Diagnostics

Graphical investigations of assumptions violations are a primary tool for assessing model problems as statistical testing approaches can be underpowered or overpowered, making them difficult to unambiguously interpret without also considering magnitude of effect. So, that is what I will highlight here. Chapter 10 of Snijders and Bosker (2012) provide a nice overview of assumption tests that are available. I will illustrate only one statistical test here, a test that the within-group variance varies randomly across groups or is a function of the predictors in the model. Formulas for this test are available in Snijders and Bosker (2012) on pp. 159-160 and Raudenbush & Bryk (2002) on p. 264. Heterogeneity problems may arise because of omitted variables, omitted effects, outliers (individuals or groups), or non-normal data.

SPSS

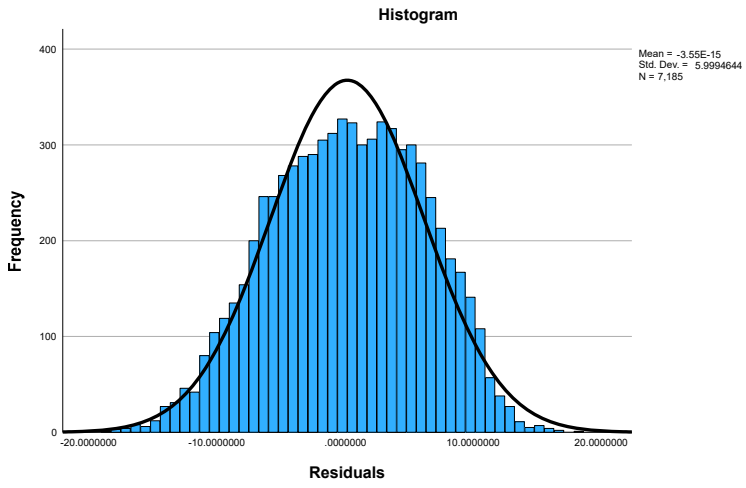
Plotting level-1 residuals. SPSS has limited ability to save values from the MIXED procedure. The SAVE subcommand can be used to save predicted values or level-1 residuals (as well as their standard errors and degrees of freedom) but no other useful diagnostic values from the model. Plots can be generated using these predicted and residual values, however. HLM outputs a number of values that could also be used for plots in SPSS (see illustrations below under the HLM section).

You can save the residuals and predicted values for Level-1 equations from the MIXED command. Here is example syntax the saves residual and predicted values using the HSB data set by adding them on the /SAVE subcommand.[1] Note that these values are added to the active data file so could be saved along with the data or called in a subsequent procedure.

```
MIXED mathach WITH ses
  /METHOD = REML
  /PRINT = SOLUTION TESTCOV
  /FIXED =  ses | SSTYPE(3)
  /RANDOM = INTERCEPT ses | SUBJECT(schoolid) COVTYPE(UN) SOLUTION
  /SAVE = PRED RESID(resid).
```

A histogram of level-1 residuals is useful for looking normality of residuals and outliers and is simple to obtain. Once the residual values are requested as in the model syntax above, they are available for subsequent procedures. There is default name used by SPSS but they can be renamed. Here I simply called them (resid) to illustrate but any name could be used in the parentheses.

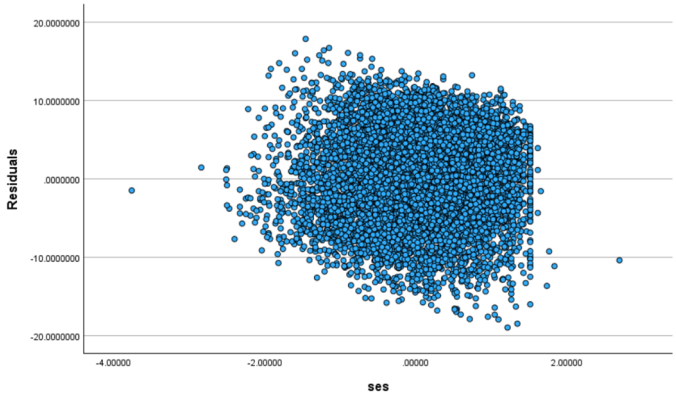
```
FREQUENCIES VARS=resid
     /HISTOGRAM=NORMAL
     /FORMAT=NOTABLE.
```



[1] Note: Standard errors and degrees of freedom can also be saved. FIXPRED PRED SEFIXP SEPRED DFFIXP DFPRED RESID are the possible keywords on the SAVE subcommand.

Below the scatter plot between the predictor (or predicted values) and the residuals can help identify nonlinearity, heteroscedasticity, or outliers. `ses` and `resid` are already available in the data set after using the /SAVE subcommand.

```
GRAPH   /SCATTERPLOT(BIVAR)= ses WITH resid.
```

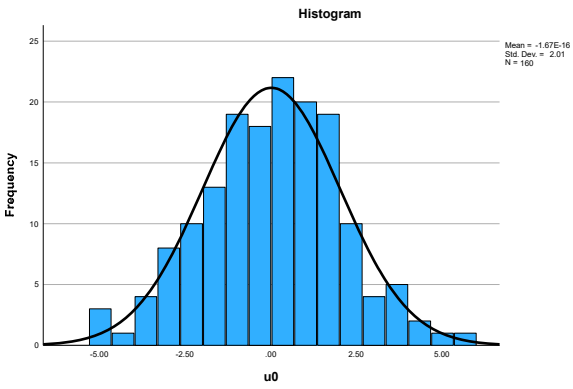


Plotting level-2 residuals. For level-2 residuals, SPSS has added an option to output what are called empirical best linear unbiased predictor (EBLUP), which are estimates for the intercepts and slopes in each group. These are model based values that are alternatively formulated (Henderson, 1950) but somewhat comparable to the EB estimates that HLM or R lme4 produces. The EBLUP estimates can be saved to a .sav data file and then read in for graphing. The first step is to add the /OUTFILE subcommand to the mixed model syntax. This requires that you also add SOLUTION to the /RANDOM subcommand—otherwise the command has the same specifications as above.

```
*random subcomannd must have SOLUTION to generate eblubs.
MIXED mathach WITH ses
  /METHOD = REML
  /PRINT = SOLUTION TESTCOV
  /FIXED =  ses | SSTYPE(3)
  /RANDOM = INTERCEPT ses | SUBJECT(schoolid) COVTYPE(UN) SOLUTION
  /SAVE = PRED RESID(resid)
  /OUTFILE = EBLUPS('c:\jason\spsswin\mlrclass\residuals.sav').
```

Then in a separate run (or below in the same syntax file), get the data file saved on the /OUTFILE subcommand. Below I calculate the intercept residuals by first calculating the average intercept (`gamma00`) and then the level-2 residual, `u0`. These residuals can be used on any subsequent procedure, but I illustrate with just a histogram to check the distribution.

```
get file='c:\jason\spsswin\mlrclass\residuals.sav'.

select if parameter eq 'Intercept'.
AGGREGATE
   /gamma00=MEAN(prediction).
compute u0=prediction - gamma00.

frequencies vars=u0
     /histogram=normal
     /format=notable.
```



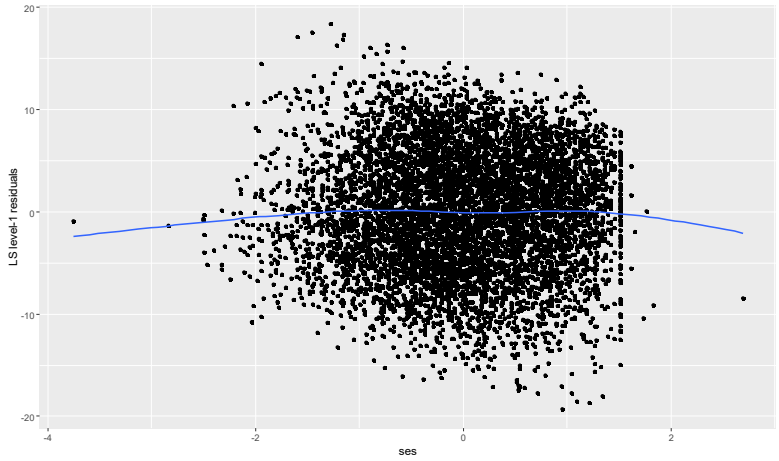
R

Here are a couple of diagnostic plots in R.

Level-1 residual plots. The `HLMdiag` package has several nice features which can be used in conjunction with `lme4`. You can print of view the residual data frame you create to see the variable names for the residuals (`.std.resid`), predicted values (`.fitted`), and other information obtained with `HLMdiag`.

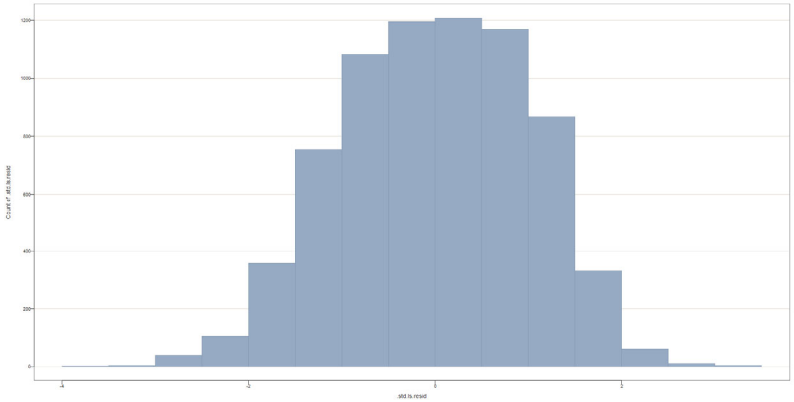
This runs the model in `lme4` then uses the residuals in a scatter plot. The residual information from `HLMdiag` must be converted to a data frame for the plots to work.

```
library(lme4)
fm1 <- lmer(mathach ~ ses + (ses|schoolid), data = mydata,REML=FALSE)

#residual plots from Loy & Hofman (2014)
library(HLMdiag)
resid1_fm1 <- as.data.frame(hlm_resid(fm1, level = 1, type = "LS", standardize = TRUE))

#level-1 residual plot
library(ggplot2)
qplot(x = ses, y = .std.resid, data = resid1_fm1,geom = c("point", "smooth")) + ylab("LS level-1
residuals")
```



Plot the level-1 residuals using the `lessR Histogram` function.

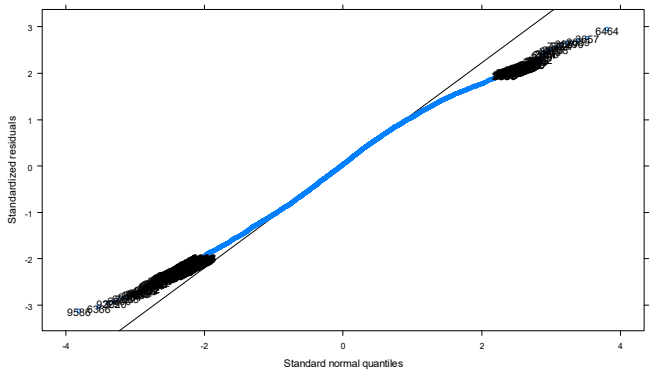
```
library(lessR)
Histogram(.std.ls.resid, data=resid1_fm1)
```



Normal probability plot of the level-1 residuals, which is just a plot of standardized residuals against the normal distribution.

```
require("lattice")
qqmath(model1, id=0.05)
if (require("ggplot2")) {
```

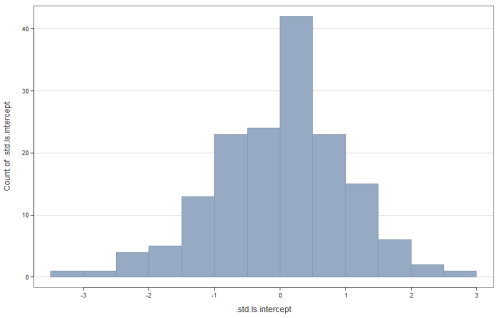
```
  ## we can create the same plots using ggplot2 and the fortify() function
  model1F <- fortify(mydata)
  ggplot(model1F, aes(.fitted,.std.ls.resid)) + geom_point(colour="blue") +
    facet_grid(.~sector) + geom_hline(yintercept=0)
  ## note: schoolids are ordered by mean mathach
  ggplot(model1F, aes(schoolid,.std.ls.resid)) + geom_boxplot() + coord_flip()
  ggplot(model1F, aes(.fitted,mathach))+ geom_point(colour="blue") +
    facet_wrap(~schoolid) +geom_abline(intercept=0,slope=1)
  ggplot(model1F, aes(ses,.std.ls.resid)) + geom_point(colour="blue") + facet_grid(.~sector) +
    geom_hline(yintercept=0)+geom_line(aes(group=schoolid),alpha=0.4)+geom_smooth(method="loess")
  ## (warnings about loess are due to having only 4 unique x values)
  detach("package:ggplot2")
}
```



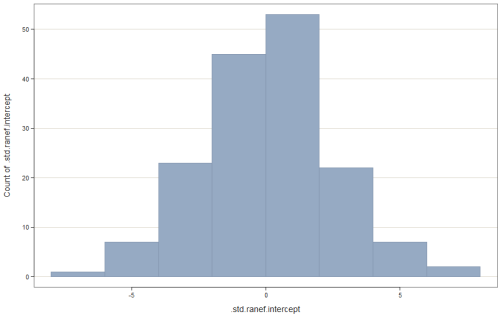
Level-2 residual plots. The HLMdiag package can be used for the level-2 residuals too. Here I just illustrate a histogram, but other plots could be generated with the values. By default, two types of residuals are generated, the least squares and the empirical Bayes. The plots are labeled `ls` for least squares, and `ranef` for empirical Bayes. Notice some shrinkage of the empirical Bayes compared with the least squares estimates!

```
#Histogram of level-2 residuals
Library(lme4)
library(HLMdiag)
fm2 <- lmer(mathach ~ ses + (1|schoolid), data = mydata,REML=FALSE)
resid2_fm2 <- as.data.frame(hlm_resid(fm2, level = 'schoolid', standardize = TRUE))
head(resid2_fm2)
library(lessR)
Histogram(resid2_fm2)
#to add density line
#Histogram(resid2_fm2,density=TRUE)
```

X Axis = std.ls.intercept for standardized least squares



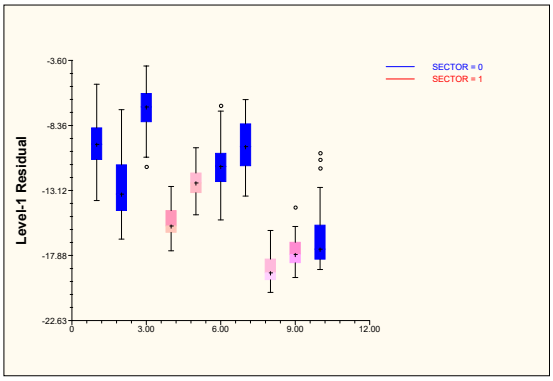
X Axis = std.ranef.intercept for the empirical Bayes



HLM

Plotting to explore distribution of level-1 residuals: linearity, heteroscedasticity, and outliers. A useful built-in graph for diagnosing assumption violations, such as nonlinearity, heteroscedasticity, or outliers is a scatter plot of the predicted and residuals. (*Graph Equations → level-1 residual v. predicted value*). I illustrate this graph using SPSS on the next page, so I do not reproduce the HLM version here.

HLM will produce some residual plots through the *Graph Equations* option under the *File menu*. For example, you can check whether residuals are equal at all values of X or appear to be equal in two groups (assumption of homogeneity) as shown in the following box-and-whiskers plot of Level-1 residuals. (Use of the sector variable—Z-focus variable—is optional).



Level-2 residuals and other plots. HLM also produces SPSS, SAS, SYSTAT, or STATA or ASCII files with residuals and other statistics that are useful for examining outliers and regression assumptions. Under the *Basic Settings* menu you can choose additional variables, file type, and location of the residual file (*Level-1 Residual File* and *Level-2 Residual File* buttons). These files allow for many other options for diagnostic graphs.

Here is a list of the variables in the Level-1 residual file by default (when I tested a model with SES as a Level-1 predictor):

l1resid = Level-1 residual; fitval = Level-1 fitted (predicted) value; sigma = square root of σ^2; ses = SES values; mathach = MATHACH values;

Here is a list of variables in the Level-2 residual file:

nj = number of cases per group; chipct = expected values on the chi-square distribution (used with Maholanobis distance for a Q-Q normal probability plot; mdist = Mahalonobis Distance of the Empirical Bayes coefficients from the fitted value (plot against chipct for normality assumption check); lntotvar = the natural logarithm of the total standard deviation within each unit; olsrsvar = the natural logarithm of the residual standard deviation within each unit based on its least squares regression; mdrsvar = the natural logarithm of the residual standard deviation from the final fitted fixed effects model; ebintrcp = Empirical Bayes intercept estimate; ebses = Empirical Bayes slope estimate for SES; olintrcp = Ordinary Least Squares intercept estimate; olses = Ordinary Least Squares slope estimate; fvintrcp = fitted (predicted) value of the intercept; fvses = fitted (predicted) value of the of the SES slope; ecintrcp = ; ecses = ; pv00, pv10, pv11, pvc00,pvc10, pvc11 = posterior variance and covariance estimates of τ_0^2 ,$\tau_{12,}$ and = τ_1^2.

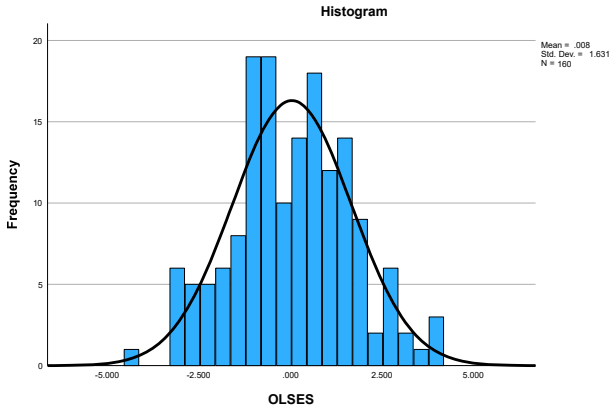
Here is a normal probability plot which is just a scatter plot of `chipct` against `mdist`.

```
get file='c:\jason\spsswin\mlrclass\resfil.sav'.
exe.
graph /scatterplot(bivar)=mdist with chipct.
```

Here is a level-2 residual histogram using the ordinary least squares residuals (`olses`). `ebses` is the empirical Bayes residuals.

```
output close *.
get file='c:\jason\hlm\mlrclass\resfil2.sav'.

frequencies vars=olses
```

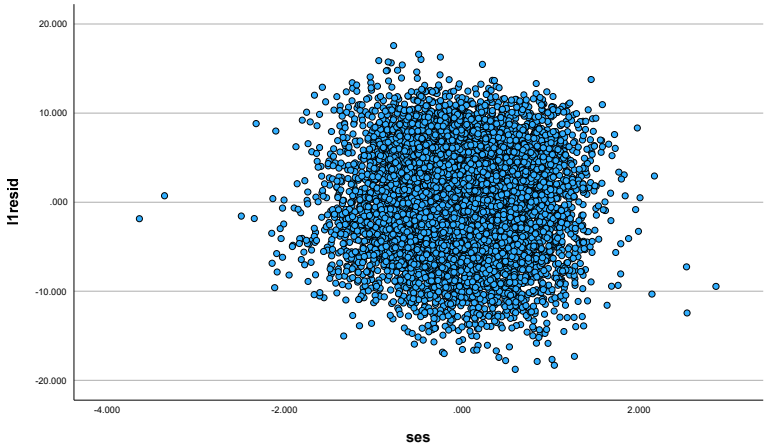
```
  /histogram=normal
  /format=notable.
```



\

Here is a level-1 scatter plot of residuals vs. predicted values (from Level-2 residual file) to examine heteroscedasticity, linearity, or outliers.

```
GRAPH
  /SCATTERPLOT(BIVAR)=ses WITH l1resid.
```



Bartlett test for homogeneity of variance. I illustrate the chi-square test of homogeneity used by Raudenbush and Bryk in the HLM package below. This test follows the Bartlett test (Bartlett & Kendall, 1946) and should be used only when data are normally distributed and there are 10 or more cases per group. I tested a model with SES (grand-centered) as a predictor of MATHACH in the HSB data. The rest of the output is the same as usual, so I have omitted it here. When you request the test (go to Other Settings -> Hypothesis Testing and then check the Test the Homogeneity of Level-1 variance box), a new, small section can be found in the output that looks like the following.

Test of homogeneity of level-1 variance

χ^2 statistic = 245.76576
degrees of freedom = 159
p-value = 0.000

A significant test indicates that the within-group variance is not equal across groups, as seems to be the case here. As with many variance tests, there can be considerable power when there are many groups, and, thus, it can be difficult to distinguish between circumstances when the violation is minor as opposed to major without using some supplemental information, such as visual exploration methods.

References

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics. 31* (2): 423–447.