Behaviour Research and Therapy 98 (2017) 19–38

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers

Andy P. Field [a, *], Rand R. Wilcox [b]

[a] School of Psychology, University of Sussex, Falmer, Brighton, BN1 9QH, UK
[b] Department of Psychology, University of Southern California, 618 Seeley Mudd Building, University Park Campus, Los Angeles, CA 90089-1061, USA

ARTICLE INFO

Article history:
Received 11 August 2016
Received in revised form
15 May 2017
Accepted 22 May 2017
Available online 26 May 2017

Keywords:
Robust statistical methods
Assumptions
Bias

ABSTRACT

This paper reviews and offers tutorials on robust statistical methods relevant to clinical and experimental psychopathology researchers. We review the assumptions of one of the most commonly applied models in this journal (the general linear model, GLM) and the effects of violating them. We then present evidence that psychological data are more likely than not to violate these assumptions. Next, we overview some methods for correcting for violations of model assumptions. The final part of the paper presents 8 tutorials of robust statistical methods using R that cover a range of variants of the GLM (*t-tests*, ANOVA, multiple regression, multilevel models, latent growth models). We conclude with recommendations that set the expectations for what methods researchers submitting to the journal should apply and what they should report.

1. Overview

The general linear model (GLM), which is routinely used in clinical and experimental psychopathology research, was once thought to be robust to violations of its assumptions. However, based on hundreds of journal articles published during the last fifty years, it is well established that this view is incorrect. Moreover, modern methods for dealing with the violations of these assumptions can result in substantial gains in power as well as a deeper, more accurate and more nuanced understanding of data. We begin with an overview of the key assumptions underlying the GLM. We then review various misconceptions about how robust the GLM is to violations of those assumptions and look at the effects that violations can have. We end the first section by looking at the evidence that psychological data, in general, are likely to violate the assumptions of the GLM.

In part 2 of the paper we overview a selection of ways to deal with violations of assumptions that fall under the headings of data transformation, adjustments to standard errors, and robust estimation. In the final part, we present 8 tutorials that use datasets relevant to this journal to show how to implement a selection of techniques (robust estimators for model parameters and standard errors) for designs common to this journal (comparing dependent and independent means, predicting continuous outcomes from continuous predictors and longitudinal designs).

2. The assumptions of the general linear model

2.1. Critical assumptions

Psychology researchers (generally) and those with interests in psychopathology (specifically) typically apply variants of the general linear model to their data. In this model, an outcome variable (Y) is predicted from a linear and additive combination of one or more predictor variables (X). For each predictor there is a parameter that is estimated from the data ($\hat{b}$) that represents the relationship between the predictor and outcome variable if the effects of other predictors in the model are held constant. There is a parameter (the constant, $\hat{b}_0$) to estimate the value of the outcome when all predictors are zero. The error in prediction is represented by the residual (ε_i), which is (for each observation, i) the distance between the value of the outcome predicted by the model and the value observed in the data (Eq. (1)). Model parameters (the $\hat{b}$s) are typically estimated using ordinary least squares (OLS) estimation, which seeks to minimize the squared errors between the predicted

* Corresponding author. School of Psychology, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH, UK.
E-mail address: andyf@sussex.ac.uk (A.P. Field).

http://dx.doi.org/10.1016/j.brat.2017.05.013
0005-7967/

and observed values of the outcome, or maximum likelihood (ML) estimation, which seeks to find the parameter values that maximise the likelihood of the observations.

$$\widehat{Y}_i = \widehat{b}_0 + \widehat{b}_1 X_{1i} + \ldots + \widehat{b}_n X_{ni} + \varepsilon_i \quad (1)$$

It is widely known that the general linear model is a flexible framework through which to predict a continuous outcome variable from predictor variables that can be continuous (often termed as 'regression' or 'multiple regression'), categorical (often referred to as 'ANOVA') or both (often referred to as 'ANCOVA'). Similarly, experimental designs containing repeated measures and longitudinal data are special cases of a multilevel linear model in which observations (level 1) are nested within participants (level 2). Despite the proliferation of terms that create artificial distinctions in the statistical models being applied, research designs that might, by many, be labelled as 'regression', 'ANOVA', 'ANCOVA', and multilevel models, are all variants of the linear model and, therefore, have a common set of underlying assumptions (see Cohen, 1968; Field, 2013; 2016, for tutorials).

The linear model has two main assumptions: (1) additivity and linearity, and (2) spherical residuals. The assumption of spherical errors implies that residuals are both independent and homoscedastic. This assumption is typically examined with respect to these two implications. Independent residuals are ones that are not correlated across observations. You would expect this assumption to be true when each observation comes from a different entity, but false when observations come from the same entities at different time points (e.g., longitudinal designs) or from different entities that share a context relevant to the outcome variable (e.g., clients being treated by the same clinician, or children taught by the same teacher). Correlation across residuals is known as *autocorrelation*. Homoscedastic residuals are ones that have the same variance for all observations. Residuals without this property are called *heteroscedastic*.

When using the general linear model researchers assume that Eq. (1) is a valid representation of the real-world process that they are trying to model. In short, they assume that the outcome variable is linearly related to any predictors and that the best description of the effect of several predictors is that their individual effects can be added together. As such, the assumption of additivity and linearity is the most important because it equates to the general linear model being the best description of the process of interest. If this assumption is not true then you are fitting the wrong model.

The assumptions of independent and homoscedastic residuals (i.e., spherical errors) relate to the Guass-Markov theorem, which states that when these conditions are met (and residuals have a mean of zero) then the linear model derived from OLS estimation will be a *best linear unbiased estimator*. In other words, it will be the unbiased linear estimator that has the least variance (i.e., is optimal).[1] 'Unbiased' means that the estimator's expected value for a parameter matches the true value of that parameter. The consequence of violating either of these assumptions is the same: the parameter estimates themselves remain unbiased, but are no longer optimal (that is, you can find estimates with lower variance). Furthermore, the formula for the variance of a parameter (b) assumes a constant variance so under heteroscedasticity this formula is incorrect. Consequently, estimates of the standard error of the parameter (which are based on the variance) are biased (Hayes & Cai, 2007). The presence of autocorrelation biases the standard errors of model parameters too.

Biased standard errors have important consequences for significance tests and confidence intervals of model parameters. For example, the test statistic, t, associated with a parameter estimate in the linear model is calculated using Eq. (2), from which a p-value is derived. If the standard error of the parameter is incorrect, then t (and the associated p) will be biased[2] and have poor power (Wilcox, 2010). Similarly, the bounds of a parameter estimates' confidence interval are constructed by adding or subtracting from the estimate the associated standard error multiplied by the quantile of a null distribution associated with the probability level assigned to the interval. For example, under normality and when the variance is known, and the goal is to compute a 95% confidence interval for the mean, the standard error of the sample mean is multiplied by 1.96, the 97.5 percentile of a standard normal distribution. Therefore, if the standard error is biased, the confidence interval will be too. Confidence intervals can be "extremely inaccurate" when the homoscedasticity assumption is violated (Wilcox, 2010).

$$t_{n-p} = \frac{\widehat{b}_{observed} - \widehat{b}_{expected}}{SE_{\widehat{b}_{observed}}} \quad (2)$$

2.2. Normality

An additional assumption that is often discussed in relation to the linear model is normality. There are three issues related to normality, the first of which is normality of residuals (the ε_i in Eq. (1)). Each case of data has a residual – the difference between the predicted and observed values of the outcome. If you inspected a histogram of these residuals for all cases, you would hope to see a normal distribution centred around 0. A residual of 0 means that the model correctly predicts the outcome value. Therefore, if the residual is zero (or close to it) for most cases, then the error in prediction is zero (or close to it) for most cases. If the model fits well, we might also expect that very extreme over- or underestimations occur rarely. A well-fitting model then would yield residuals that, like a normal distribution, are most frequent around zero and very infrequent at extreme values. This description explains what we mean by *normality of residuals*.

The Guass-Markov theorem does not assume normally-distributed residuals: even if residuals are not normally-distributed the OLS estimator will yield a model that is the best linear unbiased estimator (i.e., unbiased and optimal). In this respect, normality of residuals does not matter. If the residuals *are* normally distributed in the population, then the OLS estimator becomes the ML estimator (that is OLS and ML estimation yield identical estimates), and it will be the most accurate. That is to say, when residuals are not normally distributed, parameter estimates will be unbiased and optimal (with respect to minimizing the variance), but there may be classes of estimator (other than OLS) that are more accurate (Wilcox, 2010).

A simple example of this point is the (arithmetic) sample mean, which is an OLS estimator for the population mean. When the

[1] The Guass-Markov theorem shows that the OLS estimates of the slope and intercept are essentially a weighted mean of the outcome values. When homoscedasticity is met the OLS estimator minimizes the expected squared error relative to *other weighted means* that might be used. However, there are quite a few robust regression estimators outside of this class that result in smaller standard errors when dealing with an error term that is heavy-tailed, even under homoscedasticity (see Wilcox, 2017). The take home point is that, when using OLS, heteroscedasticity makes things worse relative to many modern robust methods.

[2] A test statistic is biased if the probability of rejecting the null is not minimized when the null is true.