pISSN: 2093-940X, eISSN: 2233-4718
Journal of Rheumatic Diseases Vol. 26, No. 1, January, 2019
https://doi.org/10.4078/jrd.2019.26.1.5

Review Article

Normality Test in Clinical Research

Sang Gyu Kwak[1], Sung-Hoon Park[2]

[1]Department of Medical Statistics and [2]Division of Rheumatology, Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu, Korea

In data analysis, given that various statistical methods assume that the distribution of the population data is normal distribution, it is essential to check and test whether or not the data satisfy the normality requirement. Although the analytical methods vary depending on whether or not the normality is satisfied, inconsistent results might be obtained depending on the analysis method used. In many clinical research papers, the results are presented and interpreted without checking or testing normality. According to the central limit theorem, the distribution of the sample mean satisfies the normal distribution when the number of samples is above 30. However, in many clinical studies, due to cost and time restrictions during data collection, the number of samples is frequently lower than 30. In this case, a proper statistical analysis method is required to determine whether or not the normality is satisfied by performing a normality test. In this regard, this paper discusses the normality check, several methods of normality test, and several statistical analysis methods with or without normality checks. **(J Rheum Dis 2019;26:5-11)**

Key Words. Normality check, Normal distribution, Normality test, Statistical analysis method

INTRODUCTION

In data analysis, given that various statistical methods assume that the distribution of the population data is normal distribution, it is essential to check and test whether or not the data satisfy the normality requirement. For example, when comparing the distribution of two independent groups, two sample t-tests, which is a parametric method, are used, if the two population data satisfy the normality requirement, and the Mann-Whitney U-test, which is a nonparametric method, if the data do not satisfy the normality requirement [1]. The two-sample t-test assumes normality and the Mann-Whitney U-test does not assume normality. If the data satisfy normality, the distribution of the two groups can be compared using a two-sample t-test using means and standard deviation. However, if normality is not satisfied, the Mann-Whitney U-test is used, which does not use the mean and standard deviation and concludes that the two groups are similar if the rankings are similar.

Although the analytical method varies depending on whether or not the normality requirement is satisfied, inconsistent results might be obtained depending on the analysis method used. Said differently, it can be concluded that two independent groups have the same distribution, although they are in fact different. On the other hand, it can be concluded that the distribution of two independent groups is the same. In order to solve these problems, it is necessary to check and test whether or not the normality requirement is satisfied.

In many clinical research papers, results are presented and interpreted without checking or testing normality. In the case when the reviewer requests the normality check or test in the review process of a thesis, the normality test is carried out to correct the contents of the submitted papers. However, when this lack of the normality check or test goes unnoticed, the results are frequently presented without a normality test. If the statistical analysis method assumes normality, a normality test should be performed to check whether or not the normality requirement is

Received : August 22, 2018, **Revised** : September 14, 2018, **Accepted** : September 14, 2018

Corresponding to : Sung-Hoon Park http://orcid.org/0000-0002-3218-5420
Division of Rheumatology, Department of Internal Medicine, Catholic University of Daegu School of Medicine, 33 Duryugongwon-ro 17-gil, Nam-gu, Daegu 42472, Korea. E-mail : yourii99@cu.ac.kr

satisfied. One of the reasons why normality tests are not performed is that the researchers' understanding of the statistical analysis methods is low.

Furthermore, the average of the collected data is the sample mean. According to the central limit theorem, the distribution of the sample mean satisfies the normal distribution when the number of samples is larger than 30 [2]. Therefore, if the number of samples is larger than 30, the analysis can be performed on the assumption that the normality is satisfied. In clinical studies, however, the number of samples is frequently lower than 30. The reasons for this scarcity of samples include cost and time restrictions during data collection. In this case, a proper statistical analysis method is required to determine whether or not the normality requirement is satisfied by performing a normality test.

The remainder of this paper is structured as follows. First, we introduce the contents of normality check, which is followed by the introduction of several methods of normality test. In addition, some statistical analysis methods that should be used when the normality requirement is or is not satisfied are described for the data analysis in clinical studies.

MAIN SUBJECTS

Normality check

There are four methods to check whether or not the collected data satisfy the normality requirement. These methods are checking the normality using plot and several statistics, such as mean, median, skewness, and kurtosis.

1) Distribution plot

A distribution plot of the collected data is useful to check normality of the data. The distribution of the data should be checked to determine that it does not deviate too much as compared to the normal distribution.

2) Difference value between mean and median

The mean is a simple arithmetic average of the given set of values or quantities. The median is a positional average and is defined as the middle number in an ordered list of values. In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal. Hence, the difference value between the mean and the median are close to zero in normal distribution. However, when the difference value between the mean and the median is big, the distribution is skewed to the right or to the left.

3) Skewness and kurtosis

Skewness is a measure of the "asymmetry" of the probability distribution, in which the curve appears distorted or skewed either to the left or to the right. In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other. When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than that on the right-hand side, and the mean is less than the mode. This situation is also referred to as negative skewness. When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also referred to as positive skewness.

Kurtosis is a measure of the "tailedness" of the probability distribution, in which the tails asymptotically approach zero or not. Distributions with zero excess kurtosis are called mesokurtic or mesokurtotic. The most prominent example of a mesokurtic distribution is normal distribution. A distribution with a positive excess kurtosis is called leptokurtic or leptokurtotic. In terms of shape, a leptokurtic distribution has fatter tails. Examples of leptokurtic distributions include the Student's t-distribution, exponential distribution, Poisson distribution, and the logistic distribution. A distribution with a negative excess kurtosis is called platykurtic or platykurtotic. Examples of platykurtic distributions include the continuous or discrete uniform distributions and the raised cosine distribution. The most platykurtic distribution is the Bernoulli distribution.

4) Q–Q plot

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on the estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions. The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two compared distributions, some quantiles may not be defined, so an in-

terpolated quantile may be plotted. If the Q–Q plot is based on the data, there are multiple quantile estimators in use. The rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

A simple case is when there are two data sets of the same size. In that case, to make the Q–Q plot, each set is ordered in the increasing order, then paired off, and the corresponding values are plotted. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.

The points plotted in a Q–Q plot are always non-decreasing when viewed from the left to the right. If the two compared distributions are identical, the Q–Q plot follows the 45° line y=x. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y=x. If the general trend of the Q–Q plot is flatter than the line y=x, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the Q–Q plot is steeper than the line y=x, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q–Q plots are frequently arced, or "S" shaped, indicating that one of the distributions is more skewed than the other one, or that one of the distributions has heavier tails than the other one. Although a Q–Q plot is based on quantiles, in a standard Q–Q plot, it cannot be determined which point in the Q–Q plot determines a given quantile. For example, it is not possible to determine the median of either of the two compared distributions b by inspecting the Q–Q plot. Some Q–Q plots indicate the deciles to enable determinations of this type.

Q–Q plots are commonly used to compare the distribution of a sample to a theoretical distribution, such as the standard normal distribution N(0,1), as in a normal probability plot. As in the case of comparing two data samples, one orders the data (formally, computes the order statistics) and then plots them against certain quantiles of the theoretical distribution.

Normality test

In the previous section, we described the methods for normality check. However, these methods do not allow us to draw conclusions whether or not the collected data satisfy the normality requirement. Only a rough guess can be made as in this respect. Therefore, to the definite answer, we have to consider a statistical test for normality. There are several methods to perform a normality test. The Kolmogorov-Smirnov test, the Shapiro-Wilk test, and the Anderson-Darling test are among the most popular methods. Specifically, the Kolmogorov-Smirnov test and the Shapiro-Wilk test are supported by IBM SPSS. All these tests follow the same procedure; 1) hypothesis set-up; 2) significance level determination; 3) test statistic calculation; 4) p-value calculation; 5) conclusion.

1) Hypothesis set-up

In general, all statistical tests have a statistical hypothesis. A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. A researcher might conduct a statistical experiment to test the validity of this hypothesis. The hypotheses typically include the null hypothesis and the alternative hypothesis. The distribution of population assumes the normal distribution in all data set. Hence, the null hypothesis (H_0) and alternative hypothesis (H_a) are follows;

H_0: The data are normally distributed.
H_a: The data are not normally distributed.

2) Significance level determination

The significance level α is the probability of making the wrong decision when the null hypothesis is true. Alpha levels (sometimes called simply "significance levels") are used in hypothesis tests. An alpha level is the probability of a type I error, or you reject the null hypothesis when it is true. Usually, these tests are run with an alpha level of 0.05 (5%); other commonly used levels are 0.01 and 0.10.

3) Test statistic calculation

Next, the test statistic for the normality test should be calculated. The calculation of the test statistic differs according to which of the normality test methods is used. The formulas for calculating the test statistic according to each statistical method are as follows.

(1) Shapiro-Wilk test statistic

The Shapiro-Wilk test tests the null hypothesis that a sample $x_1,\cdots,x_n$ comes from a normally distributed population. The test statistic is as follows (see Eq. (1)):

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (1)$$

where $x_{(i)}$ (with parentheses enclosing the subscript index i; not to be confused with x_i) is the i-th order statistic, i.e., the i-th smallest number in the sample, the sample mean $\bar{x}$ is given by Eq. (2).

$$\bar{x} = \frac{(x_1 + \cdots + x_n)}{n} \quad (2)$$

and the constants a_i are given by Eq. (3)

$$(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \quad (3)$$

where $m = (m_1, \cdots, m_n)^T$ and $m_1, \cdots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

(2) Kolmogorov–Smirnov test statistic

The Kolmogorov–Smirnov statistic for a given cumulative distribution function F(x) is computed using Eq. (4).

$$D_n = \sup_x |F_n(x) - F(x)| \quad (4)$$

where $\sup_x$ is the supremum function of the set of distances and F_n is the empirical distribution function for n i.i.d. (independent and identically distributed) in ordered observations X_i defined as shown in Eq. (5).

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty, x]}(X_i) \quad (5)$$

where $I_{[-\infty, x]}$ is the indicator function, equal to 1 if $X_i \leq x$ and to 0 otherwise. By the Glivenko–Cantelli theorem, if the sample comes from distribution F(x), then D_n converges to 0 almost surely in the limit when n goes to infinity. Kolmogorov strengthened this result by effectively providing the rate of this convergence.

(3) Anderson-Darling test statistic

The Anderson–Darling test assesses whether a sample comes from a specified distribution. It makes use of the fact that, when given a hypothesized underlying distribution and assuming the data do arise from this distribution, the CDF of the data can be assumed to follow a uniform distribution. The data can be then tested for uniformity with a distance test (Shapiro 1980). The formula for the test statistic A to assess if data $\{Y_1 < \cdots < Y_n\}$ (note that the data must be put in order) come from a CDF Φ is shown in Eq. (6).

$$A^2 = -n - S \quad (6)$$

where $S \sum_{i=1}^{n} \frac{2i-1}{n} [\ln(\Phi(Y_i)) + \ln(1 - \Phi(Y_{n+1-i}))]$

The test statistic can then be compared against the critical values of the theoretical distribution. Note that, in this case, no parameters are estimated in relation to the distribution function, Φ.

4) p-value calculation

Next, the significance value (p-value) should be calculated using the test statistic of the regularity test calculated in step 3). The significance value is the probability that a statistical value equal to or more extreme than the observed statistical value of the sample is observed, assuming that the null hypothesis is true. Said differently, the significance value is the probability of rejecting the null hypothesis despite the null hypothesis being true. Therefore the p-value is the degree of support for the null hypothesis. Since it is a probability value, it is calculated as a value between zero and one.

5) Conclusions

Finally, in order to draw conclusions of the normality test, we compare the significance level value set in step 2) and the calculated significance value (p-value) in step 4) and make the following conclusions.

If $\alpha \geq$ p-value, then the null hypothesis has to be rejected.
If $\alpha <$ p-value, then the null hypothesis is not rejected

If the null hypothesis is rejected because the significance value is smaller than the significance level value, the hypothesis that the data sample satisfies the normality requirement is rejected, and it can be said that it does not satisfy the normality requirement. If we set the probability of rejecting the null hypothesis to be 5%, we can conclude that the data sample does not satisfy the normality at the 5% significance level. Conversely, if the significance value is greater than the significance level, and the null hypothesis is not rejected, the conclusion can be drawn that "the data of the sample satisfies the normality requirement at the 5% significance level".

Table 1. Example data set

No.	Uric acid	No.	Uric acid	No.	Uric acid	No.	Uric acid	No.	Uric acid
1	3.8	6	8.1	11	5.0	16	5.8	21	6.8
2	2.8	7	7.7	12	6.2	17	5.6	22	4.8
3	9.5	8	6.1	13	5.9	18	5.4	23	5.6
4	8.0	9	7.0	14	6.0	19	5.3	24	4.9
5	7.4	10	6.2	15	6.5	20	5.0	25	7.3

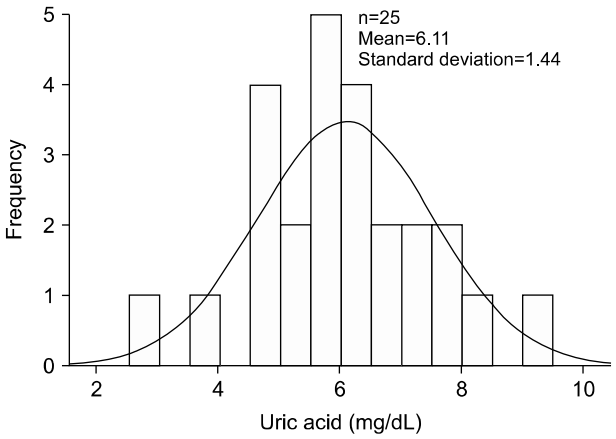

Figure 1. Histogram with normal distribution curve.

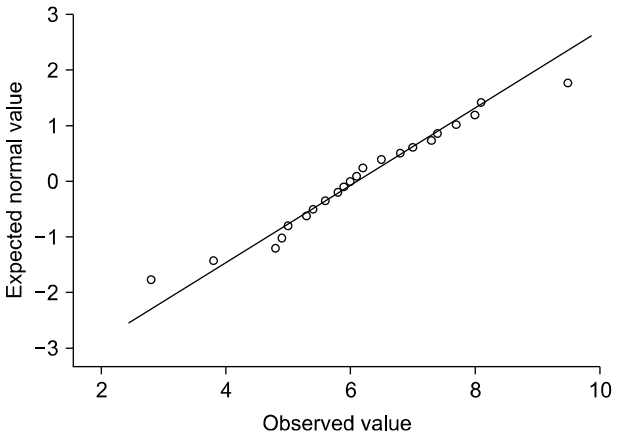

Figure 2. Q–Q plot for example data set.

Example for normality check and normality test

In this section, we illustrate the process of checking normality and testing normality using the IBM SPSS software 21.0 (IBM Co., Armonk, NY, USA) with uric acid (mg/dL) data (Table 1). First, we draw the histogram of the distribution plot with the normal distribution curve (Figure 1). The distribution plot is not much deviated from the normal distribution curve, so it can be assumed that it satisfies the normality. Second, the mean and the median are computed (6.11 and 6.00, respectively). The two values are not largely different, so it can be guessed that the data sample satisfies the normality requirement. Furthermore, the skewness and kurtosis are 0.09 and 0.68, respectively. Since both values are close to 0, the shape of the distribution can be seen as mesokurtic distribution without a shift to the left or right. Finally, we draw a Q–Q plot (Figure 2). In the Q–Q plot, the dots do not deviate much from the line, so it can be guessed that it satisfies the normality requirement.

Next, we test whether the uric acid (mg/dL) data for 25 patients satisfy the normality requirement using the Shapiro-Wilk test method and the Kolmogorov-Smirnov test method. First, we set up the hypotheses. The null hypothesis (H_0) is that the uric acid data are normally distributed, and the alternative hypothesis (H_a) is that the no uric acid data are not normally distributed. Secondly, we set the significance level to 0.05. Third, the test statistic is calculated. The test statistic according to the Shapiro-Wilk test method is 0.984, while the test statistic according to the Kolmogorov-Smirnov test method is 0.115. Fourth, we calculate the p-value. The p-value according to the Shapiro-Wilk test method is 0.949, and the p-value according to the Kolmogorov-Smirnov test method is 0.200. Finally, we and interpret the results and draw conclusions. Since the p-values according to the two normality test methods are greater than the significance level of 0.05, the null hypothesis (the uric acid data is normal distribution) is not rejected. Therefore, the uric acid data for 25 patients is considered to satisfy the normality at the 5% significance level.

Statistical analysis methods with or without normality

In selecting and using statistical analysis methods, there is a need to fully understand what statistical analysis methods are used. When establishing a hypothesis in a clinical study and analyzing collected data to test it, the most appropriate statistical analysis method should be selected and used to solve the given problem. The statistical analysis method is determined according to the

number of cases, such as the number of dependent variables, kind of dependent variable, number of independent variables, and kind of independent variable. In addition, each statistical analysis method is based on various assumptions such as normality, linearity, independence, and so on. Therefore, before using a statistical analysis method, it should be first checked whether it satisfies the assumptions of the statistical analysis method to be used; then, the selected statistical analysis method can be used. That is, if the assumption is not satisfied, the statistical analysis method could not be used. For example, when trying to compare the quantitative variables of two independent groups, the two independent t-tests that are commonly used assume normality. Therefore, an independent two-group t-test can be used only if the normality test is satisfied. If the normality is not satisfied, the Mann Whitney U-test, a statistical method other than the independent two-group t-test, should be used.

In this section, we introduce some statistical analysis methods are widely used in clinical research which assume normality. The section concludes with a discussion of statistical analysis methods that should be used when the normality requirement is not satisfied.

1) Two sample t-test

Two sample t-test is a statistical analysis method used to compare the means of two independent variables. For example, statistical analysis is used to compare the mean of serum uric acid concentrations in a group taking steroids and a group taking placebo. Two sample t-test assumes normality. Therefore, it can be used when the normality is satisfied through the normality test. In this case, the normality test should be performed for each group, and it can be said that the normality is satisfied when the normality is satisfied in both groups. Alternatively, the Mann Whitney U-test should be used [1]. The Mann Whitney U-test tests whether or not the distributions of the d data collected from two independent groups are the same; it does not compare the mean of the quantitative variables of the two independent groups.

2) Paired t-test

The paired t-test is a statistical analysis method used to compare whether or not the difference of the quantitative variables measured twice for each subject in the dependent two groups is 0 or not. This is a statistical analysis method that examines whether there is a change between two measurements. For instance, this analysis method can be used to compare the uric acid concentration in the blood measured before taking the steroid with the uric acid concentration in the blood measured after taking the steroid. Paired t-test assumes normality. Therefore, it can be used when normality is established through the normality test. In this case, the normality test should be carried out by calculating the difference between before and after the difference. If the normality is not satisfied, the Wilcoxon signed rank test should be used [3]. The Wilcoxon signed rank test tests whether or not the median of the quantitative variables differences is zero in the two dependent groups, rather than whether or not the mean of the quantitative variable differences is zero in the two dependent groups.

3) One-way ANOVA

One-way ANOVA is a statistical analysis method used to compare the means of quantitative variables over three independent groups. For example, statistical analysis is used to compare the mean of serum uric acid concentrations in a group taking steroids, a group taking steroids+vitamins, and a group taking vitamins. One-way ANOVA assumes normality. Therefore, it can be used when the regularity is satisfied through the regularity test. In this case, the normality test should be performed for each group, and it can be said that the normality requirement is satisfied when it is satisfied in all three groups. Alternatively, the Kruskal-Wallis test should be used [4]. The Kruskal-Wallis test does not compare the means of the quantitative variables over the independent three or more groups, but tests whether or not the distributions of data collected over the independent three or more groups are the same. One-way ANOVA also assumes homoscedasticity, i.e., equal dispersion. Therefore, it can be used when the homoscedasticity is satisfied through the homoscedasticity test. If the homoscedasticity is not satisfied, a Brown-Forsythe test or Welch test should be used.

4) Repeated measure one-factor analysis

Repeated measure one-factor analysis is a statistical analysis method used to repeatedly compare whether or not there is a change in the mean value of the measured quantitative variables for each subject in more than three dependent groups. This statistical analysis method examines whether or not there is a change between three or more measurements. For example, this method can be used to compare the uric acid concentration in the blood

measured before taking the steroid, 1 day after taking the steroid, and 3 days after taking the steroid. Repeated measure one-factor analysis assumes normality. Therefore, it can be used when the normality is satisfied through the normality test. In this case, the normality test should be carried out by calculating the difference between before and after each measurement point. If the regularity requirement is not satisfied, the Friedman test [5] should be used. The Friedman test tests whether or not there is a difference in the median of the quantitative variables in the dependent group, rather than compares the mean value of the quantitative variables in the dependent group.

5) Linear regression

Linear regression is a statistical analysis method used to calculate the relationship between the quantitative dependent variable and the various explanatory variables and coefficients, as well as to examine the explanatory power of the data with the estimated regression model. For example, this method can be used to examine the factors affecting the uric acid concentration in the blood. As for the factors, qualitative and quantitative variables can be set and analyzed in various ways. The regression model of linear regression assumes normality for error terms [6]. Therefore, it can be used when the normality is satisfied through the normality test. In this case, the normality test should be performed using the residual, which is an estimate of the error. If the normality requirement is not satisfied, the regression model should be modified through the model check and the data check, and the regression analysis should be performed to satisfy the normality requirement. In addition, besides normal assumption with respect to error terms, the regression model of linear regression assumes homoscedasticity, independence, and linearity. Therefore, regression analysis should be done by regression analysis that satisfies all normality, homoscedasticity, independence, and linearity by modifying the regression model through the model check and the data check.

CONCLUSION

A systematic and thorough understanding of the necessity of normality test and the method of normality test can enhance the reliability of the results reported in clinical studies. Research designs and interpretation of the findings based on the selection of the most appropriate statistical analysis method can also be considerably improved.

ACKNOWLEDGMENTS

This work was supported by research fund of the Ministry of Health and Welfare (No. 201803470002).

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Fagerland MW. t-tests, non-parametric tests, and large studies--a paradox of statistical practice? BMC Med Res Methodol 2012;12:78.
2. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017;70:144-56.
3. Kim TK. T test as a parametric statistic. Korean J Anesthesiol 2015;68:540-6.
4. Chan Y, Walmsley RP. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. Phys Ther 1997;77:1755-62.
5. Pereira DG, Afonso A, Medeiros FM. Overview of Friedman's test and Post-hoc analysis. Commun Stat - Simul Comput 2015;44:2636-53.
6. Schmidt AF, Finan C. Linear regression and the normality assumption. J Clin Epidemiol 2017;98:146-51.