic tells us whether the full model, $\hat{Y}$, is preferred to the mean, $\bar{Y}$.**
- **Say we want to test the strength of the relationship between our model and Y at the $\alpha = 0.1$ significance level…**

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ (The model is invalid) (prefer } \bar{Y})$$

$$H_a : H_0 \text{ is false (The model is valid) (prefer } \hat{Y})$$

$$\text{Test statistic : } F = \frac{MSR}{MSE} = \frac{SSR/df_R}{SSE/df_E} = \frac{130/1}{269.8/8} = 3.85$$



- **Since 3.85 falls within the rejection region, we reject H_0 and say *the full model is better than the mean as a predictor of cost.***

The t-statistic

- **This statistic tests the marginal contribution of the *independent variable* on the reduction of the unexplained variation.**
- **In other words, it tests the strength of the relationship between Y and X (or between Cost and Weight) by testing the strength of the coefficient b_1.**
- **The t-statistic is used to test the hypothesis that X and Y (or Cost and Weight) are NOT related at a given level of significance.**
- **If the test indicates that that X and Y <u>are</u> related, *then we say we prefer the model <u>with</u> b_1 to the model <u>without</u> b_1, which is the desired result.***

$$\hat{y}_x = b_0 + b_1 x$$

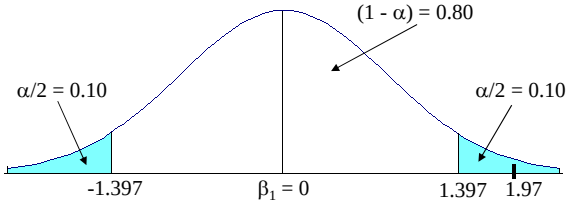
The t statistic

$H_0 : \beta_1 = 0$ (Cost is not related to weight) (prefer model without weight)

$H_a : \beta_1 \neq 0$ (Cost and weight are related) (prefer model with weight)

$$\text{Test statistic}: t_{\beta_1} = \frac{b_1 - \beta_1^{\nearrow 0}}{s_{b_1}} = \frac{b_1}{s_{b_1}} = \frac{b_1}{SE \Big/ \sqrt{\sum (X - \overline{X})^2}} = \frac{0.0831}{5.8 \Big/ 137.16} = 1.97$$

- **Say we wish to test b_1 at the $\alpha = 0.20$ significance level. Refer to a T-Table with 8 degrees of freedom…**



- **Since our test statistic, 1.97, falls within the rejection region, we reject H_0 and conclude that we prefer the model <u>with</u> b_1 to the model <u>without</u> b_1.**

F-Stats and T-Stats Summary

- **There will always be *only one F-stat* in a regression, regardless of how many independent variables there are.**
- **There will always be one t-stat for *each* independent variable.**
- **Thus, in a single variable regression, there will be exactly one F-stat and one t-stat, and the F and t-stats will be virtually identical in value.**
- **In a two (three) variable regression, there will again be only one F-stat, and there will be exactly two (three) t-stats, one for each independent variable.**

There's an Easier Way…

- **Linear Regression Results (Microsoft Excel):**

Regression Statistics	
Multiple R	0.5702
R Square	0.3251
Adjusted R Square	0.2408
Standard Error	5.8076
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	130.00	130.00	3.85	0.0852
Residual	8	269.83	33.73		
Total	9	399.82			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.477	3.823	0.648	0.535	-6.340	11.293
Weight (lbs)	0.083	0.042	1.963	0.085	-0.015	0.181

- **Now the information we need is seen at a glance.**

Important Results

- **From the Regression output, we can glean the following important results:**
  - **R^2 or Adj. R^2: The bigger the better.**
  - **Standard Error: The smaller the better.**
  - **CV: Divide Standard Error by $\bar{Y}$. The smaller the better.**
  - **Significance of F: If less than a significance level, then we prefer the model $\hat{Y}$ to the mean $\bar{Y}$. Else, vice versa.**
  - **P-value of coefficient b_1: If less than a significance level, then we prefer the model <u>with</u> b_1, else we prefer it <u>without</u> b_1.**
- **These statistics will be used to compare other linear models when more than one cost driver may exist.**
- **Discuss Regression Hierarchy**

Regression Hierarchy

- **Consider there are four cars in a lot that you can use, and you need to narrow down your choices and pick the "best" car for your needs.**
- **<u>Phase I of the Hierarchy</u>: This is similar to trying each car, to see which one actually starts and drives and brakes. The ones that do so are considered "good" or "functional" cars. Let's say one of the four does not start, so it is now excluded from further consideration (yes, even if it's a Mazzerati) (or Cost vs Weight)**
- **<u>Phase II of the Hierarchy</u>: Now that you have identified which cars start and drive, we test the "performance characteristics" (ANOVA), to see which one handles the best, drives the fastest, etc. From those metrics, we pick our "best" car that we intend to use.**

A Note About Regression

- **Ensure that when making a prediction using a regression equation, that the input of the independent variable (say Weight) is within the range of the data. If the smallest value for Weight was 100 lbs, and the largest value was 500 lbs, ensure that what you are trying to predict falls within the range of 100-500 lbs.**

- **Two Examples:**
  - **You are predicting the cost of a house based on square feet, with historical square foot data values of between 1000 to 3000 sq ft. What if predicting the cost of a 3200 sq ft house?**

  - **You are predicting the cost of an aircraft with maximum speeds of between 500 knots and Mach 2. What if predicting the cost of an aircraft with a max speed of Mach 2.3?**

Treatment of Outliers

- **In general, an outlier is a residual that falls greater than 2σ from $\overline{Y}$ or $\hat{Y}$.**
- **The standard residual is**

$$\frac{Y_i - \hat{Y}}{SE} \quad or \quad \frac{X_i - \overline{X}}{s_X} \quad or \quad \frac{Y_i - \overline{Y}}{s_Y}$$

- **Recall that since 95% of the population falls within 2σ of the mean, then in any given data set, we would expect 5% of the observations to be outliers.**

- **In general, do not throw them out unless they do not belong in your population.**

Outliers with respect to X

- **All data should come from the same population. You should analyze your observations to ensure this is so.**
- **Observations that are so different that they do not qualify as a legitimate member of your independent variable population are called outliers with respect to the independent variable, X.**
- **To identify outliers with respect to X, simply calculate $\bar{X}$ and S_X. Those observations that fall greater than two standard deviations from $\bar{X}$ are likely candidates.**
- **You *expect* 5% of your observations to be outliers, therefore the fact that some of your observations are outliers is not necessarily a problem. You are simply identifying observations that warrant a closer investigation.**

6 - 33

Example Analysis of Outliers with Respect to X

Range	$\bar{X}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$\frac{(X_i - \bar{X})}{S_X}$	
600	823	-223	49785	-0.59	
925	823	102	10379	0.27	
450	823	-373	139222	-0.99	
420	823	-403	162510	-1.07	
1000	823	177	31285	0.47	
800	823	-23	535	-0.06	
790	823	-33	1097	-0.09	
1600	823	777	603535	2.06	⇦

S_X 377.65

6 - 34

Outliers with Respect to Y

- **There are two types of outliers with respect to the dependent variable:**
  - **Those with respect to Y itself.**
  - **Those with respect to the regression model, $\hat{Y}$.**
- **Outliers with respect to Y itself are treated in the same way as those with respect to X.**
- **Outliers with respect to $\hat{Y}$ are of particular concern, because those represent observations our model does not predict well.**
- **Outliers with respect to $\hat{Y}$ are identified by comparing the residuals to the standard error of the estimate (SE). This is referred to as the "standardized residual."**

$$\frac{(Y_i - \hat{Y})}{SE} = \#\text{ of Standard Errors}$$

- **Outliers are those with residuals greater than ±2 std errors.**

Remedial Measures

- **Remember: the fact that you have outliers in your data set is not necessarily indicative of a problem. The trick is to determine WHY an observation is an outlier.**

- **Possible reasons why an observation is an outlier.**
  - **Random Error: No problem**
  - **Not a member of the same population: If so, you want to delete this observation from your data set.**
  - **You've omitted one or more other cost drivers.**
  - **Your model is improperly specified.**
  - **The data point was improperly measured (it's just plain wrong).**
  - **Unusual event (war, natural disaster).**
  - **Requires normalization.**

Remedial Measures

- **Your first reaction should not be to throw out the data point!**
- **Assuming the observation belongs in the sample, some options are:**
  - **Dampen or lessen the impact of the observation through a transformation of the dependent and or independent variables.**
  - **Develop two or more regression equations (one with and one without the outlier)**
- **Outliers should be treated as useful information.**

6 - 37

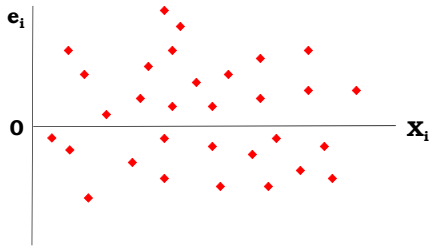
Residual Analysis

- **So what if you do a regression, and you find that the f-stats are high, the t-stats are high, R-squared's are low, and Std Errors and CV's are high? Not a good regression!**
- **Perhaps you are trying to fit a straight line to data that is not linear.**
- **How to tell? Plot the original data and see if it is linear or not.**
- **You can also check the residual plots.**

6 - 38

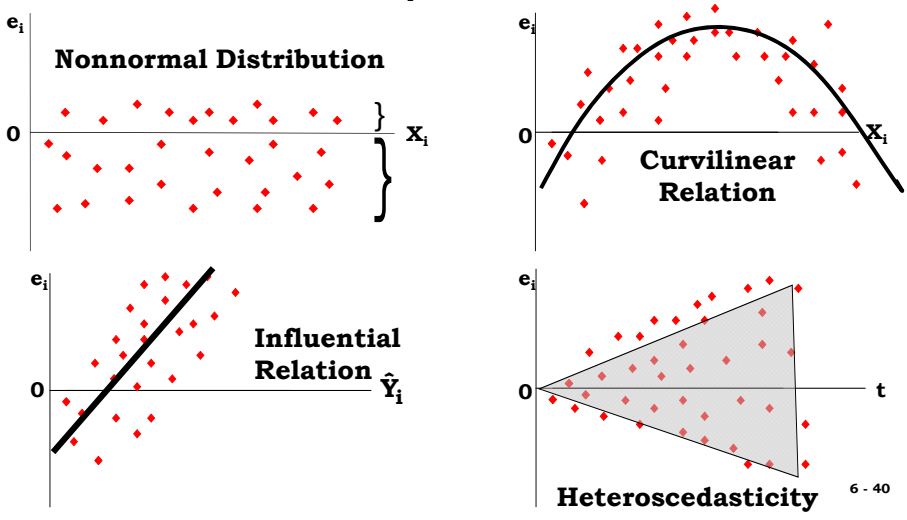
Model Diagnostics

- **If the fitted model is appropriate for the data, there will be no pattern apparent in the plot of the residuals versus X_i, $\hat{Y}_i$, etc.**
  - **Residuals spread uniformly across the range of X-axis values**



Model Diagnostics

- **If the fitted model is not appropriate, a relationship between the X-axis values and the e_i values will be apparent.**



Non-Linear Models

- **Data transformations should be tried when residual analysis indicates a non-linear trend:**

  $X' = 1/X \quad X' = 1/Y \quad X' = \log X \quad Y' = \ln Y \quad Y' = \log Y$

  - **CER are often non-linear when the independent variable is a performance parameter:**

    $$Y = aX^b$$

    $$\log Y = \log a + b \log X \Rightarrow Y' = a' + bX'$$

    - **log-linear transformation allows use of linear regression**
    - **predicted values for Y are in "log dollars" which must be converted back to dollars**

Other Concerns

- **When the regression results are illogical (i.e., cost varies inversely with a physical or performance parameter), omission of one or more important variables may have occurred or the variables being used may be interrelated**
  - **Does not necessarily invalidate a linear model**
  - **Additional analysis of the model is necessary to determine if additional independent variables should be incorporated or if consolidation/elimination of existing variables is required**

Assumptions of OLS

(1) Fixed X

- **Can obtain many random samples, each with the same X values but different Y_i values due to different e_i values**

(2) Errors have mean of 0

- **$E[e_i] = 0$**

(3) Errors have constant variance (homoscedasticity)

- **$Var[e_i] = \sigma^2$ for all I**

(4) Errors are uncorrelated

- **$Cov[e_i,e_j] = 0$ for all $i \neq j$**

(5) Errors are normally distributed

- **$e_i \sim N(0, \sigma^2)$**

Multivariate Linear Regression and Non-Linear Regression

ICEAA 2016 International Training Symposium
Bristol, 17th to 20th October 2016

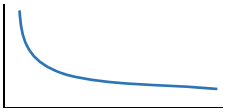
Alan R Jones
Estimata Limited
Promoting TRACEability in Estimating

Least Squares Non-Linear Regression

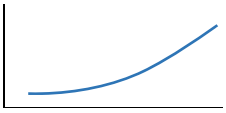
Simple Non-Linear Regression

Many Cost Estimating Relationships are not Linear

- When we plot cost data against a cost driver, it is often appears to be a curve
- For example:
  - A Learning Curve relationship curves down sharply from top-left to bottom-right



  - Escalation (normally) causes cost to curve up from the bottom left to the top-right



- Projecting curves is not as "Straight Forward" as is it for Straight Line Relationships
- This is where Logarithms often come to the rescue, where we express a number as the power to which a base value must be raised to get that number
- Why such glum faces? Does it bring back bad memories of school?

3

Logs: Just a Question of Perspective

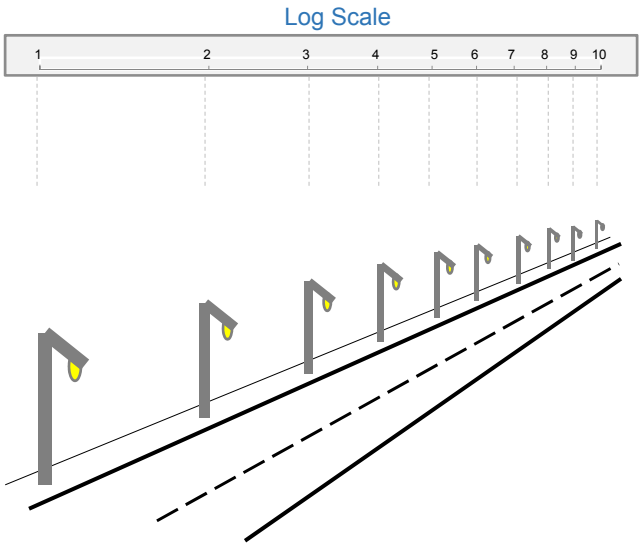
If you are not happy about using Logs, just think of them being a question of perspective

Consider a row of street lamps on a straight road:

- They are equally spaced and yet they appear to get closer together and smaller as they disappear into the distance

The same is true with integers:

- They are equally spaced but in a Log perspective they get closer together



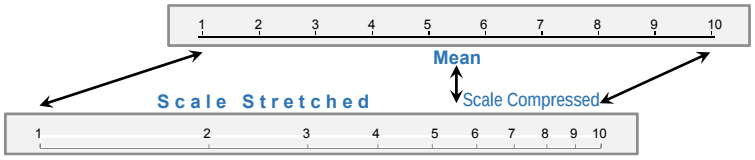
4

Logs: Just a Question of a Different Perspective

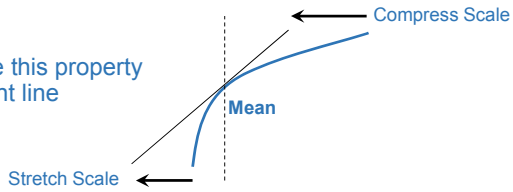
In essence, relative to the Arithmetic Mean of the data (simple average):

- Logs stretch the relative difference between equally spaced smaller values
- and compress the relative difference between equally spaced larger values

Consider, the integers from 1 to 10, their average is 5.5

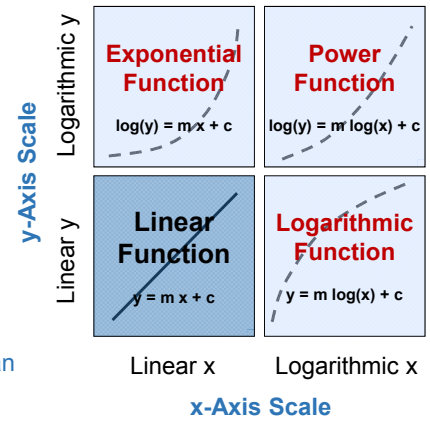


In certain circumstances we can use this property to unbend curves, giving us a straight line



Linear Transformations: Summary

There are three groups of functions that allow us to transform a relationship into a linear form



Curves and their transformations can have positive or negative gradients

Why is this important?

If by plotting any of the combinations of:

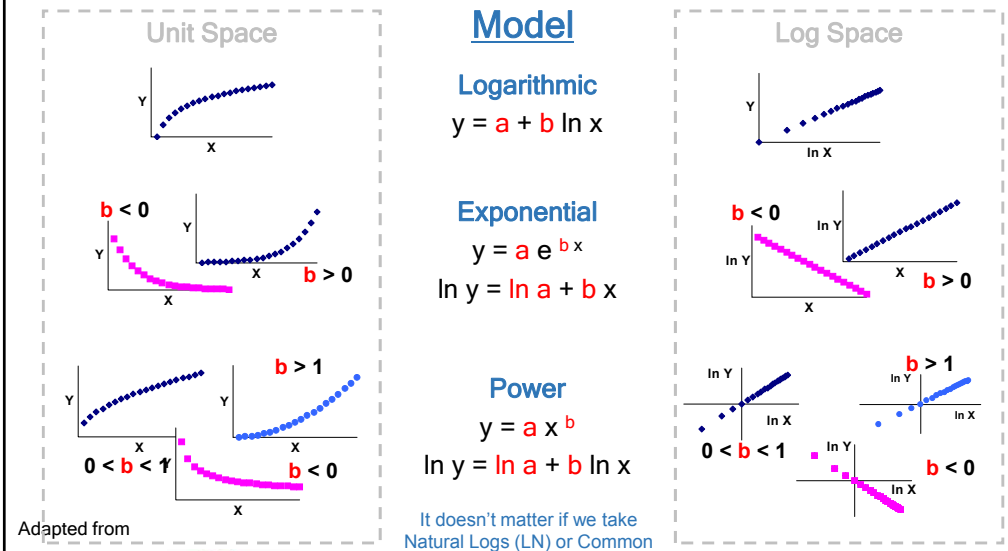
x or Log(x)

against

y or Log(y)

we get a straight line, then we can perform a linear regression on the transformed data

Linear Transformations



Adapted from

It doesn't matter if we take Natural Logs (LN) or Common Logs (base 10) or any other base

Function Types – A Word of Caution

We can use Logarithmic Transformation to convert many curved relationships into linear ones
However, there are occasions when they should not be used …

Time is sometimes used as a secondary measure or indicator of technology, or project maturity:

- We can take the Log of Elapsed Time,
  but we should **NEVER take the Log of a Date!**
  It presupposes that we know when time began

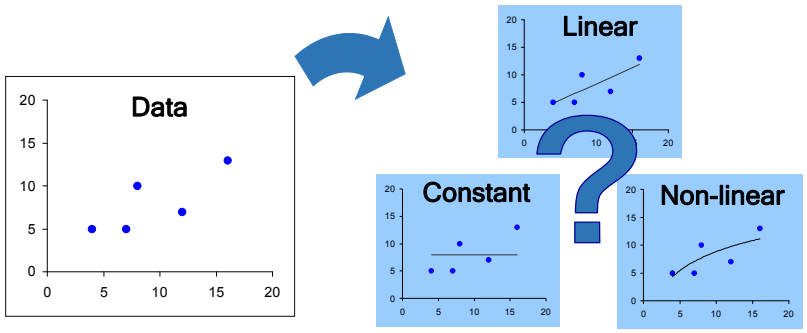
We can use a Binary Switch in Multi-variate Regression to signify whether a cost driver or cost element driver is active (1) or inactive (0)

- **NEVER take the Log of a Binary Switch!**
  Log(0) implodes – it can't be done
  Taking the Log of any Numerical Categorical Variable (zero or not) is highly questionable from a logic perspective

Creates a **#NUM!** error in Microsoft Excel

Determining the Appropriate Regression Model

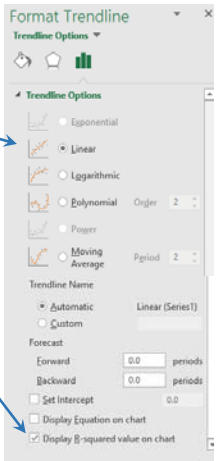
- A scatter plot should always be performed first to determine what kind of model should be tested, if any at all:
  - Specifying the wrong function for a model can lead to an incorrect interpretation of the results



Adapted from

Choosing the Appropriate Function Type

- Plot the data in Microsoft Excel with standard linear axes
- Right-click on the data. Select “Add Trendline…”
- Select the “Linear” option (default)
- Select “Display R-Squared Value on Chart”
  - Note the R-Squared Value
  - Note the scatter pattern



- Change the Trendline Option, noting the changes in the R-Squared Value and scatter pattern of each:
  1. Exponential … only if all the y-values are positive
  2. Logarithmic … only if all the x-values are positive
  3. Power … only if all the x-values and y-values are positive
     (We can’t take the Log of a non-positive number)

- As a general rule we are looking for the highest R-Squared Value and an “even” scatter around the Trendline (Homoscedastic)

Non-linear Regression Function Summary

- Before we can perform our Regression we must transform the data based on the Function Type
  - 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

It doesn't matter which base of Logs we use LOG_{10} or LN, so long as we are consistent when transforming back

Note: Other functions can be used to transform data (e.g., √x, sin x, etc.) but logarithms are the most common

- Of these, Power Functions are the most common (e.g. Learning Curves, Cost/Weight CERs)
- Time-based relationships might be Exponential Functions (e.g. escalation)
- Logarithmic Functions are less common in practice (but never say "never")
- We perform the Regression on the Transformed Data, and transform the output back to "real world" space afterwards

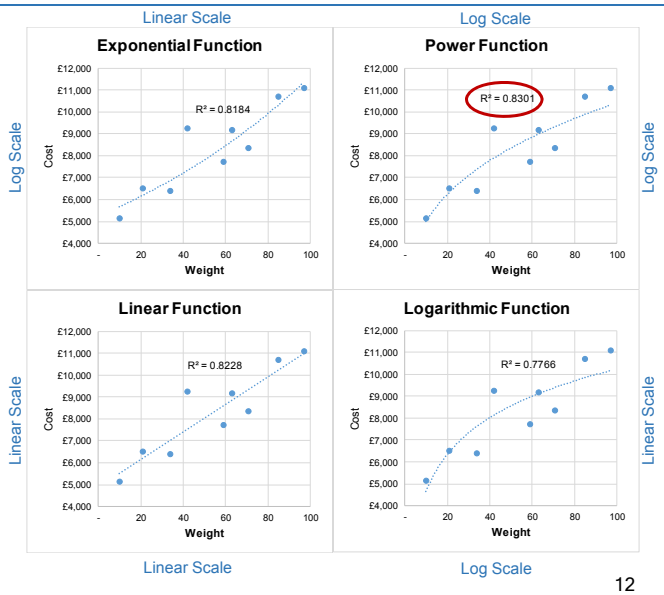
LOG10 => Transform by raising 10 to the power of the Output
LN Transform by raising "e" to the power of the Output

Adapted from:

ICEAA CEBoK

Example of Non-Linear Regression

Cost £	Weight Kg
5123	10
6527	21
6388	34
9253	42
7722	59
9182	63
8348	71
10702	85
11092	97



Plot the data as a scatter diagram and try each function type Trendline in turn

Look for the best (highest) R^2

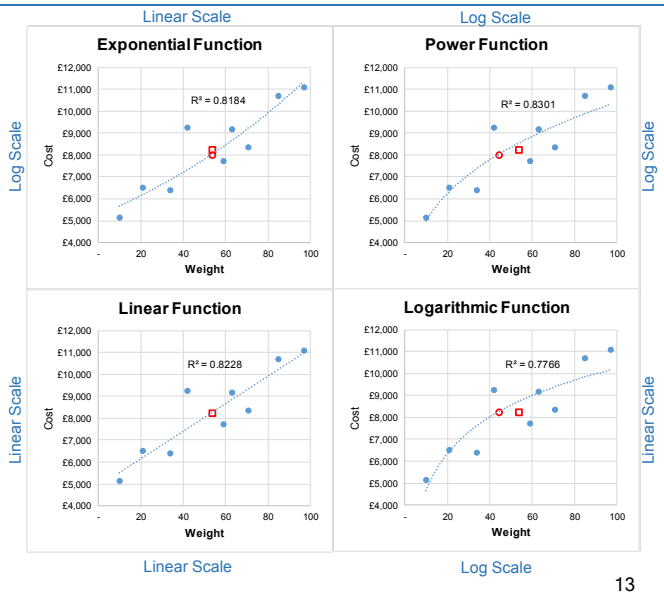
In this case the Power Function appears to be marginally better than the Linear or Exponential Functions

Example of Non-Linear Regression

Cost £	Weight Kg
5123	10
6527	21
6388	34
9253	42
7722	59
9182	63
8348	71
10702	85
11092	97

One criterion for "Best Fit" was that the line passed through the Arithmetic Mean □ of the data

The "Best Fit" now passes through the Arithmetic Mean of the Transformed Data – equivalent to the Geometric Mean O of the untransformed raw data



Example of Non-Linear Regression

Cost £	Weight Kg
5123	10
6527	21
6388	34
9253	42
7722	59
9182	63
8348	71
10702	85
11092	97

Regression

CV = 11%

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.907072211
R Square	0.822779996
Adjusted R Square	0.797462853
Standard Error	906.6035088
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	26711840.55	26711840.55	32.49892701	0.000734755
Residual	7	5753509.455	821929.9221		
Total	8	32465350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4896.171288	682.8970207	7.386020958	0.000151211	3328.668916	6463.673659
Weight Kg	62.80385562	11.0167069	5.700783017	0.000734755	36.7534833	88.85422794

Model is significant by all measures: R-Square, F and t Statistics are all high, and the CV is low.

Log Transform

Log Cost £	Log Wgt
3.7095	1.0000
3.8147	1.3222
3.8054	1.5315
3.9663	1.6232
3.8877	1.7709
3.9629	1.7993
3.9216	1.8513
4.0295	1.9294
4.0450	1.9868

Regression

CV = 1.3%

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.911100819
R Square	0.830104702
Adjusted R Square	0.805833945
Standard Error	0.049035414
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.082237376	0.082237376	34.20184657	0.000631733
Residual	7	0.016831303	0.002404472		
Total	8	0.099068679			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.381360755	0.090973005	37.16883662	2.65273E-09	3.166243781	3.596477728
Log Wgt	0.317954346	0.054367578	5.848234483	0.000631733	0.189395452	0.44651324

Model is significant by all measures. R-Square, F and t Statistics have all increased, and the Standard Error and CV have decreased.

But is it a better model?

Unit-Space Goodness of Fit Comparison

Just as we wouldn't compare linear measurements in different scales (imperial v metric) so too we cannot compare between Linear and Log Scales

Stat	Linear Model	Power - Fit Space	Power - Unit Space	How to Calculate
SSE	5753509.5	0.017	6339427.3	$SUMSQ(e_i)=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$
R^2	0.823	0.830	0.805	$1-\frac{SSE}{SST}=1-\frac{SUMSQ(e_i)}{DEVSQ(y_i)}$
Adj R^2	0.797	0.806	0.777	$1-\left[\left(1-R^2\right)\left(\frac{n-1}{(n-1)-k}\right)\right]$
SEE	906.6	0.049	951.6	$\sqrt{\frac{SSE}{(n-1)-k}}$
CV	11.0%	1.3%	11.5%	$\frac{SSE}{\bar{Y}}=\frac{s_Y}{\bar{Y}}\sqrt{1-R_a^2}$

not df-adjusted

df-adjusted

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

In this case the Linear Model is the better option

Adapted from:

ICEAA CEBoK

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

15

Least Squares Multivariate Linear Regression

16

Least Squares Multivariate Linear Regression

How does it differ from Simple Linear Regression?

- Simple Linear Regression allows the "best fit" straight line to be determined through a set of data points. It assumes that the value of the dependent variable (e.g. y) varies directly to a change in the independent variable (e.g. x)
- Multi-Variate Linear Regression allows the dependent variable (y) to vary in portion to changes in more than one independent variable (e.g. x_1, x_2, x_3 etc)
- Just as one dependent and one independent variable defines a straight line, the addition of a second independent variable defines a 2-D plane
- The addition of third independent variable defines a 3-D surface
- The addition of other independent variables … is impossible to illustrate with a physical analogy but can be done

General Equation

- Multi-Variate Linear Regression allows us to find solutions of the form:

$$y = m_1 x_1 + m_2 x_c + \ldots + m_n x_n + c$$

17

Least Squares Multi-Variate Linear Regression

When should we consider Multi-Variate linear regression techniques?

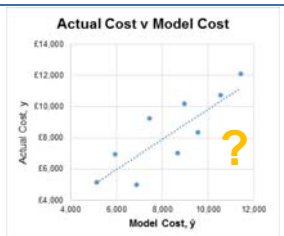
- ➢ When we suspect that the value of dependent variable (e.g. cost or effort) is dependent on the value of more than one other cost driver variable
- \+ Actual data is characterised by variations:
  - – some of which are a consequence of a change in the value of other cost driver variables
  - – others are due to errors of a more random or unpredictable nature
- \+ When you need to interpolate or extrapolate to a later or earlier value in the sequence or for a different combination of cost driver variable values

When are regression techniques not appropriate?

- ➢ When you have less data than variables (too many cost drivers, too little data)
- ➢ When you suspect that the data is from different populations but you do not have any differentiating term or factor ("*apples and oranges*")

18

Simple Linear Regression with One Cost Driver



Suppose we have a product with two potential cost drivers – weight and number of attachment interfaces

A regression against the **Weight** Cost Driver gives a result that is statistically significant on all 3 measures:
R-Square,
F and t Statistics

…but the scatter is not brilliant

	y	x1	ŷ	ε
	Cost £	Weight Kg	Model £	Error £
	5123	10	5,524.21	-401.21
	6527	21	6,215.05	311.95
	6388	34	7,031.50	-643.50
	9253	42	7,533.93	1719.07
	7722	59	8,601.60	-879.60
	9182	63	8,852.81	329.19
	8348	71	9,355.25	-1007.25
	10702	85	10,234.50	467.50
	11092	97	10,988.15	103.85
Average	8259.667	53.556	8259.667	0.00
CV	11.0%			

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.907072211
R Square	0.822779996
Adjusted R Square	0.797462853
Standard Error	906.6035088
Observations	9

ANOVA

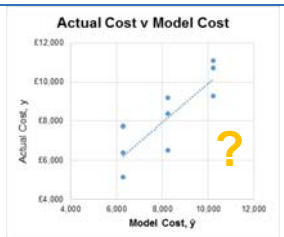
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	26711840.55	26711840.55	32.49892701	0.000734755
Residual	7	5753509.455	821929.9221		
Total	8	32465350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4896.171288	662.8970207	7.386020958	0.000151211	3328.668916	6463.673659
Weight Kg	62.80385562	11.0167069	5.700783017	0.000734755	36.7534833	88.85422794

The Coefficient of Variation, CV = 17%, **Std Err** **Average**

For a Simple Linear Regression these are always the same

Simple Linear Regression with One Cost Driver



Suppose we have a product with two potential cost drivers – weight and number of attachment interfaces

A regression against the **Attachments** Cost Driver gives a result that is statistically significant on all 3 measures:
R-Square,
F and t Statistics

…but the scatter is not brilliant

	y	x1	ŷ	ε
	Cost £	Attachments	Model £	Error £
	5123	1	6,290.67	-1167.67
	6527	2	8,259.67	-1732.67
	6388	1	6,290.67	97.33
	9253	3	10,228.67	-975.67
	7722	1	6,290.67	1431.33
	9182	2	8,259.67	922.33
	8348	2	8,259.67	88.33
	10702	3	10,228.67	473.33
	11092	3	10,228.67	863.33
Average	8259.667	2.0	8,259.67	0.00
CV	13.9%			

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.846469466
R Square	0.716510557
Adjusted R Square	0.676012065
Standard Error	1146.64629
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23261766	23261766	17.69227749	0.004004501
Residual	7	9203584	1314797.714		
Total	8	32465350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4321.666667	1011.246975	4.273601577	0.003684291	1930.447545	6712.885788
Attachments	1969	468.1163876	4.206218906	0.004004501	862.0806372	3075.919363

The Coefficient of Variation, CV = 13.9%, **Std Err** **Average**

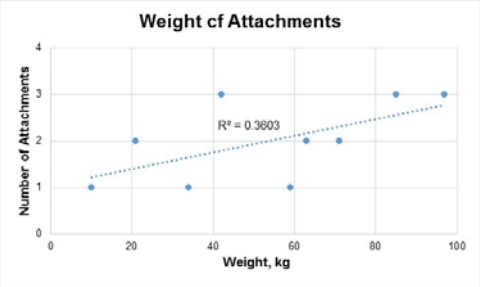
Watch Out for Collinearity Between Cost Drivers

Where we have two or more Cost Drivers, we need to ensure that any two are independent of each other

We do this by creating a scatter plot for each pair and measuring their Coefficient of Determination, R^2

Looking for a low value, certainly less than 0.5 or 50%

In our example we have a R^2 value of 0.36 which is equivalent to 60% Correlation



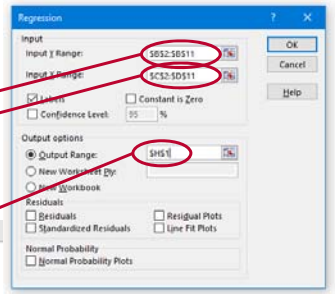
So the two are not totally independent of each other; there is an element of correlation but not sufficient necessarily to stop us from trying to create a Regression Model that uses both

If they were too closely correlated then Regression would reject one of them as being insignificant through the t-statistic P-values

Multivariate Linear Regression with Microsoft Excel

The result is statistically significant on all 3 measures:
Adjusted R-Square,
F and t Statistics

… why Adjusted R-Square this time, and ~~not just the~~ simple R-Square?



	A	B	C	D
1		Y	X1	X2
2		Cost £	Weight Kg	Attachments
3		5123	10	1
4		6527	21	2
5		6388	34	1
6		9253	42	3
7		7722	59	1
8		9182	63	2
9		8348	71	2
10		10702	85	3
11		11092	97	3
12				
13	Average	8259.667	53.556	2.0
14				
15	CV	5.2%		

Columns must be Contiguous!

The Coefficient of Variation, CV = 5.2%,

Std Err Average

Adjusted R-Square

Why do we need to use Adjusted R-Square, and what is it?

- Suppose we believe that weight is the primary cost driver
- But we are unhappy with the residual variation in that simple relationship as we believe that there is at least one secondary driver at work.
- If we add another variable to the mix, Least Squares Regression will attempt to fit any additional variable to the residual error

  … but in the process it will even sacrifice some of the best fit relationship already lined up for the primary driver in order to minimise the total error or residual

- The Regression routine will always find the Least Squares Best Fit – even where the relationship is tenuous
- Regression just a dumb calculation – no artificial intelligence involved

  … the estimator/analyst has to provide that.

- Every time we add a variable we reduce the degrees of freedom for the Fit Criteria by one
- Adjusted R-Square is a statistic that compensates for this reduction

R^2 and Adjusted R^2

R^2 or R-Square expresses the percentage of total variation in the data that can be explained by the model

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

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

SST = SSR + SSE

Adjusted R^2, or R^2_a, makes an adjustment to the Unexplained Variation to account of the degrees of freedom within the model (***n - 1 - k***)

- Can be used to compare coefficients of determination between models with different numbers of variables, ***k***
- Can be used as justification for including near-significant variables in models if those variable improve the model's performance

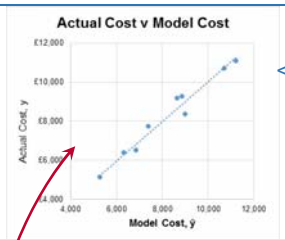
$$R^2_a = 1 - \frac{SSE}{SST}\left(\frac{n-1}{n-1-k}\right) = 1 - (1 - R^2)\left(\frac{n-1}{n-1-k}\right)$$

Penalty (> 1)

Adapted from

ICEAA CEBoK

Multivariate Linear Regression with Microsoft Excel



<<< Looking for an even scatter across model (homoscedasticity)

Regression Model
$\hat{y} = m_1 x_1 + m_2 x_2 + c$

	y	x1	x2	ŷ	ε
	Cost £	Weight Kg	Attachments	Model £	Error £
	5123	10	1	5,280.66	-157.66
	6527	21	2	6,853.81	-326.81
	6388	34	1	6,317.06	70.94
	9253	42	3	8,858.79	394.21
	7722	59	1	7,396.64	325.36
	9182	63	2	8,667.51	514.49
	8348	71	2	9,012.98	-664.98
	10702	85	3	10,715.68	-13.68
	11092	97	3	11,233.88	-141.88
Average	8259.667	53.556	2.0	8259.667	0.00
CV	5.2%				

Regression Error
$\varepsilon = y - \hat{y}$

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.982518575
R Square	0.96534275
Adjusted R Square	0.953790333
Standard Error	433.0434464
Observations	9

ANOVA

	df	SS	MS	F	Significance F
Regression	2	31340290.24	15670095.12	83.56197418	4.16277E-05
Residual	6	1125259.758	187526.6265		
Total	8	32465550			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3750.686831	391.4911862	9.575622231	7.41284E-05	2792.253025	4709.120636
Weight Kg	43.18334894	6.579372295	6.563445113	0.000599257	27.08420489	59.28249298
Attachments	1098.135796	221.0418144	4.968000283	0.002531905	557.2659612	1639.005632

m_1 m_2 c

Accepting / Rejecting the Multivarate Model

Use the F and t-Statistics to Guide us

- The F-Statistic tells us whether the model overall is valid
  - Looking for a High F => Low Sig-F
- The t-Statistic tells us whether each parameter is significant in terms of its contribution overall to the model
  - Looking for a High t => Low P-Value
- If one or more parameters are not significant, reject the least significant first (highest P-Value) and re-run the regression model
- Repeat until all remaining parameters are significant
  - This procedure is referred to as Stepwise Regression
- Rejecting the Intercept parameter is a debatable issue
- CEBoK's advice is not to reject the intercept

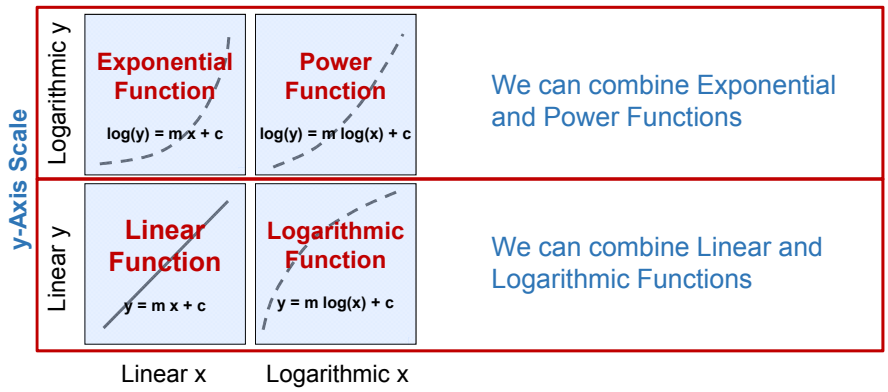
Adapted from

ICEAA CEBoK

Multivariate Models Using Linear Transformation
Selecting the Best Model

Combining Different Functional Transformations

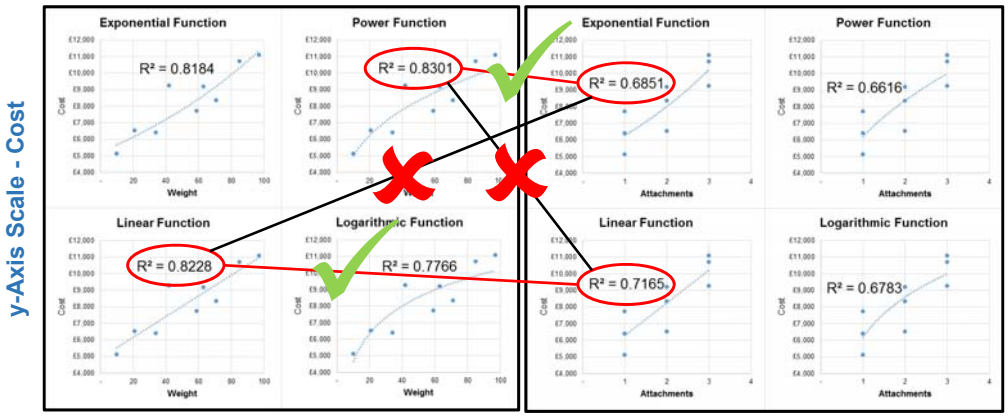
The function type of the dependent (y) variable dictates which Function Types we can combine



We cannot combine Linear and Exponential or Logarithmic and Power Functions

Combining Different Functional Transformations

Example:



Example: Multivariate Linear Model

y	x1	x2
Cost £	Weight Kg	Attachments
5123	10	1
6527	21	2
6388	34	1
9253	42	3
7722	59	1
9182	63	2
8348	71	2
10702	85	3
11092	97	3
Average	8259.667	53.556	2.0

The Coefficient of Variation, CV = 5.2%

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.982518575
R Square	0.96534275
Adjusted R Square	0.953790333
Standard Error	433.0434462
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	31340190.24	15670095.12	83.56197418	4.16277E-05
Residual	6	1125159.759	187526.6265		
Total	8	32465350			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3750.686831	391.6911862	9.575622231	7.41284E-05	2792.253025	4709.120636
Weight Kg	43.18334894	6.579372295	6.563445113	0.000599257	27.08420489	59.28249298
Attachments	1098.135796	221.0418144	4.968000283	0.002531905	557.2659612	1639.005632

The result is statistically significant on all 3 measures: Adjusted R-Square, F and t Statistics

$$Cost = c + m_1 \times Weight + m_2 \times Attachments$$

Example: Multivariate Transformed Non-linear Model

y	x1	x2
Log Cost £	Log Wgt	Attachments
3.7095	1.0000	1
3.8147	1.3222	2
3.8054	1.5315	1
3.9663	1.6232	3
3.8877	1.7709	1
3.9629	1.7993	2
3.9216	1.8513	2
4.0295	1.9294	3
4.0450	1.9868	3
Average	3.905	1.646	2.000

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.988416994
R Square	0.976968154
Adjusted R Square	0.969290872
Standard Error	0.019501009
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.096786945	0.048393472	127.2544294	1.22176E-05
Residual	6	0.002281735	0.000380289		
Total	8	0.099068679			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.411275705	0.036501084	93.45683358	1.01124E-10	3.321960771	3.500590639
Log Wgt	0.227564971	0.026096771	8.720043179	0.000125755	0.163708473	0.29142147
Attachments	0.059435938	0.009609062	6.18540461	0.000821697	0.03592341	0.082948467

The result is statistically significant on all 3 measures:
Adjusted R-Square, F and t Statistics

The Coefficient of Variation, CV = 0.5%

$$\text{Log(Cost)} = c + m_1 \times \text{Log(Weight)} + m_2 \times \text{Attachments}$$

$$\Rightarrow \quad \text{Cost} = 10^{(c + m_2 \times \text{Attachments})} \times \text{Weight}^{m_1}$$

Which is the Better Model?

Unit-Space Goodness of Fit Comparison:

Stat	Linear Model	Power – Exp Fit Space	Power – Exp Unit Space	How to Calculate
SSE	1125159.8	0.002	797338.5	$SUMSQ(e_i) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
R^2	0.965	0.977	0.975	$1 - \frac{SSE}{SST} = 1 - \frac{SUMSQ(e_i)}{DEVSQ(y_i)}$
Adj R^2	0.954	0.969	0.967	$1 - \left[(1 - R^2) \left(\frac{n-1}{(n-1)-k} \right) \right]$
SEE	433.0	0.020	364.5	$\sqrt{\frac{SSE}{(n-1)-k}}$
CV	5.2%	0.5%	4.4%	$\frac{SSE}{\bar{Y}} = \frac{s_Y}{\bar{Y}} \sqrt{1 - R_a^2}$

not df-adjusted

df-adjusted

The Non-linear Model is the better model in this case ...
Based on the Sum of Squares Error (SSE) in Unit Space
...but also all other measures in Unit Space

Adapted from:

Steps for Selecting the "Best Model"

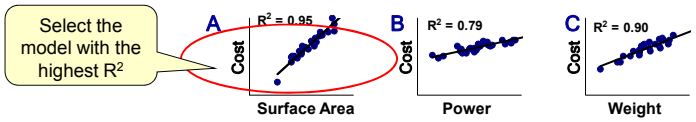
- Reject all non-significant models first
  - Where the F statistic is not significant
- Strip out all non-useful variables and made the model "minimal"
  - Variables that do not incrementally contribute to goodness of fit, overall model significance, (adjusted) variation explained, etc
  - Use t-Statistic and Adjusted R-Squre
- Select "within functional type" e.g. Linear or Power, based on:
  - Use R^2 for Simple Linear Regression (Ordinary Least Squares)
  - When comparing multivariate regression models, select based on Adjusted R^2, which compensates for the number of independent variables
- Select "across functional type", Linear v Power, based on:
  - Sum of Squares Error (SSE) across Single Variable Models
  - Standard Error Estimate (SEE) for Multivariate models

Adapted from

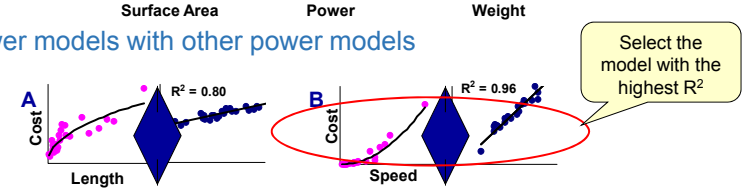
Selecting "Within Type"

- Start with only significant, "minimal" models
- In choosing among "models of a similar form", R^2 is the criterion

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

Unit III - Module 8

Selecting "Across Type"

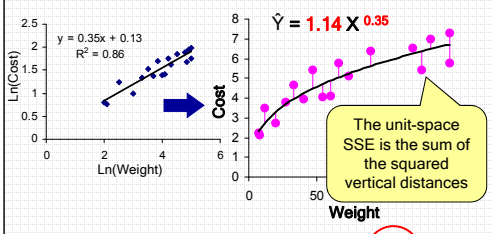
- Start with only significant, "minimal" models
- In choosing among "models of a different form":
  - the SSE in unit space is the criterion
  - SEE if degrees of freedom change;
  - CV if dependent variables changes)
- "Models of a different form" means that you will compare:
  - e.g., linear models 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

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

Selecting "Across Type" Example

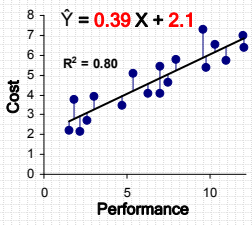
Suppose we want to choose between the following models for a method of estimating cost:

Option 1. Power Model



$$SSE_1 = \sum (Y - \hat{Y})^2 = 7.2$$

Option 2. Linear Model



$$SSE_2 = \sum (Y - \hat{Y})^2 = 8.3$$

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

What about Polynomials?

We can use the Polynomial Trendline in Microsoft Excel to create a Best Fit Curve through the data (up to a power of 6)

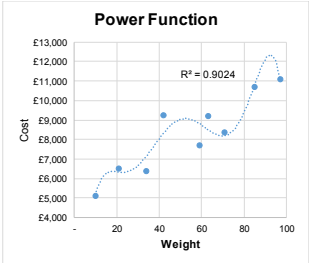
$$y = m_6x^6 + m_5x^5 + m_4x^4 + m_3x^3 + m_2x^2 + m_1x + c$$

Just because we can, it doesn't mean we should!

Polynomials should be used with great caution

Only use them where there is a sound rationale and reasonable justification for the model to be a Polynomial

For example, the Cumulative of a Straight Line is always a Quadratic through the origin



If you do get tempted to use them, do not extrapolate them outside the data range

They might turn on you unexpectedly!

Polynomial Regression … When it has been deemed valid

Each power term is used as a substitute Semi-independent variable:

e.g. x term	becomes	x1 in regression
x^2 term	becomes	x2
x^3 term	becomes	x3

The Intercept stays as the Intercept

- Then run the regression as a Multivariate Linear Model

$$y = m_6\,\mathrm{x6} + m_5\,\mathrm{x5} + m_4\,\mathrm{x4} + m_3\,\mathrm{x3} + m_2\,\mathrm{x2} + m_1\,\mathrm{x1} + c$$

- The Coefficient Parameters can then be extracted and tested for significance

$$y = m_6x^6 + m_5x^5 + m_4x^4 + m_3x^3 + m_2x^2 + m_1x + c$$

Multivariate Linear Regression and Non-Linear Regression

Any more questions?