Audio Engineering Society

Convention Paper

Presented at the 140th Convention
2016 June 4–7, Paris, France

Measurements to determine the ranking accuracy of perceptual models

Andy Pearce[1], Tim Brookes[1], Russell Mason[1], and Martin Dewhirst[1]

[1] *Institute of Sound Recording, University of Surrey, Guildford, Surrey, UK*

Correspondence should be addressed to Andy Pearce (a.pearce@surrey.ac.uk)

ABSTRACT

Linear regression is commonly used in the audio industry to create objective measurement models that predict subjective data. For any model development, the measure used to evaluate the accuracy of the prediction is important. The most common measures assume a linear relationship between the subjective data and the prediction, though in the early stages of model development this is not always the case. Measures based on rank ordering (such as Spearman's test), can alternatively be used. Spearman's test, however, does not consider the variance of the subjective data. This paper presents a method of incorporating the subjective variance into the Spearman's rank ordering test using Monte Carlo simulations, and shows how this can be beneficial in the development of predictive models.

1 Introduction

Regression modelling is a commonly used statistical method in the audio industry. It is often used to develop objective models that predict subjective ratings of an audio signal or audio device. Models such as the PEAQ, PESQ, and POLQA are all designed to predict the basic audio quality of an audio device or codec [1, 2, 3]. Models have also been created that predict perceptual factors such as punch, distraction, loudness, and spatial quality [4, 5, 6, 7].

Successful objective prediction models can reduce the need for costly and time-consuming listening tests. These models can also reveal the underlying objective parameters that affect perception, leading to a better understanding of the effects of altering objective parameters in the design of new audio products.

There are many statistical methods that can be used in the modelling process (such as linear regression, multi-linear regression, polynomial regression, etc.). Each of these methods has its own advantages and disadvantages, but this paper will focus on linear regression since this is the most common method. Linear regression also has the advantage of being mathematically simple, reducing the chance that the resulting model does not generalise well to new data.

In the design of a linear regression model, multiple objective metrics might be considered. Within these metrics, there may be parameters that can be set at different levels. For example, a model of perceived loudness may include various signal amplitude metrics with different parameters related to frequency range and integration time. In order to decide upon the most suitable objective metrics and the parameters to use, a candidate model can be created for each variation of the metrics and parameters. Depending on the number of candidate models, it may be impractical to visually inspect the performance of each. In this case, statistical measures of the prediction accuracy are often used.

The measures used to describe the performance of a model are called 'goodness-of-fit' measures. Many of these measures rely heavily on having a linear relationship between the subjective ratings and the model's output. According to these goodness-of-fit measures, a model outputting data that are nonlinearly related to the subjective ratings would not be a strong candidate; however, the model may fit the data well after a simple nonlinear transformation, such as taking the logarithm or square root of one or more of the underlying objective metrics [8].

One of the most common goodness-of-fit measures that does not rely on a linear relationship between the subjective and predicted data is the Spearman's rho [8, 9]. It may seem that assessing all candidate models using the Spearman's rho measure may remove the nonlinearity problem; however, the Spearman's rho measure does not take into consideration the variance of the subjective ratings. There are often situations where for multiple stimuli the mean subjective ratings are not statistically significantly different. Using these mean values to determine the ranks for a Spearman's rho test in these cases may be misleading, as this approach would imply an order in the subjective ratings which is not justified by the data. Consideration of the variance of the subjective data can avoid this situation.

This paper will explore one method of allowing for the variance of the subjective ratings, the Monte Carlo simulation method. The paper will also describe how this method could be applied for the selection of the most appropriate candidate model.

The paper will review the most common measures used to describe the goodness-of-fit of a linear regression model, and will discuss the advantages and disadvantages of each of these measures, in Section 2. The Monte Carlo method will then be introduced in Section 3, outlining how this could be used to alleviate problems mentioned in Section 2. A worked example that exemplifies the problems with the standard goodness-of-fit measures and shows how the Monte Carlo can be used to alleviate these will be presented in Section 4.

2 Goodness-of-fit measures

Goodness-of-fit measures are used to numerically evaluate the performance of a candidate model. This section will introduce several of the most common of these measures, will describe how they could be used for selecting the most appropriate candidate model and will outline any potential problems that may result from using only this measure to evaluate the suitability of a model.

Throughout this section the term "subjective data" refers to listener ratings of audio stimuli, i.e. the data to be modelled, where each "data point" is the mean of all listener ratings for a particular stimulus; "objective metrics" are the sets of data extracted directly from the audio stimuli in order to characterise these stimuli objectively; the objective metrics are then combined into a regression model which outputs the "predicted data". In a good model, the predicted data will closely match the subjective data.

2.1 Root mean square error

The Root Mean Square Error (RMSE) is a measure of the difference between the subjective data and the predicted data. This can be calculated as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (x_n - y_n)^2}{N}}, \tag{1}$$

where N is the total number of stimuli, and x_n and y_n are the n^{th} values of the observed and predicted datasets respectively [8].

The RMSE measure has two main limitations: 1) the measure assumes a linear, one-to-one relationship between the subjective data and the predicted data, and 2) the variance of the subjective data is not taken into consideration. If the subjective data and predicted data are not linearly related, this will result in high levels of RMSE of the model, indicating a poor fit. A nonlinear relationship can be seen by visual inspection of a residuals plot.

2.1.1 Residuals plots

A residual, r_n, is the difference between the n^{th} subjective data point, x_n, and the n^{th} predicted value, y_n. This can be calculated as:

$$r_n = x_n - y_n. \tag{2}$$

There is a residual for each data point. The RMSE is the root mean square of the residuals. A nonlinear relationship can be seen by plotting the residuals against the predicted data. If a pattern can be seen in the plot, there is most likely some form of nonlinearity between the subjective data and predicted values. If the plot of the residuals against the predicted values forms an 'n' or 'u' shaped distribution, this is a strong indicator of a simple nonlinear relationship which may be removed by transforming the objective metrics nonlinearly prior to regression modelling [8].

Unfortunately, determining a pattern in the residuals plot can be difficult without visual inspection of the plot. Even if a pattern can be discerned mathematically, it can be difficult to predict the type of transformation required in order to linearise the objective metrics. Visual inspection of all residual plots may not be practical or possible if there are a large number of candidate models.

2.2 Epsilon insensitive RMSE

The Epsilon-insensitive RMSE (RMSE*) is a modification of the RMSE measure which does take into consideration the variance of the subjective data. For RMSE* the error is only considered when the predicted value lies outside the 95% confidence intervals of the corresponding subjective data point. The error for any predicted value that does lie outside the 95% confidence interval of the subjective data is calculated from the closest 95% confidence interval.

The error, E_n, for the n^{th} predicted value can be expressed as:

$$E_n = \begin{cases} 0, & \text{if } I_L <= y_n <= I_H \\ \min\begin{cases}(x_n - I_L) - y_n, \\ (x_n - I_H) - y_n, \end{cases} & \text{otherwise.} \end{cases} \tag{3}$$

where x_n and y_n are the n^{th} subjective and predicted data points respectively, and I_L and I_H are the upper and lower bounds of the confidence intervals for the n^{th} subjective data point [5]. The RMSE* is then calculated as:

$$\text{RMSE*} = \sqrt{\frac{\sum_{n=1}^{N} E_n^2}{N}}. \tag{4}$$

Although, in allowing for the variance in this way, the RMSE* is better than the RMSE for the assessment of subjective data, this measure still relies on the subjective and predicted data being linearly related. If there is a nonlinear relationship, the RMSE* will be lower than the RMSE measure but will still indicate a poor fit.

2.3 Pearson's *r*

Pearson's r is one of the most common measures to describe the goodness-of-fit of a model. This describes the correlation between the subjective and predicted data. Pearson's r can be calculated as:

$$r = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2}\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}}, \tag{5}$$

where N is the total number of stimuli, x_n and y_n are the n^{th} data points in the subjective and predicted datasets respectively, and $\bar{x}$ and $\bar{y}$ are the mean of the subjective and predicted datasets respectively. Unlike the RMSE and RMSE* measures which have no bounds, Pearson's r has values ranging only from -1.0 to +1.0, where +1.0 indicates a perfect linear correlation, -1.0 indicates a perfect negative correlation, and 0 indicates no correlation.

Pearson's r is a measure of the linear relationship between the subjective and predicted. Therefore, as with RMSE and RMSE*, if the predicted values were nonlinearly related with the subjective data, Pearson's r may indicate that the model provides a bad fit. Additionally, Pearson's r does not take into consideration the variance of the subjective data; however, the distance of each predicted value from the corresponding subjective data point is taken into consideration.

2.4 Spearman's rho

All of the goodness-of-fit measures discussed this far have the same problem, that there needs to be a linear relationship between the subjective data and the predicted values. Spearman's rho, however, is a measure of the rank order of the stimuli and thus does not assume a linear relationship. Spearman's rho can be calculated as:

$$\rho = 1 - \frac{6\sum_{n=1}^{N}(RX_n - RY_n)^2}{N(N^2-1)}, \tag{6}$$

where RX and RY is the rank order of the stimuli, according to the subjective and predicted data respectively, and N is the total number of stimuli [9]. As with Pearson's r, the Spearman's rho measure ranges from -1.0 to +1.0, with +1.0 meaning perfect rank order of the stimuli, -1.0 representing perfect inverse rank order of the stimuli, and 0 being no relationship between the rank orders of the stimuli according to the subjective data and predicted values.

Although the Spearman's rho does not rely on a linear relationship between the subjective and predicted data, this is an ordinal measure, and violations of the rank order are treated equally regardless of the extent to which the rank order is violated. This can be an issue for subjective data where the subjective variance is not considered.

For example, if a dataset exists where several of the subjective data points are very similar and have confidence intervals that overlap, the rank ordering according to the subjective data is potentially misleading; the true rank ordering could be any permutation. Hence, this may unfairly penalise any candidate models that fail to predict the rank ordering, but would do so if the the variance was taken into consideration.

This problem can be solved with the Monte Carlo simulation method [10, 11].

3 Monte Carlo simulation

In the Monte Carlo method, multiple simulated datasets are generated from the original subjective dataset. Each data point in each simulated dataset is randomly shifted within its standard deviation. Each data point can be calculated as:

$$\hat{x}_n = x_n + (\mathscr{G} \times \Delta x_n), \tag{7}$$

where x_n is the n^{th} data point of the observed dataset, Δx_i is the standard deviation of the n^{th} data point, and $\mathscr{G}$ represents a random number drawn from the Gaussian distribution with a mean of 0 and a width of 1, meaning that 95% of the random numbers will range between -1.0 and 1.0.

By using this method, each new simulated dataset will contain the subjective data randomly varied with respect to the standard deviation. Therefore, if two data points are similar and have confidence intervals that overlap, there is a probability that at least one of the simulated datasets will reverse the rank order of the corresponding stimuli. In general, it is advised to generate a large number of new simulated datasets ($M > 1000$) [11].

3.1 Monte Carlo Spearman's rho

The Spearman's rho measure can be calculated for each of the simulated datasets. It can then be useful to examine the maximum, minimum and mean of the resulting set of coefficients.

Examining the maximum value of the Monte Carlo Spearman's rho allows identification of candidate models that best predict the rank stimulus order for the data variations most conducive to modelling. Using this method to select candidate models would be beneficial over the measures outlined in Section 2 since this method: 1) does not rely on a linear relationship between the subjective data and predicted values, and 2) takes into consideration the variance of the subjective data.

Examining the minimum Monte Carlo Spearman's rho allows selection of candidate models using a *minimax* method.

3.2 Minimax selection

The minimax method is in game and decision theory. Under the minimax paradigm, rather than selecting the model that can best predict the data in its optimal permutation, the model is selected that performs the best under the most adverse of conditions.

This can be achieved by examination of the minimum Monte Carlo Spearman's rho measure for each candidate model. Candidate models that have high values of minimum Monte Carlo Spearman's rho can be considered as having the best rank order with the least favourable permutation of the subjective data (taking into consideration the variance of the data).

4 Worked Example

To help explain the usefulness of the Monte Carlo method, a dummy dataset has been created that may be typical of the results from a listening test. This dummy dataset was created to mimic twelve stimuli rated on a 100 point scale, shown in Figure 1. The synthesised data had a range of mean values (data points) and a variance for each stimulus to represent the variance of real listener data, shown by the 95% confidence intervals in Figure 1. These results are representative of those that may arise from a multiple stimulus comparison.

From visual inspection of the dummy data, it can be seen that the data points at the extremes of the scale (stimuli 1–3 and 9–12) do not have overlapping confidence intervals, whereas the middle data points (stimuli 4–7) have confidence intervals that do overlap. Therefore, it is reasonable to suggest that a good model of this data could predict the rank order for the end data points well compared to the intermediate data points whose confidence intervals all overlap.

To illustrate the usefulness of the Monte Carlo Spearman's rho as a measure to select the most appropriate model, three candidate models were created. The predicted data against the dummy data are shown in Figure 2. Candidate model 1, Figure 2a, was the dummy data points with a random offset added to each. Candidate model 2, Figure 2b, has larger offsets added to the data points corresponding to the highest three and lowest three rated stimuli (stimuli 1–3 and 9–12), causing the rank order of these to be altered. Candidate model 3, Figure 2c, has random offsets added only to the middle six data points, altering the rank ordering of stimuli 4–8.

The line of $y = x$, that represents an ideal model is plotted on all of these graphs. This line passes through all of the confidence intervals for candidate model 3, Figure 2c, yet very few of the confidence intervals for candidate models 1 and 2, Figures 2a and 2b. It is clear from visual inspection of Figure 2 that candidate model 3 should be the most suitable to predict the dummy data.

Candidate models 2 and 3 were then nonlinearly transformed to produce candidate models 2' and 3'. These two new transformed models, as well as candidate model 1 are shown in Figure 3. These three models were analysed using the goodness-of-fit measures discussed in Section 2. For each measure, the best-performing model is identified. An ideal measure would identify model 3' as the best since, with the application of a nonlinear transform, it becomes model 3 and fits the dummy data extremely well (Figure 2c).

The results of the goodness-of-fit measures for all three tested models are shown in Table 1, with the model that performs best in each measure highlighted in bold.

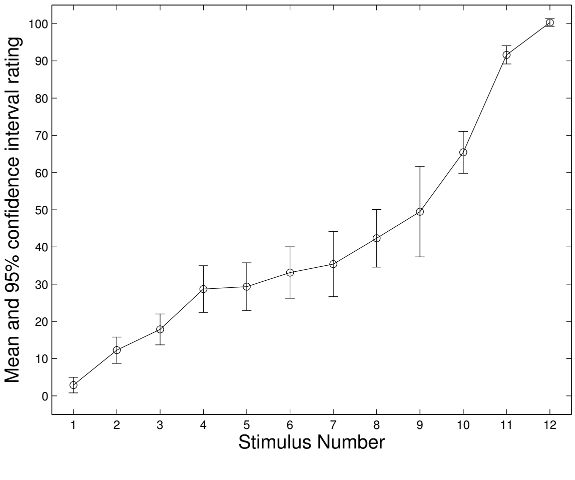
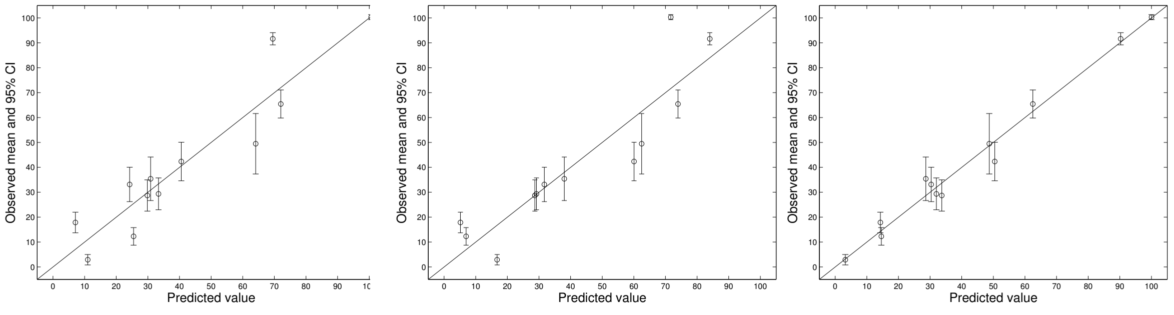


Fig. 1: Mean and 95% confidence intervals for the dummy dataset.



(a) Model 1 **(b)** Model 2 **(c)** Model 3

Fig. 2: Three candidate models created to predict the dummy data.

4.1 RMSE analysis

According to the RMSE, model 1 is the best candidate model. This is because the RMSE measure depends on a linear relationship between the dummy data and predicted values.

Using RMSE to select a model would therefore lead to model 3' (ultimately the best) being discarded.

4.2 RMSE* analysis

According to the RMSE* model 1 again performs best. The RMSE* values are lower than the RMSE, because the 95% confidence intervals of the dummy data are taken into account. However, using RMSE* to select a model would again lead to model 3' (ultimately the best) being discarded.

4.3 Pearson's *r* analysis

According to Pearson's *r*, model 1 is again the best performing. As with the RMSE and RMSE*, this is because the Pearson's *r* measure relies on a linear relationship between the dummy data and predicted values. The poor rating of models 2' and 3' can be explained by the nonlinear relationship.

Using Pearson's *r* would again lead to model 3' (ultimately the best) being discarded.

Spearman's rho analysis

The Spearman's rho metric does not rely on the linear relationship between the dummy data and predicted values, meaning that candidate models 2' and 3' might be rated higher. Indeed, according to Spearman's rho model 2' performs best, followed by models 3' and 1. It is interesting to note that model 1 performed best according to the RMSE, RMSE*, and Pearson's *r* measures yet the worst according to Spearman's rho.

However, using Spearman's rho would still lead to model 3' (ultimately the best) being discarded.

Monte Carlo simulation

For the Monte Carlo simulation analysis, 5000 simulated datasets were generated from the dummy data. For each of the candidate models (1, 2', and 3'), the Spearman's rho was calculated. To compare these models, the maximum and minimum Monte Carlo Spearman's rho are shown in Table 1.

The maximum value of the Monte Carlo Spearman's rho indicates the goodness-of-fit of each model to the best possible rank ordering of the stimuli (allowing for the variance of the dummy data). According to this measure, candidate model 3' provides the best fit, having a perfect rank ordering of the data.

The minimum Monte Carlo Spearman's rho shows how a candidate model performs when the variance of the dummy data is considered but the rank order is shifted in the way least beneficial to the predicted values. This measure is used in the minimax method discussed in Section 3.2. According to the minimum Monte Carlo Spearman's rho and the minimax method model 3' performs best.

Model Number	RMSE	RMSE*	Pearson's r	Spearman's rho	Maximum Monte Carlo Spearman's rho	Minimum Monte Carlo Spearman's rho
Model 1	**10.5414**	**5.9447**	**0.9364**	0.9091	0.9790	0.6154
Model 2'	12.8486	9.9238	0.9038	**0.9510**	0.9650	0.6783
Model 3'	11.8205	7.1028	0.9193	0.9161	**1.000**	**0.7133**

Table 1: Goodness-of-fit measures between the dummy dataset and candidate models 1, 2', and 3'. numbers in bold indicate the best-performing model according to each measure.

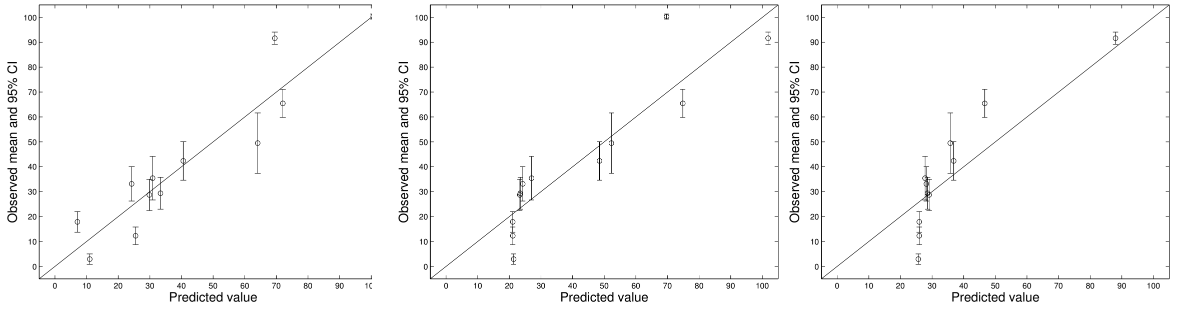
5 Summary

Regression modelling is commonly used to create objective measurement models that predict subjective qualities of audio. This paper has identified the weaknesses of common goodness-of-fit measures and proposed a new measure that may prove useful in the assessment of candidate models, the Monte Carlo Spearman's rho. Taking the Spearman's rho of multiple Monte Carlo simulated datasets provides a more appropriate means to evaluate candidate models that: 1) do not yet produce predicted values having a linear relationship with the subjective data, and 2) take into consideration the variance of the subjective data.

Although this measure is more appropriate than other for nonlinear candidate models, it does not provide any information regarding the nature of the nonlinearity, nor does it indicate whether or not the relationship can be linearised. It is recommended that this measure be used purely as a method to select a suitable shortlist of models from a large selection. With this shortlist, visual analysis of residuals plots and other statistical measures can be used to select and develop the best candidate.

Section 4 provided a worked example of the candidate model selection process using the common goodness-of-fit measures and the Monte Carlo Spearman's rho. This example demonstrated that the standard measures of goodness-of-fit will sometimes lead to the best candidate model being discarded. In this example, only the Monte Carlo Spearman's rho measure led to the selection of the model that would ultimately provide the best fit to the target data. However, it should be noted that the candidate models in this paper were artificially created to exemplify the problems with the standard measures.

Using the minimum Monte Carlo Spearman's rho, candidate models can be selected using the minimax method; this favours models that perform the best under the least favourable variation of the subjective data. This may aid in the selection of candidate models that are not overfitted, and that will be more robust and therefore more likely to perform well with subsequent validation data.



(a) Model 1 **(b)** Model 2' **(c)** Model 3'

Fig. 3: Candidate models 1, 2, and 3 used for analysis, with nonlinear transformations applied to models 2 and 3.

6 Acknowledgements

The work reported in this paper was supported by the UK's Engineering & Physical Sciences Research Council (EPSRC) and Cirrus Logic.

The data underlying the findings presented in this paper are available from doi 10.5281/zenodo.47245. Further project information can be found at http://iosr.uk/micQuality.

References

[1] ITU-R, "ITU-R BS.1387-1 Method for Objective Measurements of Perceived Audio Quality," Recommendation BS.1387-1, International Telecommunication Union, Geneva, 2001.

[2] ITU-T, "ITU-T P.862 Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Recommendation P.862, International Telecommunication Union, Geneva, 2001.

[3] ITU-T, "ITU-T P.863 Perceptual objective listening quality assessment," Recommendation P.863, International Telecommunication Union, Geneva, 2011.

[4] Fenton, S. and Hyunkook, L., "Towards a Perceptual Model of "Punch" in Musical Signals," in *139th Convention of the Audio Eng. Soc.*, New York, USA, 2015.

[5] Francombe, J., *Perceptual Evaluation of Audio-on-Audio Interference in a Personal Sound Zone System*, PhD thesis, University of Surrey, Guildford, Surrey, UK, 2014.

[6] ITU-R, "ITU-R BS.1387-1 Method for Objective Measurements of Perceived Audio Quality," Recommendation BS.1387-1, International Telecommunication Union, Geneva, 2015.

[7] Conetta, R. and Brookes, T. and Rumsey, F. and Zielinski, S. and Dewhirst, M. and Jackson, P. and Bech, S. and Meares, D. and George, S., "Spatial Audio Quality Perception (Part 2): A Linear Regression Model," *J. Audio Eng. Soc*, 62(12), pp. 847–860, 2015.

[8] Field, A., *Discovering Statistics using IBM SPSS Statistics*, SAGE Publications, 4th edition, 2013.

[9] Argyrous, G., *Statistics For Research : With A Guide To SPSS*, SAGE Publications, London, 2nd edition, 2005.

[10] Mooney, C., *Monte Carlo Simulation*, SAGE Publications, 0 edition, 1997.

[11] Curran, P., "Monte Carlo error analyses of Spearman's rank test," in *available at http://arxiv.org/abs/1411.3816*, 2015.