CHAPTER

10

Duxbury Thomson Learning

Making Hard Decision

Third Edition

Using Data

A. J. Clark School of Engineering •Department of Civil and Environmental Engineering

FALL 2003

By
Dr . Ibrahim. Assakkaf

ENCE 627 – Decision Analysis for Engineering

Department of Civil and Environmental Engineering

University of Maryland, College Park

Reliability of the Regression Equation

- Criteria to be assessed:
  - Correlation coefficient.
  - The standard error of estimate.
  - The *F* statistics for the analysis of Variance
  - The rationality of the coefficients and the relative importance of the predictor variable.
  - The degree to which the underlying assumptions of the regression model are met.

Reliability of the Regression Equation

- Correlation Coefficient
  - The correlation coefficient *R* is an index of the degree of linear association between two random variables.
  - The magnitude of *R* indicates whether the regression provides accurate predictions of the criterion variable.
  - The square of the correlation coefficient R^2 equals the percentage of the variance in the criterion variable that is explained by the predictor variable.

Reliability of the Regression Equation

- Correlation Coefficient

$$\mathrm{TV} = \mathrm{EV} + \mathrm{UV}$$

$$1 = \frac{\mathrm{EV}}{\mathrm{TV}} + \frac{\mathrm{UV}}{\mathrm{TV}}$$

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$R = \frac{\mathrm{EV}}{\mathrm{TV}}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}}$$

Reliability of the Regression Equation

- Standard Error of Estimate
  - In absence of additional information, the mean is the best estimate of the criterion variable.
  - The standard deviation S_Y of Y is an indication of the accuracy of the prediction.
  - If Y is related to one or more predictor variables, the error of prediction is reduced from S_Y to the standard error of estimate S_e.

Reliability of the Regression Equation

- Standard Error of Estimate
  - The standard error of estimate equals the standard deviation of the errors

$$S_e = \sqrt{\frac{1}{\nu}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where

$$\nu = n - p - 1$$

$$n = \text{number of sample points}$$

$$p = \text{number of predictor variables}$$

Reliability of the Regression Equation

- Standard Error of Estimate
  - In terms of separation of variation, the standard error of estimate equals the square root of the ratio of the unexplained variation (UV).
  - It is important to know that S_e is based on (n – p –1) degrees of freedom, while S_Y is based on (n – 1) degrees of freedom.
  - If $S_e \approx S_Y$, the regression is not good
  - If $S_e << S_Y$, the regression has improved the prediction

Reliability of the Regression Equation

- Standard Error of Estimate

$$\mathrm{TV} = \mathrm{EV} + \mathrm{UV}$$

$$S_Y^2 = \frac{\mathrm{TV}}{n\text{-}1}$$

$$R^2 = \frac{\mathrm{EV}}{\mathrm{TV}}$$

$$S_e^2 \frac{\mathrm{UV}}{n-p-1}$$

$$\longrightarrow \quad \boxed{S_e^2 = S_Y^2\left(\frac{n-1}{n-p-1}\right)\left(1-R^2\right)}$$

Reliability of the Regression Equation

- Standard Error of Estimate
  - Relationships between S_e and S_Y

$$S_e = S_Y\sqrt{1-R^2} \qquad \text{(approximate formula)}$$

$$S_e = \sqrt{\left(\frac{n-1}{n-p-1}\right)S_Y^2\left(1-R^2\right)} \qquad \text{(exact formula)}$$

Multiple Regression Analysis

- Multiple regression are used to improve the accuracy of predictions if the accuracy from a bivariate regression is still not sufficient for the design problem.
- Several predictor variables may provide sufficient prediction accuracy.
- One reason to use multivariate models rather than a bivariate model is to reduce the standard error of estimate.

Multiple Regression Analysis

- The same least-squares objective function is used to calibrate the regression coefficient.
- Bivariate correlation are still computed.
- The major difference between multivariate and bivariate analyses is the necessity to account for interdependence (correlation) of the predictor variables.

Multiple Regression Analysis

- Correlation Matrix
  - After graphical analysis, the bivaraite correlation coefficients should be computed for each pair of variables; this includes
    1. The correlation between the criterion variable and each predictor variable.
    2. The Correlation between each pair of predictor variables.

Multiple Regression Analysis

- Correlation Matrix

	X_1	X_2	X_3	...	X_p	Y
X_1	1	r_{12}	r_{13}	...	r_{1p}	r_{1Y}
X_2		1	r_{23}	...	r_{2p}	r_{2Y}
X_3			1	...	r_{3p}	r_{3Y}
.					.	.
.				1	.	.
.					.	.
X_p					1	r_{pY}
Y						1

Multiple Regression Analysis

- Properties of Correlation Matrix
  - The matrix includes p predictor variables (X_i, I = 1, 2, …,p) and the criterion variable Y.
  - The matrix is symmetric, that is, $r_{ij} = r_{ji}$.
  - The elements in the principal diagonal equal 1.0.
  - The matrix is $(p + 1) \times (p + 1)$.

Multiple Regression Analysis

- Example: Evaporation Data
  - Correlation Matrix for Evaporation Data

	X_1	X_2	X_3	X_4	Y
X_1 : temperature (^0F)	1.000	-0.219	0.578	0.821	0.581
X_2 : wind speed (mi/day)		1.000	-0.261	-0.304	-0.140
X_3 : radiation			1.000	0.754	0.578
X_4 : vapor pressure deficit				1.000	0.635
Y : pan evaporation (inches)					1.000

Multiple Regression Analysis

- Calibration of the Multiple Linear Model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

– The Objective Function

$$F = \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} \left(b_0 + \sum_{j=1}^{p} b_j x_{ij} - y_i \right)^2$$

Multiple Regression Analysis

- Example: Multiple Regression
  - Consider the case where p = 2, thus

$$\overline{Y} = b_0 + b_1 X_1 + b_2 X_2$$

$$F = \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} \left(b_0 + b_1 x_{i1} + b_2 x_{i2} - y_i\right)^2$$

  - The resulting derivatives are

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_o} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_{i1} + b_2 x_{i2} - y_i\right)(1) = 0$$

Multiple Regression Analysis

- Example (cont'd): Multiple Regression

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_{i1} + b_2 x_{i2} - y_i\right)(x_{i1}) = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_{i1} + b_2 x_{i2} - y_i\right)(x_{i2}) = 0$$

The following set of normal equations can be obtained:

Multiple Regression Analysis

- Example (cont'd): Multiple Regression

$$nb_0 + b_1\sum_{i=1}^{n} x_{i1} + b_2\sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$

$$b_0\sum_{i=1}^{n} x_{i1} + b_1\sum_{i=1}^{n} x_{i1}^2 + b_2\sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i$$

$$b_0\sum_{i=1}^{n} x_{i2} + b_1\sum_{i=1}^{n} x_{i1}x_{i2} + b_2\sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i$$

The solution of the three simultaneous equations yields values for b_0, b_1, and b_2

Multiple Regression Analysis

- Example:

The following table provides values for the criterion variable Y and two predictor variables X_1 and X_2. The sample consists of 6 observations. Find the partial regression coefficients b_0, b_1, and b_2.

Multiple Regression Analysis

- Example (cont'd):

	Y	X_1	X_2	X_1^2	X_1X_2	X_1Y	X_2^2	X_2Y
	2	1	2	1	2	2	4	4
	2	2	3	4	6	4	9	6
	3	2	1	4	2	6	1	3
	3	5	5	25	25	15	25	15
	5	4	6	16	24	20	36	30
	6	5	4	25	20	30	16	24
Σ	21	19	21	75	79	77	91	82

Multiple Regression Analysis

- Example (cont'd):
  - Using Excel to estimate regression coefficients

Regression Statistics	
Multiple R	0.742867
R Square	0.551852
Adjusted R Squa	0.253086
Standard Error	1.420094
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	7.45	3.725	1.847107	0.300008
Residual	3	6.05	2.016667		
Total	5	13.5			

	Coefficients	*standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	1.3	1.378249	0.943226	0.415151	-3.086207	5.686207	-3.086207	5.686207
X Variable 1	0.75	0.584408	1.283351	0.289522	-1.109848	2.609848	-1.109848	2.609848
X Variable 2	-0.05	0.538042	-0.09293	0.931818	-1.762292	1.662292	-1.762292	1.662292

Multiple Regression Analysis

- Example (cont'd):

$$nb_0 + b_1\sum_{i=1}^{n} x_{i1} + b_2\sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$

$$b_0\sum_{i=1}^{n} x_{i1} + b_1\sum_{i=1}^{n} x_{i1}^2 + b_2\sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i$$

$$b_0\sum_{i=1}^{n} x_{i2} + b_1\sum_{i=1}^{n} x_{i1}x_{i2} + b_2\sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i$$

$$\rightarrow \quad \begin{aligned} 19b_0 + 75b_1 + 79b_2 &= 77 \\ 21b_0 + 79b_1 + 91b_2 &= 82 \\ 6b_0 + 19b_1 + 21b_2 &= 21 \end{aligned}$$

$$\boxed{\hat{Y} = 1.30 + 0.75X_1 - 0.05X_2}$$

Multiple Regression Analysis

- Example (cont'd):
  - Using Excel to estimate regression coefficients

Regression Statistics	
Multiple R	0.742867318
R Square	0.551851852
Adjusted R Squa	0.25308642
Standard Error	1.420093894
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	7.45	3.725	1.847107	0.300007705
Residual	3	6.05	2.016667		
Total	5	13.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*
Intercept	1.3	1.37824885	0.943226	0.415151	-3.086207075	5.686207	-3.086207
X Variable 1	0.75	0.584407613	1.283351	0.289522	-1.109847593	2.609848	-1.109848
X Variable 2	-0.05	0.53804205	-0.09293	0.931818	-1.76229154	1.662292	-1.762292

Regression Analysis of Nonlinear Models

- Advantages of Linear Models
  - Simple
  - Easily applied
  - Statistical reliability is easily assessed
- Disadvantages of Linear Models
  - May be rejected because of theoretical considerations or empirical evidence.

Regression Analysis of Nonlinear Models

- Common Nonlinear Alternatives
  - Bivariate
    - Polynomial

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + \ldots + b_p X^p$$

    - Power

$$\hat{Y} = b_0 X^{b_1}$$

Regression Analysis of Nonlinear Models

- Common Nonlinear Alternatives
  - Multivariate
    - Polynomial

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1^2 + b_4 X_2^2 + b_5 X_1 X_2$$

    - Power

$$\hat{Y} = b_0 X^{b_1} X_2^{b_2} \ldots . X_3^{b_p}$$

Regression Analysis of Nonlinear Models

- Calibration of Polynomial Models
  - The power and polynomial models are nonlinear forms that can transformed in a way that it is possible to use the principle of least squares.
  - Although the transformation to a linear structure is desirable, it has important consequences in terms of assessing the goodness of fit.

Regression Analysis of Nonlinear Models

- Calibration of Polynomial Models

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + \ldots + b_p X^p$$

– Let $W_i = X^i \quad \text{for } i = 1,2,\ldots,p$

$$\boxed{\hat{Y} = b_0 + b_1 W_1 + b_2 W_2 + \ldots + b_p W_p}$$

Regression Analysis of Nonlinear Models

- Fitting a Power Model

Let $\hat{Y} = b_0 X^{b_1}$

$\hat{Z} = \ln \hat{Y}, \quad c = \ln b_0, \quad \text{and } W = \ln X$

Then, the following bivariate model can be obtained :

$$\hat{Z} = c + b_1 W$$

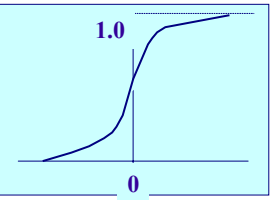
The Regression Approach

Example: *E(Sales\ Ad., Price, Compt. Price)*

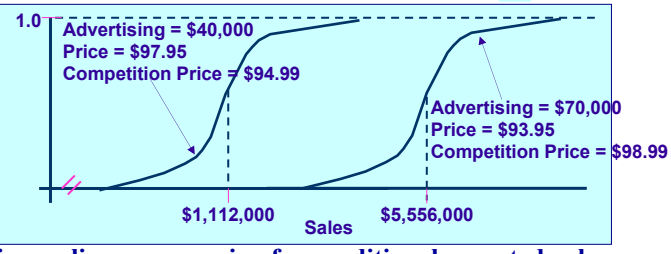
= 2000 + 14.8(Ads.) –500(Price) + 500(Compt. Price)

= 2000 + 14.8(40) –500(97.95) + 500(94.99)
= 1112($1000s).

A CDF for error in the sales example.



Conditional CDFs for sales in two scenarios.



Our two assumptions-a linear expression for conditional expected value and the constant shape for the conditional distribution-take us quite a long way toward being able to use data to study relationships among variables.

Estimation: The Basics

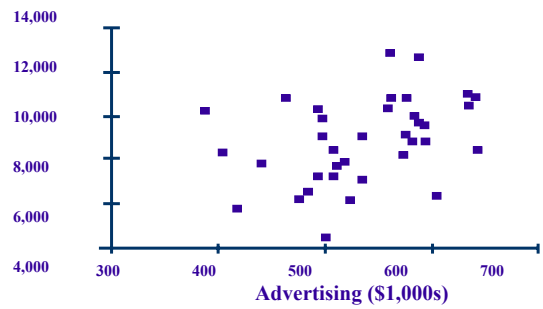
- We assume that we know the β coefficients and the distribution of the errors.
- We may be able to make subjective judgments of these quantities, but if data are available, we may want to use that information in a systematic way to estimate the β's and to construct a data-based distribution of errors.

Estimation: The Basics

- The linear-regression assumptions.

$$E(Y \mid X_1) = \beta_0 + \beta_1 X_1$$

A scatter plot of advertising versus sales.

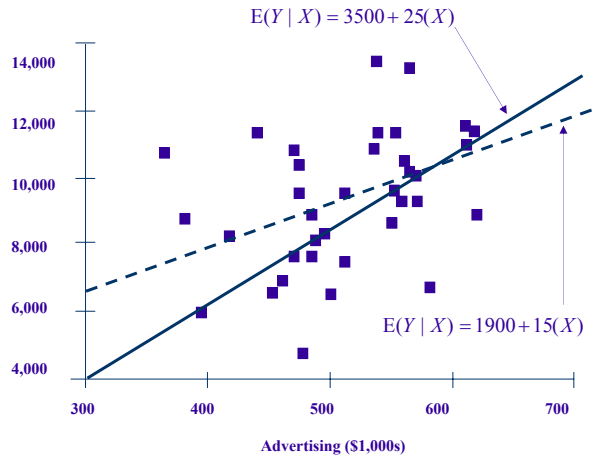


Estimation: The Basics

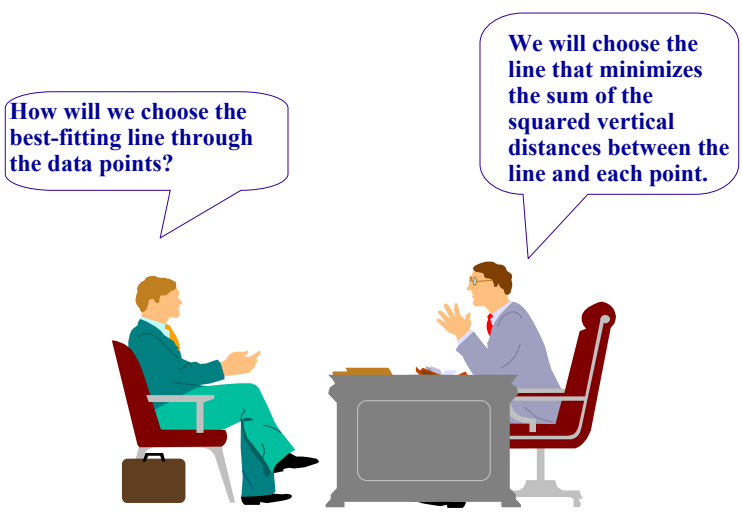
- We use our data to estimate β_0 and β_1.
- Estimating β_0 and β_1 amounts to finding a line that passes through the cloud of points in the scatter plot.
- No single line can pass precisely through all of the points at once, but we would like to find one that in some sense is the “best fitting” line.
- There are many reasonable estimates for β_0 and β_1.

Estimation: The Basics

Two possible lines relating expected sales and advertising.

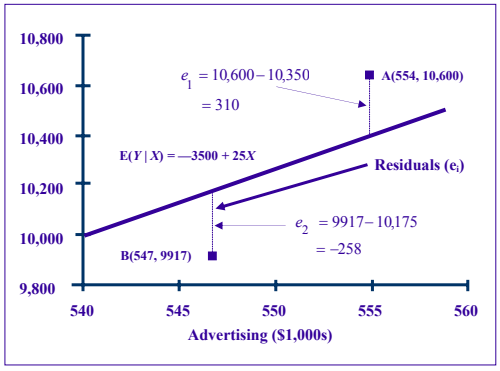


Calculating Residuals



Calculating Residuals

Calculating residuals



$$e_i = y_i - (b_0 + b_1 x_i)$$

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

Calculating Residuals

- How can we use the data to construct a model of the distribution of errors?
  - For every data point we have a residual, which can be thought of as an estimate of the error associated with that particular point.
  - The distribution of the residuals should be a good approximation to the distribution of the errors.

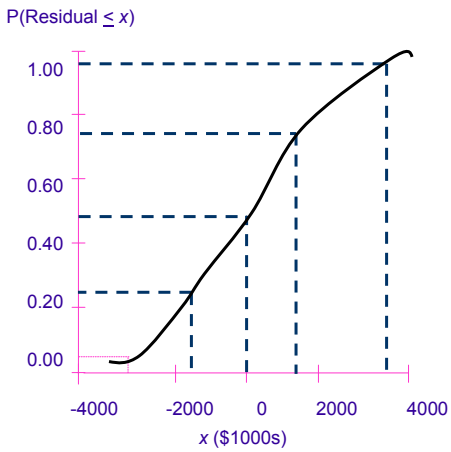
Calculating Residuals

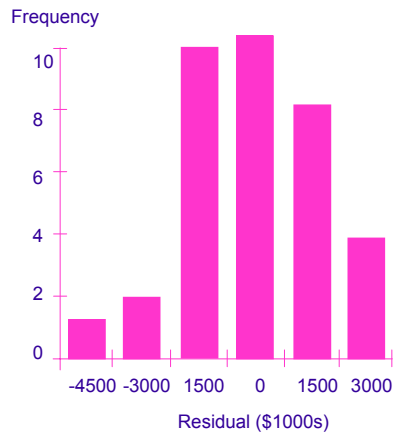
- The CDF can be used to make probability statements about the errors. We can also use the CDF to construct a discrete approximation.
- The CDF can also be used as a basis for representing the uncertainty.
- Typically in regression analysis, a normal distribution is fit to the residuals, but virtually any continuous distribution that makes sense for the situation could be used.

Calculating Residuals

- The residuals could be used as data in Best Fit to find the best-fitting distribution.
- With the coefficient estimates and the distribution of errors, we have a complete, if simplified, model of the uncertainty.

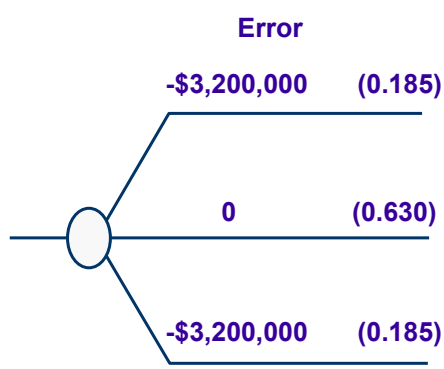
CDF and Histogram of Residuals





Approximation for Error Distribution

Extended Pearson-Tukey approximation for error distribution.



Regression Analysis and Modeling: Some Do's and Don'ts

- We have seen a number of approaches to creating uncertainty models:
  1. Subjective assessment. The use of theoretical distributions.
  2. Statistical model and analysis, when based on an appropriate data set of adequate size, is very persuasive. The drawback: Data collection can take time and resources and may not always be feasible. In such cases, an analyst might be better off relying on theoretical distributions and expert subjective judgments.

Regression Analysis and Modeling: Some Do's and Don'ts

3. Data. The use of data can be very powerful.

– Creating an uncertainty model with regression can be quite powerful. Some important limitations:

1. The data set must have "adequate" number of observations. A conservative rule of thumb would be to have at least 10 observations for each explanatory variable and never less than 30 observations total.

Regression Analysis and Modeling: Some Do's and Don'ts

2. Even with an adequate data set and a satisfactory model and analysis, there remains an important limitation. Our regression model is a linear combination of the explanatory variables. Our model may be a terrific approximation of the relationship for the variables in the neighborhood of the data that you have. But if you try to predict the response variable outside of the range of your data, you may find that your model performs poorly.

Regression Analysis and Modeling: Some Do's and Don'ts

3. You may try to predict the response variable for a combination of the explanatory variables that is poorly represented in the data. Even though the value of each explanatory variable falls within its own range, the combination for which you want to create a forecast could be very unusual

Regression Analysis and Modeling: Some Do's and Don'ts

- Note:
  - Although there is no simple way to avoid this problem and to ensure that the point for which you wish to predict the response variable lies within the data cloud in each of the scatter plots.

Summary

- We have seen some ways in which data can be used in the development of probabilities and probability distributions for decision analysis.
- The basis of constructing histograms and empirically based CDFs.
- Use of data to estimate parameters for theoretical distributions.