Introducing Area Approximated (A New Linear Normed Goodness of Fit Metric) and Comparing It Versus R^2 a.o. Using Fully Generalized Normal Error - Version 0.9 -

presented by Heiko Schlingmann

November 6, 2017

as PhD Thesis

to receive a Doctor's Degree

at

t.b.d. University

t.b.d. Faculty

The author is temporarily handicapped, so that he cannot perform the usual PhD study process. By the pre-publication of the thesis *the author is still looking for an alma mater. (Conditional) alma mater invitations may be send to mail@heiko-schlingmann.com.*

Also, graduated academics active in university research are invited to question one or more results of the thesis. Such questions will typically be answered within one week, except for Christmas holiday season and summer holiday season (mid of July to mid of August).

Contents

List of Tables

List of Figures

Listings

Chapter 1

Introduction

Today, there is a surplus of prediction models available, as well "classical" statistical prediction models like the regression family, as well as modern machine learning models. Typically each prediction model brings its own goodness of fit statistic (GOF statistic). But when the statistician has fitted different types of prediction models on the data, then *universal* goodness of fit statistics, that enable inter-model-type comparisons, are rare. Especially, there are only very few *normed* universal goodness of fit metrics, that additionally would allow for data-set independent goodness-of-fit bench-marking, either in statistical software packages and either in statistical theory. Thus, practicing statisticians building different types of prediction models still use the good old GOF statistics R-square and Pearson correlation for inter-model-type comparisons on interval scale data. This mainly for the following reasons. First, these metrics easily provide data-set independent goodness of fit comparisons by their normed measurement. Especially they provide an intuitive goodness of fit bench marking to third persons, that are not familiar with the data-set. $R^2 = 0.80$ can be interpreted as the prediction model has captured 80% of the target variable's squared deviations from the mean. Second, they are very popular, thus broadly accepted.

However, R-square and correlation originally have not been developed for universal inter-model-type comparisons. Especially, the squaring of the elementary errors in these two goodness of fit statistics may cause deformed - especially influential - GOF measuring. For this reason the author has developed a new linear and normed goodness of fit statistic, called "Area Approximated", which also provides well intuitive interpretation. The introduction of the new Area Approximated GOF metric and its comparison to R-square based on study cases are the major subjects of this work. Also Pearson correlation is included in the first comparisons. But it soon turns out, that it is not a proper GOF statistic. Thus, Pearson correlation is only presented in the basic analysis, but not researched any further.

The distribution of the residual error may be a big issue for fair objective GOF measuring. Conclusively, appropriate test data also must include prob-

lematic residual errors, especially skew and kurtosed residual errors. These skew and kurtosed errors are ideally produced by one distribution generator to be able to flexibly combine them at any degree where necessary. For this purpose the author has developed the first fully generalized normal distribution, which can generate random variables, that are kurtosed (kurtosis different from 3) and as well skew (skewness different from 0) at the same time. As the full term "a fully generalized normal distribution version 1" is not handy, the new distribution is shortly called "S-distribution", which enables to state random error generation in error function notation and will also provide distinction from further fully generalized normal distributions in the future.

Beside the development of the new fully generalized normal distribution "S-distribution" there has been some further tool work necessary. To gain a straight scientific presentation, that strictly constructs bottom-up, all this basic research and its resulting tools are aggregated in chapter 2 "Tools". Generally all over this work, special author developments are marked with " * " in the section title. These little markings shell draw special attention to sections, that may be important for a proper understanding of this work and as well important for a scientific validation of this work.

Chapter 3 turns over to normed universal goodness of fit measuring for interval scale data. This starts with the definition of the terms "goodness of fit", "universal goodness of fit" and "normed goodness of fit". Next, the two known normed universal GOF statistics Pearson product-moment correlation and R-square are recapitulated. In the section "Goodness of Fit Measuring Insights" the author lays out, that a loss function is also inherent in goodness of fit measuring and further, that the definition of the loss function is central for the preference of statistical models - respectively its optimization technique - by goodness of fit metrics. In following section the new "Area Approximated" GOF metric gets developed, which is driven by the idea to avoid influential squared error accounting.

Chapter 4 is about the quality comparison of the new "Area Approximated" GOF metric. The quality criteria definition is begun by the search for the true loss function of interval scale data. It turns out, that it is (linear) distance. This conclusion may be a little controversial at first. However, the author's scientific procedure to look a critical issue the up in the definition - here looking up the critical error power in the definition of interval data scale - is a probed and widely accepted scientific method. Following three quality criteria are defined, namely consistence, calibration and linearity. In the next step 24 study cases, ordered in 6 test series out of 4 study cases each, are developed. In last section of this chapter the quality comparison of the new "Area Approximated" GOF metric is performed, based on the study cases.

As common, the final chapter 5 contains a summary of the findings of this work. The major result of this work is the strong indication, that there are no weaknesses in the new Area Approximated goodness of fit statistic, but unexpectedly the traditional R-square seems to perform largely positively inflated goodness of fit measuring, when accepting the author's finding, that the true neutral loss function for objective goodness of fit measuring is absolute

loss. Secondary, the discovery of the first fully generalized normal distribution "S-distribution" is a nice side-effect, although the distribution is - from a mathematical point of view - still in the construction process. However, the author expects, mathematicians will plug in the missing components, soon.

One last remark concerning the form of presentation. As reproducibility has gained a major role in scientific research, the R Code of central parts of this work is presented in line.

Chapter 2

Tools

2.1 Equal Distance in Metric Scales of Measurement

Following Steven's typology a number a quantity is classified interval scale, if the difference between two numbers and the same difference between any other two numbers reference equal distance. An example for an interval scale quantity is temperature in degrees Celsius. While its difference is well defined, its ratio is not. Following Steven's typology a quantity, that *additionally* to equal distance property of interval scale also has a well defined ratio, is classified ratio scale. A defined ratio is understood this way, that a value of x references x-times the unit value, in the means of x-times much or x-times many. The x-times interpretation requires, that the quantity has an objective non-arbitrary zero value. Often, the zero value also references the absolute minimum of the quantity. E.g. a temperature in degrees Celsius is classified interval scale, because the reference to water (freezing point and boiling point) is arbitrary, but a temperature in Kelvin is classified ratio scale, as 0° Kelvin is the lowest possible temperature, thus 20° Kelvin must be 20-times 1° Kelvin.

However, for this work only the equal distance property is relevant, which is - further following Steven's typology - inherent in both metric scales of measurement, interval scale and ratio scale.

2.2 Error Definition, Error Aggregation and Loss Function

Generally, an error is the discrepancy between a true or desired value (target value) and the corresponding computed value. In statistics discrepancies are also called deviations, especially when they arise from a central tendency, such as the mean.

In approximation theory discrepancies are mainly measured as absolute error, namely absolute value of the difference.[1] Discrepancies on quantities with a huge value range are measured as relative error in approximation theory, namely absolute error divided by the absolute value of the target.[2]

In statistics a discrepancy on interval scale quantities is measured as difference.[3] Namely, the discrepancy between an interval scale target value y_i and the corresponding interval scale estimate $\hat{y}_i$ is defined the difference $y_i - \hat{y}_i$; it is also referenced as elementary error. Unfortunately, there is no generally-accepted, specific elementary error definition for ratio scale quantities in statistics. Following, discrepancies in ratio scale quantities are usually treated as interval scale errors in statistics, as there is no generally-accepted specific discrepancy definition for them.

The synopsis of elementary error definition is little confusing. Approximation theory does not define distinct levels of measurement for metric quantities, but has two elementary error definitions, absolute error and relative error. Statistics has two distinct levels of measurement for metric quantities, however the distinction is academic, as the elementary error is defined difference for both measurement levels.

In numerical mathematics, especially in approximation theory, the errors of an one-dimensional vector - particularly computed by an approximation function - , are aggregated by its maximum error. This means, an approximation - typically an approximation of the target function by another less complex or more common approximation function, is qualified to be always better than the guaranteed maximum error limit.[4] Additionally, this basic one-dimensional error aggregation concept can be extended to N-dimensional vectors by replacing the absolute value with an N-norm.[5] Systematically, the basic error aggregation concept substantially consists out of the compression of an error vector to its maximum error, which is appropriate in approximation theory, as all errors are deterministic and fully causally result from the approximation function.

In statistics, mathematical optimization and related disciplines however, we typically deal with vectors, that are at least partially random/ stochastic. Following, the maximum error aggregation concept is not appropriate any more, as a maximum error would be largely influenced by the stochastic part. Thus, in statistics we are basically interested in the mean error, calculated by the error sum divided by the number of observations. However, when the number of observations is given as datum for the optimization, which is common in statistics, the number of observations becomes irrelevant for the determination of the optimum. Conclusively, the mean error can often be equivalently represented by the error sum, only.

The function, that aggregates the elementary errors/ deviations - typically to a single value - , is mostly referenced as loss function. A loss function gets

[1] XYZ.
[2] XYZ.
[3] XYZ.
[4] XYZ.
[5] XYZ.

minimized in optimization. Other terms for loss functions are cost function (machine learning, economics) and fitness function (neural networks, evolutionary algorithms). An objective function (continuous data) or a score function (nominal data) [6] can be either a loss function or more typical its negative. In the typical second case they are maximized in optimization. The term error function seems to be outdated, as it with collides with the (Gauss) error function in special functions theory and the error function notation in statistics.[7], [8]

Although there is huge number of loss functions, the squared loss functions is dominantly common.

$$L = \sum (y - \hat{y})^2 \tag{2.1}$$

also quoted

$$L = \sum \frac{1}{2}(y - \hat{y})^2 \tag{2.2}$$

Another popular loss function is the absolute loss function.

$$L = \sum |y - \hat{y}| \tag{2.3}$$

The squared loss function has the big advantage, that it is globally continuous and particularly also globally differentiable. Global differentiation capability is highly desired, as popular optimization algorithms are based on differentiation[9] and alternative optimization algorithms free of differentiation are less performant.[10] On the other hand the squared loss function has the disadvantage to be dominated by outliers, also referenced as outlier influenced. When summing over a set of squared errors in $\sum (y - \hat{y})^2$, the errors of outlying observations are larger and grow much more by squaring than smaller errors. Thus, the squared loss function cannot precisely represent the mean error. However, the squared loss function - when dividing it by the number of observations - always measures more error than the true mean error. Thus, unnormed error metrics like mean squared error or root mean squared error can be considered conservative.

The absolute loss function has the strength, that it measures the mean error exactly, when dividing it by the number of observations. On the other hand it has the weakness, that it is not differentiable at $y - \hat{y} = 0$. Conclusively,

[6] The term score function is also used in generalized linear models for the (approximate) first derivation of the log-likelihood to the parameters. [Gill (2000)].

[7] Definitions taken from XYZ, XYZ, XYZ, XYZ.

[8] For error function notation compare 2.10.1 on page 32.

[9] Such as least squared error (OLS regression), maximum likelihood (generalized linear model) and back-propagation (neural network). Compare [Fox (1997)], [Gill (2000)] and [Herve, Valentin et al. (1999)].

[10] [Eubank, Kupresanin (2012), Spall (2003)].

the majority of statistical and machine learning models for interval scale data cannot be applied on the foundation of an absolute loss function, as their the optimization algorithms are based on differentiation. Thus, the choice of the loss function is not arbitrary but very restrictive. [Klebanov, Rachev et al. (2009)] suggest to choose the loss function by its desirable properties for each application individually.

2.3 The Gaussian Integral and Its Generalizations

Although there is no elementary indefinite integral for

$$\int exp(-x^2)dx$$

the definite integral can be evaluated.

$$\int_{-\infty}^{\infty} exp(-x^2)dx = \sqrt{\pi}$$

Known Generalizations:

$$\int_{-\infty}^{\infty} exp\left(-0.5\left(\frac{x-\mu}{\sigma}\right)^2\right)dx = \sigma\sqrt{2\pi} \tag{2.4}$$

$$\int_{0}^{\infty} exp(-ax^2)dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}$$

$$\int_{0}^{\infty} exp(-ax^k)dx = \frac{\Gamma(\frac{1}{k})}{ka^{\frac{1}{k}}} \tag{2.5}$$

Further generalization suggested by the author following repeated equivalence to numerical integration:

$$\int_{-\infty}^{\infty} exp\left(-a\left(\frac{|x-c|}{z}\right)^k\right)dx = \frac{2\,z\,\Gamma(\frac{1}{k})}{k\,a^{\frac{1}{k}}} \tag{2.6}$$

- [PROOF]

2.4 Probability Density Function (PDF)

2.4.1 Definition

Consulting [Fahrmeir, Künstler et al. (2004)] a probability density function (smoothly) enumerates the relative frequency of a continuous random variable.

2.4.2 PDF of the Normal Distribution

Standard normal distribution's PDF

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-0.5\,x^2) \tag{2.7}$$

Where

$f(x)$	Probability density function (PDF)
x	Random Variable
μ	Construction mean
σ	Construction standard deviation
$\exp(u)$:	e^u

Normal distribution's PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.8}$$

Where

$f(x)$	Probability density function (PDF)
x	Random Variable
μ	Construction mean
σ	Construction standard deviation
$\exp(u)$:	e^u

2.4.3 Probability Density Function Axioms

Regarding to the three probability density function axioms a valid probability density function $f(x)$ of a continuous random variable x must satisfy the following conditions.[11]

1. $f(x)$ is a continuous function.
2. $f(x) \geq 0$
3. $\int_{-\infty}^{+\infty} f(x)dx = 1$

[11] [Fahrmeir, Künstler et al. (2004)], p. 88.

Where

$f(x)$	Probability density function (PDF)
x	Random Variable
$\int_{-\infty}^{+\infty} dx$	Definite Integral with the integration limits $-\infty$ and $+\infty$.

2.4.4 Frequency Norming inside a PDF

To fulfill the probability density function axiom no. 3, that the the PDF has to sum up to 1, a PDF is usually designed this way, it is build out of two parts, the norming constant and the core frequency function. For instance, the PDF of the normal distribution consists out of

the norming constant

$$\frac{1}{\sigma\sqrt{2\pi}}$$

and the core frequency function

$$f^*(x) = \exp(-0.5\,(\frac{x-\mu}{\sigma})^2)$$

Remembering equation 2.4 from section 2.3 on page 13 we are discovering, that the norming constant exactly counts 1 divided by the integral of the core frequency function.

Thus, we can more generally express the PDF of the normal distribution as

$$f(x) = \frac{1}{\int_{-\infty}^{+\infty} f^*(x)\,dx} \cdot f^*(x) \tag{2.9}$$

Where

$f(x)$	Probability density function (PDF)
$f^*(x)$	(Unnormed) Frequency Function
x	Random Variable
$\int_{-\infty}^{+\infty} dx$	Definite Integral with the integration limits $-\infty$ and $+\infty$.

We have just explicated the design logic inherent in the PDF of the normal distribution.

Later on in this work we will use this design logic to build up new generalized versions of the normal distribution. Especially, when the core frequency function is altered, the new function must again sum up to 1 again to fulfill PDF axiom no. 3. Thus, we have to solve the norming integral

$$\int_{-\infty}^{+\infty} f^*(x)\,dx$$

to state a compact norming constant. This frequency norming will become very central in the sections 2.6.3 on page 21 and 2.7 on page 25. And, as a modification of the normal distribution is concerned, it is directly linked to the Gaussian integral in section 2.3 on page 13.

2.5 Statistical Indicators

2.5.1 Moments

2.5.1.1 Mean

Sample Mean The very well known the sample mean $\overline{x}$ is calculated

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where

n	Number of Observations
i	Iterator
x	Random Variable

Theoretic Distribution Mean The theoretic distribution mean $\mu$is calculated

$$\mu = \int_{-\infty}^{\infty} f(x)\,x\,dx$$

Where

$\int_{-\infty}^{\infty}$	Definite Integral with the integration limits $-\infty$ and $+\infty$
$f(x)$	Probability Density Function
x	Random Variable

2.5.1.2 Variance

Sample Variance The well known the sample variance s^2, the 2nd central moment, is calculated

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Where

n	Number of Observations
i	Iterator
x	Random Variable
$\overline{x}$	Sample Mean of x

Theoretic Distribution Variance The theoretic distribution varianceσ^2 is calculated

$$\sigma^2 = \int_{-\infty}^{\infty} f(x) \, (x - \mu)^2 \, dx$$

Where

$\int_{-\infty}^{\infty}$	Definite Integral with the integration limits $-\infty$ and $+\infty$
$f(x)$	Probability Density Function
x	Random Variable
μ	Population Mean

2.5.1.3 Skewness

Sample Skewness The moment coefficient of skewness γ_S, the 3rd central moment, is the most common metric of skewness. The core equation is common statistical knowledge. However, there may exist different versions of sampling corrections. Thus, it is good to know, which formula is used in this work.

$$\gamma_S = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Where

n	Number of Observations
i	Iterator

x	Random Variable
$\overline{x}$	Sample Mean of x
s	Sample Standard Deviation of x

Interpretation

$\gamma_S < 0$	Left skewed distribution
$\gamma_S = 0$	Symmetric distribution (without skewness)
$\gamma_S > 0$	Right skewed distribution

Theoretic Distribution Skewness For a theoretic distribution the moment coefficient of skewness γ_S is calculated

$$\gamma_S = \int_{-\infty}^{\infty} f(x) \left(\frac{x - \mu}{\sigma} \right)^3 dx$$

Where

$\int_{-\infty}^{\infty}$	Definite Integral with the integration limits $-\infty$ and $+\infty$
$f(x)$	Probability Density Function
x	Random Variable
μ	Population Mean
σ	Population Standard Deviation

2.5.1.4 Kurtosis

Sample Kurtosis The moment coefficient of kurtosis γ_K, also known as 4th central moment, is the most common metric of kurtosis. The core equation is common statistical knowledge, too. However, there definitely exist different versions of sampling corrections. Thus, it is important to know, which formula is used in this work.

$$\gamma_K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4$$

Where

n	Number of Observations
i	Iterator
x	Random Variable

$\overline{x}$	Mean of x
s	Sample Standard Deviation

Interpretation

$\gamma_K < 3$	Less peaked and less fat-tailed than the normal distribution
$\gamma_K = 3$	As peaked and as fat-tailed as the normal distribution
$\gamma_K > 3$	More peaked and more fat-tailed than the normal distribution

Theoretic Distribution Kurtosis For a theoretic distribution the moment coefficient of kurtosis γ_K is calculated

$$\gamma_K = \int_{-\infty}^{\infty} f(x) \left(\frac{x - \mu}{\sigma} \right)^4 dx$$

Where

$\int_{-\infty}^{\infty}$	Definite Integral with the integration limits $-\infty$ and $+\infty$
$f(x)$	Probability Density Function
x	Random Variable
μ	Population Mean
σ	Population Standard Deviation

2.5.2 Other

2.5.2.1 Absolute Deviation

The absolute deviation D between the single outcomes of two random variables a_i and b_i is calculated

$$D = |a_i - b_i| \tag{2.10}$$

The mean absolute deviation $\bar{D}$ between of two random variables a and b is calculated

$$\bar{D} = \frac{1}{n} \cdot \sum_{i=1}^{n} |a_i - b_i| \tag{2.11}$$

Although obvious, the mean absolute deviation $\bar{D}$ between the outcomes of a random variable a and a constant$\bar{c}$ is calculated

$$\bar{D} = \frac{1}{n} \cdot \sum_{i=1}^{n} |a_i - \bar{c}|$$

2.5.2.2 Euclidean Distance

The euclidean distance is the ordinary straight-line distance between two points. Precisely, the euclidean distance between two points p and q in an N-dimensional (euclidean) space is the length of the line segment $\overline{pq}$. The N-dimensional euclidean distance $d(p, q)$ computes

$$d(p,q) = \sqrt{\sum_{j=1}^{N}(p_j - q_j)^2} \tag{2.12}$$

Where

N	Number of Dimensions
j	Dimension iterator

One-dimensional euclidean distance between two points p and q on the real line computes

$$d(p,q) = \sqrt{(p-q)^2}$$

$$\Leftrightarrow d(p,q) = |p-q| \tag{2.13}$$

Thus, one-dimensional euclidean distance is equivalent to absolute deviation.

2.6 Generalized Normal Distributions

2.6.1 Generalized Normal Distribution Version 2

[Hosking, Wallis (1997)] invented a generalized normal distribution generalized for skewness, which is nowadays half-officially referenced as generalized normal distribution version 2. This generalization was not involved in the development of S-distribution. It has only been mentioned for completeness.

2.6.2 Generalized Normal Distribution Version 1

[Nadarajah (2005)] published a generalized normal distribution generalized for kurtosis, which is nowadays half-officially referenced as generalized normal distribution version 1. This generalization has been very helpful in the development of the S-distribution. There also exist some variations of the original generalized normal distribution version 1 by Nadarajah, which however did not have impact on the development of the S-distribution. The probability density function of the original generalized normal distribution version 1 by Nadarajah is

$$f_{GND-V1}(x) = \frac{k}{2\,z\,\Gamma(\frac{1}{k})} \cdot \exp\left(-\left(\frac{|x-c|}{z}\right)^k\right) \tag{2.14}$$

Legend in the wording of S-distribution

$f(x)$	Probability density function (PDF)
x	Random variable
c	Construction center (corresponding to normal distribution's μ)
z	Construction deviation (corresponding to normal distribution's σ)
k	Power coefficient
$\Gamma(\ )$	Gamma function
$\exp(t)$	e^t

The concentrated reader may have recognized, that the 0.5 is missing in the exponent of Nadarajah's generalized normal distribution version 1. [Nadarajah (2005)] does not say, why he changed the factor in the exponential power. He just mentions, that his generalized normal distribution version 1 - respectively its PDF - is "a natural generalization of the normal distribution". As the Gaussian integral is needed to state the frequency norming of an PDF within a short formula, which is a complicated one in this case, as there is no elementary indefinite integral for the Gaussian, it can be guessed, that the latest generalization of the Gaussian integral may not been known in 2005. This little exponent imperfection leads to the fact, that Nadarajah's generalized normal distribution version 1 cannot exactly reproduce the normal distribution. Proof:

$$f_{GND-V1}(x = 0,\ c = 0,\ z = 1,\ k = 2) = 0.28209$$

$$f_{Normal}(x = 0,\ \mu = 0,\ \sigma = 1) = 0.39894$$

Following mathematical logic a single non-matching case is sufficient to refuse an - implicit - proposal.

2.6.3 Kurtosed Normal Distribution *

To heal the little exponent imperfection in the generalized normal distribution version 1 we have to add the factor 0.5 in the exponent of the PDF. To be able to state a short norming constant [12] for the modified core frequency function again, the author had to drive the latest known generalization of the Gaussian integral a little further.[13] Using the author's further generalization of the Gaussian integral - equation 2.6 - on the modified frequency formula establishes a new normed PDF.

$$f(x) = \frac{k\,0.5^{\frac{1}{k}}}{2\,z\,\Gamma(\frac{1}{k})} \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right) \tag{2.15}$$

[12] Review section 2.4.4 on page 15.
[13] Review section 2.3 on page 13.

Where

$f(x)$	Probability density function (PDF)
x	Random variable
c	Construction center (corresponding to normal distribution's μ)
$z > 0$	Construction deviation (corresponding to normal distribution's σ)
$k > 0$	Power coefficient
$\Gamma(\,)$	Gamma function
$\exp(t)$	e^t

To distinguish the new distribution, which is now straight-forward from the normal distribution and 100% downward-compatible with it, without any misunderstanding from Nadarajah's generalized normal distribution version 1, it is called kurtosed normal distribution. Further, as the generalization causes the effective parameters to deviate from their construction pendants, the names of the parameter used for construction get the addition "construction", especially construction center c and construction deviation z. It is obvious, that the construction center c corresponds to the normal distribution's population mean μ and the construction deviation z corresponds to the normal distribution's population standard deviation σ. When kurtosis parameter $k = 2$, the normal distribution gets reproduced exactly. However, if $k < 2$, the distribution gets more wide, so that the effective deviation is greater than the construction deviation ($\sigma > z$). And vice versa, if $k > 2$. Later, it will be same with the effective population mean μ in the S-distribution, when further generalizing for skewness. Thus, the parameter is - looking forward to the S-distribution - already named in S-distribution wording to avoid double naming for the same thing, although the construction center c of the kurtosed normal distribution does not deviate from the effective mean, so far.

2.6.3.1 Density Graphs

A new distribution always gets more intuitive, when drawing some exemplary density graphs. But first, it is shown, that the kurtosed normal distribution can exactly reproduce the normal distribution, as the gained full compatibility to the normal distribution has been reason for its development.

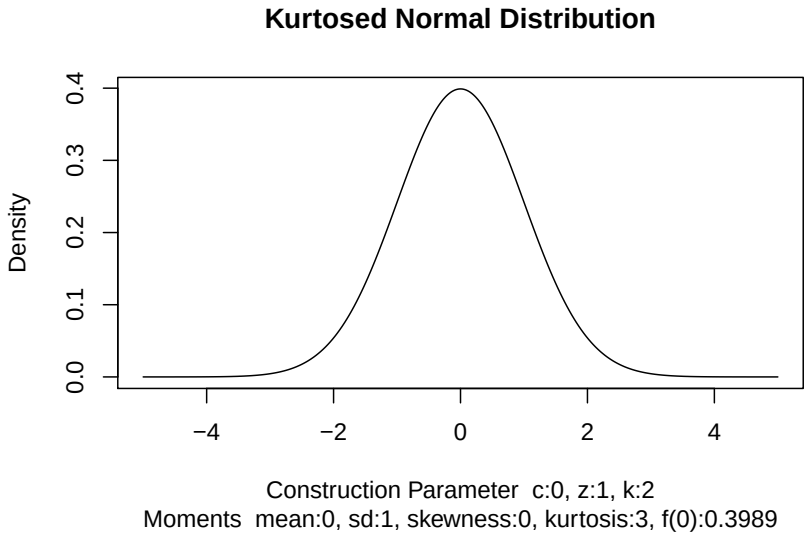

Figure 2.1: Kurtosed Normal Distribution Reproducing the Normal Distribution

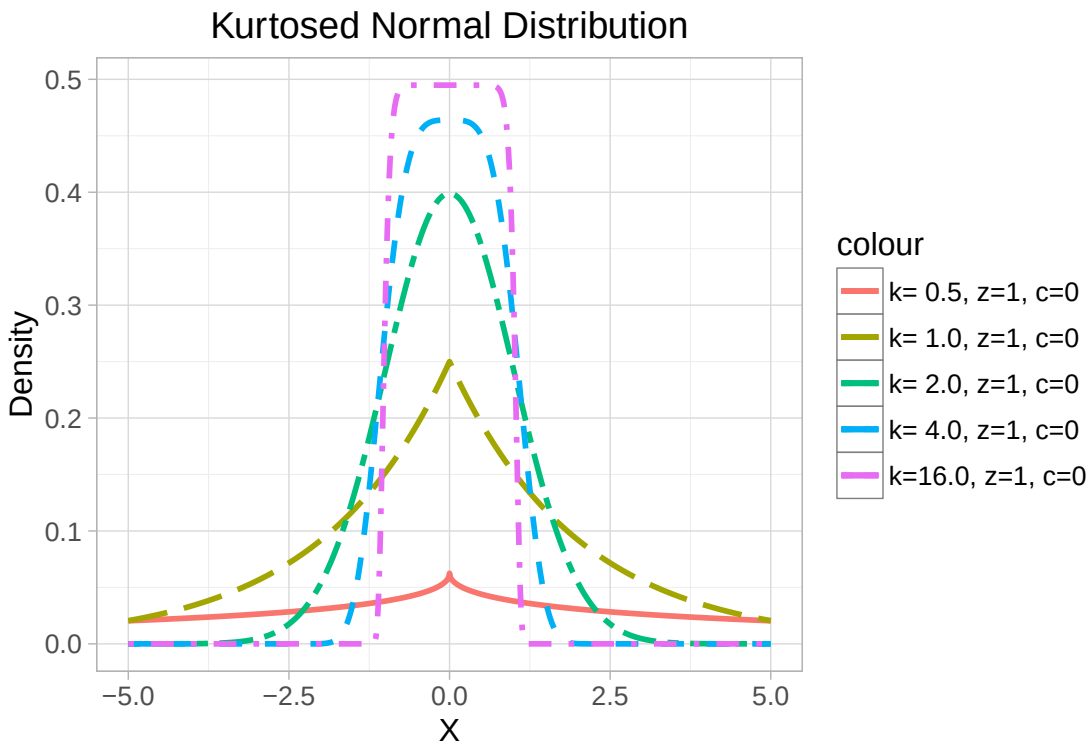

Figure 2.2: Kurtosed Normal Distribution with Excess-Kurtosis

2.6.3.2 Fulfilling the PDF Axioms

The axioms of a valid PDF have been mentioned in section 2.4.3 on page 14. Recapitulating, the probability density function $d(x)$ of a new (continuous) distribution variant must at minimum fulfill the axioms:

1. $f(x)$ is a continuous function.
2. $f(x) \geq 0$.
3. $\int_{-\infty}^{+\infty} f(x)dx \stackrel{!}{=} 1$

So we have to make sure, that the probability density function of the kurtosed normal distribution

$$f(x) = \frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right)$$

satisfies these axioms.

@ Axiom no. 1: $f(x)$ is a continuous function.

- $\exp(-u) = e^{-u}$ is a continuous function $\forall\, u \in \mathbb{R}$.
- $(\frac{|x-c|}{z})^k$ is a continuous function $\forall\, x,\, c,\, z,\, k \in \mathbb{R}$.
- The norming $\frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})}$ is a constant without any effect on continuity.

Thus, $f(x) = \frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right)$ is a continuous function.

@ Axiom no. 2: $f(x) \geq 0$.

- $\exp(-u) = e^{-u} \geq 0\, \forall\, u \in \mathbb{R}$
- $(\frac{|x-c|}{z})^k \geq 0\, \forall\, x,\, c,\, z,\, k \in \mathbb{R}$.
- The constant norming $\frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \geq 0$.

Thus, $f(x) = \frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right) \geq 0$.

@ Axiom no. 3: $\int_{-\infty}^{+\infty} f(x)dx \stackrel{!}{=} 1$

Given the latest generalization of the Gaussian integral[14]

$f^*(x) = \exp(-0.5\, (\frac{|x-c|}{z})^k)$
and
$\int_{-\infty}^{+\infty} f^*(x)\, dx = \frac{2\, z\, \Gamma(\frac{1}{k})}{k\, 0.5^{\frac{1}{k}}}$

$$\Rightarrow \int_{-\infty}^{+\infty} f(x)\, dx = \frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \cdot \frac{2\, z\, \Gamma(\frac{1}{k})}{k\, 0.5^{\frac{1}{k}}} = 1 \text{ for } k > 0.$$

Thus, $\int_{-\infty}^{+\infty} f(x)\, dx = \int_{-\infty}^{+\infty} \frac{k\, 0.5^{\frac{1}{k}}}{2\, z\, \Gamma(\frac{1}{k})} \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right) dx = 1 \text{ for } k > 0.$

2.7 S-Distribution - A Fully Generalized Normal Distribution *

The S-distribution has been derived from the kurtosed normal distribution. Basically, it is quite difficult to derive a fully generalized normal distribution from the kurtosed normal distribution. For instance, if we tried to set up a fully generalized normal distribution by multiplying the kurtosed normal distribution by a scalable logistic-function to redistribute the density to the left or the right side, we would experience, that this kind of generalized normal distribution -

[14] Review equation 2.6 on page 13.

in cases of extended skewness - would have local extremums in its probability density function (which are undesirable). Local extremums typically result from opposite slope of two function terms in the same amount at a certain location. As the kurtosed normal distribution covers a large range of power by its power coefficient k, it is not possible to introduce further power terms into the PDF without causing a power-conflict at some location resulting in a local extremum (but an asymmetric power modification is necessary to create skewness). Thus, reassembling the left branch and the right branch of two kurtosed normal distributions with different power coefficient k to a new compounded distribution is a very good option to generate a fully generalized (skew and kurtosed) normal distribution free of local extremums. As left and right branches with different power k are different steep, this automatically generates skewness. Fortunately, at $x - c = 0$ the core term

$$\exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^{k}\right)$$

always counts 1 regardless of the power coefficient k. Thus, the location $x - c = 0$ is naturally designated as the glue point, where to patch two kurtosed normal distributions together.

The norming constant [15] of the new assembled distributions also needs to be adapted. When assembling at $x - c = 0$ each branch has the same x-length . So, we can simply equal weight the two normings of each branch to get the appropriate norming for the composed distribution. We get the PDF formula:

$$f(x) = \begin{cases} \left(0.5 \cdot \frac{k\,a\,0.5^{\frac{1}{k\,a}}}{2\,z\,\Gamma(\frac{1}{k\,a})} + 0.5 \cdot \frac{\frac{k}{a}\,0.5^{\frac{a}{k}}}{2\,z\,\Gamma(\frac{a}{k})}\right) \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^{k\,a}\right) & for\, x - c < 0 \\ \left(0.5 \cdot \frac{k\,a\,0.5^{\frac{1}{k\,a}}}{2\,z\,\Gamma(\frac{1}{k\,a})} + 0.5 \cdot \frac{\frac{k}{a}\,0.5^{\frac{a}{k}}}{2\,z\,\Gamma(\frac{a}{k})}\right) \cdot \exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^{\frac{k}{a}}\right) & for\, x - c \geq 0 \end{cases} \tag{2.16}$$

Where

$f(x)$	Probability density function (PDF)
x	Random variable
c	Construction center (corresponding to normal distribution's μ)
$z > 0$	Construction deviation (corresponding to normal distribution's σ)
$k > 0$	Power coefficient
$0 < a < \infty$	Asymmetry coefficient
$\Gamma(\,)$	Gamma function

[15] Review section 2.4.4 on page 15.

$\exp(t)$ e^t

The constructed fully generalized normal distribution is shortly called S-distribution for the reason to have a short handy name easy to write and remember, which even could be distinguished from other future distributions of this kind, and especially for the reason to be able to state random error generation in error function notation. The construction parameter a is called the asymmetry coefficient, as it determines the distribution's skewness. When asymmetry parameter $a = 1$, the kurtosed normal distribution gets reproduced exactly. When $a > 1$, the distribution gets right skew, and vice versa when $a < 1$, the distribution gets left skew.

The inverse linking of the effective powers of both kurtosed normal distributions branches by the asymmetry parameter a is not very restrictive for skewness and kurtosis, when generating S-distributed random variables, because there is a lot of flexibility provided by the power coefficient k. The author hopes, that an efficient estimation technique for the S-distribution can be found easier, when there is a linking asymmetry parameter, instead of two independent power coefficients.

2.7.1 Density Graphs

Again, it is shown first, that the kurtosed normal distribution can exactly reproduce the normal distribution, which is one of its major features.

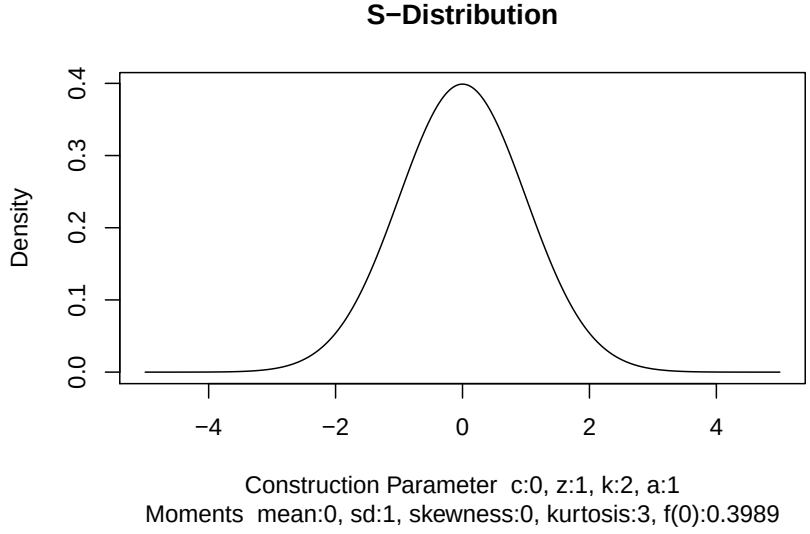

Figure 2.3: S-Distribution Reproducing the Normal Distribution

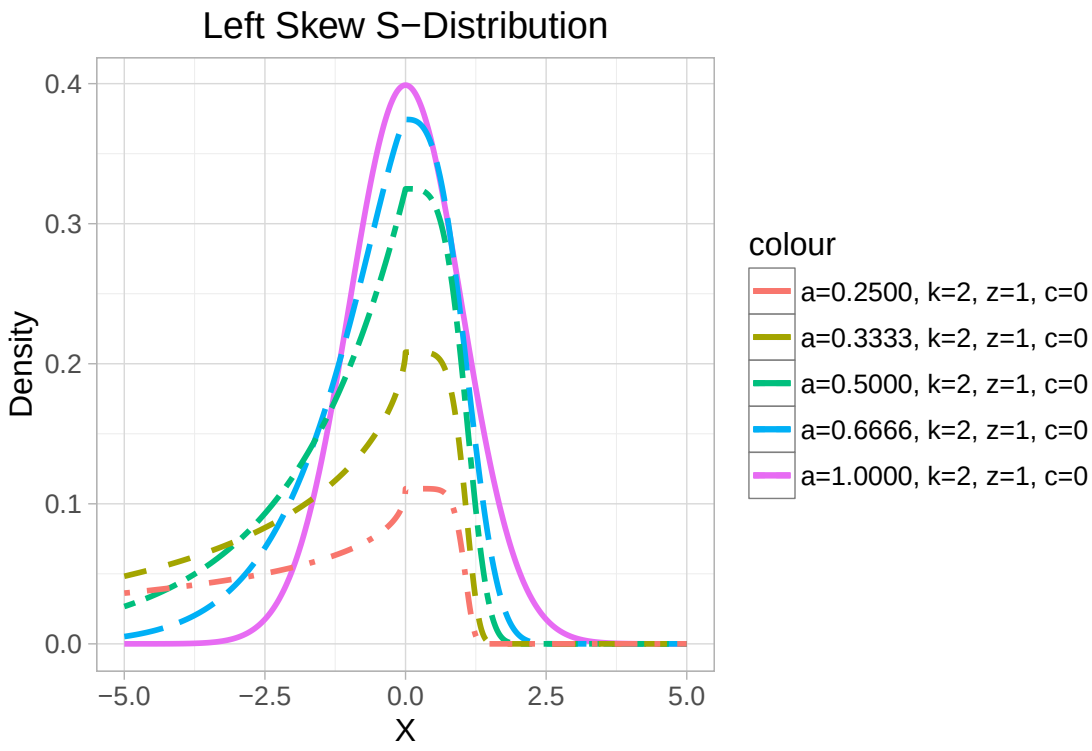

Figure 2.4: S-Distribution with Left Skewness

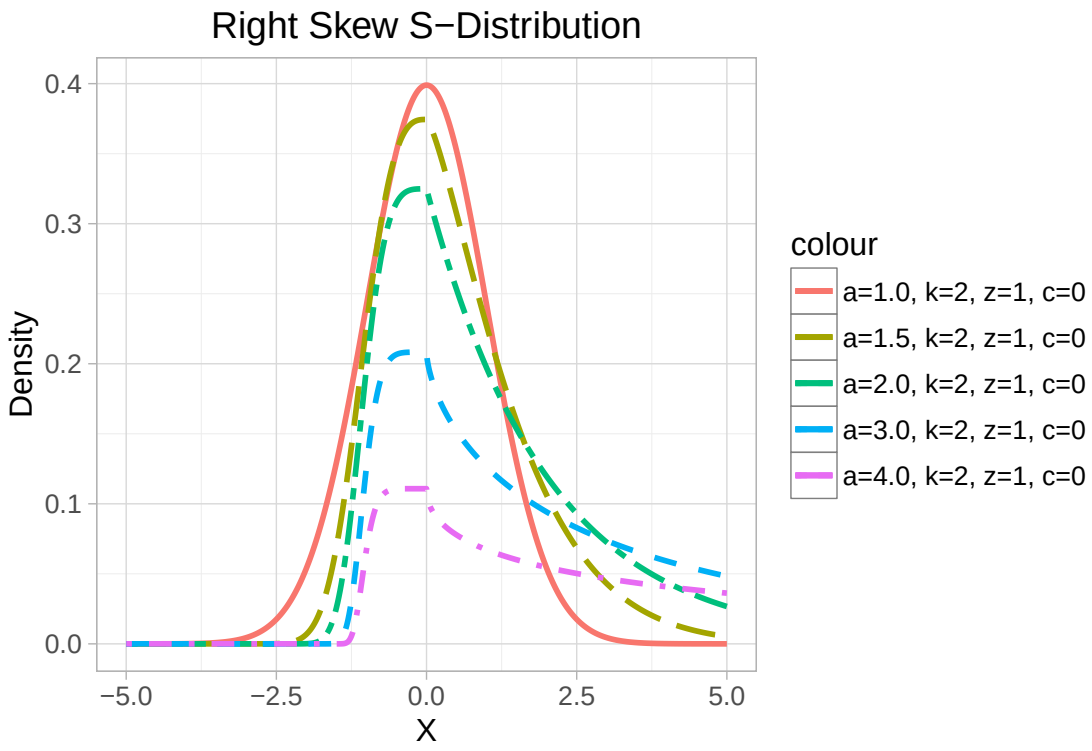

Figure 2.5: S-Distribution with Right Skewness

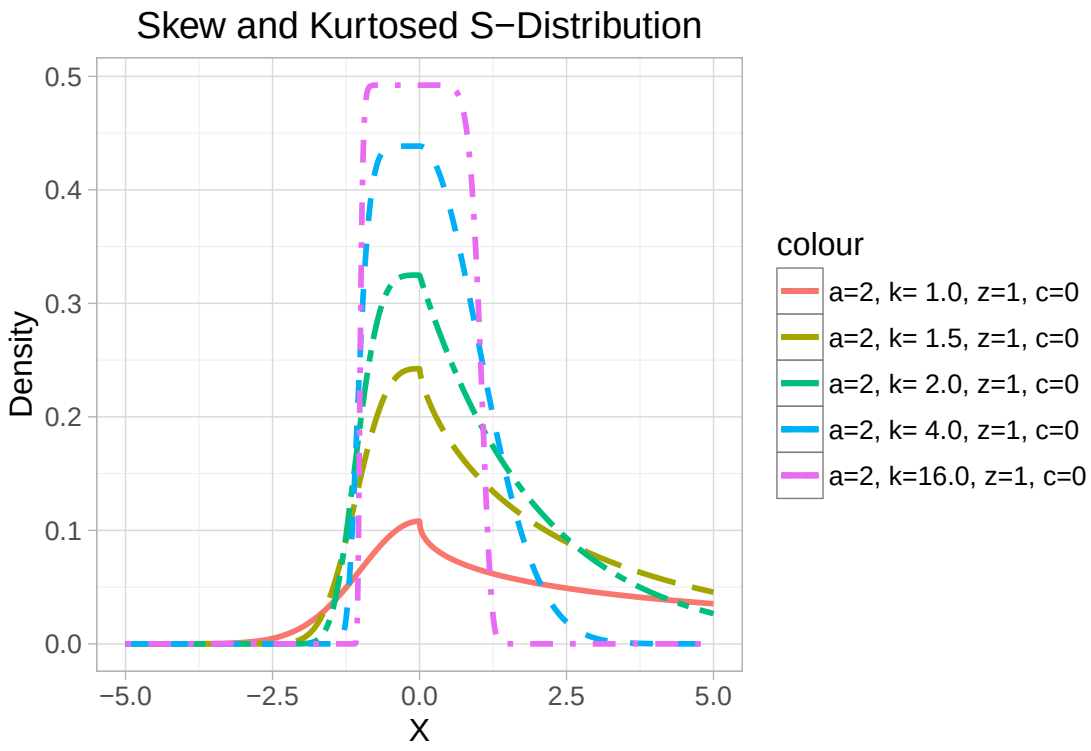

Figure 2.6: S-Distribution with Skewness and Alternative Power

Last, we have the skew and kurtosed S-distribution, more precisely the S-distribution with skewness and alternative power, as a skew S-distribution's kurtosis is already different from 3.

2.7.2 Fulfilling the PDF Axioms

The axioms of a valid PDF have been mentioned in section 2.4.3 on page 14. Recapitulating, the probability density function $d(x)$ of a new (continuous) distribution variant must at minimum fulfill the axioms:

1. $f(x)$ is a continuous function.
2. $f(x) \geq 0$.
3. $\int_{-\infty}^{+\infty} f(x)dx \stackrel{!}{=} 1$

So we have to make sure, that the probability density function of the kurtosed normal distribution satisfies these axioms.

@ Axiom no. 1: $f(x)$ is a continuous function.

The S-distribution is an assembled kurtosed normal distribution. At $x - c = 0$ the core term

$$\exp\left(-0.5 \cdot \left(\frac{|x-c|}{z}\right)^k\right)$$

counts 1 regardless of the value of the power coefficient k. Following, $f(x)$ is continuous at the glue location $x - c = 0$. Thus, if the kurtosed normal distribution fulfills the PDF axiom no. 1, which has been proven, the S-distribution fulfills the axiom no. 1, either.

@ Axiom no. 2: $f(x) \geq 0$.

The S-distribution is an assembled kurtosed normal distribution. Thus, if the kurtosed normal distribution fulfills the PDF axiom no. 2, which has been proven, the S-distribution fulfills the axiom no. 2, either.

@ Axiom no. 3: $\int_{-\infty}^{+\infty} f(x)dx \stackrel{!}{=} 1$

The S-distribution is an assembled kurtosed normal distribution. Each kurtosed normal distribution branch has the same x-length. Thus, we can simply equal weight the two normings to get the appropriate norming for the composed distribution. Thus, if the kurtosed normal distribution's PDF adds up to 1, which has been proven, the S-distribution's PDF adds up to 1, either.

2.8 R Programming Language

In this work the statistical R programming language is used for random number generation, generating test data, calculating statistical indicators and creating graphical figures. As needed, the authors also has coded mathematical operators, statistical indicators and the new fully generalized distribution together with a related sampling procedure in R. The author is pretty aware, that basically an introduction into R programming would increase the understanding of this work. But for the three heavy weighting reasons, that, first, R is the most popular statistical programming language, second, a half introduction into a programming language causes more new questions than it helps to understand and, last, a detailed introduction would be far out of the scope of this work, an introduction into the R programming is not provided here. In case, an imho fabulous introduction to the R programming language can be found in [Adler (2009)]. Anyway, the R code used in this work is mostly easy to understand, even without R-programming skills.

2.9 General Setup Program (Required for Every Program Listing)

- [PRG Listing]

2.10 Random Number Generation

2.10.1 Error Function Notation

In statistics the error function notation is used to define a random variable in mathematical formula language, that follows a certain theoretical distribution. E.g. the statistical error function

$$\varepsilon \sim N(\,0,\,4)$$

defines a normal distributed random variable ε. The first normal distribution parameter is the mean μ. In case of the normal distribution it is a frequent - bizarre - convention to quote the variance as second parameter, although there is no variance variable in the PDF of the normal distribution. However, some statistical authors reference a normal distribution with the *variance* 4, *not* standard deviation 4, by the expression $\varepsilon \sim N(\,0,\,4)$.

The defined stochastic variable ε can be used as any other variable now, for instance to state a formula for a variable y, that consists out of a deterministic and a stochastic part.

$$y = a + bx + \epsilon$$

As in a statistical model on y to determine the parameters a and b the stochastic variable ε would appear as residual error, a stochastic variable definition of the form

$$\varepsilon \sim N(\,0,\,4)$$

is also called error function.

For clarity regarding the new S-distribution, its error function notation is laid out now. There is no such bizarre convention as in normal distribution. For defining a random variable ε following S-Distribution using error function notation, simply its PDF construction parameters are referenced. E.g.

$$\varepsilon \sim S(c = 0,\, z = 2,\, k = 2,\, a = 1)$$

to define exactly the same normal distributed ε again, as above.

Where

$S(\,)$	S-distribution
c	Construction center (corresponding to normal distribution's μ)
z	Construction deviation (corresponding to normal distribution's σ)
k	Power coefficient
a	Asymmetry coefficient

For short statement, the construction parameter order c, z, k, a is made convention.

$$\varepsilon \sim S(0,\ 2,\ 2,\ 1)$$

$$\Leftrightarrow \varepsilon \sim S(c = 0, z = 2, k = 2, a = 1)$$

2.10.2 Random Number Generation Error

The process of creating a random number following a certain distribution is called random number generation. However, uniform random number generation algorithms and the transformation from uniform distribution to another distribution are far out of the scope of this work. But a random number generation error, namely the random variable's discrepancy from the desired distribution, has an impact on this work. Here, the random number generation error is assessed by the effective deviations from the theoretical distribution's true 4 central moments. The following program exemplary compares the 4 central moments of a random number generated by R's standard random number generator versus the true 4 central moments of the normal distribution (computed by numerical integration).

- [PRG Listing]

The program exemplary shows, that there is an error in the central moments of the random numbers created by R's standard random number generator.

2.10.3 Density Dump *

Obviously, random variables with significant random number generation errors are not appropriate to be shown as exemplary study cases. The test data for goodness of fit measuring cases must be generated very precisely. Especially the random part's number generation error, namely the discrepancy from the desired theoretic distribution, has to be this small, that it can be neglected. As shown in the previous section, R's standard random number generator is not this precise. Therefore, the author has developed a new random number generation procedure, called density dump.

The density dump process works as follows. Initially the absolute theoretic frequency is calculated by multiplying the PDF density of each x by M. Second, the absolute theoretic frequency is cumulated. Third, the algorithm allocates the frequency - for the left and the right side of the distribution separately - to an x-value. For both sides the allocation is done by iteratively moving along the cumulated theoretic frequency from the outer to the center. In each iteration step, the rounded frequency is dumped and deducted from the cumulated frequency ahead. Effectively, this is only a redistribution of the decimal frequency parts. The decimal frequency parts are carried in the cumulated frequency, until they accumulate to 0.5 or above. Forth, x-data is created by drawing

each x-value times the dumped frequency. Last but not least, the x-data is randomly permutated. Over all the density dump procedure can be interpreted as creating a pseudo population by multiplying the PDF with a larger number and then drawing a full random sample without replacement out of the pseudo population.

The following program again exemplary computes the 4 central moments of a random number following the standard normal distribution, respectively following the normal distribution, however generated by density dump this time.

- [PRG Listing]

Random variables generated by density dump are highly precise, as far the number base (the set of numbers where the random number is drawn from) has an appropriate range, the number base step length is granular and the multiplier M is big enough. In cases of heavy skewness, it can easily happen, that some proportion of the density lies outside the number base range. That would theoretically lead to the loss of some x-data. However, this error is easily detected, as R does not process a vector shorter than expected any further. In the setup, random number base step length equal 0.001 and the density multiplier $M = 1000$, the author has repeatedly observed, that the precision for the third and the forth central moment is at least +/- 1 digit on the 4th decimal. Especially the moment errors are not of stochastic nature, but deterministic As the goodness of fit statistics R-square and Area Approximated have lower powers than the third and fourth central moments, namely the power 2 and the power 1, the author assumes, that the two goodness of fit metrics in the same setup have at least the same precision (conservative assumption).

Summarizing, density dumps are very feasible to generate highly precise data for exemplary statistical study cases.

Chapter 3

Normed Universal Goodness of Fit Metrics

3.1 Goodness of Fit (GOF)

Defining, a goodness of fit metric measures, how well a set of values created by a statistical model fits a set of observations. Therefor, goodness of fit metrics evaluate the discrepancies between the observed values and the values expected by the model in consideration; typically the evaluation is condensed to a single statistical figure. There are two use cases of goodness of fit. In the first case one wants to know the prediction-quality of a prediction model. E.g. a regression has made an estimation. Then, to evaluate the quality of the estimation, the goodness of fit of the estimated values is calculated. For completeness only, in the second case, one wants to know, by which degree a set of observations follows a certain theoretical distribution. E.g. it is assumed, that a set of observations is normal distributed. Then, the frequency of each outcome value is compared to the frequency expected by the statistical distribution. In this work, we only deal with the first case, goodness of fit for prediction models.

3.1.1 Universal Goodness of Fit

Defining, a *universal* goodness of fit statistic has the capability to fairly evaluate the prediction quality of predictions by statistical models of different kinds. A universal goodness of fit measurement can either be hindered by software, especially when the goodness of fit metric is exclusively bundled with the statistical model. E.g. a GLM regression only offers maximum likelihood goodness of fit measurement and an alternative neural network only offers Akaike information criterion. Or the universal goodness of fit measurement can also be hindered by theoretical reasons. E.g. the maximum likelihood goodness of fit calculated for an GLM regression is based on some distribution assumption, that cannot be used for a fair evaluation of the predictions by a neural network model.

3.1.2 Normed Goodness of Fit

Defining, a figure is called *normed*, when one or more certain points of its range are linked to another objective fact. E.g. in physics the temperature of 0° Kelvin references the absolute minimum temperature. In the context of goodness of fit, a normed goodness of fit statistic is primarily expected to have a defined maximum fit value, typically a maximum of 1 (100%), indicating that the estimator is fully equivalent to the observed target variable, respectively that the target variable is fully explained by the estimator. Secondary, a normed goodness of fit statistic usually also has a defined neutral value. Normed goodness of fit metrics with a defined maximum value and a defined neutral value, do provide data-set independent stand-alone interpretation. In contrast unnormed goodness of fit metrics like mean squared error or root mean squared error only provide reasonable interpretation for different models on the same data set. E.g. a mean squared error of 0.1 may be a good result for one data-set, but a bad one for another data-set.

3.2 Known Normed Universal Goodness of Fit Metrics

3.2.1 Pearson Product-Moment Correlation

Although Pearson correlation is no official goodness of fit statistic, it is often used for this purpose in practice. There exist the following equivalent formulations of Pearson product-moment correlation R_P.

$$\mathrm{R}_P = \frac{Cov(x,\, y)}{s(x) \cdot s(y)}$$

$$\Leftrightarrow \mathrm{R}_P = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s(x) \cdot s(y)} \tag{3.1}$$

$$\Leftrightarrow \mathrm{R}_P = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s(x)} \cdot \frac{(y_i - \bar{y})}{s(y)}$$

Where

Cov	Covariance
$(x,\, y)$	A set of paired observations
x	An observed random variable
y	An observed random variable
$\bar{x}$	Mean of x
$s(x)$	Sample Standard Deviation of x

i	Index
n	Total Number

3.2.2 R^2

There are two versions of the well known R-square metric, the residual version and the explained version. The dominating residual R-square version is

$$R^2_{res} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.2}$$

The sum $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$is called the sum of squared residual deviations (SSR) and the sum $\sum_{i=1}^{n}(y_i - \bar{y})^2$ is called sum of squared total deviations (SST).

$$\Leftrightarrow R^2_{res} = 1 - \frac{SSR}{SST} \tag{3.3}$$

The explained R-square version is

$$R^2_{expl} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

The sum $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is referenced as sum of squared explained deviations (SSE)

$$\Leftrightarrow R^2_{expl} = \frac{SSE}{SST} \tag{3.5}$$

Where

y	Target Variable
$\hat{y}$	Estimator for y
$\bar{y}$	Mean of y
i	Index
n	Total Number
SSR	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
SST	$\sum_{i=1}^{n}(y_i - \bar{y})^2$
SSE	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

The sum of squared residual deviations $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the loss function of R-square residual version, while R-square explained version contains a to maximize objective function $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ instead of a loss function.[1]

The R-square statistic has specific properties. First, if we use the naive expected value estimation $\bar{y}$ as estimator $\hat{y}_i$, the estimator $\hat{y}_i$ and the mean $\bar{y}$ count the same. This thus the whole metric accounts 0, for both the explained and the residual version. Following, the whole metric (both versions) accounts 0 for the naive expected value estimation (mean estimation) $\bar{y}$. Thus, the R-square statistic has a normed neutral value, which is calibrated the way, that the R-square count of 0 is linked to the naive expected value estimation (mean estimation).

Second, as the name R-*square* already tells us, it is a squared goodness of fit metric. However, the power of two can also be interpreted as the multiplication of the deviations $\hat{y}_i - \bar{y}$ respectively $y_i - \hat{y}_i$ by themselves, as they were weights. Thus, R-square also is considered a goodness of fit statistic with self-weighted errors.

3.3 Goodness of Fit Insights

Beside Pearson product moment correlation/ R-square, there are no further *normed* and *universal* goodness of fit metrics, neither in statistical software packages nor in statistical theory, that possess "generally-accepted" status. Here, the term *universal* references the capability for inter-prediction-model-type comparisons. Instead, goodness of fit measuring currently is mainly provided as a control statistic of the optimization performed inside the prediction model. E.g. OLS regressions use R-square goodness of fit and generalized linear models typically provide GOF measuring by summed deviance.[2] Following, there is a shortage in normed *universal* goodness of fit metrics (that allow for inter-model-type comparisons). This shortage among others leads to some kind of normed GOF measuring paradox in robust regression. Although it can be shown, that robust regression models are more stable, especially less sensitive to influential outliers than OLS regression,[3] an OLS regression model is always evaluated best by the R-square GOF metric. Effectively, this is no real paradox, as it is well known, that OLS-regression and R-square are similarly constructed.

As shown in the previous section the loss function in R-square residual version is

$$L_{R^2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The OLS-regression loss function[4] is

[1] For loss function definition review section 2.2 on page 10.
[2] [Gill (2000)].
[3] [Andersen (2008)].
[4] Compare 6.2 on page 63

$$L_{OLS} = \sum_{i=1}^{n}(y_i - \beta X_i)$$

Where

$$\hat{y}_i = \beta X_i$$

$$\Leftrightarrow L_{OLS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

As shown, the loss functions of OLS-regression and R-square residual version are identical. *Thus, when evaluating an OLS estimation by R-square, a mathematical tautology is generated, that the OLS estimation is evaluated quite good mainly for the reason, it is measured exactly the same way again, it has been optimized (thus naturally must be quite good). This mathematical tautology particularly gains weight, when using R-square to compare OLS estimates to other estimates, that cannot gain this equal loss function advantage.*

As in R-square, any goodness of fit metric naturally must apply some definition of elementary errors and has to perform some kind of aggregation. Thus, a loss function must also be inherent in any goodness of fit metric, not only in optimization techniques.[5] This insight is essential. Summarizing, inter-model-type goodness of fit measuring shell be performed out-of-optimization-technique. Otherwise, the goodness of fit statistic will favor that estimation, that has been optimized by the most likely loss function.

Last, we are remembering, that the squared loss function used in R-square and OLS regression is considered to be outlier influenced, as large errors of outlying observations square up much more.[6]

3.4 Area Approximated (AA) *

Area Approximated has been designed as a normed goodness-of-fit metric, which

1. is universal, particularly suitable for inter-prediction-model-type comparison,
2. does not favor OLS regression estimates and
3. grants every error/ every observation the same influence/ weight.

3.4.1 Derivation

Area Approximated is derived by improving R-square residual version. R-square's loss function is squared loss

$$L_{R^2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

[5] For loss function review 2.2 on page 10.
[6] Compare 2.2 on page 10.

Obviously, the unbalanced error and observation influence in squared loss is caused by the self-weight with the error.[7] To gain balanced error measurement we want to unweight the squared loss function by drawing the square root.

$$L' = \sum_{i=1}^{n} \sqrt{(y_i - \hat{y}_i)^2}$$

$$\Leftrightarrow L' = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

We get the absolute loss function. From section 2.2 we already know, it is not very suitable for optimization. However, when goodness of fit needs to be performed, the optimization is already done. So, it is fine to use the absolute loss function for a new goodness of fit statistic. Especially, as we can expect more exact measurement of the mean error.

Further, the popular R-square goodness of fit statistic is calibrated this way, that the naive expected value prognosis - aka mean prognosis - is linked to 0.[8] So, we naturally want the same calibration for the Area Approximated metric, too. In R-square residual version this is applied by dividing the loss function by the norming term

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

So, to achieve an appropriate norming for our modified loss function L', we simply have to apply the same square root operation also on the R-square norming term N_{R^2}. We get

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

$$N' = \sum_{i=1}^{n} |y_i - \bar{y}|$$

Finally, as in R-square residual version, we deduct from one. Putting it all together, we have derived the new Area Approximated metric.

$$\text{Å} = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|} \tag{3.6}$$

Recapitulating, in Area Approximated the absolute residual deviations

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

are set into relation to the total absolute deviations from the mean

$$\sum_{i=1}^{n} |y_i - \bar{y}|$$

[7] For self-weight compare section 3.2.2 on page 37.
[8] Review section 3.2.2 on page 37.

. Further the whole fraction is deducted from 1. It is exactly the same norming logic as in R-square residual version to get a normed goodness of fit metric. However - in contrast to R-square - all elementary errors/ deviations remain unweighted. As the deviations in Area Approximated are not squared any more, the effective power of the whole formula is one. This means, Area Approximated is a linear goodness of fit statistic. The qualities of the new linear goodness of fit metric will get examined and compared in chapter 4.

3.4.2 Expressed In Mean Absolute Deviation

Adding $\frac{1}{n}$ on both sides of the fraction in formula 3.6 we can express Area Approximated in mean absolute deviation (MAD).

$$\text{Å} = 1 - \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|}{\frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \bar{y}|}$$

$$\Leftrightarrow \text{Å} = 1 - \frac{MAD(y,\, \hat{y})}{MAD(y,\, \bar{y})} \tag{3.7}$$

Where

$MAD(a,\, b) = \frac{1}{n} \cdot \sum_{i=1}^{n} |a_i - b_i|$

y	Target Variable
$\hat{y}$	Estimator for y
$\bar{y}$	Mean of y
i	Index
n	Total Number

3.4.3 Graphical Interpretation

In the following, a simple graphical interpretation for Area Approximated is provided - which also has been the reason its the name. In the following example the target and the

estimation are for ease provided by positive functions.

Let us assume a target given by the positive function

$$v(x) = 0.5x^2 - 5x$$

and an estimator given by the positive function

$$w(x) = 30x$$

to be goodness of fit measured in the interval [0, 100].

The more the area under the curve of the estimation $w(x)$

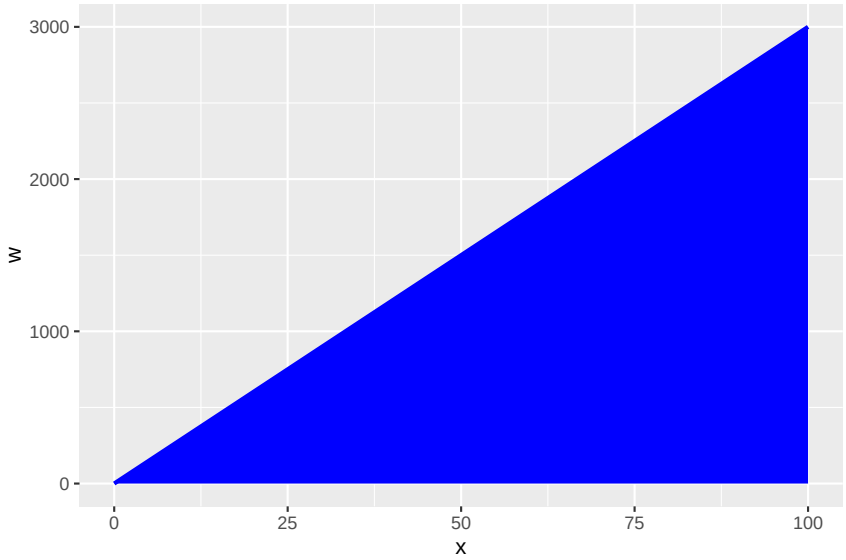

Figure 3.1: Graphical Interpretation of Area Approximated: Estimation can approximate the area under the curve of the target $v(x)$,

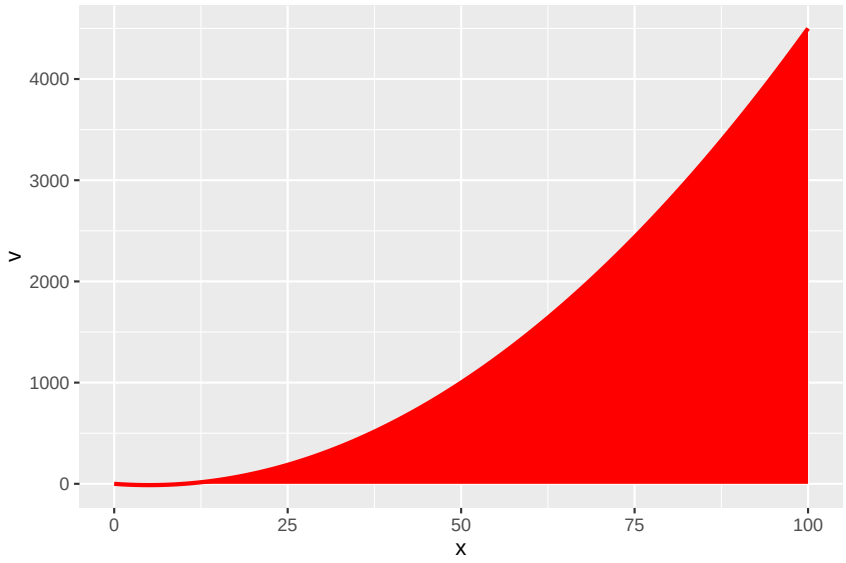

Figure 3.2: Graphical Interpretation of Area Approximated: Target

the lower gets the loss function $\sum_{i=1}^{n} |y_i - \hat{y}_i|$, which can be visualized by the area between the two curves.

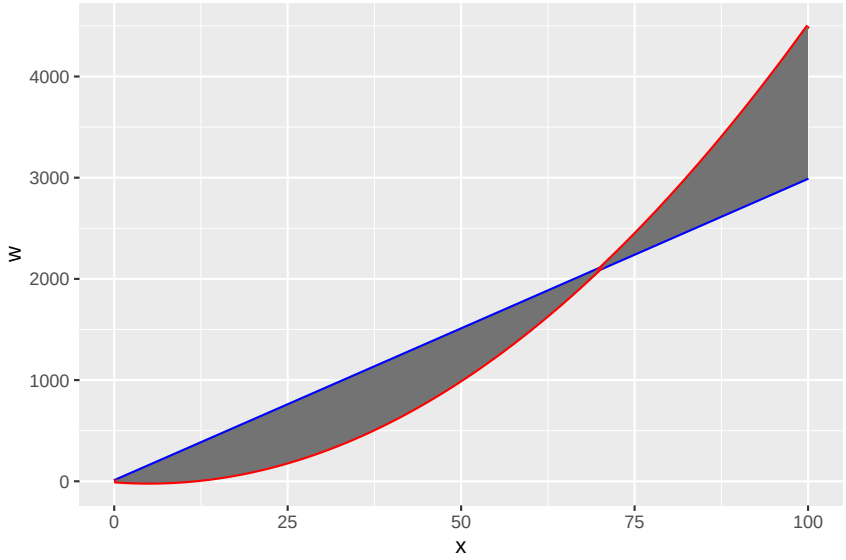

Figure 3.3: Graphical Interpretation of Area Approximated: Loss Function

For the purpose of calibration, the total amount of the loss function is expressed as proportion of the norming term $\sum_{i=1}^{n} |y_i - \bar{y}|$, that would result from an naive expected value/ mean estimation. The norming error sum can be visualized by the area between the target and the mean of the target.

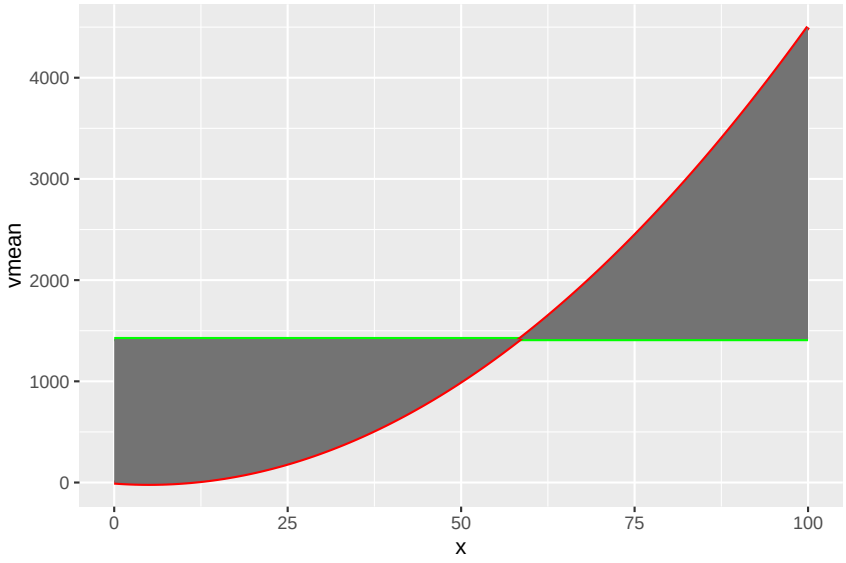

Figure 3.4: Graphical Interpretation of Area Approximated: Norming Term

Chapter 4

Quality Comparison

4.1 True Loss Function *

Quality criteria for objective goodness of fit comparison cannot be fairly developed, without considering the true loss function, if there is one. We have already researched in section 3.3 on page 38, that even a goodness of fit statistic contains a loss function. And we have also discovered, that a fair loss function should be independent from the optimization technique used inside the prediction model.

But of which power is a true objective loss function for universal goodness of fit measuring? The fully generalized S-distribution makes up a unlimited set of possible powers, even the power of the right side of the distribution is not necessarily the same as in the left side any more. Thus, there is an unlimited set of possible powers, that residual errors might have. Why should the power of 2 be the true power for independent goodness of fit measuring? Simply for the reason, that squared GOF-measuring is easy, as two variances sum up, when two random variables are added? Of course not for the reason, that squared estimators also have good (squared) variance properties![1]

From the author's point of view, we have to look for a true neutral loss function. But where from could a neutral loss function neutrally be derived? The most neutral way to to derive it, is to look up the data scale definition. Interval scale is defined equal distance.[2] Precisely, equal distance means equal one-dimensional euclidean distance. One-dimensional euclidean distance is defined the absolute value of the difference, which is equivalent to absolute deviation.[3] Following, the elementary error definition regarding true neutral goodness of fit must be absolute deviation. Thus, the true neutral loss function of an one-dimensional interval scaled error vector is absolute loss (the sum of the absolute deviations).

[1] The fact, that squared estimators automatically have good variance properties is considered a mathematical tautology. Review section 3.3 on page 38.

[2] Review section 2.1 on page 10.

[3] Review section 2.5.2.2 on page 20.

Would this be a contradiction with the Gauss-Markov theorem? The Gauss-Markov theorem states, given linearity and further given independent and variance homogeneous errors, that the squared linear estimator is the most efficient linear estimator, namely the one producing linear parameter estimates with the least variance. Effectively, this is not a contradiction. First, the theorem only proposes about estimator's parameter quality not about the overall goodness of fit. Second, the theorem only states about most efficient estimator regarding to the squared variance metric. Measuring the squared loss function parameter estimates by squared variance, will most probably be the mentioned goodness of fit tautology again.[4]

Last, the authors needs to limit his statement. The found absolute loss function is only stated to be the true loss function, as far the equal distance property of the target variable is intact. E.g. it is common in finance to transform price series to (percentage) return series and treat these return series as interval scale. However, a decrease of - 50% return followed by an increase of +50% return does not regain 100% (0.5 * 1.5 = 0.75). The author does not state about cases like these, where the equal distance property of interval scale is not met any more.

4.2 Quality Criteria

4.2.1 Consistence

If a goodness of fit statistic has a maximum value, it should count lower than the maximum, when the estimation is not fully equivalent to the target variable.

4.2.2 Calibration

We are testing, if the goodness of fit statistic is properly calibrated in the aspect of norming.[5] For the reason R-square is the predominant goodness of fit metric for interval scale data theses days,[6] we are testing, if a goodness of fit statistic also has the two R-square calibration properties.[7] Especially, we are examining whether the goodness of fit statistic accounts 1 (100%) for an estimation, that is fully equivalent to and the target variable. Further, we are examining whether the goodness of fit metric accounts 0 (100%) for a naive expected value estimation (mean estimation).

4.2.3 Linearity

A goodness of fit statistic is considered linear, if it is a linear transformation of the true loss function of interval scale, which is absolute loss. To test it,

[4] Review section 3.3 on page 38.

[5] Review section 3.1.2 on page 36.

[6] Compare www.google.com: "goodness of fit interval scale" and scholar.google.com: "goodness of fit interval scale.

[7] Review section 3.2.2 on page 37.

the delta of the goodness of fit metric is divided by the delta of absolute loss function.

4.3 Study Cases

The study cases are designed this way, that they basically represent errors, that might occur in estimations of prediction models. Deriving from the quality criteria possible types of prediction errors have been categorized and important error categories have been made a test series out of 4 cases. As the goodness of fit statistics are tested for universal - prediction model independent - goodness of fit measuring capability, also error types have been included, that cannot be produced, when familiar loss functions would be used for optimization (e.g. squared loss function or absolute loss function). Typically, the magnitude of the particular error is elevated from case to case to detect, how the goodness of fit metrics evaluate this kind of error.

4.3.1 Calibration Series

Case	Target y	Estimator y_E
1	$y = x$	$y_E = y$
2	$y = x$	$y_E = \frac{1}{n} \sum_{i=1}^{n} y_i$
3	$y = x^2$	$y_E = \frac{1}{n} \sum_{i=1}^{n} y_i$
4	$y = x^{0.5}$	$y_E = \frac{1}{n} \sum_{i=1}^{n} y_i$

- [PRG Listing]

4.3.2 Bias Series

Case	Target y	Estimator y_E
5	$y = x$	$y_E = -1 + x$
6	$y = x$	$y_E = -2 + x$
7	$y = x$	$y_E = -3 + x$
8	$y = x$	$y_E = -4 + x$

- [PRG Listing]

4.3.3 Deviating Slope Series

Case	Target y	Estimator y_E
9	$y = x$	$y_E = 1 + 0.8x$
10	$y = x$	$y_E = 2 + 0.6x$
11	$y = x$	$y_E = 3 + 0.4x$
12	$y = x$	$y_E = 4 + 0.2x$

- [PRG Listing]

4.3.4 Normal Error Series

Case	Target y	Estimator y_E	Distribution of Random Variable u
13	$y = x$	$y_E = x + 1u$	$u \sim N(\,0,\,1)$
14	$y = x$	$y_E = x + 2u$	$u \sim N(\,0,\,1)$
15	$y = x$	$y_E = x + 3u$	$u \sim N(\,0,\,1)$
16	$y = x$	$y_E = x + 4u$	$u \sim N(\,0,\,1)$

Where

$u \sim$ Random Variable u following a distribution

$N(\,\mu,\,\sigma)$ Normal distribution with Mean $\mu$and Standard Deviation σ

- [PRG Listing]

4.3.5 Kurtosed Error Series (incl. Uniform)

Case	Target y	Estimator y_E	distribution of Random Variable u
17	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0,\,1/2.825,\,1,\,1)$
18	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0,\,1,\,2,\,1) \;\Leftrightarrow u \sim N(\,0,\,1)$
19	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0,\,1/0.6909,\,4,\,1)$
20	$y = x$	$y_E = x + 1u$	$u \sim U\left(-\frac{1}{\sqrt{1/3}},\,\frac{1}{\sqrt{1/3}}\right)$

Where

$u \sim$ Random Variable u following a distribution

$N(\,\mu,\,\sigma)$ Normal distribution with Mean $\mu$and Standard Deviation σ

$S(c, z, k, a)$ S-distribution with Construction Center c , Construction Deviation z , Power Coefficient k and Asymmetry Coefficient a

$U(\,\alpha,\,\beta)$ Uniform distribution with Lower Limit α and Upper Limit β

- [PRG Listing]

4.3.6 Skew Error Series

Case	Target y	Estimator y_E	distribution of Random Variable u
21	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0.1966075,\,1/1.06878,\,2,\,1.25)$
22	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0.3473039,\,1/1.254319,\,2,\,1.50)$
23	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0.4588214,\,1/1.565017,\,2,\,1.75)$
24	$y = x$	$y_E = x + 1u$	$u \sim S\,(\,0.5370607,\,1/2.040715,\,2,\,2)$

Where

$u \sim$ Random Variable u following a distribution

$S(c, z, k, a)$ S-distribution with Construction Center c , Construction Deviation z , Power Coefficient k and Asymmetry Coefficient a

- [PRG Listing]

4.4 Case Study Results

	Case	1	2	3	4
		X, X	X, Mean(X)	X^2, Mean(X^2)	$X^{0.5}$, Mean($X^{0.5}$)
Goodness of Fit	Å	1.000	0.000	0.000	0.000
	R^2_{res}	1.000	0.000	0.000	0.000
	R^2_{expl}	1.000	0.000	0.000	0.000
	R_p	1.000	NA	NA	NA

Table 4.1: Calibration Series, Results

All measures pass the calibration probe by the cases 1, 2, 3, and 4, with the only exception of Pearson correlation, which cannot compute for the cases 2, 3 and 4.

	Case	5	6	7	8
	bias	-1.0	-2.0	-3.0	-4.0
Goodness of Fit	Å	0.600	0.200	-0.200	-0.600
	R^2_{res}	0.880	0.520	-0.080	-0.920
	R^2_{expl}	1.120	1.480	2.080	2.920
	R_p	1.000	1.000	1.000	1.000

Table 4.2: Bias Series, Results

In all cases R-square explained version and Pearson correlation are not consistent, as they evaluate 1 and above for estimations, that are not fully equal to the target. We can observe, that Area Approximated always decreases by 0.4 for any decrease in bias by 1. Controversially, R-square residual version decreases by a different amount in each case.

	Case	9	10	11	12
	slope	0.8	0.6	0.4	0.2
Goodness of Fit	Å	0.800	0.600	0.400	0.200
	R^2_{res}	0.960	0.840	0.640	0.360
	R^2_{expl}	0.640	0.360	0.160	0.040
	R_p	1.000	1.000	1.000	1.000

Table 4.3: Deviating Slope Series, Results

Pearson correlation is not consistent in all cases, as it evaluates 1 for estimations, that are not fully equal to the target. We can observe, that Area Approximated always deducts 0.2 for any decrease in slope by 0.2. R-square residual version and R-square explained version evaluate heterogeneous in each case. Interestingly, the arithmetic mean of both can recompute Area Approximated in this series.

Interim wise summarizing, the GOF statistics Area Approximated and R-square residual version have been without any obvious errors this far. But they measure differently. Thus, we want to perform a linearity analysis on them. Although, for Area Approximated it has already been shown in section 3.4, that it is linear. So, this will only be a redundant linearity check in case of Area Approximated.

In section 4.1 we have derived, that the true loss function for interval scale is absolute loss (AL).

$$AL = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

The derivation of the true loss function for interval data scale has revealed the same linear loss function as used in Area Approximated (which however has been deducted by unweighting R-square's loss function). As absolute loss has the power of one, we know for sure, it is linear. Thus, it is perfectly feasible to analyze for linearity. To perform this, we are dividing the change of the GOF statistic by the change in absolute loss.

Case	1	5	6	7	8
AL	0	1000001	2000002	3000003	4000004
ΔAL		1000001	1000001	1000001	1000001
Å	1.000	0.600	0.200	-0.200	-0.600
ΔÅ		-0.400	-0.400	-0.400	-0.400
ΔÅ/ΔAL		-4.00E-07	-4.00E-07	-4.00E-07	-4.00E-07
R^2_{res}	1.000	0.880	0.520	-0.080	-0.920
ΔR^2_{res}		-0.120	-0.360	-0.600	-0.840
$\Delta R^2_{res}/\Delta AL$		-1.20E-07	-3.60E-07	-6.00E-07	-8.40E-07

Table 4.4: Bias Series, Linearity Analysis

Case	1	9	10	11	12
AL	0	500001	1000002	1500003	2000004
ΔAL		500001	500001	500001	500001
Å	1.000	0.800	0.600	0.400	0.200
ΔÅ		-0.200	-0.200	-0.200	-0.200
ΔÅ/ΔAL		-4.00E-07	-4.00E-07	-4.00E-07	-4.00E-07
R^2_{res}	1.000	0.960	0.840	0.640	0.360
ΔR^2_{res}		-0.040	-0.120	-0.200	-0.280
$\Delta R^2_{res}/\Delta AL$		-8.00E-08	-2.40E-07	-4.00E-07	-5.60E-07

Table 4.5: Deviating Slope Series, Linearity Analysis

In the both series, the bias series and as well the deviating slope series, the R-square residual version does *not* linearly account goodness of fit. Not surprisingly - as the loss functions are identical - Area Approximated does evaluate linearly. Thus, as we have proven Area Approximated's linearity a second and a third time now, we now can more comfortably compare versus Area Approximated directly to analyze for linearity.

Using this information gain, we graph R-square residual version over Area Approximated to show its deformation from linear goodness of fit measuring. For best comparative visualization we include Area Approximated's identity graph in the figure.

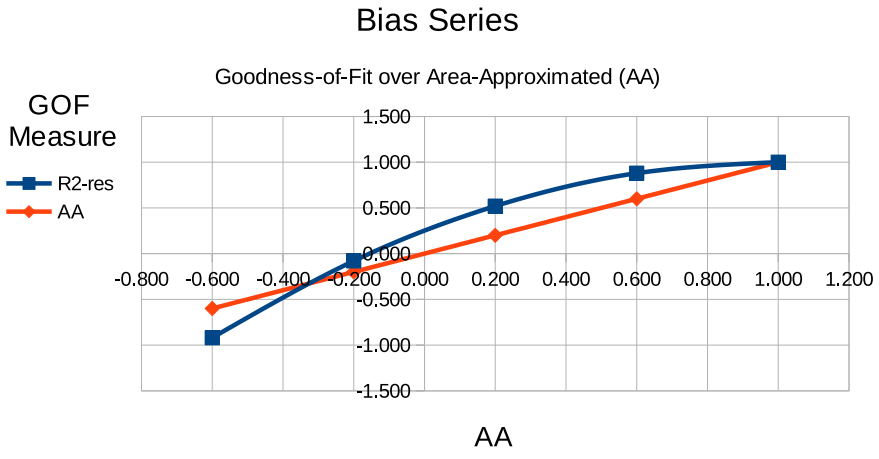

Figure 4.1: Bias Series, R-Square over Area Approximated

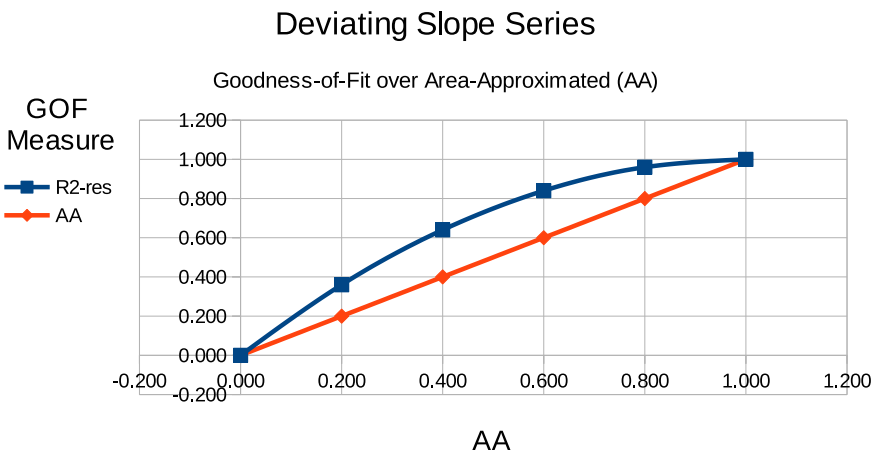

Figure 4.2: Deviating Slope Series, R-Square over Area Approximated

We can see, that R-square (residual version), with the exceptions of perfect match and mean prediction, respectively some negative bias values in the bias series, always accounts more goodness of fit than Area Approximated. Applying interpolation, R-square (residual version) may be most inflated in relation to linear Area Approximated in an Area Approximated range around 0.5. This would roughly correspond to R-square (residual version) in the range around $1/\sqrt{2}$.

We are turning over to the normal error series, now.

	Case	13	14	15	16
	sd	1.0	2.0	3.0	4.0
Goodness of Fit	Å	0.693	0.449	0.290	0.195
	R^2_{res}	0.893	0.675	0.480	0.342
	R^2_{expl}	0.893	0.676	0.481	0.343
	R_p	0.945	0.822	0.693	0.585

Table 4.6: Normal Error Series, Results

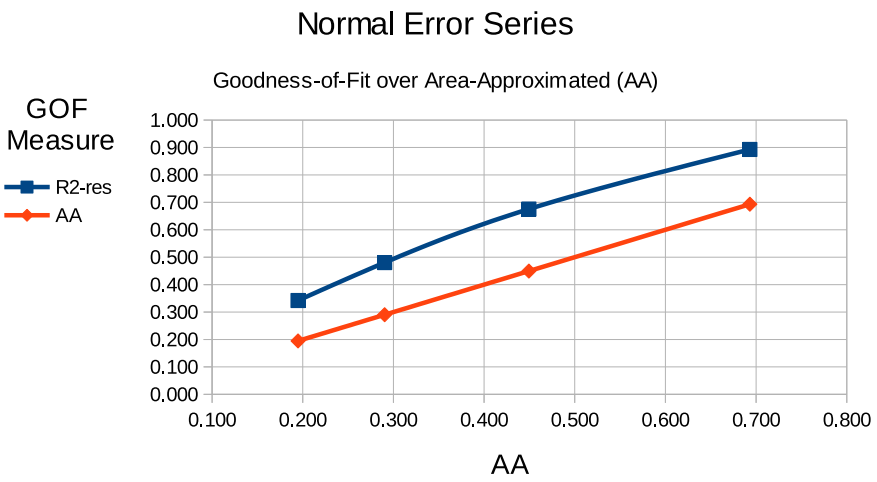

Figure 4.3: Normal Error Series, R-Square over Area Approximated

Basically, the normal error series results the same as the two series before. R-square (residual version) counts more goodness of fit than Area Approximated. Applying some over the thumb interpolation, R-square (residual version) is most inflated, when Area Approximated ranges around 0.5. This roughly correspond to levels of R-square (residual version) in the range around $1/\sqrt{2}$.

Last, but not least the kurtosed error series and the skew error series.

	Case	17	18	19	20
	k	1.0	2.0	4.0	inf
	Kurtosis	5.998	3.000	2.188	1.801
Goodness of Fit	Å	0.728	0.693	0.676	0.667
	R^2_{res}	0.893	0.893	0.893	0.893
	R^2_{expl}	0.892	0.892	0.893	0.892
	R_p	0.945	0.945	0.945	0.945

Table 4.7: Kurtosed Error Series, Results

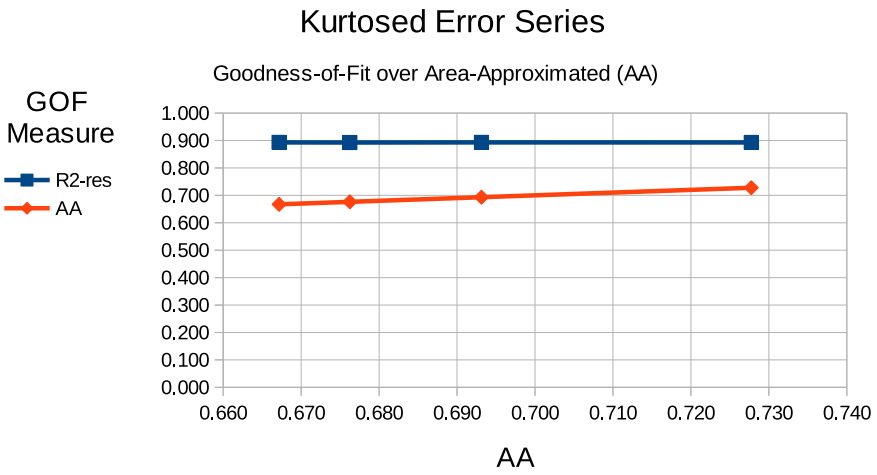

Figure 4.4: Kurtosed Error Series, R-Square over Area Approximated

	Case	21	22	23	24
	a	1.25	1.50	1.75	2.00
	Skewness	0.563	1.051	1.489	1.885
Goodness of Fit	Å	0.696	0.702	0.709	0.717
	R^2_{res}	0.893	0.893	0.893	0.893
	R^2_{expl}	0.893	0.893	0.892	0.892
	R_p	0.945	0.945	0.945	0.945

Table 4.8: Skew Error Series, Results

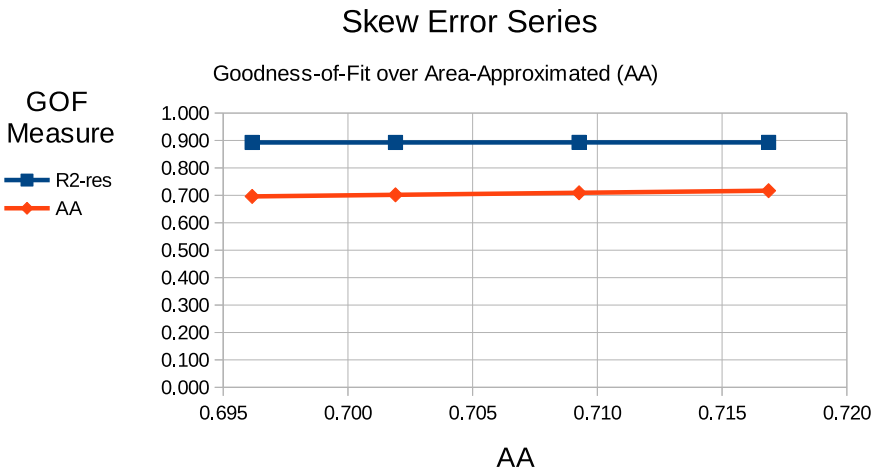

Figure 4.5: Skew Error Series, R-Square over Area Approximated

To evaluate the results of the kurtosed error series and the skew error series, we have to keep their construction in mind. Alternative levels of kurtosis and skewness in the error have been achieved by altering the power coefficient k, respectively the asymmetry coefficient a, of the S-distribution. However, changes in theses parameters also effect the total deviation of the distribution. To isolate the net effects of kurtosis and skewness respectively, the deviation effects have been recompensated this way, that its standard deviations count 1 again. Thus, the spread of goodness of fit in theses cases is low, especially R-square is constant.

Summarizing from the cases, the effect of the error's 3rd and 4th moments, namely the error's shape away from its total deviation, on goodness of fit measuring is pretty low compared to the huge total deviation effect in the three series before.This is quite surprising, as most statisticians think, R-square would rather prefer normal distributed errors to errors distributed another way. But the opposite is true. From a linear point of view R-square has only a relatively little preference for low kurtosis (in skewed S-distributions the part with lower effective k overcompensates the other part). But R-square has a huge preference for high total error deviation in relation to the deviation inherent in the deterministic part of the test data. All this evaluated from a linear point of view using Area Approximated.

Chapter 5

Results and Prospects

5.1 Area Approximated

1. Over all there have no weaknesses been found in the new Area Approximated goodness of fit statistic during this work.

2. Area Approximated provides a truly objective goodness of fit measuring. The search for an improved goodness of fit metric was started with elimination of influential GOF accounting by the squared loss function in R-square. The downgrading of the power of the elementary errors to the exponent of one has led to exactly the same absolute loss function as the true loss function, that has been derived from the equal distance property in metric scales of measurement.

3. Transparent case studies support the thesis, too, that the new Area Approximated metric evaluates goodness of fit much more exactly than R-square.

4. Like R-square Area Approximated is an intuitive goodness of fit metric. As R-square, it has been calibrated this way, that a perfect match is linked to the GOF-count of 1 and the evaluation of an expected value/ mean prediction counts 0.

5. In contrast to R-square Area Approximated is a linear goodness of fit statistic. The combination of linearity with the intuitive calibration makes AA even more intuitive than R-square. An AA count of X is equivalent to an explanation degree in deviations from the mean of $X * 100$ percent, while in R-square a GOF count of X expresses the proportion of explained *squared* deviations from the mean.

6. Concluding from the case studies Area Approximated does not have problems with unconventional error distributions (e.g. bias series, slope series), where R-square had problems..

7. Area Approximated has not shown any consistence errors.

8. Thus, Area Approximated seems to be a sophisticated, normed and universal goodness of fit statistic. Although not complicated designed it seems to be a big information gain to R-square.

9. However, independent research is required to validate the new Area Approximated GOF statistic and the concept of linear goodness of fit measuring fully.

5.2 R-Square

1. Briefly summarizing, this work delivers strong indication, that R-square generally performs largely positively inflated goodness of fit measuring, when accepting the author's finding, that the true objective loss function has the power of one, which is equal to demanding linear goodness of fit measuring. In the case studies, applying some interpolation the highest inflation of goodness of fit could roughly over the thumb be located at moderate goodness of fit levels ($AA \approx 0.5$, $R - Square \approx 1/\sqrt{2}$)

2. The author's scientific procedure to look a critical issue the up in the definition - here looking up the critical error power in the definition of interval data scale - is a probed and widely accepted logical method. Thus, the result of this work may have some weight. Anyway, this finding will be little surprising for robust regression researchers, remembering that the R-square metric is designed similar to the ordinary-least-squares regression. [1]

3. In R-square surprisingly not the distribution of the error has been the main problem, but the non-linear norming of the errors, especially the non-linear inflation of the total deviations by squaring.

4. R-square has problems with measuring unconventional error distributions (bias series, slope series). Luckily, these cases do not appear when using unbiased estimation techniques like OLS regression. But when using other prediction models like machine learning models (smaller) errors like these might occur, depending on their inner optimization technique.

5. The author believes, that R-square has just been the gold-standard for normed GOF measuring so long, because absolute deviations do not have this nice addition property, that (squared) Variance has. Thus, till today, there was no normed linear GOF statistic, which could provide a neutral linear evaluation of errors. Consequently, the new neutral linear insight into R-square deviation evaluation by Area Approximated instantly arises the question, why the squared error should be the true loss function.

[1] Compare section 6.2 on page 63.

6. As some statisticians have evaded from R-square to Pearson product-moment correlation, it has to be stated, that Pearson product-moment correlation in the case studies has shown even more measurement problems than R-square. For this reason, Pearson correlation has not been researched further.

7. Last but not least the author is very proud, it has been possible to reduce the uncovering of R-square's and Pearson correlation's measurement problems to these few transparent study cases, so that even non-statisticians can easily understand the weaknesses inherent in R-square and Pearson correlation as goodness of fit metrics.

5.3 S-Distribution

Although the S-distribution is not an explicit research subject of this work, the discovery of the first fully generalized normal distribution may be worth a few points regarding its *assumed* advantages.

1. On the first look a patch-work distribution does not appeal very elegant. But for the reason, that all other tried modifications of a kurtosed normal distribution have produced at least local extremums in the probability density function[2], which are undesirable, the assembling of two kurtosed normal distribution with different power coefficients seems to be a quite good choice. If there were so many other possibilities, there would be other fully generalized normal distributions already, but by now there is not even one other.

2. The S-distribution has a high flexibility in skewness and kurtosis. Especially it can produce skewness levels above 1. And it combines alternative kurtosis and alternative skewness at the same time. As a result it should basically be able to approximate any other continuous distribution, at least roughly, even including the uniform distribution.

3. Further, the S-distribution has full downward compatibility to the normal distribution. Currently, the normal distribution is the benchmark of continuous statistical distributions. Thus, when creating non-normal distributed random variables, a fluent transition possibility to the normal distribution is a big plus. In this context it is much more efficient to generate any desired uni-modal deformation from the normal distribution in a scalable degree out of one distribution (instead out of a wild bunch of incompatible piece-work distributions).

4. The distribution behavior of S-distribution is easy to understand, as the core frequency function is basically still the same as in normal distribution. During this work it was always easy to check, whether a sample

[2] Review section 2.7 on page 25.

had been a generated correctly or not, by looking up, if the new effective moments were reasonable with the altered construction parameter(s). When a surprising result came up during this work, it usually could be plausibilized easily. In some very tricky cases a full validation by a semi-automatic reproduction of the two corresponding halves of the kurtosed normal distribution was performed. Such plausibilizations had been a terrible penalty work with an early predecessor version of the S-distribution (a kurtosed normal distribution multiplied by a scalable logistic function).

5. In the end the S-distribution tool had been a little bit over-sophisticated for the goodness of fit research - which, however, is a result impossible to know before having it tested. But overall the S-distribution has been very helpful for stress-testing the Area Approximated and the R-square GOF-Measure and especially made this research much more efficient. You may imagine more study cases have been tested than presented in this work.

5.4 Prospects

Regarding the Area Approximated goodness of fit statistic the author expects, once some research and independent validation will have been performed and statisticians will have psychologically recovered from the shock having over decades trusted - from a linear point of view - largely positively inflated goodness of fit measuring by R-square, that the statistical community will constantly gain trust in Area Approximated and it will become the new gold-standard for universal - namely prediction model and optimization technique independent - goodness of fit measuring of interval scale data, till another more sophisticated universal goodness of fit statistic will take its place in future. However, the author does not believe, Area Approximated will be adopted by the whole statistical community at the same speed, but much more rapidly in areas, where there is a big need for linear goodness of fit measuring. Especially the author expects, that robust regression researchers will be one of those, who adopt it first.

Additionally, the author believes, that the implicit loss function of future goodness of fit statistics will be researched much more intensively and future goodness of fit metrics will be stress-tested much harder, before they will earn the same trust of R-square again. This will also apply to the new Area Approximated goodness of fit statistic, although it has been developed to cover this weakness up.

Last, the author hopes, that out-of-model goodness of fit measuring, respectively out-of-optimization-technique GOF measuring, will one day be applied as automatically as out-of-sample goodness of fit measuring for machine learning models.

Concerning S-distribution. Although its central moments can provided by a calculation program, the introduced S-distribution is - from a mathematical

point of view - still in the construction process, especially as the moments are not described by a closed formula, till now. However, the author expects, mathematicians will add the currently missing moment formulas soon. As there is no other fully generalized normal distribution so far, the author is quite sure, it will be used frequently to generate kurtosed and as well skew errors, especially when a fluent transition to a normal distributed error is required.

Bibliography

[Adler (2009)] J. Adler. "R in a Nutshell", O'Reilly Media, 2009.

[Andersen (2008)] R. Andersen. "Modern Methods for Robust Regression", Sage Publications, 2008.

[Eliason (1993)] S. R. Eliason. "Maximum Likelihood Estimation - Logic and Practice", Sage Publications, 1993.

[Eubank, Kupresanin (2012)] R. L. Eubank, A. Kupresanin. "Statistical Computing in C++ and R", CRC Press, 2012.

[Fahrmeir, Künstler et al. (2004)] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz. "Statistik - Der Weg zur Datenanalyse", 5th Ed., Springer-Verlag, 2004.

[Fox (1997)] J. Fox. "Applied Regression Analysis, Linear Models And Related Methods", Sage Publications, 1997.

[Gill (2000)] J. Gill. "Generalized Linear Models - A Unified Approach", Sage Publications, 2000.

[Herve, Valentin et al. (1999)] A. Herve, D. Valentin, B. Edelman. "Neural Networks", Sage Publications, 2008.

[Hosking, Wallis (1997)] J. R. M. Hosking, J. R. Wallis. "Regional Frequency Analysis: An Approach Based on L-Moments", Cambridge University Press, 1997.

[Kaas, Goovaerts et al. (2008)] R. Kaas, M. Goovaerts, J. Dhaene, M. Denuit. "Modern Actuarial Risk Theory", 2nd Ed., Springer, 2008.

[Klebanov, Rachev et al. (2009)] B. Klebanov, S. Rachev, F. J. Fabozzi. "Robust and Non-Robust Models in Statistics", Nova Scientific Publishers, 2009.

[Nadarajah (2005)] S. Nadarajah. "A Generalized Normal Distribution", in Journal of Applied Statistics, Vol. 32, No. 7 September, 2005.

[Spall (2003)] J. C. Spall. "Introduction to Stochastic Search and Optimization", John Wiley, 2003.

Chapter 6

Appendix

6.1 Moments Calculator

[PRG Listing]

6.2 OLS Regression Loss Function

The derivation of the loss function in ordinary least squares (OLS) regression has been adopted from the maximum likelihood derivation of the OLS estimator in [Eliason (1993)].

Given the error model

$$y_i = \beta \cdot X_i + \varepsilon_i$$

$$\Leftrightarrow \varepsilon_i = y_i - \beta \cdot X_i$$

Where

y	Target variable
β	Parameter vector
X	Predictor matrix
ε	Random variable

If the elements ε_i of ε are independent, the joint distribution of ε can be formulated as the product of marginal distributions.

$$\Lambda(\beta) = \prod_{i=1}^{n} f(\varepsilon_i, \beta)$$

Where

$\Lambda(\beta)$	likelihood function
$\prod$	Product operator
f	Probability density function

The parameter vector β can be found by maximizing $\Lambda(\beta)$. As maximizing a product results in considerable computational difficulties, instead the natural logarithm of the product is maximized to solve for β.

$$ln[\Lambda(\beta)] = \sum_{i=1}^{n} ln[f(\varepsilon_i, \beta)] \stackrel{!}{=} max$$

If ϵ is normal distributed

$$\varepsilon \sim N$$

Where

$\varepsilon \sim N$	Random variable ε follows a normal distribution

and the probability density function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

the to maximize log-likelihood $ln[\Lambda(\beta)]$ can be stated as

$$ln[\Lambda(\beta)] = \sum_{i=1}^{n} ln\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{y_i - \beta \cdot X_i}{\sigma}\right)^2\right)\right] \stackrel{!}{=} max$$

$$\Leftrightarrow -\sum_{i=1}^{n}(y_i - \beta \cdot X_i) \stackrel{!}{=} max$$

The simple inversion of the sign provides the OLS regression loss function

$$\Leftrightarrow \sum_{i=1}^{n}(y_i - \beta \cdot X_i) \stackrel{!}{=} min$$