Validity_practical

Based on Najera et al (Forthcoming): https://hectornajera83.github.io/book/preface.html

Dr. Hector Najera

18 July 2019

Criterion Validity

Criterion or predictive validity holds when there is a correlation between an scale and an alternative measure on the cause or effects of the construct of interest. In poverty research, this kind of validation has been used in the empirical literature (Gordon 2010; Guio, Gordon, and Marlier 2012; Nandy and Pomati 2015). We will again use our simulated multidimensional measure to illustrate how a validation exercise can be undertaken and to underline some issue researchers might find in practice.

```
library(MplusAutomation)

library(lavaan)

library(ggplot2)

#############################################

setwd("C:/Proyectos Investigacion/Lectures/Bristol Poverty Workshop 2019/Data")
```

Fitting a regression model to assess the relationship between a proposed index and an alternative measure is a common approach to assess predictive validity. To illustrate how this kind of validation works, we will use the simulated data ("Rel_MD_data_1_1.dat"). This data set contains the nine manifest variables (x1-x9) plus the two unreliable indicators(x10-x11). Three variables were simulated as alternative measures. One is a "perfect" measure of the resources available for each household in the sample. So in principle, this measure ranks the households according to their potential to fulfil their needs. The measure is expressed in monetary terms to facilitate the interpretation. Education years of the household head and occupation (skill scale) are two predictors of the living standards of the households. These two variables reflect the often common case where the survey was not designed with a validator in mind. We will use the variable "hh_members" to adjust the estimates.

```
library(plyr)
Rel_MD_1<-read.table("Rel_MD_data_1_1.dat")
Rel_MD_1$ds<-rowSums(Rel_MD_1[,c(1:9)])
colnames(Rel_MD_1)<-c("x1","x2","x3","x4","x5","x6",
                      "x7","x8","x9","x10","x11",
                      "resources","educ_yr","occupation",
                      "hh_members","class","ds")
Rel_MD_1[1:5,1:11]
##   x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11
## 1  1  1  1  1  1  0  0  0  0   0   0
```

```
## 2   0   0   0   0   0   0   0   0   0    0    0
## 3   0   0   0   1   0   0   0   0   0    0    0
## 4   1   1   0   0   0   0   1   0   0    0    0
## 5   1   0   0   0   0   0   0   0   0    1    1
Rel_MD_1[1:5,12:15]
##    resources educ_yr occupation hh_members
## 1   3276.687       6          4           5
## 2   7508.982      15          2           1
## 3   7183.707       8          5           2
## 4   1574.356       6          2           7
## 5   2210.297       9          5           4
```

One way to conduct the validation analysis consists in estimating the association between the manifest variables of our index with the validator. This can be simply done by fitting a series of regression models. Because deprivations are binary variables, we need to use a Generalised Linear Model (GLM) with the appropriate distribution. Relative Risk Ratios (RRR) are easier to interpret, so we will fit a Poisson model with log link to obtain the RRRs. Of course, there is no problem in estimating odd-ratios as here we are interested in looking at the association between variables.

In total we have 11 dependent variables (x1-x11) and, thus 11 models. In principle, x1-x11 resulted unreliable and should have been dropped from the analysis but we will keep them just to discuss some connections between reliability and validity. We will create a simple function `lms()` below to loop across the deprivation indicators. We will also transform the resources to get a more sensible metric.

```
Rel_MD_1$resources<-Rel_MD_1$resources*.01


lms<-function(index)

{

  fit<-glm(Rel_MD_1[,index] ~ Rel_MD_1$resources +

                              Rel_MD_1$hh_members,

          family=poisson(link="log"))

  exp(cbind(OR = coef(fit), confint(fit)))

}


coefs<-lapply(1:11,lms)


coefs[[1]]
```

We could check each of the outputs in list but it is easier to plot the RRRs of resources for each one of the 11 variables. We will not show the code here but one could just simply extract the coefficients and use to produce the graph. The coefficients are displayed with 95% confidence intervals in plot~. The null hypothesis in this model is that there is no relationship between resources and deprivation. For items x1-x9 we see that the difference seems to be different from zero and that the estimates are likely to be less than one. This suggests the higher the resources

and lower the chances of being deprived. This is in line with our expectation. For items x10 and x11, however, we found no relationship at all. This is an indication that both items are unreliable and invalid. This reinforces our previous suspicion that these two items are not useful to measure poverty.

```r
coefs<-lapply(coefs, function(x) unlist(x[2,]))

coefs<- as.matrix(matrix(unlist(coefs), nrow=length(coefs), byrow=T))

coefs<-data.frame(rbind(coefs[,c(1,2,3)]))


coefs$item <- rep(c("x1","x2","x3","x4","x5","x6",
                    "x7","x8","x9","x10","x11"),1)

coefs$var<-c(rep("Resources (*100)", 11))

coefs
```

We can simply plot the coefficients of each variable using the object `coefs` and `ggplot2()` as follows:

```r
library(ggplot2)

p<- ggplot(coefs, aes(x=item,y=X1)) + geom_point() +

  geom_errorbar(aes(ymin=X2, ymax=X3)) +

 theme_bw() + scale_y_continuous(trans = 'log10', limits = c(.9, 1.1))

p + facet_grid(. ~ var) + labs(y="Relative Risk Ratios") + geom_hline(yintercept=1,
linetype="dashed",

                color = "red", size=2)
```

This plot shows the Relative Risk Ratios for the resources variable, adjusted by the household size. Having more resources reduces the risk of being deprived of the item x, as expected.

Now we will go through the case of the lack of a validator. Most of the time researchers will lack a validator that was designed a priori. In these circumstances researchers need to use variables that predict poverty. Education attainment of the household head and occupation status are one of the two best predictors of poverty. We will rewrite our `lms()` function to fit a series of models using both education and occupation. All models adjusted by the household size. Again we will fit a GLM to obtain relative risks.

```r
lms<-function(index)

{

  fit<-glm(Rel_MD_1[,index] ~ Rel_MD_1$occupation +

                             Rel_MD_1$educ_yr +

                             Rel_MD_1$hh_members,

        family=poisson(link="log"))

  exp(cbind(OR = coef(fit), confint(fit)))

}


coefs<-lapply(1:11,lms)

## Waiting for profiling to be done...
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
coefs[[1]]
##                           OR      2.5 %     97.5 %
## (Intercept)         0.6866791 0.5503763 0.8554189
## Rel_MD_1$occupation 1.0531859 1.0377362 1.0688189
## Rel_MD_1$educ_yr    0.9206710 0.9051449 0.9364577
## Rel_MD_1$hh_members 1.0163837 0.9939741 1.0391996
coefs<-lapply(coefs, function(x) unlist(x[2:3,]))
coefs<- as.matrix(matrix(unlist(coefs), nrow=length(coefs), byrow=T))
coefs<-data.frame(rbind(coefs[,c(1,3,5)],coefs[,c(2,4,6)]))


coefs$item <- rep(c("x1","x2","x3","x4","x5","x6",
                    "x7","x8","x9","x10","x11"),2)
coefs$var<-c(rep("Occupation (Skill) scale", 11), rep("Education years", 11))
coefs
##           X1        X2        X3 item                 var
## 1  1.0531859 1.0377362 1.0688189   x1 Occupation (Skill) scale
## 2  1.0691741 1.0489816 1.0896804   x2 Occupation (Skill) scale
## 3  1.0781765 1.0510322 1.1058950   x3 Occupation (Skill) scale
## 4  1.0590273 1.0433546 1.0748872   x4 Occupation (Skill) scale
## 5  1.0721107 1.0517452 1.0927967   x5 Occupation (Skill) scale
## 6  1.0993761 1.0713526 1.1280058   x6 Occupation (Skill) scale
## 7  1.0610757 1.0448192 1.0775352   x7 Occupation (Skill) scale
## 8  1.0728834 1.0512460 1.0948827   x8 Occupation (Skill) scale
## 9  1.1084601 1.0802115 1.1373136   x9 Occupation (Skill) scale
## 10 1.0001156 0.9817638 1.0187286  x10 Occupation (Skill) scale
## 11 1.0087886 0.9905182 1.0273171  x11 Occupation (Skill) scale
## 12 0.9206710 0.9051449 0.9364577   x1         Education years
## 13 0.8977545 0.8780443 0.9178932   x2         Education years
## 14 0.8758993 0.8500184 0.9025356   x3         Education years
```

```
## 15 0.9311522 0.9153057 0.9472680   x4          Education years
## 16 0.9052537 0.8851807 0.9257679   x5          Education years
## 17 0.8979686 0.8708921 0.9258545   x6          Education years
## 18 0.9150086 0.8987985 0.9315038   x7          Education years
## 19 0.9029371 0.8816859 0.9246844   x8          Education years
## 20 0.8974289 0.8704803 0.9251801   x9          Education years
## 21 0.9848866 0.9650072 1.0051781   x10         Education years
## 22 0.9958716 0.9759022 1.0162522   x11         Education years
```

Once the models have been fitted, we could proceed to inspect the parameters. To inspect them we produce two plots shown in figure~@ref(fig:val2). The plot show the RRRs for both education and occupation adjusted by the household size. There is no evidence to support an association between items x10 and x11 and both predictors of poverty. In contrast, education and occupation predict an decrease and increase in the likelihood of being deprived of items x1-x9. On this basis we could conclude that our scale has criterion validity.

```
p<- ggplot(coefs, aes(x=item,y=X1)) + geom_point() +

  geom_errorbar(aes(ymin=X2, ymax=X3)) +

 theme_bw() + scale_y_continuous(trans = 'log10', limits = c(.8, 1.2))
p + facet_grid(. ~ var) + labs(y="Relative Risk Ratios") + geom_hline(yintercept=1,
linetype="dashed",

               color = "red", size=2)
```

This plot shows the Relative Risk Ratios for each item using two validators (adjusted by the total household members)

Construct Validity

Validity now is seen under a unified approach that looks at different aspects of the extent to which our scale can be interpreted as it is supposed to- a measure of poverty. Predictive validity might be a useful way to check the predictive validity at item-level. However, such kind of validation tells nothing about the structure of the measure. In section @ref{} we mention that modern poverty research should walk toward the specification of measurement models so that researchers make their assumptions better. We have mention that our scale is a higher-order scale with three dimensions, each one measured by three items. Construct validity concerns with the assessment of the structure of our scale. We will address several hypothesis about our scale:

- Are three dimensions a sensible way to arrange our indicators?
- Is a higher order factor present in our scale?
- Is the contribution to the explanation of the variance of each item equal or unequal?

We will focus on the first two question for now. To assess the validity of our measure we will use CFA to assess whether our measurement model is an adequate representation of poverty given these data. A CFA explicitly asks the question about the capacity of a model to reproduce the observed data. The first step, thus, consists in specifying our model. We have done already this in section @ref{} when we estimated the reliability statistics ω and ω_h. We will fit again the model using the `lavaan` R-package and Mplus. We will start with `lavaan` by specifying the MD_{model}. As can be appreciated we are assuming three factors (f1 to f3) and a higher order factor h. We are also stating that the indicators are manifest of one factor, i.e. we do not see x1 in f2 or f3.

Then we can simply use the `sem()` function and tell that our items are categorical. We will store the output in the fit object.

```
MD_model <- ' f1  =~ x1 + x2 + x3

              f2 =~ x4 + x5 + x6

              f3  =~ x7 + x8 + x9

               h =~ f1 + f2 + f3

'


fit <- sem(MD_model,
           data = Rel_MD_1,ordered=c("x1","x2","x3","x4","x5",
                                     "x6","x7","x8","x9"))
```

Once the model has been estimated, we can request the global statistics of fit of our model saved in the fit object. To extract the statistics we will use the function `fitmeasures()`. We will request the χ_2 test (absolute fit), the CFI and TLI values and RMSEA (relative fit). The p-value of the χ_2 test suggest that we reject the hypothesis that the model does not reproduces the observed data. That means that dimensions, classification of the indicators and the presence of the higher order factor do a good job in representing the structure of the data. CFI, TLI and RMSEA point in the same direction.

```
chisq<-fitmeasures(fit,fit.measures = c("chisq","df","pvalue"))

relfit<-fitmeasures(fit,fit.measures = c("tli","cfi"))

rmsea<-fitmeasures(fit,fit.measures = c("rmsea", "rmsea.ci.lower",

                                        "rmsea.ci.upper", "rmsea.pvalue"))

chisq

##  chisq     df pvalue

## 17.717 24.000  0.817

relfit

## tli cfi

##   1   1

rmsea

##         rmsea rmsea.ci.lower rmsea.ci.upper    rmsea.pvalue

##         0.000          0.000          0.007           1.000

#install.packages("semPlot")

library(semPlot)

## Warning: package 'semPlot' was built under R version 3.5.3

semPaths(fit, residuals=F,sizeMan=7,"std",edge.label.cex=1)
```

A joint assessment: Criterion and construct validity

Ideally, we would like to move toward a unified validation of scales. This involves examining both criterion and construct validity in the same model. Previously, we discussed that our full model looks like figure @ref(fig:valmdmimic). This is called a MIMIC model. This moves us from the

world of CFA into Structural Equation Modelling (SEM) but still the focus is on measurement and not so much on explanation. Again we will use `lavaan()` and Mplus to fit the model. In `lavaan()` we just need to create a new model that includes a new path. We would like to assess whether the higher-order factor (h) is associated with resources, adjusting by the total of household members. This can be simply achieved by adding a new line with a regression of h on the variables resources and hh_members. We fit and save the model in the fit object.

```
MD_model <- ' f1  =~ x1 + x2 + x3

              f2 =~ x4 + x5 + x6

              f3  =~ x7 + x8 + x9

               h =~ f1 + f2 + f3

               h ~ resources + hh_members

'


fit <- sem(MD_model,
          data = Rel_MD_1,ordered=c("x1","x2","x3","x4","x5",
                                    "x6","x7","x8","x9"))
```

Construct validity is assessed on the same terms. We will look at the overall fit of our model, which now know includes a new path, using the same statitsics: χ_2, CLI, TLI and RMSEA. We find that our measurement model still holds.

```
chisq<-fitmeasures(fit,fit.measures = c("chisq","df","pvalue"))

relfit<-fitmeasures(fit,fit.measures = c("tli","cfi"))

rmsea<-fitmeasures(fit,fit.measures = c("rmsea", "rmsea.ci.lower",
                                        "rmsea.ci.upper", "rmsea.pvalue"))

chisq

##  chisq    df pvalue

## 26.066 40.000  0.956

relfit

## tli cfi

##   1   1

rmsea

##        rmsea rmsea.ci.lower rmsea.ci.upper   rmsea.pvalue

##            0              0              0              1
```

Now we can check criterion validity by looking at the parameters of the regression part of our model. To extract the values of the parameters we will use the function `parameterEstimates()`, which is applied to the object fit. This is save in the slope object, which has all the estimated parameters in our model. For simplicity we will only show the slope h on resources by selecting the appropriate row. We observe that indeed there is a relationship between the factor and our parameters. What is the meaning of the reported value? The factor scores are presumed to follow a normal distribution. The higher the values of the factor, the higher the severity of poverty and vice versa. Therefore, we see that higher resources predict a decrease in the factor score, which is the expected behaviour in our measurement model.

```
slope<-as.data.frame(parameterEstimates(fit))
```

```
slope[13,]

##    lhs op      rhs          est             se       z pvalue    ci.lower
## 13   h  ~ resources -0.02989634 0.0009216571 -32.4376      0 -0.03170276
##      ci.upper
## 13 -0.02808993
#install.packages("semPlot")
library(semPlot)
semPaths(fit, residuals=F,sizeMan=7,"std",edge.label.cex=1)
```

With some code we could request the estimate of the slope as we did with the `lavaan()` model. However, we will stress the importance of visualising our measurement models by looking at the standardised parameters on a diagram of our model. Figure (fig:valsem1) shows the standardised estimates of our model. We can see that our validator *resources* predicts poverty - latent factor- (following Townsnend's theory representation in this case) and this constitutes a validation of our measure.This variable has the expected sign- an increase in resources reduces the latent severity of poverty.

```
knitr::include_graphics("val_sem_1.png")
```

This is a MIMIC model were a higher-order factor model loads into three dimensions and there is one path to examine criterion validity (resources and hh members)

Gordon, D. 2010. "Metodología de Medición Multidimensional de La Pobreza a Partir Del Concepto de Privación Relativa." In *La Medicion de La Pobreza Multidimensional En México*, edited by M. Mora, 401–98. El Colegio de México. CONEVAL.

Guio, A., D. Gordon, and E. Marlier. 2012. "MEASURING Material Deprivation in the Eu: Indicators for the Whole Population and Child-Specific Indicators." EUROSTAT.

Nandy, Shailen, and Marco Pomati. 2015. "Applying the Consensual Method of Estimating Poverty in a Low Income African Setting." *Social Indicators Research* 124 (3): 693–726. https://doi.org/10.1007/s11205-014-0819-z.