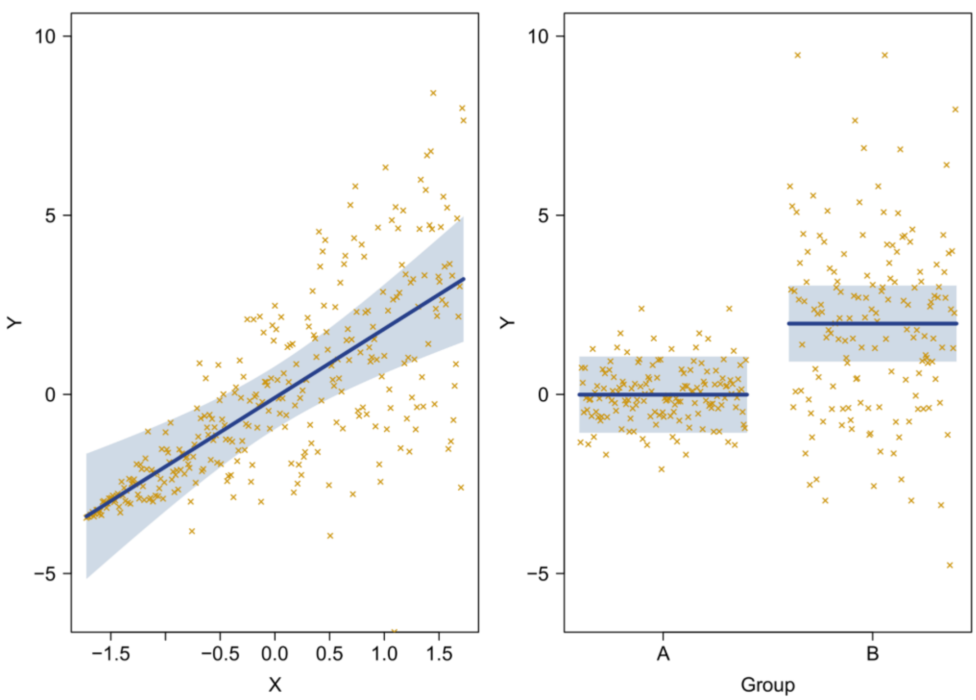
STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

Heteroscedasticity

E-mail distributed on 22-02-2023

Dear all,

An important assumption in *F*- and *t*-tests is homogeneity of variances, also known as homoscedasticity. That is, the variance of the outcome Y is assumed to be constant across levels of the predictor variables (e.g., across categorical groups and/or continuous covariates). This assumption is important to obtain valid standard errors (SE) for test statistics. When violated, the variance is said to be heteroscedastic (see attached examples), and corrections need to be applied to SEs.



How to test it?

For a simple one-way ANOVA, most researchers are familiar with Levene's test, available as `leveneTest` in package `car`. However, most software do not output the original Levene test but instead default to the robust modification proposed by Brown and Forsythe. A significant *p*-value for the Levene/BF test indicates a significant departure from equal variances (we reject homoscedasticity), and requires a corrected test. For models that are complex (e.g., multiple regression with categorical and continuous predictors), the Breusch-Pagan test is available and can be run on the model's residuals to diagnose heteroscedasticity. Package `lmtest` has an implementation with `bptest`. Once again, a significant *p*-value indicates significant departure from homoscedastic residuals. However, be warned that statistical tests on assumptions are generally not very reliable, sometimes breaking down in *both* small and large samples! Some authors therefore suggest that unequal variances should be assumed by default, as does R's `t.test` function.

How NOT to correct?

It is a common misconception that non-parametric tests can correct for heteroscedasticity, when in fact they do not. Rank- and permutation-tests also assume equal variances across groups and predictor levels. Correcting for a violation of variance assumptions always requires **additional parametric assumptions**, typically producing larger standard errors and lower (often fractional) denominator degrees of freedom (DDF).

How to correct?

For a simple one-way ANOVA, historical alternatives include the Brown-Forsythe *F** test, and the Welch test. The latter is widely known for the special case of 2 independent groups although its original formulation was an *F*-test. **Caution:** the Brown-Forsythe *F** test should not be confused with the Brown-Forsythe test for equal variances! Both the *F** and Welch test have been implemented in the R library onewaytests, as `bf.test` and `welch.test`. For example:

```
library(MASS)
library(car)
library(onewaytests)

model <- lm(Mg~Site, data=Pottery)
Anova(model)                                #DDF= 22
bf.test(Mg~Site, data=Pottery)              #DDF= 13.734
welch.test(Mg~Site, data=Pottery)           #DDF= 4.147

  Welch's Heteroscedastic F Test (alpha = 0.05)
-------------------------------------------------------------
  data : Mg and Site

  statistic  : 399.4961
  num df     : 3
  denom df   : 4.146845
  p.value    : 1.487798e-05

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

These functions will return *F*-tests with down-corrected fractional DDF, which can be very low for the Welch test compared to the original *F*-test! Maxwell and Delaney (2004) give an excellent comparison of strengths and weakness of the two corrected tests and the uncorrected test:

Data	*F*	*F**	*W*
Equal sample size			
Equal variances	Appropriate	Slightly conservative	Robust
Unequal variances	Robust, but liberal for strongly unequal variances	Robust, but liberal for very unequal variances	Robust
Unequal sample size			
Equal variances	Appropriate	Robust	Robust, but slightly liberal for very unequal variances
Large samples paired with large variances	Conservative	Robust, but slightly liberal for very unequal variances and sample sizes	Robust, but slightly liberal for very unequal variances and sample sizes
Large samples paired with small variances	Liberal	Robust, but slightly liberal for very unequal variances and sample sizes	Robust, but slightly liberal for very unequal variances and sample sizes

For more complex models, there is no similar correction. Instead, analysts either weight observations for their empirical residual variance (weighted least squares), or replace the ordinary least-squares (OLS) covariance matrix with a robust alternative. The oldest such estimator is known popularly as the White estimator, and later as HC0 in the expanded family of Heteroscedasticity-Concistent (HC) estimators. Today HC3 is recommended because of its good small-sample performance. It can be used as simple as:

```
model <- lm(Height~Volume+Girth, data=trees)
Anova(model, white.adjust="hc3")
```

Caution: Unfortunately, the DDF reported in the `Anova` table are from the uncorrected *F*-test. Studies have shown that this produces Type I error rates consistently above the nominal level (Imbens and Colesar, 2016; Dobriban & Su, 2018), especially in small samples! Although there has been some work on DDF calculation for HC-estimators, it remains an obscure subject with few software implementations. Fortunately, a method is available in package `dfadjust`, with the function `dfadjustSE` based on the HC2 estimator:

```
model <- lm(Height~Volume+Girth,data=trees)
Anova(model,white.adjust="hc2")
dfadjustSE(model)

             Estimate HC1 se HC2 se Adj. se    df p-value
(Intercept)     83.30  10.02  10.15   11.07 14.64  >0.001
Volume           0.58   0.22   0.22    0.25 11.15   0.024
Girth           -1.86   1.21   1.23    1.35 14.05   0.154
```

Once again the DDF are dramatically different from the nominal ones (28). Note also that each parameter has a different DDF correction! If you report heteroscedasticity-corrected analyses in your paper, I strongly recommend to report the adjusted DDFs and *p*-values.

Best,
Ben

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79

References

- Imbens, G.W., & Kolesár, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, *98(4)*, 701–712. https://doi.org/10.1162/REST_a_00552
- Long, J. S., & Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, *54(3)*, 217–224. https://doi.org/10.2307/2685594
- Maxwell, S.E., & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd Edition)*. Mahwah, NJ: Larwence Erlbaum Associates.