General Psychiatry

Homoscedasticity: an overlooked critical assumption for linear regression

Kun Yang,[1] Justin Tu,[2] Tian Chen[3]

To cite: Yang K, Tu J, Chen T. Homoscedasticity: an overlooked critical assumption for linear regression. *General Psychiatry* 2019;**32**:e100148. doi:10.1136/gpsych-2019-100148

Received 18 September 2019
Accepted 19 September 2019

[1]Department of Family Medicine and Public Health, University of California System, San Diego, California, USA
[2]PGY-2, Physical Medicine and Rehabilitation, University of Virginia Health System, Charlottesville, Virginia, USA
[3]Department of Mathematics and Statistics, University of Toledo, Toledo, Ohio, USA

Correspondence to
Dr Tian Chen, Department of Mathematics and Statistics, University of Toledo, Toledo, OH 43606, USA;
tian.chen@utoledo.edu

SUMMARY

Linear regression is widely used in biomedical and psychosocial research. A critical assumption that is often overlooked is homoscedasticity. Unlike normality, the other assumption on data distribution, homoscedasticity is often taken for granted when fitting linear regression models. However, contrary to popular belief, this assumption actually has a bigger impact on validity of linear regression results than normality. In this report, we use Monte Carlo simulation studies to investigate and compare their effects on validity of inference.

INTRODUCTION

Linear regression (LR) is arguably the most popular statistical model used to facilitate biomedical and psychosocial research. LR can be used to examine relationships between continuous variables, and associations between a continuous and a categorical variable. For example, by using one binary independent variable, LR can be used to compare the means between two groups, akin to the two independent samples t-test. If we have a multilevel categorical independent variable, LR yields the analysis of variance (ANOVA) model. Although the t-test for unequal group variance is often used as an alternative for comparing group means when large differences in group variances emerge, the same homoscedasticity assumption underlying ANOVA is often taken for granted when this classic model is applied for comparing more than two groups. For ANOVA, much of the focus is centred on normality, with little attention paid to homoscedasticity.

Contrary to popular belief, the homoscedasticity assumption actually plays a more critical role than normality on validity of ANOVA. This is because the *F*-test, testing for overall differences in group means across all the groups (omnibus test), is more sensitive to heteroscedasticity than normality. Thus, even when data are perfectly normal, *F*-test will generally yield incorrect results, if large group variances exist. Although the Kruskal-Wallis (KW) test is applied when homoscedasticity is deemed suspicious,[1] this test is less powerful than the *F*-test, since it discretises original data using ranks, a sequence of natural numbers such as 1, 2 and 3 to represent ordinal differences in the original continuous outcomes. An even more serious problem with the KW test is its extremely complex distribution of the test statistic and consequently limited applications in practice.[2]

Over the past 30 years, many new statistical methods have been developed to address the aforementioned limitations of the classic LR and associated alternatives. Such new models apply to cross-sectional and longitudinal data, the latter being the hallmark of modern clinical research. Semiparameter statistical models are the most popular, since they require one of the distribution assumptions and apply to continuous outcomes without changing the continuous scale.[3] In this report, we use the Monte Carlo simulation study to investigate and compare results when one of the two assumptions is violated, and to show the importance of homoscedasticity for valid inference for LR. We will discuss and perform head-to-head comparison of power between the classic KW test and modern semiparametric models in a future article.

LR MODEL

We start with a brief overview of the classic LR. Consider a continuous outcome of interest, Y, and a set of p independent variables, X_1, X_2, X_p. We are interested in modelling the relationship of Y with the independent variables. Given a sample of n subjects, the classic LR models this relationship as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \mathrm{L} + \beta_p \mathrm{X}_{ip} + \varepsilon_i,\ \varepsilon_i \sim N(0,\ \sigma^2),\ 1 \le i \le n, \quad (1)$$

where i indexes the subjects, $\beta_0, \beta_1, \beta_2, \mathrm{K}\ldots, \beta_p$ are the regression coefficients (parameters), ε_i is the error term, $N(\mu,\ \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The LR in equation (1) posits a linear association between the outcome (dependent variable) Y and each of the independent variables. The latter have been called different names such as predictors, covariates and explanatory variables.

The first part of LR,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \mathrm{L} + \beta_p \mathrm{X}_{ip},\ 1 \le \mathrm{i} \le \mathrm{n}\,, \quad (2)$$

is called the conditional (population) mean of Y_i given the independent variables $X_1, X_2, \mathrm{K}, \mathrm{X_p}$. On estimating the regression coefficients, this conditional mean can be calculated to provide an estimate of Y_i for each subject. In addition to the assumed linear relationship, there are two additional assumptions in equation (1): (A) normal distribution and (B) homoscedasticity, or constant variance σ^2 for all subjects.

All three assumptions play an important role in obtaining valid inference for regression coefficients. For example, if the association of Y with a particular independent variable X_1 is quadratic, the linear model in equation (1) must also include X_{i1}^2, since otherwise estimates of β_1 will generally be biased. Likewise, if the error term ε_i is not normally distributed, SEs of estimated coefficients may be incorrect. Both the linearity and normality have been receiving great coverage in the literature.

In contrast, the impact of homoscedasticity on statistical inference of regression coefficients has received much less attention. Most publications in the biomedical and psychosocial literature do not even acknowledge this assumption for their applications of LR. Contrary to popular belief, inference about regression coefficients is actually more sensitive to departures from homoscedasticity than normality. In fact, normality actually does not matter at all when sample size is relatively large. In contrast, homoscedasticity remains an issue regardless of how large the sample size becomes. Below we illustrate these facts using Monte Carlo (MC) simulated data. For ease of exposition, we focus on one-way ANOVA, but the same conclusions apply to general LR as well.

ANOVA MODEL

One particularly popular special case of LR is the ANOVA model. This occurs when the independent variables $X_1, X_2, \mathrm{K}, X_p$ are binary indicators, representing different levels of a categorical or ordinal variable for multiple groups. The conditional mean of Y in equation (2) becomes the group mean. For example, if there are three groups, we may use group 1 as the referent and the other two independent variables X_1, X_2 to represent groups 2 and 3:

$$\mathrm{X}_{i1} = \begin{cases} 1 & \text{if subject } i \text{ is in group 2} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i2} = \begin{cases} 1 & \text{if subject } i \text{ is in group 3} \\ 0 & \text{otherwise} \end{cases}$$

In this case, the LR in equation (1) becomes:

$$\mathrm{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \varepsilon_i & \text{if subject } i \text{ is from group 1} \\ \beta_0 + \beta_1 + \varepsilon_i & \text{if subject } i \text{ is from group 2} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if subject } i \text{ is from group 3} \end{cases}$$

Thus, the regression coefficients become the group mean of Y:

$$\mu_1 = \beta_0, \mu_2 = \beta_0 + \beta_1, \mu_2 = \beta_0 + \beta_2,$$

where μ_k denotes the mean of Y for group k $(1 \le k \le K)$.

Because of the relationship of the coefficient with the group mean, the ANOVA is often simply expressed as:

$$Y_{ki} = \mu_k + \varepsilon_{ki}, \varepsilon_{ki} \sim N\left(0, \sigma^2\right), 1 \le i \le n_k, 1 \le k \le K, \quad (3)$$

where Y_{ki} denotes the outcome from the i th subject within the k th group, $\mu_k = E\left(Y_{ki}\right)$ is the (population) mean of the kth group, and K is the total number of groups. For the one-way ANOVA in equation (3) the linearity assumption does not apply, as μ_k represents the group mean and no linear or any relationship is assumed between the group means. The normality and homoscedasticity become easier to interpret and check as well, as they apply to distributions of Y_{ki} within each group.

Under ANOVA, comparisons of group means across all groups are readily expressed by a null, H_0, and alternative H_a hypothesis as:

$$\begin{aligned} H_0&: \mu_i = \mu_k \textit{ for all } 1 \le i < k \le I \textit{ v.s.} \\ H_a&: \mu_i \ne \mu_k \textit{ for at least one pair } i \textit{ and } k, 1 \le i < k \le I. \end{aligned} \quad (4)$$

Under the null hypothesis H_0, all groups have the same mean. If H_0 is rejected, post hoc analyses are followed to determine the groups that have different group means. We focus on the hypothesis in equation (4) for overall group difference below, but the same conclusions apply to post hoc pairwise group comparisons as well.

ANOVA uses F-tests for testing the null hypothesis of no group difference in equation (4). This omnibus test is defined by elements of a so-called ANOVA table:

Source	df	Sum of squares (SS)	Mean squares (MS)
Groups	$K-1$	$SS(R) = \sum_{k=1}^{K} n_k \left(\bar{Y}_{k+} - \bar{Y}_{++}\right)^2$	$MS(R) = SS(R)/(K-1)$
Error	$N-K$	$SS(E) = \sum_{k-1}^{k} \sum_{i=1}^{n_k} \left(Y_{ki} - \bar{Y}_{k+}\right)^2$	$MS(E) = SS(E)/(N-K)$
Total	$N-1$	$SS(Total) = \sum_{k=1}^{k} \sum_{i=1}^{n_k} \left(Y_{ki} - \bar{Y}_{++}\right)^2$	

In the above table,

$$\bar{Y}_{k+} = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k}, \bar{Y}_{++} = \frac{\sum_{k=1}^{I} \sum_{i=1}^{n_m} Y_{ki}}{N}, N = \sum_{k=1}^{K} n_k, 1 \le k \le K,$$

where $\bar{Y}_{k+} = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k}$ is the sample mean of group k, $\bar{Y}_{++} = \frac{\sum_{k=1}^{I} \sum_{i=1}^{n_k} Y_{ki}}{N}$ is the sample mean of the entire sample (grand

mean), N is the total sample size from all groups, $\sum_{i=1}^{n} A_i$ denotes sum of all A_k, and $\sum_{k=1}^{I}\sum_{i=1}^{n} A_{ki}$ denotes sum of A_{ki} over both indices k and i. The terms $SS(R)$, $SS(E)$ and $SS(Total)$ are called the regression, error and total sum of squares, respectively, and $MS(R)$ and $MS(R)$, obtained by dividing $SS(R)$ and $SS(E)$ by their respective df, are called the mean regression and the mean error sum of squares mean. The three sums of squares are related as:

$$SS\left(Total\right) = SS\left(R\right) + SS\left(E\right),$$

where $SS\left(R\right)$ and $SS\left(E\right)$ form a partition of $SS\left(Total\right)$.

Under the null of no group difference, the sample group mean $\bar{Y}_{k+} = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k}$ will be close to the sample grand mean, $\bar{Y}_{++} = \frac{\sum_{k=1}^{I}\sum_{i=1}^{n_k} Y_{ki}}{N}$, in which case $SS\left(R\right)$ will be close to 0 and $SS\left(E\right)$ will be close to $SS\left(Total\right)$. Otherwise, $SS\left(R\right)$ will be different from 0 with its magnitude reflecting differences between the group means and $SS\left(E\right)$ will account for a smaller portion of $SS\left(Total\right)$ in this case. Thus, relationships between $SS(R)$ and $SS(E)$ indicate if there is evidence of differences in group means. By normalising the two by number of groups and sample size, the F-test is defined by their normalised counterparts, $MS(R)$ and $MS(R)$.

Under the null hypothesis H_0, the F statistic, or the ratio $\frac{MS(R)}{MS(E)}$, follows the F distribution:

$$F = \frac{MS(R)}{MS(E)} \sim F_{K-1,N-K} \quad (5)$$

where $F_{I-1,N-I}$ denotes the F distribution with K–1 (numerator) df and N–K (denominator) df. As a smaller $MS(R)$ relative to $MS(E)$ indicates evidence supporting the null and vice versa, this is consistent with the fact that a larger value of the F statistic leads to rejection of the null and vice versa.

As the F-test is derived under ANOVA in equation (3), validity of the F distribution for the F statistic in equation (5) depends on (A) normality and (B) homoscedasticity. Under both assumptions, the outcomes Y_{ki} do follow a normal distribution for each group, and all groups have the same variance. As noted earlier, the extant literature has largely focused on normality with little discussion on homoscedasticity. Next, we use MC simulation to examine the performance of the F-test under departures from each assumption.

SIMULATION STUDY

To evaluate the performance of the F-test when either normality or homoscedasticity is violated, we use MC simulated data to create multiple samples from an ANOVA. By repeatedly testing the null of no group mean difference using each simulated sample and comparing the per cent of times when the null is rejected with a prespecified type I error, we can see how each assumption impacts the performance of the F-test. We start with specifying an ANOVA from which samples will be drawn using MC simulation.

For brevity and without loss of generality, we consider three groups with a common mean μ and group size n, in which case the ANOVA in equation (3) simplifies to:

$$Y_{ki} = \mu + \varepsilon_{ki}, \varepsilon_{ki} \sim N\left(0,\ \sigma^2\right), 1 \le i \le n, 1 \le k \le 3. \quad (6)$$

We set $\mu = 0$, $\sigma^2 = 1$ and consider testing the null of no group difference,

$$\begin{aligned} H_0&: \mu_1 = \mu_2 = \mu_3 = 0, \\ H_a&: \mu_1 \neq 0 \ or \ \mu_2 \neq 0 \ or \ \mu_2 \neq 0. \end{aligned} \quad (7)$$

Thus, the null hypothesis H_0 is true for the ANOVA in equation (6). If we set type I error level to, say, α=0.05, and repeatedly test the null hypothesis using data Y_{ki} sampled from the ANOVA in equation (6) using MC simulation, we expect to reject the null H_0 5% of the times. We will refer to per cent of times when the null is rejected through repeated testing from multiple samples (MC simulated samples in the current setting) as empirical type I errors.

For evaluating the performance of the F-test under non-normal distribution, we simulate the error terms ε_{ki} in equation (6) from (A) a centred and rescaled χ^2 with 1 df and (B) a centred and rescaled Weibull distribution with both shape and scale equal to 1, so the resulting distributions in both cases have mean 0 and variance 1. In this case, the simulated Y_{ki} have the same mean (μ=0) and variance (σ=1) across the three groups. However, as the Y_{ki} no longer follow the normal distribution, the F-test in equation (5) may not follow the F distribution, in which case we may not reject the null H_0 5% of the times if we set the nominal level as α=0.05.

To evaluate the impact of homoscedasticity on type I errors, we consider two scenarios: (a) all three groups have different variances, and (b) one group has a different variance than the other two groups:

$$\text{Scenario (a)}: \sigma_1^2 = 1, \sigma_2^2 = 3k\sigma_1^2, \sigma_3^2 = 9k\sigma_1^2, k = 1\ 10\ 100,$$

$$\text{Scenario (b)}: \sigma_1^2 = 1, \sigma_2^2 = 3k\sigma_1^2, \sigma_3^2 = 3k\sigma_1^2, k = 1\ 10\ 100.$$

In each scenario above, we consider three sets of variances indexed by k (=1, 10, 100). As differences between the variances become larger as k varies from 1 to 10 to 100, this setting will show if increased degree of heteroscedasticity will have a larger effect on type I errors.

To reduce the sampling variability, we set MC sample size to M=1000. We consider the three sample sizes (A) n = 10, (B) n = 100 and (C) n = 1000 to see if and how sample size will affect the performance of the F-test.

Shown in table 1 are empirical type I errors for testing the null of no group mean difference by the F-test from 1000 MC simulated outcomes under the ANOVA in equation (6) with no violation and violation of each of the two assumptions. If none of the assumptions is violated, empirical p values are quite close to their nominal counterparts. When normality is not met, there is downward bias in empirical p values for sample size n=10, but the bias seems to have disappeared when n=100 and n=1000. When homoscedasticity is violated, however, there is clearly upward bias in

Table 1 Empirical type I errors for testing the null of no difference in group mean across three groups in equation (7) from the *F*-test based on 1000 Monte Carlo simulated samples from the ANOVA in equation (6) under (A) no violation of normality and homoscedasticity, (B) violation of normality and (C) violation of homoscedasticity for sample size n=10, 100 and 1000 with nominal type I error α=0.05 and α=0.01

	n=10		n=100		n=1000	
	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05
No violation						
Normality and homoscedasticity	0.007	0.043	0.014	0.057	0.01	0.051
Violation of normality						
χ^2 distribution	0.005	0.031	0.011	0.053	0.006	0.054
Weibull distribution	0.009	0.037	0.009	0.053	0.01	0.054
Violation of homoscedasticity						
Scenario (i), k=1	0.019	0.056	0.013	0.061	0.032	0.073
Scenario (i), k=10	0.025	0.061	0.014	0.063	0.034	0.076
Scenario (i), k=20	0.025	0.059	0.013	0.063	0.034	0.077
Scenario (ii), k=1	0.012	0.043	0.01	0.054	0.016	0.059
Scenario (ii), k=10	0.018	0.05	0.012	0.06	0.022	0.066
Scenario (ii), k=20	0.017	0.051	0.012	0.057	0.022	0.065

ANOVA, analysis of variance.

both scenarios. Also, the bias becomes larger as the degree of heteroscedasticity increases under both scenarios. The first scenario has a larger bias than the second, which is expected since it has a higher degree of heteroscedasticity than the second scenario. Moreover, unlike the case of violation of normality, the upward bias becomes larger as sample size increases from 10 to 100 to 1000. An increased empirical type I error means the likelihood of rejecting the null hypothesis is higher than the specified nominal level, yielding a higher rate of false positives.

DISCUSSION

In this report, we examined the performance of the *F*-test in one-way ANOVA using MC simulation when data violate the (1) normality and (2) homoscedasticity assumption. Our simulation results show that although there is downward bias in type I errors for extremely small sample size (n=10), it virtually disappears for moderate (n=100) and large (n=1000) sample sizes. The diminishing bias as sample size increases is not a coincidence and is actually the working of a mechanism known as the central limit theorem in statistics.[3] Thus, unless sample size is small (eg, less than n=50), normality can be ignored when applying ANOVA.

In contrast, the *F*-test is more sensitive to departures from homoscedasticity. Our simulation study results show that type I errors are upwardly biased when groups have different variances. Moreover, unlike violation of normality, the amount of bias persists and actually increases as sample size increases. Thus, one needs to pay close attention to this assumption when applying ANOVA. If violation of this assumption is suspected either by formal testing[4] or rule of thumb (eg, the ratio of the largest to the smallest variance is larger than 2), one may need to apply other tests for comparing group means. For example, the KW test may be used for inference. However, as this test generally has lower power, one may opt for modern alternatives such as the semiparametric models, which provide inference for linear and more general regression models without imposing any distributional distribution.[3] Unlike the KW test, semiparametric models use original continuous outcomes and thus provide more power than the KW test. We will compare the KW test and applications of semiparametric models to ANOVA in a future article.

Contributors KY conceived the initial idea, searched the literature on related topics, performed analyses involving Monte Carlo simulations and assisted in manuscript preparation. JT conceived the initial idea, participated in the discussion of the statistical problems and their implications in biomedical research, wrote parts of the manuscript and helped finalise the manuscript. TC researched the statistical issues, directed simulation studies, drafted parts of the manuscript and finalised the manuscript. All authors provided critical feedback and helped shape the research, analysis and the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Kruskal WH, Wallis WA. Use of ranks in One-Criterion variance analysis. *J Am Stat Assoc* 1952;47:583–621.
2 Choi W, Lee JW, Huh M-H, *et al*. An algorithm for computing the exact distribution of the Kruskal–Wallis test. *Commun Stat Simul Comput* 2003;32:1029–40.

3 Tang W, He H, Tu XM. Applied categorical and count data analysis, 2012. Chapman & Hall/CRC, FL. Available: http://www.crcpress.com/product/isbn/9781439806241

4 Bartlett MS. Properties of sufficiency and statistical tests". *Proceedings of the Royal Statistical Society, Series A* 1937;160:268–82.

Kun Yang graduated from Huazhong Agricultural University in 2016. He has been reading for a master's degree of bioinformatics in Huazhong Agricultural University since 2016. Now he is studying and working at University of California, San Diego, as a visiting student. His research interests include bioinformatics, biostatistics and machine learning.