

Estimating Relationships I: Factors and Simple Models

Philip M. Lurie

February 14, 2008

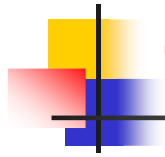


INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive, Alexandria, Virginia 22311-1882



Outline

- Simple parametric models
- Identifying and estimating models
- Assessing model fit
- Conditional and simultaneous tests
- Predicting cost of new system
- Confidence intervals
 - on model parameters
 - on cost of new system



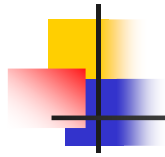
Statistics in Cost Analysis

- Why use statistics in cost analysis?
 - cost analysis frequently requires making inferences in the presence of uncertainty
- There is much uncertainty in estimating the cost of a new system
 - things rarely work out the way they are planned
 - requirements may change
 - subcontractors may be late
 - technology may not be available
 - funding instability



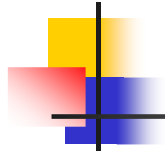
Estimating Total Cost

- Work Breakdown Structure
 - hierarchical system of subordinate level cost elements that are directly related to activities that define a project under development or production
 - cost is determined directly by summing all sub-components
- Cost Estimating Relationship (CER)
 - models cost in terms of observable explanatory factors (cost drivers) that serve as proxies for unobserved relationships
 - e.g., aircraft speed is a proxy for labor and materials required to develop or produce engine
 - relies on historical data to determine relationships



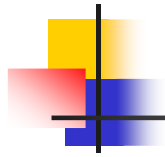
Work Breakdown Structure: Unmanned Space Vehicle

- **Spacecraft**
 - Structure, Interstage/Adapter
 - Thermal Control
 - Attitude Determination Control System (ADCS)
 - Attitude Determination (AD)
 - Reaction Control System (RCS)
 - Electrical Power Supply (EPS)
 - Power Generation
 - Power Storage
 - Power Conditioning and Distribution (PCD)
 - Telemetry, Tracking, and Command (TT&C)
 - Transmitter
 - Receiver/Exciter
 - Transponder
 - Digital Electronics (Signal/Data Processor)
 - Analog Electronics
 - Antennas
 - RF Distribution
 - Propulsion-Apogee Kick Motor (AKM)



Work Breakdown Structure: Unmanned Space Vehicle (cont.)

- **Communications Payload**
 - Transmitter
 - Receiver/Exciter
 - Transponder
 - Digital Electronics (Signal/Data Processor)
 - Analog Electronics
 - Antennas
 - RF Distribution
- **Integration, Assembly, and System Test (IA&T)**
- **Program Level**
 - Program Management
 - Systems Engineering
 - Data



Estimating Total Cost

- Work Breakdown Structure
 - hierarchical system of subordinate level cost elements that are directly related to activities that define a project under development or production
 - cost is determined directly by summing all sub-components
- Cost Estimating Relationship (CER)
 - models cost in terms of observable explanatory factors (cost drivers) that serve as proxies for unobserved relationships
 - e.g., aircraft speed is a proxy for labor and materials required to develop or produce engine
 - relies on historical data to determine relationships



Simple Factor Model

- The cost of a new system or product is estimated by multiplying the cost of a related system by a fixed (usually predetermined) factor
- Useful only when characteristics of the system remain unchanged
- Example 1:
 - Initial spares cost = $f \times \text{Flyaway Cost}$ ($f \approx .1$ to $.2$)
- Example 2:
 - DoD wants to implement certain process improvements adopted by industry to manage the depot maintenance system, which currently costs about \$12B
 - a review of case studies in the commercial sector finds that when industry implements these process improvements, total costs are reduced by 8%
 - cost of improved system is estimated to be $.92 \times 12 = \$11\text{B}$



Analogy Model

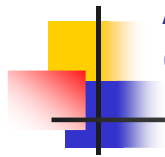
$$\frac{Cost_N}{Cost_O} = \left(\frac{X_N}{X_O}\right)^k \Rightarrow Cost_N = Cost_O \left(\frac{X_N}{X_O}\right)^k$$

- X_N and X_O are characteristics of the new and existing systems (e.g., weight) and k is a predetermined constant
- Used when a system similar to the one being costed exists and sufficient historical data are lacking
- Example:
 - It is known from previous experience that the cost of software written in FORTRAN increases exponentially with lines of code, where $k = 1.2$
 - Want to know the cost of reprogramming a 1000 KLOC software project in ADA, where code written in ADA requires about 10% more lines of code than FORTRAN. Previous code cost \$100M.
 $Cost = \$100M(1.1)^{1.2} = \$112M$



Parametric Model

- Uses historical data to estimate relationships between cost and characteristics of the system
- Factor and analogy models are special cases of a parametric model
- In general, form of the model is unknown and cost depends on several (or many) cost drivers
- Steps needed to identify model
 - graphical display to suggest functional form
 - estimate model parameters
 - diagnostics to assess model fit
 - test significance of model parameters

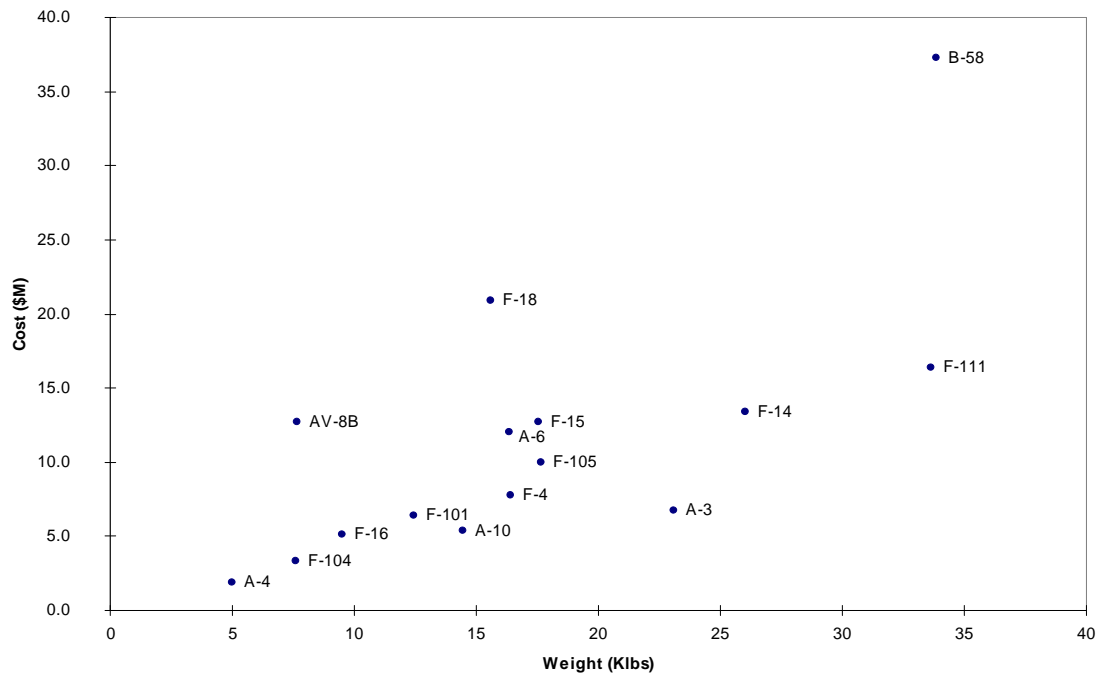


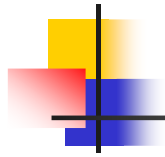
Airframe Cost and Characteristics Data

<u>Aircraft</u>	<u>Weight (Lbs)</u>	<u>Speed (Knots)</u>	<u>Advanced Materials (%)</u>	<u>Weight Complexity Factor</u>	<u>100th Unit Cost (\$M)</u>
A-10	14,439	389	1	0.40	5.4
A-3	23,104	545	1	0.41	6.7
A-4	4,987	577	1	0.32	1.9
A-6	16,359	561	1	0.56	12.0
AV-8B	7,662	533	34	0.70	12.7
B-58	33,833	1,147	30	0.56	37.3
F-101	12,436	875	1	0.57	6.4
F-104	7,635	1,150	1	0.70	3.3
F-105	17,668	1,195	1	0.33	10.0
F-111	33,625	1,262	35	0.45	16.4
F-14	26,038	1,170	24	0.44	13.4
F-15	17,574	1,434	28	0.54	12.7
F-16	9,503	1,184	1	0.61	5.1
F-18	15,573	1,029	23	0.58	20.9
F-4	16,440	1,150	1	0.49	7.8



Graphical Display of Data





Fitting a Model to the Data

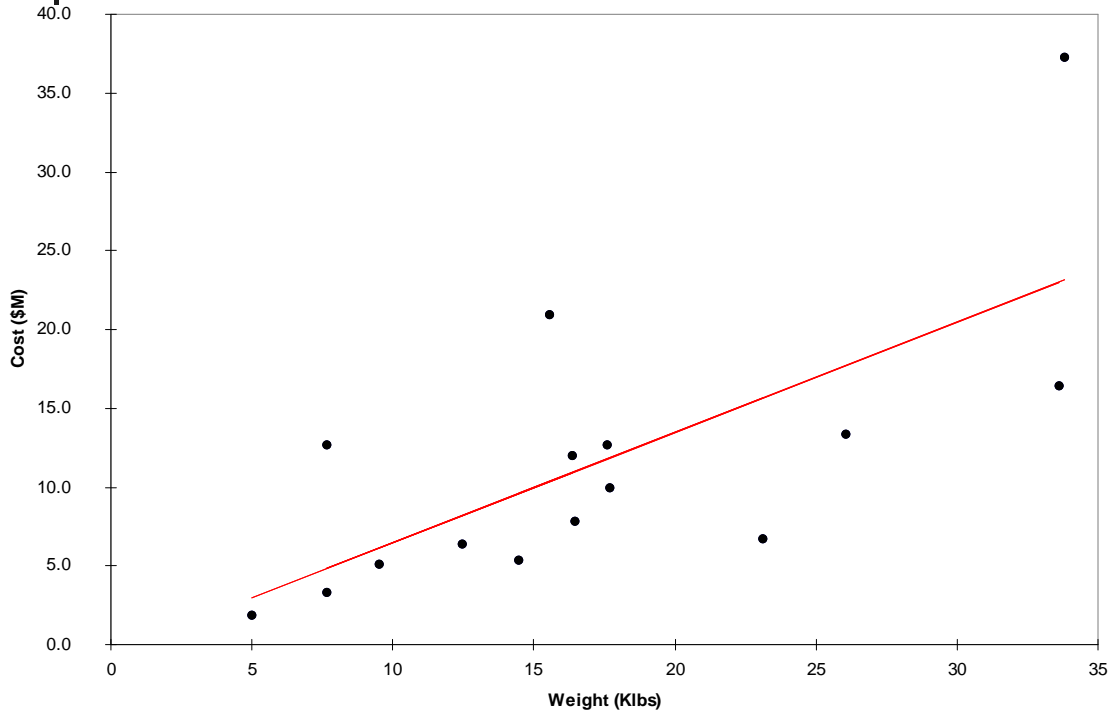
- Graphical display suggests cost may be a linear function of weight
- Try fitting $Cost_i = \beta_0 + \beta_1 Weight_i + \varepsilon_i$
- The error is assumed to be additive and the variance of ε_i is assumed to be constant
- How do we estimate β_0 and β_1 ?
- Traditional least squares approach:
Minimize $SSE = \sum_{i=1}^n (Cost_i - \beta_0 - \beta_1 Weight_i)^2$
- When the model is linear, least squares estimators
 - are unbiased
 - have smaller variance than any other linear estimator



Matrix Representation of Linear Regression

- $Y = X\beta + \varepsilon$
- In the preceding example, the first column of X is a vector of 1's and the second column is a vector of airframe weights
- The least squares estimator of β is
$$\hat{\beta} = (X'X)^{-1}X'Y$$
- The mean and variance of $\hat{\beta}$ are
$$E(\hat{\beta}) = \beta$$
$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$
where σ^2 is the variance of the error term (ε)

Regression Fit



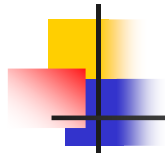
Examining Model Fit: The R^2 Statistic

- A useful summary measure of the regression fit is the R^2 statistic

- Define $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

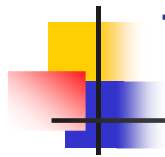
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

- When the model is linear with a constant term, $SST = SSR + SSE$
- R^2 is defined as SSR/SST and is often referred to as the "percentage of variance explained"
- $0 \leq R^2 \leq 1$
- The R^2 statistic for the regression of *Cost* against *Weight* is .49



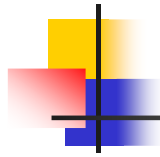
The R^2 Statistic

- When the model is linear with a constant term, there are several equivalent definitions of R^2 including
 - $1 - (SSE/SST)$
 - square of the correlation between y and \hat{y}
- Use caution when interpreting R^2 . In a non-linear model or in a linear model with no constant term
 - R^2 can no longer be interpreted as the percentage of variance explained
 - alternative definitions of R^2 are no longer equivalent
 - R^2 may turn negative
- The preferred definition of R^2 for use in all types of modelling situations is $1 - (SSE/SST)$



The Adjusted R^2 Statistic

- R^2 will never decrease when more variables are added to the regression equation
- As more variables are added to the equation, R^2 will continue to rise but the parameter estimates become progressively less precise
- The adjusted R^2 statistic was devised to allow comparisons between regressions with different numbers of independent variables
- The adjusted R^2 statistic is defined as
$$\bar{R}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2)$$
- The adjusted R^2 may decline when a variable is added to the model, or may even be negative



Examining Model Fit: Residual Plots

- The estimate of the error variance is

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}$$

and the estimate of the standard error is s

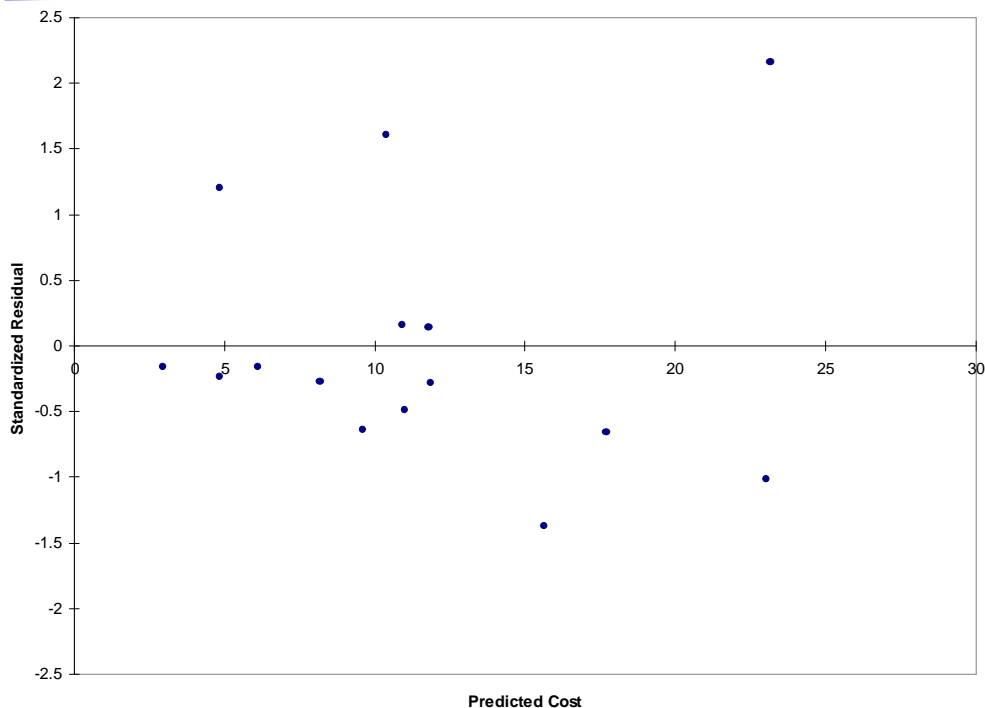
- Define standardized residuals as

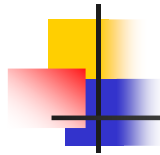
$$e_i = \frac{y_i - \hat{y}_i}{s}$$

- Expect most residuals to fall within ± 2
- Plot standardized residuals against predicted value and against independent variables

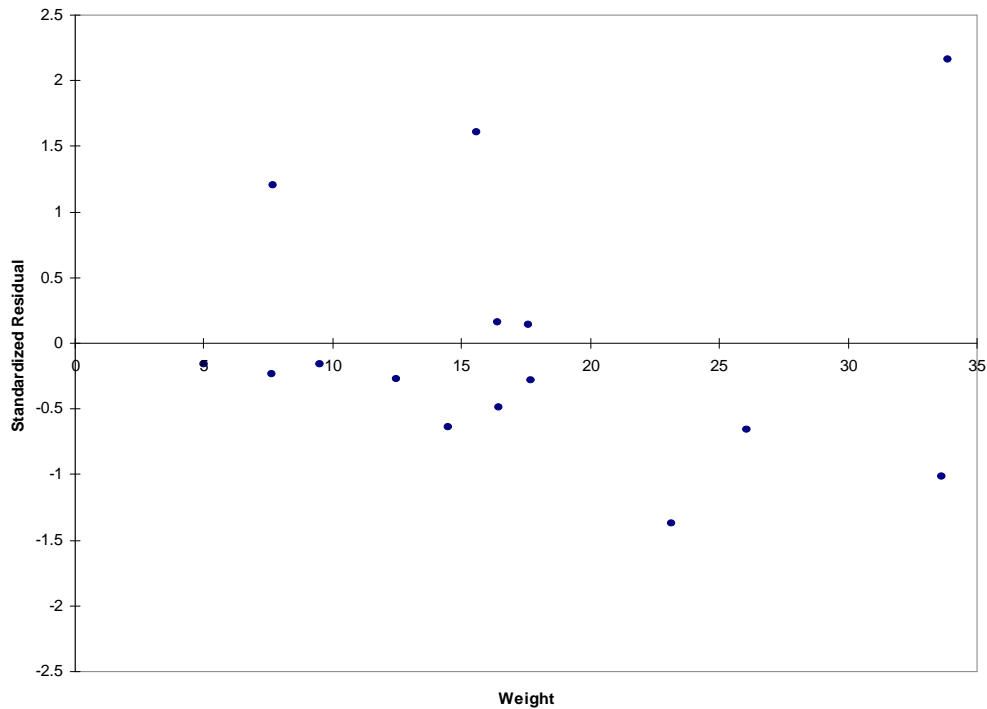


Residuals vs. Predicted Values



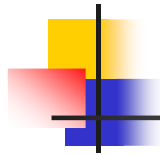


Residuals vs. Weight

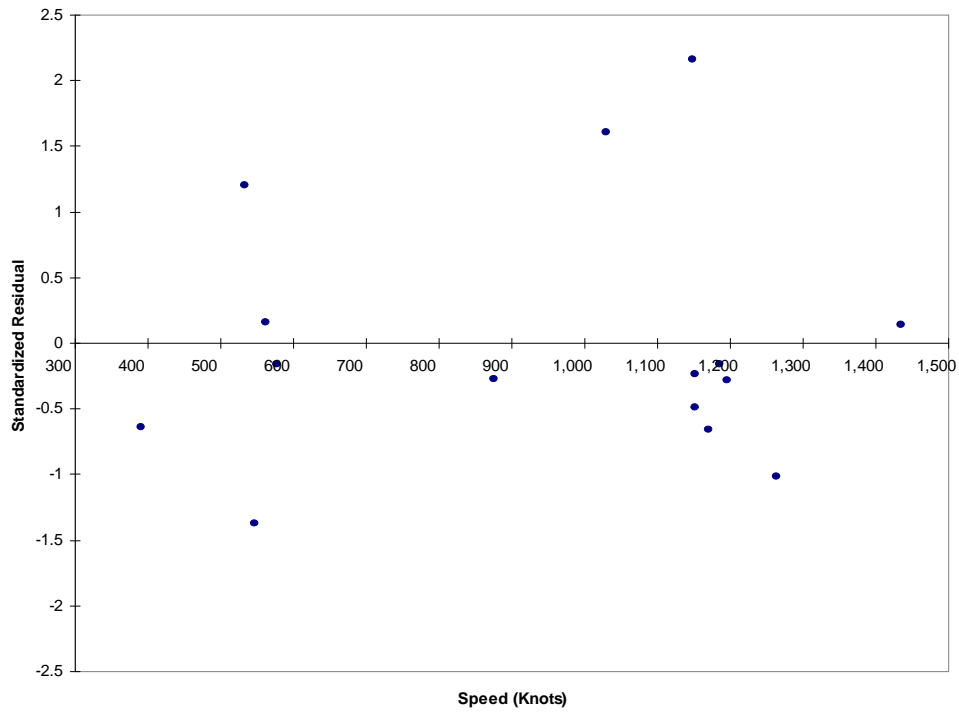


Results of Residual Plots

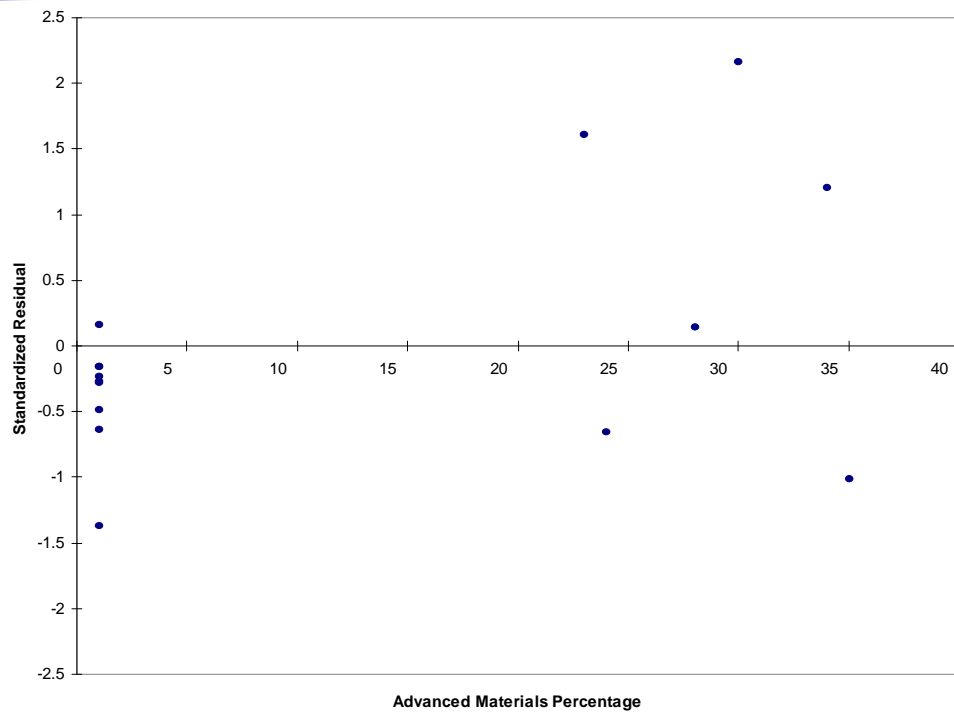
- Both residual plots appear to exhibit random scatter but not a very tight fit
- May be another variable or variables not accounted for in model
- Consider using another aircraft characteristic as an explanatory variable
- Plot residuals against aircraft characteristics



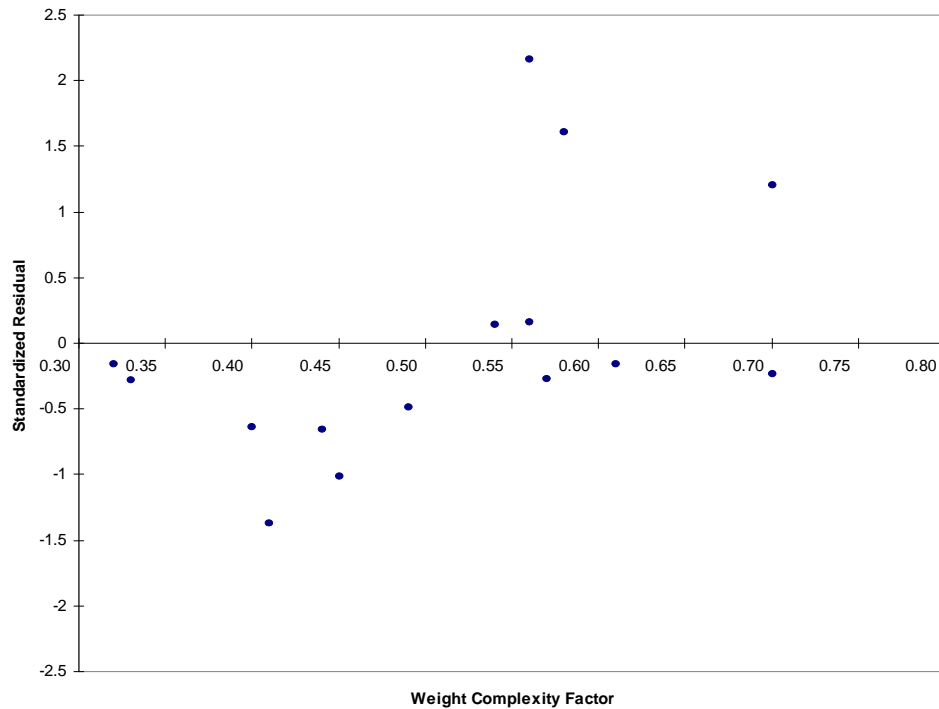
Residuals vs. Speed



Residuals vs. Advanced Materials Percentage



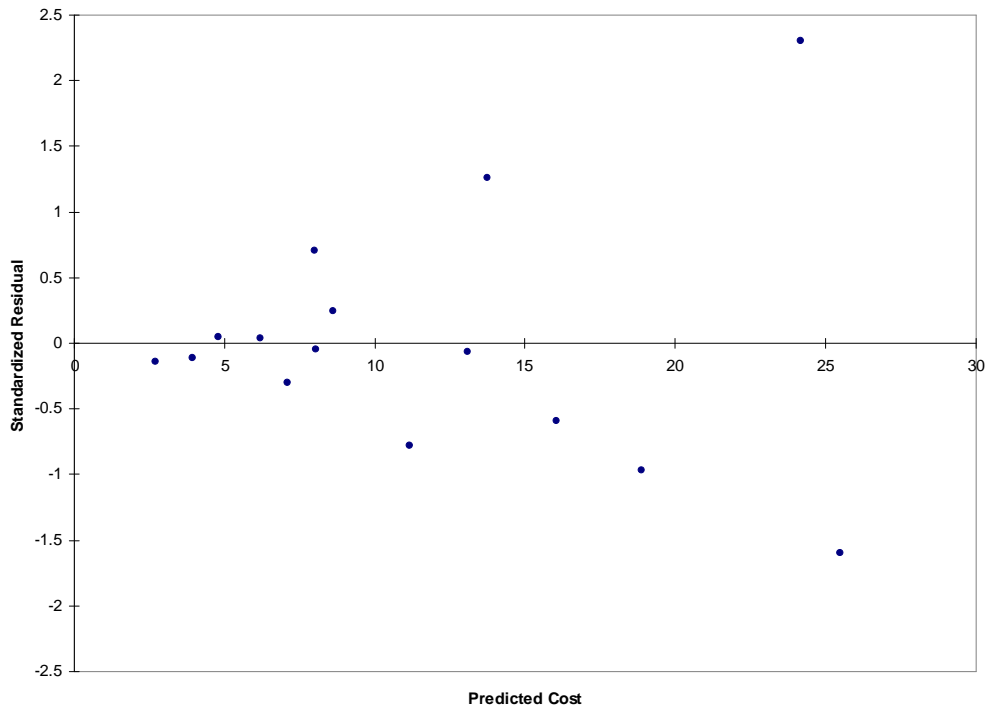
Residuals vs. Weight Complexity Factor



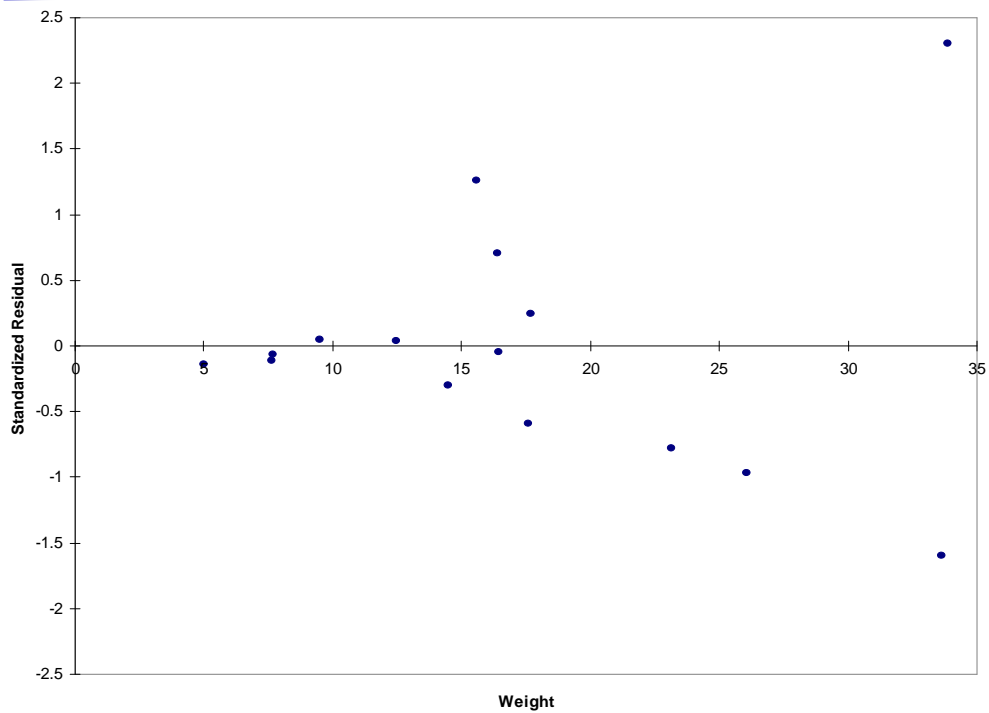
New Regression Results

- Plots indicate a positive relationship between *Cost* and both *Advanced Materials Percentage* and *Weight Complexity Factor*
- The regression of *Cost* against *Weight* and *Advanced Materials Percentage* produced the best results
- $R^2 = .64$
- Regression fit appears to have improved substantially (R^2 increased from .49 to .64)

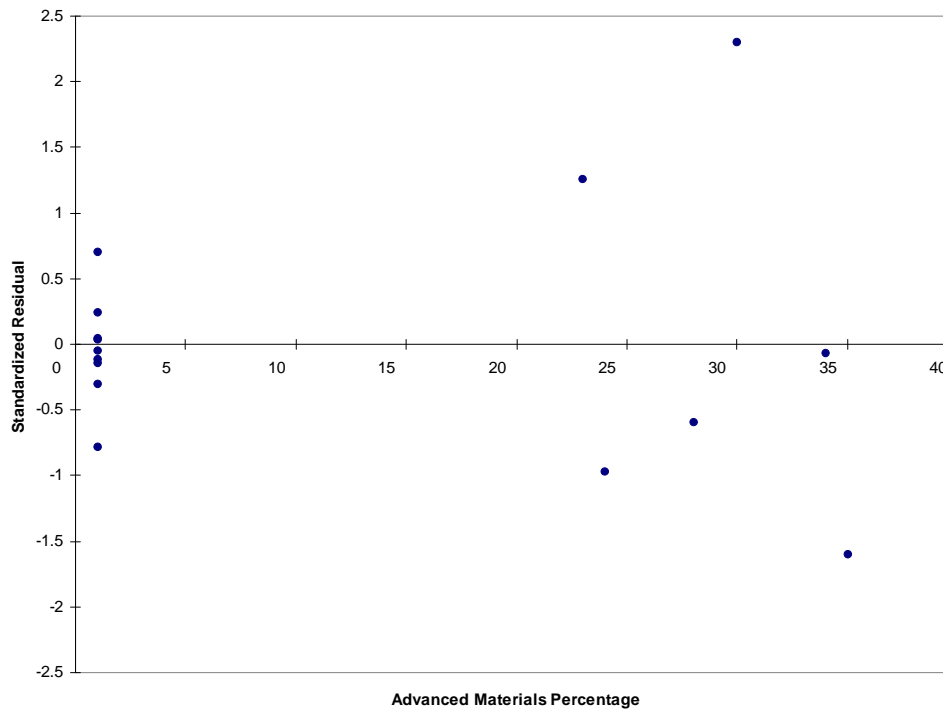
New Residuals vs. New Predicted Values



New Residuals vs. Weight



New Residuals vs. Advanced Materials Percentage



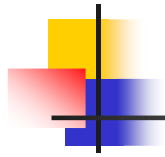
Checking Normality of Residuals

- Compute empirical percentiles
 - order residuals from smallest to largest
 - for the i th ordered residual, the estimated percentage less than or equal to that value is i/n
 - usually use $(i-1/2)/n$ to avoid percentile equal to 1 at largest value

- Compute corresponding percentiles of standard normal distribution, i.e.,

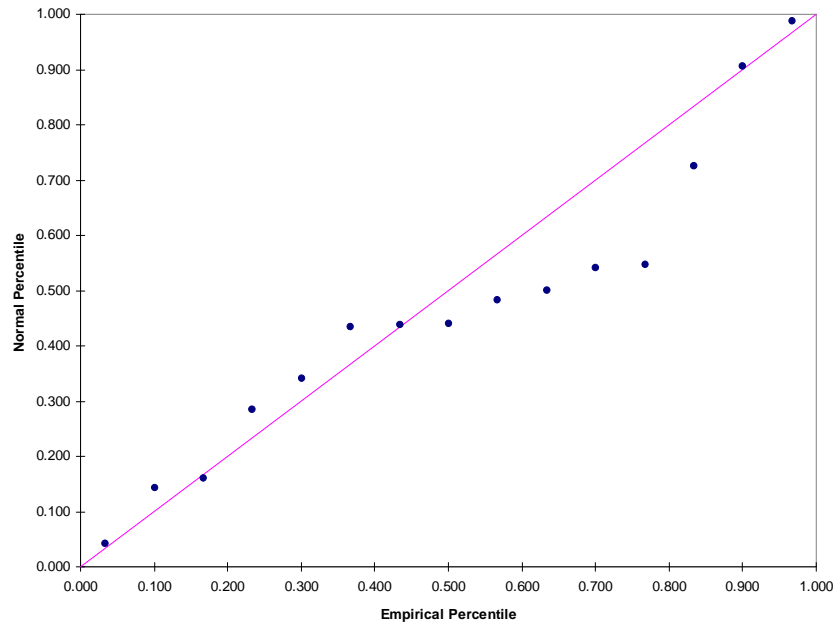
$$p_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r_i} e^{-1/2 t^2} dt$$

- Plot empirical percentiles against normal percentiles — normal probability plot
- Adherence to straight line indicates residuals are normally distributed



Normal Probability Plot

Linear Regression			
i	Residual	ECDF	NCDF
1	-1.718	0.033	0.043
2	-1.064	0.100	0.144
3	-0.992	0.167	0.161
4	-0.567	0.233	0.285
5	-0.410	0.300	0.341
6	-0.164	0.367	0.435
7	-0.151	0.433	0.440
8	-0.148	0.500	0.441
9	-0.040	0.567	0.484
10	0.002	0.633	0.501
11	0.105	0.700	0.542
12	0.118	0.767	0.547
13	0.603	0.833	0.727
14	1.324	0.900	0.907
15	2.262	0.967	0.988

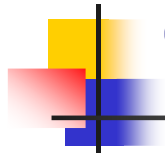


Testing Model Parameters

- Want to test the hypotheses:
$$H_0: \beta_i = 0$$

vs.
$$H_A: \beta_i \neq 0$$

for each i
- Need to distinguish two types of hypotheses
 - conditional tests of one coefficient at a time
 - all other parameters held constant
 - simultaneous tests of some or all model parameters
- Standard tests assume error term is normally distributed with constant variance



Conditional Tests

- To test $\beta_i = 0$, compute the test statistic

$$t_i = \frac{\hat{\beta}_i}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}}$$

- t_i has a t-distribution with $n-p$ degrees of freedom, where n is the number of observations and p is the number of estimated parameters
- Compare t_i with $(1-\alpha/2)$ percentile of t-distribution, where α is the significance level desired, to determine whether to accept or reject hypothesis



Example of Conditional Tests

- The regression of *Cost* against *Weight* and *Advanced Materials Percentage* produced the following results:

<u>Variable</u>	<u>Coefficient</u>	<u>Std. Error</u>	<u>t-Value</u>	<u>Sig. Level</u>
Intercept	0.10346	3.31701	0.03119	0.97563
Weight	0.46633	0.20132	2.31631	0.03903
AdvMat	0.27683	0.12216	2.26611	0.04274

- The effects of *Weight* and *Advanced Materials Percentage* are significant but the constant term (intercept) is not significant



Simultaneous Test

- To test $\beta_i = 0$ for a subset of k coefficients simultaneously, first compute the quantities

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \quad \text{and} \quad \mathbf{e}_* = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_*$$

where $\hat{\boldsymbol{\beta}}_*$ is the estimate of $\boldsymbol{\beta}$ when the k coefficients are set to zero

- Next compute the statistic

$$F = \frac{(\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e}) / k}{\mathbf{e}'\mathbf{e} / (n - p)}$$

- Compare F with $(1-\alpha)$ percentile of F-distribution with k and $n-p$ degrees of freedom, where α is the significance level desired



Example of Simultaneous Test

- In regression of *Cost* against *Weight* and *Advanced Materials Percentage*, test that both β_1 and β_2 are equal to zero
- With both an intercept and airframe characteristics in the model, the error sum of squares is 389.05 with 12 degrees of freedom (15 observations - 3 parameters)
- With only a constant term in the model, the error sum of squares is 1087.29 with 14 degrees of freedom
- The F statistic is therefore

$$F = \frac{(1087.29 - 389.05) / 2}{389.05 / 12} = 10.77$$

- The 95th percentile of the F distribution with 2 and 12 degrees of freedom is 3.89, i.e., reject hypothesis



Additional Notes on the *F* Statistic

- The *F* statistic displayed in most regression packages tests the regression against a model with only a constant term
- The *F* statistic is usually displayed as the outcome of an analysis of variance:

	DF	Sum of Squares	Mean Square
Regression	2	698.24692	349.12346
Residual	12	389.04641	32.42053

F = 10.76859 Signif *F* = .0021

- In the case where all the regression parameters (except the constant term) are tested simultaneously, the *F* statistic reduces to:

$$F = \frac{SSR / (p - 1)}{SSE / (n - p)} = \frac{MSR}{MSE}$$



Predicting the Cost of a New System

- Use regression results to predict costs of new systems
- The regression coefficients and characteristics of the new systems are:

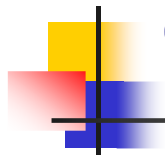
<u>Variable</u>	<u>Coefficient</u>	<u>F-X</u>	<u>F-Y</u>
Constant	0.10346	1	1
Weight	0.46633	19	12
AdvMat	0.27683	65	20

- Predicted costs
 - F-X: \$27.0M
 - F-Y: \$11.2M



Confidence Intervals

- On parameters
 - conditional intervals on one coefficient at a time
 - all other parameters held constant
 - simultaneous intervals on some or all model parameters
 - confidence ellipse
- On fitted regression line
- On cost of new system



Confidence Intervals on Parameters

- Compute variance of parameters:
 - $$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$
 - $$V(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$
 - $$V(\hat{\beta}_i) = \sigma^2 (\mathbf{X}'\mathbf{X})_{ii}^{-1}$$
- Estimate σ^2 with $s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)$
- Confidence interval on parameter is then

$$\hat{\beta}_i \pm t_{n-p; 1-\alpha/2} \sqrt{V(\hat{\beta}_i)}$$



Example of Confidence Intervals on Parameters

- The matrix \mathbf{X} has 3 columns
 - a column of 1's (for the constant term)
 - a column of *Weights*
 - a column of *Advanced Materials Percentages*

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.33937 & -0.01667 & 0.00104 \\ -0.01667 & 0.00125 & -0.00039 \\ 0.00104 & -0.00039 & 0.00046 \end{bmatrix}$$

- $s^2 = 32.42$ and $t_{12;.975} = 2.179$
- The 95% confidence intervals are therefore:

$$\hat{\beta}_0: .103 \pm 2.179\sqrt{(32.42)(0.33937)} = (-7.125, 7.331)$$

$$\hat{\beta}_1: .466 \pm 2.179\sqrt{(32.42)(0.00125)} = (.027, .905)$$

$$\hat{\beta}_2: .277 \pm 2.179\sqrt{(32.42)(0.00046)} = (.011, .543)$$



Confidence Interval on Fitted Regression Line

- To compute variance of fitted value, recall that

$$V(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

where σ^2 is estimated by $s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)$

- Therefore

$$V(\hat{Y}_0) = V(X_0\hat{\beta}) = \sigma^2 X_0 (\mathbf{X}'\mathbf{X})^{-1} X_0'$$

- Confidence interval on prediction is then

$$\hat{Y}_0 \pm t_{n-p;1-\alpha/2} \sqrt{V(\hat{Y}_0)}$$



Example: Confidence Interval on Fitted Regression

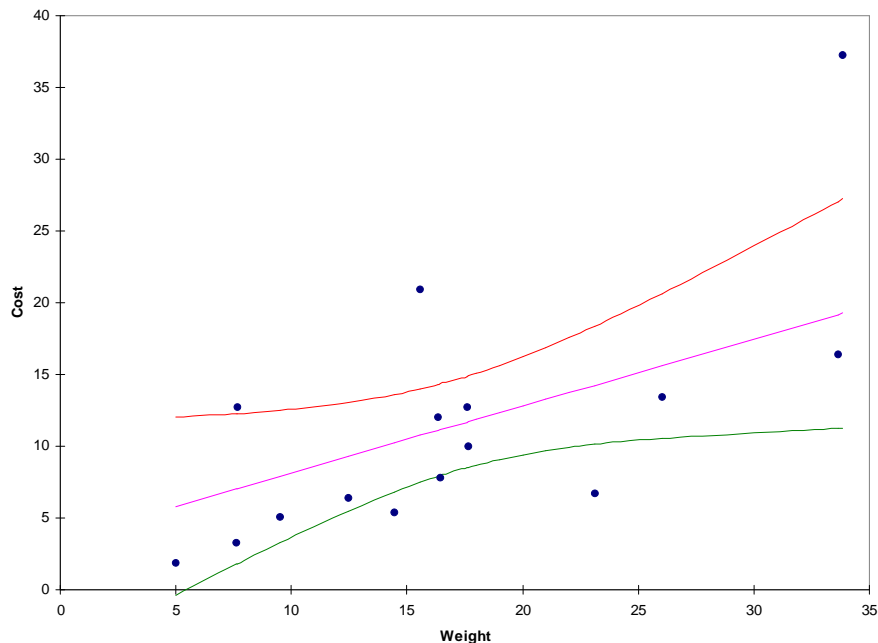
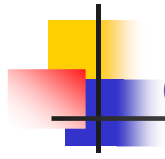
- Recall that the vector of predictor variables for the new aircraft systems are:

	Constant	Weight	AdvMat
F-X:	1	19	65
F-Y:	1	12	20

- Recall also that $s^2 = 32.42$, $t_{12;.975} = 2.179$, and the predicted costs of the new systems are \$27M for the F-X and \$11.2M for the F-Y
- Substituting the above values into the formulas on the previous slide, we obtain the following confidence intervals for the expected cost:
 - F-X: (13.0, 41.0)
 - F-Y: (6.3, 16.1)

95% Confidence Interval on Cost Regression

(Holding Advanced Materials Percentage Fixed at its Mean)





Confidence Interval on Cost of New System

- Recall from earlier slide:

$$V(\hat{Y}_0) = V(X_0\hat{\beta}) = \sigma^2 X_0(\mathbf{X}'\mathbf{X})^{-1} X_0'$$

- The cost of a new system (a future value of Y) can be represented as: $C = \hat{Y}_0 + \varepsilon$

- The estimated cost of the new system is $\hat{C} = \hat{Y}$

- The variance of this cost is therefore

$$\begin{aligned} V(C) &= V(\hat{Y}_0 + \varepsilon) = \sigma^2 X_0(\mathbf{X}'\mathbf{X})^{-1} X_0' + \sigma^2 \\ &= \sigma^2 (X_0(\mathbf{X}'\mathbf{X})^{-1} X_0' + 1) \end{aligned}$$

and the confidence interval is $\hat{C} \pm t_{n-p, 1-\alpha/2} \sqrt{V(C)}$



Example: Confidence Interval on Cost of New System

- Recall that the vector of predictor variables for the new aircraft systems are:

	Constant	Weight	AdvMat
F-X:	1	19	65
F-Y:	1	12	20

- Recall also that $s^2 = 32.42$, $t_{12, .975} = 2.179$, and the predicted costs of the new systems are \$27M for the F-X and \$11.2M for the F-Y
- Substituting the above values into the formulas on the previous slide, we obtain the following confidence intervals for the costs of the new systems:
 - F-X: (8.3, 45.7)
 - F-Y: (0.0, 24.6)