On the scattering in normal distribution of the peak ground acceleration residuals in Alborz region.

H. Ghasemi[1], M. Zare[2] and Y. Fukushima[3]

[1] *Research Student, Earthquake Research Institute, University of Tokyo, Tokyo. Japan*
[2] *Associated Professor, Dept. of Seismology, International Institute of Seismology, Tehran, Iran*
Email: hghasemi@gmail.com

ABSTRACT :

The scattering of residuals has great influence on probabilistic seismic hazard assessment which is fairly upgraded from its original method by including the scatter in the integration scheme and several methodologies for desegregation. This article gives the results of the scattering study in the obtained residuals based on the recently developed attenuation relationship for Alborz region. Several statistical tests have been applied to show to what extend the scatter in residuals is truly represented by the normal distribution. As revealed by the results of the present study the lognormal distribution is valid just in the range of three standard deviations around the median value. Thus the truncation in PSHA studies must be considered in the given range. The applicability of the proposed relationship for Alborz region as well as several other relations, developed for shallow crustal environments, is also studied by means of statistical tools. The results clearly reflect the significance of using region-specific strong motions in developing attenuation relations suitable for the specific region.

KEYWORDS: Alborz, seismic hazard, Goodness-of-fit tests

1.INTRODUCION

Nowadays PSHA is a common method to deal with seismic hazard in seismically active regions. A key part of such studies is attenuation relations, which are empirically or theoretically equations to show the relationship between desired strong motion parameter and several factors influencing the random nature of the ground motion e.g. distance, magnitude etc. Using such relations the scattering for the predicted parameter would be quite significant for example the standard deviations for such equations might be in the range of 0.5 to 0.7 in logarithmic scale (Douglas, 2003). In addition, because of the presumed lognormal distribution for ground motion parameter, most of them have a great drawback: there is no upper limit for residuals in other words the ground motion parameter could take any value. To overcome this, common procedure is to truncate the probability density function to a certain number of standard deviations. But the question is to what extend the PDF must be truncated. The answer to this question will have strong influence on the final seismic hazard curves. Truncating in the range where lognormal distribution is valid would be reasonable answer to the mentioned question. In this regard statistical tests must be applied to the database first to validate the presumed lognormal distribution and second to determine the range where the PDF must be truncated.
In the present study the scatter in residuals of recent attenuation relation developed for Alborz region is investigated using strong motions recorded in this region. This area is located within the Alpine-Himalayan active mountain belt. Many active faults affect the Alborz, most of which are parallel to the range and accommodate the present day oblique convergence across it (Jafari, 2007). This region has been affected several times by historical and recent earthquakes that confirm the importance of seismic hazard assessment through it. A review of selected statistical measurements and tests are presented after a brief description of the accelerometric data-bank, considered in this study. Finally the proposed attenuation model for Alborz region is compared with several attenuation relations developed for shallow crustal environments, following the scheme proposed by Scherbaum et al. (2004) and the results are presented.

2.DATA

Data used in the present study include ground motions recorded in Alborz region by ISMN[1] network. Distribution of magnitude and distance (hypocentral distance) for records, included in this catalogue, is shown in Fig. 1. The data-base is recently used to develop empirical attenuation model to predict peak ground acceleration in this region (Sinaeian, 2006). This model, containing the near-fault amplification saturation term, is in accordance with the geometrical spreading and intrinsic attenuation for spherical body waves. In this model, the site conditions are classified as rock and soil sites. The sites of shear wave velocities greater than 760 m/s are defined as rock sites and those of lower than the 760 m/s are soil ones.

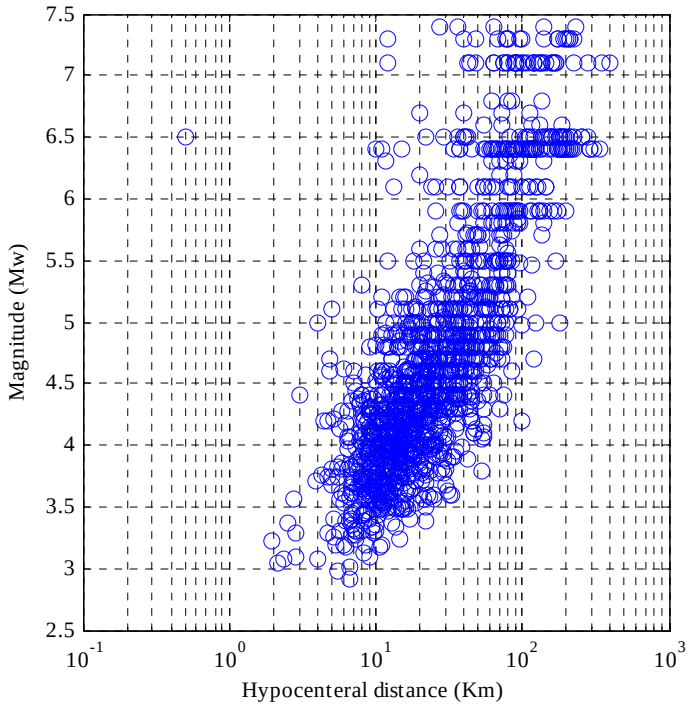

Figure 1. Distribution of strong motion records with respect to magnitude and hypocenteral distance

The residuals from a fitted model are the differences between the responses, i.e. peak horizontal acceleration, observed at each combination values of the explanatory variables and the corresponding prediction of the response computed using the regression function. In this study the residuals are calculated using (Eq. 1). When the model fits the data well, the independent random errors are approximated by using residuals.

$$R = \log(Y_{Observed}) - \log(Y_{Predicted}) \quad (1)$$

In general, ground motion parameters are usually assumed to be log-normally distributed. Extending this assumption, it is accepted that the logarithm residuals of peak ground motion parameters have also normal distribution (Bommer, 2001).

3.STATISTICAL MEASUREMENTS

Despite the existence of many statistical tools for model validation, the graphical residual analysis (NIST, 2006) is the primary one in most modeling processes. There are also several numerical residual analysis measurements and tests, useful for model validation purpose (NIST, 2006). Regarding the graphical techniques, one easy way to test whether or not obtained residuals fit to normal distribution is to compare the PDF and CDF of the sample data with the theoretically ones. Figure 2 and Figure 3 show the results of such comparison. The fitted normal distributions (using maximum likelihood approach) are shown as continues line while the PDF and CDF of the sample data are shown as histograms. As can be seen the normal distribution seems to be reasonable.

[1] Iranian Strong Motion Network

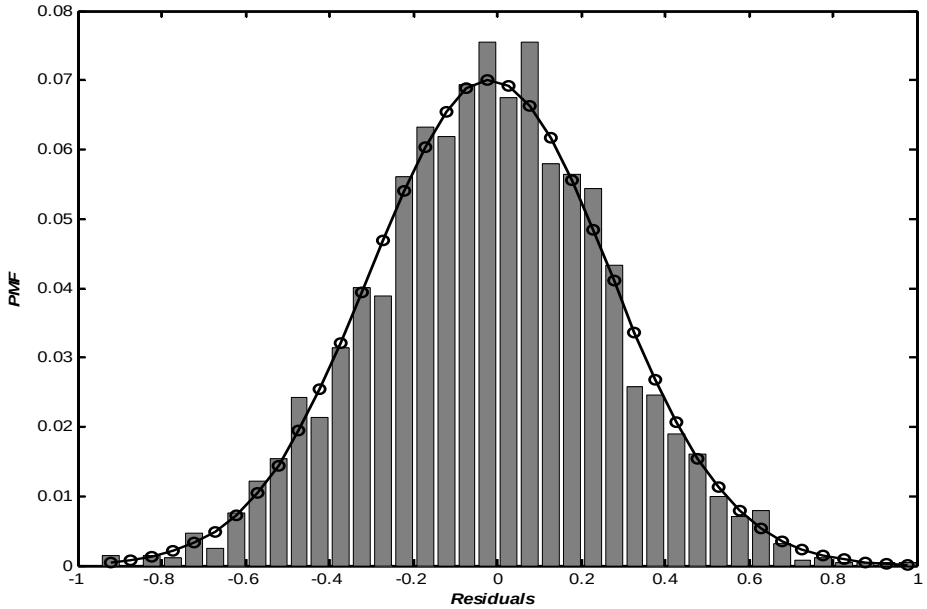

Figure 2. Probability mass function for the original data and the fitted normal distribution (solid line)

The next graphical technique, quantile-quantile plot (q-q plot), shown in Fig. 4, is used to check whether or not the residuals are normally distributed in the candidate model. It displays the residual quantiles versus theoretical quantiles from a normal distribution. If the residual sets do come from the normal distribution, the plot will be linear. The plot has the quantiles of the residuals displayed with the plot symbol '+'. A reference line, a line joining the first and third quartiles of each distribution, is superimposed on the plot. For the candidate ground motion model it can be seen that the normal distribution fits the data well between the ranges of 3 standard deviations around the median.

Regarding the numerical methods, there are several statistical measurements and tests, suitable to gain an insight into the "goodness" of a fit by the candidate model. The main goal of such measurements is testing null hypothesis (H0), i.e. the residual sets fit a normal distribution with zero mean and specified variance, against the alternative hypothesis (Ha) in which the residual sets do not fit the specified distribution. The first considered GOF test is chi-square test which can be applied on a residual set to test if a sample of data came from a population with a specific distribution, in our case normal distribution. An advantage of the chi-square goodness-of-fit test is that it can be applied to any unvaried distribution for which it is possible calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data thus the value of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid. The test statistic is defined as (Eq. 2).

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} \tag{2}$$

Where O_i is the observed frequency for bin *i* and E_i is the expected frequency for bin *i*. The expected frequency is calculated by (Eq. 3).

$$E_i = N(F(Y_u) - F(Y_l)) \tag{3}$$

where F is the cumulative Distribution function for the distribution being tested, Y_u is the upper limit for class *i*, Y_l is the lower limit and *N* is the sample size.

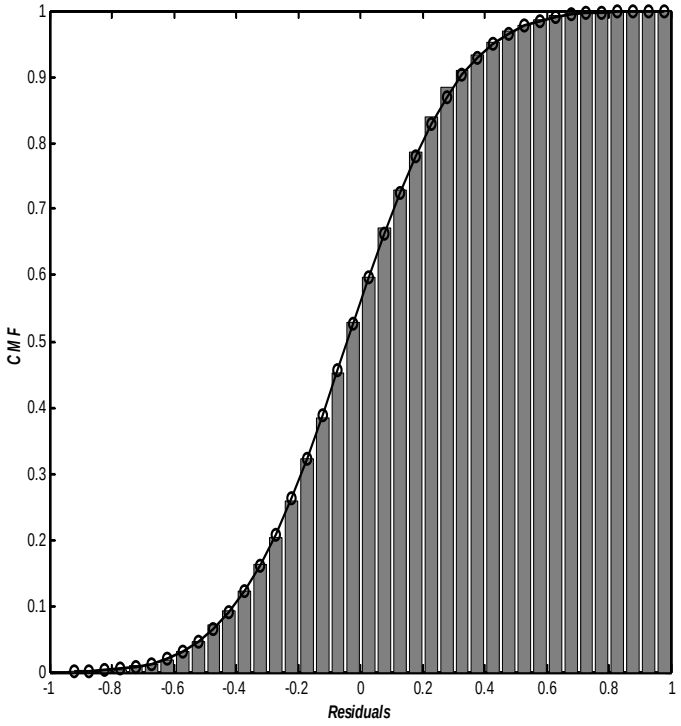

Figure 3.Cummulative mass function for the original data and the fitted normal distribution (solid line)

To accept the null hypothesis, to any desired significance level, the results of the test statistics must be lower than the value of the chi-square percent point function. The chi-square test statistic for the residuals of PGA and chi-square percent point function at 10%, 5% and 1% significance levels are tabulated in Table 1. According to the results the test statistic is greater than the chi-square percent point function for a significance level of 10% but is lower for 5% and 1% significance levels. In this regard, the null hypothesis that the residuals follow the normal distribution must be rejected for 10% significance level.

The next GOF test, Kolmogorov-Smirnov test, is used to decide if a sample comes from a normal distribution with zero mean and standard deviation equal to the one determined for *the selected attenuation model. The test is based on the maximum distance between the empirical cumulative* distribution function and the standard normal cumulative distribution function. Mathematically, this can be written as (Eq. 4).

$$Max(|G(x)-F(x)|) \tag{4}$$

where $G(x)$ is the proportion of X values less than or equal to x and $F(x)$ is the normal cumulative distribution function evaluated at x. The Kolmogorov-Smirnov test statistic for the residuals of PGA and critical values at 10%, 5% and 1% significance levels are tabulated in Table 1. Considering the results, the null hypothesis should not be rejected at considered significance levels.

Table 1. Results for goodness-of-fit tests

Chi-Square	PGA	Kolmogorov-Smirnov	PGA	Lilliefors	PGA
χ^2	11.7950	D	0.0145	L	0.0145
$\chi^2_{10\%}$	10.6446	$D_{10\%}$	0.0773	$L_{10\%}$	0.0152
$\chi^2_{5\%}$	12.5916	$D_{5\%}$	0.0862	$L_{5\%}$	0.0167
$\chi^2_{1\%}$	16.8119	$D_{1\%}$	0.1033	$L_{1\%}$	0.0208

The Lilliefors test is used to evaluate the hypothesis of a residual set having normal distribution with unspecified mean and variance. In this test the samples, having the same mean and estimated rather than primarily specified variance, are compared for empirical and normal distribution. The results of Lilliefors test

are listed in Table 1. According to the results the residuals follow the normal distribution and the null hypothesis is accepted.

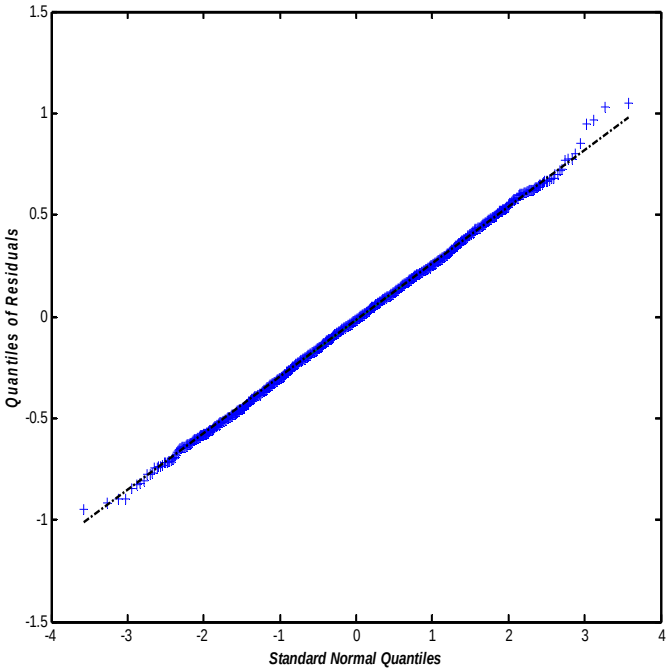

Figure 4. Quantile-quantile plot of residuals versus standard normal

4.COMPARISON WITH OTHER RELATIONSHIPS

In seismic hazard studies, the aleatory uncertainty is represented by the standard deviation of the PDF for the attenuation equations, and the epistemic uncertainty is represented by means of the formulation of the logic tree, Abrahamson 2000. In this approach each attenuation relation is assigned a weighting factor that is interpreted as the relative likelihood of that relation being correct. In this regard, selecting and ranking properly the ground motion models, will have great influence on final results of the seismic hazard assessment; hence a quantitative, data-driven scheme to assign logic tree weights is desirable. Recently, Scherbaum et al. (2004) proposed a simple yet effective scheme to select and rank appropriate ground motion models. This classification scheme is based on the statistical analysis of normalized residuals, which are the differences between the logarithms of the data values and logarithmic model predictions, divided by the corresponding standard deviations of the logarithmic model. Ideally, this should result in residuals that are normally distributed with zero mean and unit variance. They introduced a new likelihood based, measure called LH in their ranking scheme. LH measure is suitable to quantify not only the model fit, but also the underlying statistical assumptions. The median LH value is defined as the statistical test, mainly due to its stability regarding outliers. This range measure of 0 to 1 with a value of 0.5 shows a perfect accordance with standard normal distribution. Applying this scheme, even a rather small data-set of registered ground motion records in the target region can be helpful in selecting and ranking the proper candidate ground motion models systematically and quantitatively, Scherbaum et al. (2004) and Drouet et al. (2007).

The NGA Ground motion models, considered in this study (i.e. Boore and Atkinson 2007; Campbell and Bozorgnia 2007 and Chiou and Youngs 2006), are developed based on strong motions, recorded in seismically active shallow crustal environments which is similar to the current tectonic regime of Alborz region. These relationships can be applied in all earthquakes relevant to the shallow crustal ones occurred in California, Abrahamson 2007. The database was common for all members of NGA, but their data selecting criteria, parameters, and functional forms were different, Campbell and Bozorgnia (2006). Using VS30 for the site condition and, inclusion the factors of rupture depth, hanging wall and soil depth are the key changes, applied by the NGA members to their previous models. All models include nonlinear site response effects. The main

characteristics of considered models are summarized in Table 2.

Table 2. Main characteristic of the selected ground motion models.

Model	Horizontal component definition	M type	M range	R type	R range
Sinaeian (2006)	Independent	Mw	3.0-7.4	Rhypo, Rjb	4-250
Boore and Atkinson (2007)	GMRotI50**	Mw	5-8	Rjb	0-250
Campbell and Bozorgnia (2007)	GMRotI50	Mw	5-8	Rrup, Rjb	0-250
Chiou and Youngs (2006)	GMRotI50	Mw	5-8	Rrup, Rjb	0-250

In order to check the applicability of the mentioned relationships as well as the one considered in this study, the 120 strong motions recorded during 2004 Kojour (Mw 6.4) earthquake, are considered. The supplementary detailed studies on rupture process and causative fault geometry are available for this event, Tatar et al. (2007). Such information is essential in estimating the necessary inputs of NGA ground motion models. The ranking scheme proposed by Scherbaum et al. (2004) is based on LH median value together with the mean, median and standard deviation of the residuals. If a median LH value of at least 0.2, with the absolute value of mean and median of the normalized residuals, and their standard deviation smaller than 0.75, are determined for a normalized residual set, the ground motion model should be ranked as class “C”. The sample standard deviation also should be smaller than 1.5. The median LH value of at least 0.3, with the absolute value of mean and median of the normalized residuals, and their standard deviation of smaller than 0.5, and sample standard deviation of smaller than 1.25 should be considered in class “B”. In case of corresponding normalized residuals, a median LH value of at least 0.4, with the absolute value of mean and median of the normalized residuals, and their standard deviation smaller than 0.75 and sample standard deviation smaller than 1.125, the rank “A” will be assigned to the models. A model that does not satisfy the mentioned criteria for any of these classes should be ranked as class “D”.

For residual sets of considered ground motion models, the median LH values, the maximum likelihood estimates of the central tendency parameters, and corresponding standard deviations of these parameters are tabulated in Table 3. The determined standard deviations of each measure are calculated using bootstrap technique through data re-sampling. According to the scheme criteria, the proposed model for Alborz region should be ranked ‘A’, and Campbell and Bozorgnia, relation is of ‘C’ rank.

Table 3. Results of the applied ranking scheme

Model	LH	sigma	mean	sigma	median	sigma	std	sigma	Rank
Boore and Atkinson (2007)	0.089	0.027	-0.190	0.166	0.269	0.362	2.075	0.092	D
Campbell and Bozorgnia (2007)	0.351	0.014	0.078	0.054	0.212	0.115	1.271	0.038	C
Chiou and Youngs (2006)	0.028	0.016	0.659	0.213	1.148	0.320	2.250	0.094	D
Sinaeian (2006)	0.412	0.049	0.031	0.105	0.211	0.152	1.106	0.063	A

5.CONCLUDING REMARKS

1. Comparing the PDF and CDF of the residual set determined for Sinaeian (2006) model with the theoretically ones, indicates that normal distribution of residuals is reasonable.
2. According to the quantile-quantile plot for residual set determined for Sinaeian (2006) model, the normal distribution fits the data well between the ranges of 3 standard deviations around the median.
3. The results of goodness-of-fit tests indicate that the residuals determined for Sinaeian (2006) model follow the normal distribution.
4. According to the results of this study, the most proper model to be used in seismic hazard projects in Alborz region is the one developed based on only ISMN data, despite its simple functional form.

REFERENCES

Abrahamson, N. (2007). “Results and Implications of the Next Generation Attenuation (NGA) Ground Motion Project (Abstract)”. SMIP07 Seminar on Utilization of Strong-Motion Data, pp. 107 - 108.

Boore, D. M., and Atkinson, G. (2007). "Boore-Atkinson NGA ground motion relations for the geometric mean horizontal component of peak and spectral ground motion parameters". PEER, Rep. No. 2007/01.

Campbell, K., and Bozorgnia, Y. (2006). "Next Generation Attenuation (NGA) empirical ground motion models: can they be used in Europe". 13th ECEE.

Campbell, K., and Bozorgnia, Y. (2007). "Campbell-Bozorgnia NGA ground motion relations for the geometric mean horizontal component of peak and spectral ground motion parameters". PEER, Rep. No. 2007/02.

Chiou, Y., and Youngs, R. (2006). "Chiou-Youngs NGA ground motion relations for the geometric mean horizontal component of peak and spectral ground motion parameters". PEER, Interim report for USGS review.

Douglas, J. (2003). “Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectra ordinates”. *Earth Sci Rev* 61. pp. 43–104

Drouet, S., Scherbaum, F., Cotton, F., and Souriau, A. (2007). "Selection and ranking of ground motion models for seismic hazard analysis in the Pyrenees", *Journal of Seismology*, Vol. 11, pp. 87-100.

Jafari, M. (2007). “Time independent seismic hazard analysis in Alborz and surrounding area”, Natural Hazards, Vol. 42, pp. 237-252.

NIST/SEMATECH e-Handbook of Statistical Methods, (2006). http://www.itl.nist.gov/div898/handbook

Scherbaum, F., Cotton, F. and Smit, P. (2004). "On the use of response spectral reference data for the selection and ranking of ground-motion models for seismic hazard analysis in regions of moderate seismicity: the case of rock motion", *Bull. Seismol. Soc. Am.*, Vol. 94, pp. 1-22.

Sinaiean, F. (2006). "*Study on Iran Strong Motion Records*", Ph.D. Thesis, International Institute of Earthquake Engineering and Seismology, Tehran, Iran.

Tatar, M., Jackson, J., Hatzfeld, D., and Bergman, E., (2007). “The 2004 May 28 Baladeh earthquake (Mw 6.2) in the Alborz, Iran: overthrusting the South Caspian Basin margin, partitioning of oblique convergence and the seismic hazard of Tehran”, *Geophys. J. Int.*, Vol 170, pp. 249–261