Supplementary Material

Professor: Daniel Egger

October 14, 2015

1 Binary Classification

1.1 Diagram of Confusion Matrix with Basic Definitions

In the confusion matrix, you have two classifications, “Positive” or “Negative,” with the underlying, true conditions being either “+” or “-”.

Examples of Binary Conditions:
Radar - Enemy Bomber/No Enemy Bomber
Diagnostics - Cancer/No Cancer
Credit Scoring - Borrower defaults/does not default

Table 1: Confusion Matrix.

				Classification/Test,X	
				“Test Positive”	“Test Negative”
				c	d
Outcome	Y	“+”	a	e	f
		“-”	b	g	h

Where,
(a,b) is the Marginal Probability Distribution of Condition $p(Y)$. Note that the rarer of the two is traditionally assigned to “+” and the probability $p(a)$ is called the “incidence” of a.
(c,d) is the Marginal Probability Distribution of the Classification $p(X)$
(e,f,g,h) is the Joint Probability Distribution of the Condition and the Classification, $p(X, Y)$.

Another way of representing the confusion matrix is:

Table 2: Confusion Matrix, another representation.

				Classification/Test,X	
				“Test Positive”	“Test Negative”
				c	d
Outcome	Y	”+”	a	”True Positive” (TP)	”False Negative” (FN)
		”-”	b	”False Positive” (FP)	”True Negative” (TN)

1.2 Important Measures of the Confusion Matrix

True Positive (TP) Rate: The conditional probability that someone with the condition has a positive test.

$$\frac{e}{a} = p(TestPositive|+)$$

False Negative (FN) Rate: The conditional probability that someone with the condition has a Negative test.

$$\frac{f}{a} = p(TestNegative|+)$$

Note that $TP\ rate + FN\ rate = 1$.

False Positive (FP) Rate: The conditional probability that someone who does not have the condition has a positive test.

$$\frac{g}{b} = p(TestPositive|-)$$

True Negative (TN) Rate: The conditional probability that someone who does not have the condition has a negative test

$$\frac{h}{b} = p(TestNegative|-)$$

Note that $FP\ rate + TN\ Rate = 1$.

Positive Predictive Value (PPV): The conditional probability that someone who has a positive test, has the condition.

$$\frac{e}{c} = p(+|TestPositive)$$

$$1 - PPV = \frac{g}{c}$$

1.3 Receiver Operating Characteristic (ROC) Curve

Known Data x about each unique item (radar image, medical test subject, potential borrower) is converted through some scoring function $s(x)$ into a single "score" for that item, s.

A "threshold" value t is chosen so that all scores where $s > t$ are assigned to the "Test Positive" category, and all scores where $s \leq t$ are assigned to the "Test Negative" category.

Holding the scoring function $s(x)$ constant, and changing the threshold value t results in a Confusion Matrix for each threshold.

A scoring function and threshold value produces on a given set of data produce a unique Confusion Matrix. The FP rate and TP rate can be plotted as an ordered pair (x, y), where $x = FP\ rate$ and $y = TP\ rate$.

Possible values range from $(0, 0)$ to $(1, 1)$. The "classification" that assigns "Test Negative" to every score has FP rate of 0 and TP rate of 0, represented as point $(0, 0)$. The "classification" that assigns "Test Positive" to every value has FP rate of 1 and TP rate of 1, represented as the point $(1, 1)$.

Joining all known (x, y) points by straight lines, the area under the resulting empirical ROC curve is known as the Area Under the Curve or **AUC**. The AUC ranges from 0.5, for a test no better than chance, to 1, for a perfect test.

The Area Under the Curve is calculated by the summation of the average height multiplied by the width, as follows:

$$\left(\frac{TP(n) + TP(n+1)}{2}\right)\left(\frac{1}{(FP(n+1) - FP(n))}\right)$$

See Excel Spreadsheets for examples of calculating each point on an empirical ROC Curve.

2 Information Measures

2.1 Probability Review

2.1.1 Basic Probability Definitions

Joint probability: $p(X, Y)$
The probability that both X and Y are true. Joint probability is commutative: $p(X, Y) = p(Y, X)$.

Conditional probability: $p(X|Y)$
The probability that X is true, given that Y is true.

Note that "Rates," PPV and NPV are Conditional Probabilities: True Positive Rate = $p(TP\ Test|+)$,

False Negative Rate = $p(FN\ Test|+)$,
False Positive Rate = $p(FP\ Test|-)$,
True Negative Rate = $p(TN\ Test|-)$,
Positive Predictive Value = $p(+|Test\ Positive)$,
Negative Predictive Value = $p(-|Test\ Negative)$.

Marginal probability: $p(X)$
The probability that X is true, independent of the value of Y.

Product distribution: $p(X)p(Y)$
The product of two marginal distributions.

Independence: $p(X, Y) = p(X)p(Y)$
Two distributions X and Y are independent if and only if their joint distribution is equal to their product distribution.
Note that an equivalent statement of independence is that if X and Y are independent, $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$.
If two distributions are not independent, they must be dependent.

2.1.2 Basic Probability Theorems

Sum rule
The marginal probability $p(X)$ is equal to the sum of the joint probabilities $p(X, Y) = p(X|Y)p(Y)$ over all possible values of Y.

Product rule
The joint probability $p(X, Y)$ is equal to the product of the conditional probability $p(X|Y)$ and the marginal probability $p(Y)$.
In other words, $p(A, B) = p(A|B)p(B)$.

Bayes' Theorem
Given that $p(A, B) = p(B, A)$ the product rule gives $p(A|B)p(B) = p(B|A)p(A)$.
Dividing both sides of the above equation by $p(B)$ gives:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

2.2 Entropy Measures for Discrete Random Variables

2.2.1 Definition of Entropy

The entropy, $H(X)$ of a discrete random variable X is defined by:

$$H(X) = -\sum_{x \in X} p(x) \log(P(x))$$

From the perspective of Bayesian logical inference, each $p(x_i)$ is a measure of **degree of belief** in an individual proposition x_i, and $H(X)$ is a measure of the **degree of uncertainty** represented by the probability distribution as a whole.

2.2.2 Units

By convention, $H(X)$ is expressed in units of bits, unless otherwise specified.
Other common units for H(X) are:
The *nat*: $\log_e(p) = \ln(p)$
The *ban*: $\log_{10}(p)$
The *deciban*: $10\log_{10}(p)$

Example in base 2

$$p(X) = (\frac{1}{2}, \frac{1}{2})$$

$$H(X) = \frac{1}{2}\log(2) + \frac{1}{2}\log(2) = 1 \; bit$$

Note that when a condition is known with certainty, $p(X) = 1, 0$, and $H(X) = 0$. Since $H(X)$ goes to zero when the result is certain, it can be considered as a measure of missing information.

2.2.3 Change-of-Base Formula

To convert units, use the change-of-base formula,

Example

For example, to convert an entropy expressed in bits into an entropy expressed in *nats*,

$$\log_e(2) = \frac{\log_2(2)}{\log_2(e)}$$

$$= \frac{1}{1.443} = 0.6931$$

or,

$$= \frac{1 \; bit}{1.443 \; bits/nat} = 0.6931 \; nats$$

2.2.4 Joint Entropy Definition H(X,Y)

The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $P(x,y)$ is defined as:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x,y))$$

2.2.5 Conditional Entropy Definition

$H(Y|X)$, $H(X|Y)$:

$$H(Y|X) = -\sum_{i=1} np(x_i)H(Y|X = x_i)$$

$$H(X|Y) = -\sum_{i=1} np(y_i)H(X|Y = y_i)$$

2.2.6 The Chain Rule

$$H(X,Y) = H(X) + H(Y|X)$$

The joint entropy of X and Y equals the entropy of X plus the conditional entropy of Y given X.

2.2.7 Mutual Information Definition

$$I(X;Y) = H(X) - H(X|Y)$$

The mutual information $I(X;Y)$ is the reduction in the uncertainty of the random variable X due to the knowledge of outcomes in Y.

Similarly,

$$I(Y;X) = H(Y) - H(Y|X)$$

The mutual information $I(Y;X)$ is the reduction in the uncertainty of Y due to the knowledge of X.

By symmetry, $I(X;Y) = I(Y;X)$.

Note that by the definition of independence, when X and Y are independent, $p(X|Y) = p(Y)$ and $p(Y|X) = p(Y)$. Therefore $I(X;Y) = 0$.

2.2.8 Relative Entropy Definition

Relative entropy, is a measure of the "distance" between two probability mass functions $p(x)$ and $q(x)$, *taken in that order.*

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Note the following facts about relative entropy:

1. It is not a symmetric, or true, distance measure: It is often the case that $D(p||q) \neq D(q||p)$.
2. It is always the case that $D(p||q) \geq 0$. This Non-Negativity is a useful property.
3. $D(p||q) = 0$ if and only if $p = q$.
4. For any symbol $x \in X$, if $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

2.2.9 Use of Relative Entropy to Calculate Mutual Information

Given two discrete random variables X, Y with probability mass function $p(x, y)$ and marginal probability mass functions $p(x), p(y)$:

$$I(X;Y) = D(p(x, y,)||p(x)p(y))$$

In other words, the mutual information I(X;Y) equals the relative entropy between the joint distribution p(x,y) and the product distribution $p(x)p(y)$.

This is consistent with the definition of Independence: When $p(x, y) = p(x)p(y)$ then

$$I(X;Y) = D(p(x, y,)||p(x)p(y)) = 0$$

2.2.10 Summary of Useful Information Equalities for Joint Entropy H(X,Y)

1. When X and Y are independent, $H(X, Y) = H(X) + H(Y) \rightarrow I(X;Y) = I(Y;X) = 0$.
2. When X and Y are dependent, $H(X, Y) = H(X) + H(Y)–I(X;Y) \rightarrow I(X;Y) = I(Y;X) > 0$.
3. $H(X, Y) = H(Y, X)$, symmetry of joint information
4. $H(X, Y) = H(X) + H(Y|X)$
5. $H(X, Y) = H(Y) + H(X|Y)$
6. $H(X, Y) = H(X|Y) + H(Y|X) + I(X;Y)$

2.2.11 Summary of Useful Information Equalities for Mutual Information I(X:Y)

1. $I(X;Y) = H(X)–H(X|Y)$
2. $I(X;Y) = H(Y)–H(Y|X)$
3. $I(X;Y) = H(X) + H(Y)–H(X, Y)$
4. $I(X;Y) = D(p(X, Y)||p(X)p(Y))$

5. $I(X;Y) = I(Y;X)$, symmetry of mutual information

2.3 Linear Regression

2.3.1 Standardization of a Set of Numbers

Values $x_i \in X$ will be in units such as miles per hour, yards, pounds, dollars, etc. It is often convenient to standardize units by converting each value xi into “standard units.” Standard units are expressed in “standard deviations from the mean of X.”

A data point represented in standard units is also known as a **Z-Score**. To convert a data point into its corresponding Z-score, subtract the mean of the data set from the individual value, then divide by the standard deviation of the data set.

In other words, the Z-score of x_i is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

Note that individual values larger than the mean will be positive, and values less than the mean will be negative. The mean has a Z-Score of 0.

Note also that when a set of values $X = x_1, x_2, ...x_n$ is expressed in Standard Units as Z-scores, so long as the mean $\bar{x}$ and standard deviation σ_x are known, all information about the original values is preserved, and can be recovered at any time.

A data set can be converted into standard units by using the Excel function **Standardize**.

2.3.2 Basic Regression Definitions

In the linear regression formula $\hat{y_i} = \alpha + \beta x_i$, the model is a function on known x_i that generates $\hat{y_i}$, where the “hat” on y_i indicates an “estimate”, or “forecast”, of the true value y_i,

- α is the y-intercept of the best fit regression line,
- β is the slope of the best fit regression line, and
- α and β are the two “parameters” of the model.

Linear regression sets model parameters α and β so as to minimize “root mean square error,” defined below.

For one ordered pair $(x_i$,$y_i)$, the model error, or “residual,”

$$y_i - \hat{y_i}$$

$$= y_i - \alpha - \beta x_i$$

The residual equals the distance between the true value y_i and a point on the regression line $\hat{y_i} = \alpha + \beta x_i$ which is the "point estimate" of y_i.

Root mean square error is calculated as follows for a set of ordered pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3)$:

1. Each residual is squared,

$$(y_i - \alpha + \beta x_i)^2$$

2. The squared errors are added together

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

3. The mean squared error is calculated by dividing by n,

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

4. The square root of the resulting mean is taken,

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}$$

This value is the "root mean square" ("r. m. sq.") error of the model on a particular set of n ordered pairs.

The regression line that minimizes root mean square error is known as the "best fit" line.

Note that because taking the square root and dividing by n are both strictly increasing functions, it is sufficient to minimize the sum of squares:

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

to determine the model parameters α and β for the best fit line.

It can be demonstrated that when α and β are chosen to minimize the root mean square residual, the mean residual $(\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2) = 0$. Therefore, the root mean square residual is equal to the standard deviation of residuals, σ_e.

2.3.3 Determining the Parameters α and β

We use elementary calculus methods for calculating the minima of a function to solve for the values α and β that minimize the r. m. sq. error. Note that $\bar{x}_n$ and $\bar{y}_n$ are means over the set of n ordered pairs.

Take the first derivative of $Q = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ with respect to β, and set it equal to 0.

$$\frac{dQ}{d\beta} = 0$$

Solving for β gives:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}$$

Second, take the first derivative of $Q = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ with respect to α, and set it equal to 0.

$$\frac{dQ}{d\alpha} = 0$$

$$n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

Solving for α,

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$$

These two equations are known as the “normal equations” for a regression line.

The notation $\hat{\beta}$ and $\hat{\alpha}$ with “hats” signifies that these values are calculations based on a particular finite set of n ordered pairs, $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$. Adding more ordered pairs will change the values for $\hat{\beta}$ and $\hat{\alpha}$. When we can assume the existence of a “stationary” process relating two dependent random variables $\mathbf{X}$ and $\mathbf{Y}$, then, at the limit, as the size of our “sample” of ordered pairs n gets very large, $\hat{\beta}$ will approach the “true” value β, and $\hat{\alpha}$ will approach the “true” value α.

2.3.4 Using the Normal Equations on Standardized Data

It is often convenient in data mining to “standardize” a set of values $x_1, x_2, ..., x_n$, converting them into to Z-scores, by subtracting the mean from each value and dividing by the standard deviation.

$$z_{x_1} = \frac{x_1 - \bar{x}}{\sigma_x}$$

The resulting set has a number of convenient properties, including mean $\bar{z_x} = 0$ and standard deviation $\sigma_{z_x} = 1$.

If we standardize each of the x and y values in a set of n ordered pairs, so that
$\bar{z_x} = 0$ and $\bar{z_y} = 0$
$\sigma_{z_x} = 1$ and $\sigma_{z_y} = 1$

Then the normal equations become:

$$\hat{\alpha} = 0$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{Cov_{XY}}{Var_X} = R$$

The correlation coefficient which is called by definition R. The best fit regression line for standardized ordered pairs passes through the origin (0,0) and its slope equals the correlation.

2.3.5 Useful Properties of Gaussians

For parametric models we are interested in the family of Gaussian (also called "normal") probability density functions of the form:

$$f(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ and σ^2 are parameters, and $\sigma \neq 0$. These functions are all continuous and differentiable.

2.3.6 Moments of Gaussians with Mean = 0

(Note: when mean = 0, "raw" moments = moments around the mean)

The 0th Moment

$$E(X^0) = \int_{-\infty}^{+\infty} x^0 f(x)dx = 1$$

the area under the curve.

the 1st Moment

$$E(X^1) = \int_{-\infty}^{+\infty} x f(x)dx = \mu$$

the mean.

The 2nd Moment

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx = \sigma^2$$

the variance.

Note that μ is also the median and mode of a Gaussian. The square root of the variance, σ, is known as the "standard deviation."

2.3.7 Notation for Gaussians

The notation $\sigma(m, v)$ always represents a Gaussian distribution with mean m and variance v.

2.3.8 Cumulative Normal Function

The function $F(x) = p$ represents the probability that a random variable falls in the interval $(-\infty, x]$.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note that the derivative of the cumulative normal function $F'(x) = f(x)$.
When adding two independent Gaussians, the resulting distribution is also a Gaussian, with mean equal to the sum of the means, and variance equal to the sum of the variances.

$$\sigma_1(m_1, v_1) + \sigma_2(m_2, v_2) = \sigma_3(m_3, v_3)$$

The converse is also true: if the sum of two independent random variables is a Gaussian, both of those random variables must be Gaussians.
Linear transformations of Gaussian random variables are also Gaussians.
If $X = \sigma(\mu, \sigma^2)$,

$$aX + b = \sigma(a\mu + b, a^2, \sigma^2)$$

2.3.9 Standard Normal Distribution

The Gaussian with mean $\mu = 0$ and variance $\sigma^2 = 1$, $\sigma(0, 1)$, is known as the "standard normal" distribution. It has cumulative normal function:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x)^2}{2}}$$

2.3.10 Gaussian Models With Linear Regression

Models that assume data are drawn from known distributions, with known mean and variance, are called "parametric" models. The most common parametric models assume that observed data are draws from random variables with a normal, or Gaussian, distribution.

One parametric model often used is to assume that the residuals, $y_i - \alpha - \beta x_i$ are normally distributed with mean = 0 and Standard Deviation = σ_ϵ.

Each value y_i can then be represented as:

$$y_i = \alpha + \beta x_i + \epsilon,$$

where ϵ is an independent random draw from $Z = \sigma(0, \sigma_\epsilon^2)$, a normal distribution with mean = 0 and variance = σ_ϵ^2.

If we further assume that x values are drawn from a random variable **X** with normal distribution, mean = $\bar{x}$ and standard deviation = σ_x, then **Y**, as the sum of two normal distributions, must also be a normal distribution.
Note that **Z** is independent of **X**.
Therefore the variances add, so that:

$$\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_\epsilon^2$$

By the formula for adding means and variances of two normal distributions, **Y** is a normally-distributed random variable, with mean = $\beta\bar{x} + \alpha$, and variance = $\beta^2\sigma_x^2 + \sigma_\epsilon^2$.

2.3.11 The Relationship Between Correlation and the Root Mean Square Residual for Parametric Models

It should be apparent that the larger the variance of the residual (relative to the variance of **Y**), the smaller the absolute value of the correlation R.

When **X** and **Z** are normally-distributed, the exact relationship between σ_ϵ and R can be determined by substitution for β, which is related to R by the formula:

$$\beta = R\frac{\sigma_y}{\sigma_x}$$

Since,

$$\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_\epsilon^2$$

$$\sigma_y^2 = \frac{R^2 \sigma_y^2 \sigma_x^2}{\sigma_x^2} + \sigma\epsilon^2$$

$$= R^2 \sigma_y^2 + \sigma_\epsilon^2$$

Rearranging terms,

$$\sigma_\epsilon^2 = \sigma_y^2 - R^2 \sigma_y^2$$

$$R^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}$$

or

$$R = \sqrt{1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}}$$

When using standardized variables, $\sigma_y = 1$, and $R = \sqrt{1 - \sigma_\epsilon^2}$.

2.3.12 Differential Entropy and Entropy of a Gaussian

For continuous random variables, the "differential entropy" $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

$$h(X) = -\int_S f(x) \log f(x) dx$$

where subscript S indicates that the domain is limited to the support set of the random variable, that part of the real line where $f(x) \neq 0$.

Differential entropy can be interpreted as a measure of the uncertainty about the value of a continuous random variable, or a measure of missing information.

Random variables "collapse" to $h(X) = 0$ once they are associated with a known outcome or event x.

2.3.13 Entropy of a Gaussian (Calculated in Nats with Conversion to Bits)

Start with a Gaussian with mean = 0 and standard deviation σ. Entropy is defined by:

$$h(X) = -\int_S f(x) \ln f(x) dx$$

$$= -\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} [\frac{-x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}] dx$$

$$= -(\int f(x)[\frac{-x^2}{2\sigma^2}] dx - \ln \sqrt{2\pi\sigma^2})$$

$$= (\frac{1}{2\sigma^2} \int x^2 f(x) dx) + \ln \sqrt{2\pi\sigma^2}$$

By definition of the second moment of a Gaussian, this equals:

$$= \frac{E(X^2)}{2\sigma^2} + \frac{1}{2} \ln(2\pi) + \ln \sigma$$

$$= \frac{1}{2} + 0.9189 + \ln \sigma$$

$$= 1.4189 + \ln \sigma$$

$$conversion\ bits$$

2.3.14 The Relationship Between Correlation and Mutual Information

Assume a Gaussian random variable **Y** which is the sum of two independent Gaussian random variables **X** and **Z**.

As mentioned above, every value of y, $y_i = \alpha + \beta x_i + \epsilon$,

where ϵ is an independent random draw from $\mathbf{Z} = \phi 0, \sigma_{\epsilon^2}$, a normal distribution with mean = 0 and variance = σ_ϵ^2

Note that $(\mathbf{Y} \mid \mathbf{X}) = \mathbf{Z}$.

In other words, the uncertainty remaining in **Y**, when **X**, and the *linear model that relates the dependent part* of **Y** to **X**, $\hat{y_i} = \alpha + \beta x_i$, are known, is equal to the residual component ϵ.

By definition,

$$I(X;Y) = H(Y)–H(Y|X)$$

$$= H(Y)–H(Z)$$

$$= (1.42 + ln\sigma_y) - (1.42 + ln\sigma_\epsilon)$$

$$= -ln\frac{\sigma_\epsilon}{\sigma_y}$$

By substitution from $R^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}$ above,

$$= -\frac{1}{2}ln(1 - R^2)$$

or, in terms of entropy to the base 2,

$$= -\frac{1}{2}log\frac{1}{1 - R_{xy}^2}$$

Note that, unlike discrete entropy, differential entropy in infinite (undefined) when $R^2 = 1$.

Note that the linear model, plus knowledge of X, leaves H(Y|X) = residual Z; it is possible that a better-than-linear model exists. For this reason, when R is known we know that the mutual information I(X;Y) is

$$\geq \frac{1}{2} \log \frac{1}{1 - R_{xy}^2}$$

with equality when the best model is the linear model.

2.3.15 Converting Linear Regression Point Estimates to Probabilistic Forecasts (When Data are Parametric)

Each value x_i, when combined with the linear model $\hat{y_i} = \alpha + \beta x_i$, and the known root mean square residual σ_ϵ can be thought of as providing either,

1. a "point" forecast $\hat{y_i}$, or
2. a probabilistic forecast in the form of Gaussian probability distribution with mean = $\hat{y_i}$ and standard deviation = σ_ϵ.

Example
Assume $\hat{y_i} = 2$ and σ_ϵ of the linear model = 3. Assume the errors have a Gaussian distribution.

The "true" value of y_i is unknown. Suppose you need to know the probability that the true value of $y_i > 5$.

This probability equals 1-(the cumulative normal distribution from $-\infty$ to 5) of the Gaussian function with mean = 2 and standard deviation = 3.

This is equal to the cumulative standard normal distribution from $-\infty$ to -1. In Excel, this is "=norm.s.dist(-1, true)" and is equal to 15.67%.

Suppose you need to know the probability that the true value of $y_i < 0$. This probability is equal to the cumulative standard normal probability distribution from $-\infty$ to -2/3. In Excel this is "=norm.s.dist(-.66667, true)" and is equal to 25.25%.

A more precise conversion to probabilities can be adopted that assumes that the model's values for α and β are themselves estimates, derived from a set of n known ordered pairs. This more advanced model is beyond the scope of the present discussion.

2.3.16 Adjustment of Linear Regression Model Error

Note: Adjustment of the root mean square error of a point estimate when linear regression is calculated on a sample of small size.

We typically assume that our data set for linear correlation forecasting is parametric – meaning that we assume that our ordered pairs (x, y) of unstandardized or standardized data are actually drawn from underlying Gaussian Probability Distributions X and Y – with constant "true" standard deviations and covariances, and consequently correlation and root mean square error.

We've observed that for any finite sample of ordered pairs drawn from the above random variables, the values for beta, the slope of a regression line, for alpha, the y-intercept, and for the linear correlation R, can and will all differ from the "true" values for the random variables.

Similarly, the observed root mean square error of the best-fit line will also differ from the "true" error as a function of the number of observations n. Interestingly, the true error is also changed by the z-score of each individual x value – the farther the x value is from the mean, the greater the true error. We can adjust the confidence interval for individual point estimates to take these variations into account. However, in general, the difference between the observed error and the "true" error is small if n is larger than 100 and z is between -3 and 3. The formula for the True Error of a point estimate, adjusted for n and sample size, is

$$RMS * \sqrt{\frac{(z^2+1+n)}{n}}$$

For values of n greater than 100 and z-scores between $(-1.5to1.5)$, the theoretical adjustment is always less than 2%, and can safely be ignored.

On samples with small n and large z-scores the adjustment is worthwhile. For example, assume the observed root mean square error is 0.74, sample size of $n = 50$ ordered pairs, and a particular $x(i)$ has a $z - score = 2$.

The best estimate for the "true" error of the y point estimate is

$$0.74\sqrt{\frac{4+1+50}{50}} = (.74)(1.049) = .78.$$

This increase of approximately 5% in the error is also accompanied by a change of the distribution away from a pure Gaussian shape. However, when n is larger than 100 a Gaussian is a very good approximation for the distribution of the true error.