Heteroscedastic ANOVA breakdown for lm/lmer

E-mail distributed on 24-03-2025

Dear all,

In a previous post I have discussed simple *t*-tests and *F*-tests with a correction for unequal variances (a.k.a., heteroscedasticity), for example the Brown-Forsythe test and Welch *F*-test. Both of these tests are available in the R package `onewaytests` (Dag, Dolgun & Konar, 2018), with the eponymous drawback that they only allow a single grouping variable. For the general multiple regression case, there is the family of so-called HC-estimators, which replace the ordinary least squares covariance matrix of the parameters with a "robust" alternative, to produce robust standard errors for the model parameters.[1] R package `car` (Fox & Weisberg, 2019) for example allows these HC estimators to be plugged into the `vcov` argument of its `Anova` function (with HC2 or HC3 recommended). In fact, many packages for model fitting support these custom replacements.

Unfortunately, heteroscedastic estimators have some practical drawbacks: **(a)** they require large samples to be reliable (e.g., $N > 500$), **(b)** they may not be applicable to less traditional models (e.g., multilevel regression), **(c)** there is no adjustment to the denominator degrees of freedom (DDF) of the *F*-test, and/or **(d)** they cannot perform omnibus tests on multiple parameters simultaneously. While some R packages have implemented DDF adjustments (`dfadjust`; Kolesar, 2024), they are restricted to single parameters and thus cannot be used for categorical effects with more than 2 levels. Ideally, we would like to have a function that performs a full ANOVA breakdown for our regression model with Type II heteroscedastic *F*-tests. For this posting, I am sharing exactly such a function, `AnovaHC`.

1. Function output

`AnovaHC` is basically a wrapper for the function `Wald_test` from the R package `clubSandwich` (Pustejovsky, 2024).[2] This package concerns the estimation of so-called "cluster robust" standard errors, which is a variation of HC estimators where heteroscedasticity is linked to a known cluster variable, as commonly occurs in repeated measures data (e.g., subjects, class rooms, cities). A CR estimator can be plugged into a fitted multilevel regression with the random intercept as its cluster variable, and produce parameter tests that are robust against a misspecified random effects structure. The following example is data from an unpublished study on fear regulation in virtual reality (VR), where participants walked over a virtual track once at 0 meter (control phase), and once at 50 meter (fear

[1] Sometimes called empirical standard errors

[2] What gives with the sandwich, you ask. This name refers to the general mathematical form of heteroscedastic estimators, where a matrix M is "sandwiched" by two inverted matrices B^{-1}, in the equation $\text{Cov}(\beta) = B^{-1}MB^{-1}$.

phase). Moreover, they were assigned to one of five regulation conditions to reduce their fear (placebo, reappraisal, remotivation, respiration control, muscle control). A basic multilevel model would be:

```
library(lmerTest)
library(clubSandwich)
fear <- read.table("https://drive.switch.ch/index.php/s/rVgtv1UJawhnY8g/
  download",header=TRUE,as.is=FALSE)

>    ID   phase height   condition gender age appr_danger mot_freeze feel_fear
> 1 DUJ    fear     50      muscle      F  22           5          4         4
> 2 DUJ control      0      muscle      F  22           0          0         0
> 3 DXL    fear     50     placebo      F  25           5          5         4
> 4 DXL control      0     placebo      F  25           0          1         1
> 5 ECI    fear     50 respiration      F  20           6          0         5
> 6 ECI control      0 respiration      F  20           0          0         1

model <- lmer(feel_fear~phase*condition+gender+age+(1|ID), data=fear)
```

Where we predict self-reported fear by PHASE and CONDITION, while controlling for gender and age. Let us compare the traditional multilevel ANOVA output with the cluster-robust output:

```
anova(model,type=2)
AnovaHC(model, random="ID")

> Type II Analysis of Variance Table with Satterthwaite's method
>                  Sum Sq Mean Sq NumDF DenDF  F value  Pr(>F)
> phase           260.741 260.741     1    82 135.7792 < 2e-16 ***
> condition        24.746   6.186     4    80   3.2215 0.01666 *
> gender            0.407   0.407     1    80   0.2121 0.64641
> age               1.787   1.787     1    80   0.9304 0.33766
> phase:condition  15.291   3.823     4    82   1.9907 0.10358
> ---
> Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> Heteroscedastic Type II ANOVA breakdown:
> CR2 estimator with Satterthwaite's correction for DDF
>
>  Cluster variable: ID
>
>            Effect test       Fstat     delta df_num  df_denom        p_val sig
> 1           phase  HTZ 136.6628018 1.0000000      1 81.730871 3.977335e-19 ***
> 2       condition  HTZ   3.1820079 0.9394494      4 46.545336 2.161332e-02   *
> 3          gender  HTZ   0.1856896 1.0000000      1 28.452785 6.697757e-01
> 4             age  HTZ   1.7975774 1.0000000      1  5.746923 2.305665e-01
> 5 phase:condition  HTZ   2.0457839 0.9413846      4 48.181051 1.027449e-01
```

Comparing the two outputs, we get fairly similar results, except for the reduced DDF in the heteroscedastic model. How does the function achieve this? Heteroscedastic tests on single model parameters can be obtained with the function `coef_test` from `clubSandwich`.

```
coef_test(model,vcov="CR2")

>                               Estimate    SE     t  DF(Satt) p(Satt)  Sig
> (Intercept)                       1.60 0.670  2.38     10.4  0.0371    *
> phasefear                         2.16 0.480  4.51     17.0  <0.001  ***
> conditionplacebo                 -0.13 0.431 -0.31     32.7  0.7544
> conditionreappraisal             -0.42 0.431 -0.99     34.4  0.3259
> conditionremotivation             0.06 0.478  0.12     32.7  0.9008
> conditionrespiration              0.54 0.518  1.04     32.9  0.3039
> genderM                          -0.17 0.411 -0.43     28.4  0.6698
> age                              -0.02 0.022 -1.34      5.7  0.2306
> phasefear:conditionplacebo        0.06 0.695  0.09     32.8  0.9220
> phasefear:conditionreappraisal   -0.44 0.636 -0.69     34.0  0.4900
> phasefear:conditionremotivation   1.30 0.666  1.95     32.8  0.0590    .
> phasefear:conditionrespiration    0.53 0.681  0.79     32.8  0.4342
```

This is like the usual summary output of parameter estimates, but with robust SEs and DDFs adjusted by Satterthwaite's correction for heteroscedasticity. Unfortunately, it does not tell us the omnibus effect of CONDITION, or its interaction with PHASE. For this purpose, the `Wald_test` function can do multi-parameter tests, e.g., for the interaction:

```
Wald_test(model, constraints=constrain_zero(c("phasefear:conditionplacebo",
  "phasefear:conditionreappraisal","phasefear:conditionremotivation",
  "phasefear:conditionrespiration")),
  vcov="CR2", test="HTZ")

> test Fstat df_num df_denom p_val sig
>  HTZ  2.05      4     48.2 0.103
```

Within `Wald_test`, we can specify which parameters to constrain to zero simultaneously, thus yielding an omnibus test, using CR2 for cluster-robust SEs, and Hotelling's T^2 as an approximation of the null distribution of the test statistic, as recommended by studies (Bell & McCaffrey, 2002; Imbens & Kolesar, 2016; Pustejovsky & Tipton, 2017). The same approach will work for any other effect in the model, with one important caution. That is, constraining a subset of parameters to be 0 while leaving all others unconstrained is a Type III ANOVA approach, and will not return marginal effects when higher-order effects remain in the model. To obtain the Type II equivalent, one can either refit the model and omit higher-order effects when testing lower-order effects, or one can recode all variables to be centered on their mean. For categorical variables, this requires sum-coding (-1/1) of dummy variables rather than treatment coding (0/1).

2. Generalization to lm models

The above works great for multilevel models but `Wald_test` does not extend automatically to ordinary regression models. When attempting to plug an `lm` object into the function it demands a

cluster variable to be specified. However, one can simply use `1:nrow(data)` instead, to obtain the HC2 equivalent of the CR2 estimator (Graham, Arai, & Hagströmer, 2016). `AnovaHC` will do this by default when entering `lm` objects, e.g., for the fear data in the fear condition:

```
model <- lm(feel_fear~condition+gender+poly(age,2), data=fear,
  subset=phase=="fear")
AnovaHC(model)

> Heteroscedastic Type II ANOVA breakdown:
> CR2 estimator with Satterthwaite's correction for DDF
>
>  Cluster variable: case-wise
>
>         Effect test      Fstat     delta df_num  df_denom     p_val sig
> 1    condition  HTZ 3.41811343 0.9394494      4 46.545336 0.0156756   *
> 2       gender  HTZ 0.04639164 1.0000000      1 28.452785 0.8310020
> 3 poly(age, 2)  HTZ 0.40618972 0.7369642      2  2.801764 0.7000017
```

This example also illustrates that `AnovaHC` can handle data subsetting and transformations inside the model formula (e.g., quadratic effect of age). The output now prints that clustering was handled on a "case-wise" basis.

3. Recommendations and cautions

When running an analysis, it is good practice to do a sensitivity check to see if the effects persist under more robust conditions. This could mean running non-parametric models, models reducing the influence of outliers, and/or models that permit heteroscedasticity as a back-up check. By default, this should also be a part of your analysis plan. The `AnovaHC` function facilitates this check by allowing heteroscedastic ANOVA breakdowns for both `lm` and `lmer` models, which makes it broadly applicable to most analyses in the social sciences. Nevertheless, some important cautions need to be observed when using this function:

- **Power and reliability:** You will note in the above output that the reduction in degrees of freedom under heteroscedasticity can be extreme. This means that the analysis can be substantially underpowered compared to the ordinary ANOVA. If you intend for a heteroscedastic test to be your main analysis, you should therefore anticipate a large sample to ensure reliability and adequate power.
- **Clusters:** If your data are clustered you should *always* use the cluster variable for `AnovaHC`, and never the case-wise clustering. For `lmer` models, using the cases would produce overly optimistic DDFs for *F*-tests.
- **Multiple clusters:** At present `AnovaHC` cannot handle more than one source of clustering.[3] For `lmer` models with multiple random intercepts, you will therefore need to choose one intercept for the output, or combine intercepts in the case of hierarchical data (e.g., classrooms combined with participants).
- **Paradoxical effects:** It may occur that `AnovaHC` produces strongly significant effects that were non-significant in the ordinary ANOVA. This is particularly likely for categorical variables with

[3] But see R package multiwayvcov for an approach (Graham, Arai, & Hagströmer, 2016)

many levels in small samples. Here as well, some caution is advised with running heteroscedastic tests in samples that are too small.

- **Centering:** `AnovaHC` reproduces Type II ANOVA by running a Type III ANOVA on centered data. This should be considered approximate, however, and works best when the data are perfectly balanced across all categorical levels (including subjects in `lmer` models). When unbalanced, `AnovaHC` is weighting categorical levels proportional to their frequency. For `lmer`s, large imbalances in repeated measures (e.g., some subjects have 100 measurements, others only 2) will distort naive centering of subject-level covariates. Ideally, such covariates are centered at the subject-level, not at the measurement-level. If your data has these characteristics and you wish to avoid biases you may want to "manually" run the Type II ANOVA breakdown using the `Wald_test` function from `clubSandwich`.

References

Bell, R.M., and McCaffrey, D.F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2), 169–181

Dag, O., Dolgun, A., Konar, N. (2018). onewaytests: An R package for one-way tests in independent groups designs. *The R Journal*, 10(1), 175–199. doi:10.32614/RJ-2018-022

Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression (Third edition)*. Sage, Thousand Oaks CA.

Graham, N., Arai, M., and Hagströmer, B. (2016). *multiwayvcov: Multi-Way Standard Error Clustering*. R package version 1.2.3, <https://CRAN.R-project.org/package=multiwayvcov>

Imbens, G. W., & Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4), 701–712. doi:10.1162/rest_a_00552

Kolesár M (2024). *dfadjust: Degrees of Freedom Adjustment for Robust Standard Errors*. R package version 1.1.0, <https://CRAN.R-project.org/package=dfadjust>.

Pustejovsky J (2024). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.5.11, <https://CRAN.R-project.org/package=clubSandwich>

Pustejovsky, J. E. & Tipton, E. (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business and Economic Statistics*, 36(4), 672–683. doi:10.1080/07350015.2016.1247004

--

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

```
#########################################################
## ANOVAHC
## Heteroscedastic Type II ANOVA for regression models
##
## Ben Meuleman, 2025
#########################################################

## REQUIRED PACKAGES
library(clubSandwich)
library(foreach)
library(insight)

## FUNCTION
AnovaHC <- function(model,random=NULL) {

  ## MODEL INFORMATION
  modinfo <- terms(model)
  data <- get_data(model)

  #predvars <- attr(modinfo,"term.labels")[attr(modinfo,"order")==1]
  predvars <- find_predictors(model)$conditional
  classes <- sapply(data[,predvars],is.numeric)

  options(contrasts=c("contr.sum","contr.sum"))
  tempdata <- data
  tempdata[,predvars[which(classes)]] <- scale(tempdata[,predvars[which(classes)]])

  tempmod <- lm(formula(modinfo),data=tempdata)
  effects <- attr(model.matrix(tempmod),"assign")
  parnames <- names(coef(tempmod))
  cluster <- if(is.null(random)) { random <- "case-wise" ; cluster <- 1:nrow(tempdata) } else {
cluster <- data[,random] }

  ## CYCLE THROUGH EFFECTS
  out <- foreach(i=1:max(effects),.combine="rbind") %do% {
    wtest <- Wald_test(tempmod, constraints = constrain_zero(parnames[which(effects==i)]), vcov =
"CR2", cluster=cluster, test = "HTZ")
    data.frame(Effect=attr(modinfo,"term.labels")[i],as.data.frame(wtest))
  }
  rm(tempdata,tempmod)
  out <- as.data.frame(out)
  out
  out$sig <- cut(out$p_val,breaks=c(0,0.001,0.01,0.05,0.1,1),labels=c("***","**","*",".",""))
  options(contrasts=c("contr.treatment","contr.treatment"))

  ## OUTPUT
  cat("Heteroscedastic Type II ANOVA breakdown:","\nCR2 estimator with Satterthwaite's correction
for DDF","\n\n","Cluster variable:",random,"\n\n")
  out
}

#########################################################
```