Tema 21: Tamaño de la muestra para modelos lineales

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
John W. Tuke

□ Peter B. Mandeville

Para mí, determinar el número de repeticiones para el análisis de modelos lineales es un problema cotidiano. Los modelos tienen una variable de respuesta continua y factores (ANOVA), covariables (regresión múltiple) o factores y covariables (ANCOVA) como variables explicativas. Si no se tienen los resultados de un estudio piloto, entonces no existe más remedio que utilizar las recomendaciones de Feinstein de un mínimo de 20 repeticiones por término.[1] Feinstein basa su recomendación sobre simulaciones del número de repeticiones para conseguir un buen estimador de los coeficientes de regresión parciales poblacionales.[2]

Si se tienen los resultados de un estudio piloto, entonces se puede efectuar el cálculo exacto. La hipótesis de nulidad sigue una distribución F, y la hipótesis alterna sigue una distribución F no central.[3,4] Una de los medidas de tamaño del efecto más común es la diferencia entre medias estandarizadas, d, definido como $d = \frac{M_t - M_c}{SD}$, donde *Mt* y *Mc* son las medias de los grupos de tratamiento y control, respectivamente, y *SD* es la raíz cuadrada de la media ponderada de las varianzas, *pooled standard deviation*. Un efecto pequeño se define como d=0.20, un efecto mediano como d= 0.50, y un efecto grande como d=0.80.[3,4]

Otro enfoque muy general es estimar el porcentaje de la varianza (PV), explicado por los varios efectos incluidos en el modelo. El porcentaje de varianza (PV) asociado con cada efecto en un modelo lineal que proporciona una medida muy general de los efectos de los tratamientos, si son grandes o pequeños. Hay un número de estadísticas específicas utilizadas en estimar PV, notablemente η^2 y R^2, que típicamente están en los contextos de análisis de varianza y regresión múltiple, respectivamente.[3,4]

La correspondencia entre los diferentes tamaños de los efectos, ES, son:[3,4]

Efecto	d	PV	r	f2
pequeño	0.20	0.01	0.10	0.02
mediano	0.50	0.10	0.30	0.15
grande	0.80	0.25	0.50	0.35

Se debe notar que Cohen's ,

$$f^2 = R^2/(1-R^2) = \eta^2/(1-\eta^2) = PV/(1-PV)$$

donde

$$\eta^2 = SS_{tratamiento} / SS_{total}$$

. Si se considera que los datos *birthwt*[6] son los resultados de un estudio piloto, se cargan los datos con *bwt* como la variable de respuesta.

```
library(MASS)
data(birthwt)
```

Se puede ver una descripción de los datos con:

```
help(birthwt)
```

Se declaran los factores:

```
birthwt$race <- as.factor(birthwt$race)
birthwt$smoke <- as.factor(birthwt$smoke)
birthwt$ht <- as.factor(birthwt$ht)
birthwt$ui <- as.factor(birthwt$ui)
```

y se define el modelo máximo:

```
mod <- lm(bwt~age+lwt+race+smoke+ptl+ht+ui+ftv,data=birthwt)
```

El siguiente paso es determinar si existe multicolinearidad, esto es, falta de independencia entre las variables explicativas con la función vif.[6] Si los factores de inflación de las varianzas, *vif*, son menores que 5, entonces no hay evidencia de multicolinearidad.

```
> library(car)
> vif(mod)
        GVIF Df GVIF^(1/2Df)
age   1.155069  1     1.074741
lwt   1.252135  1     1.118988
race  1.335314  2     1.074969
smoke 1.207031  1     1.098650
ptl   1.125009  1     1.060665
ht    1.087725  1     1.042941
ui    1.087875  1     1.043012
ftv   1.077053  1     1.037811
```

Se determina cuáles variables son significativas. En este ejemplo se utilizó un análisis de regresión múltiple escalada hacia atrás:

```
stp <- stepAIC(mod,trace=F)
```

se muestran los resultados utilizando ANOVA simultánea:

```
> Anova(stp)
Anova Table (Type II tests)
Response: bwt
          Sum Sq  Df F value    Pr(>F)
lwt       2674229   1  6.4093 0.0121981 *
race      6630123   2  7.9452 0.0004919 ***
smoke     4950633   1 11.8652 0.0007099 ***
ht        3584838   1  8.5918 0.0038100 **
ui        6353218   1 15.2268 0.0001341 ***
Residuals 75937505 182
—
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

y se nota que todos los términos son significativos. Se muestra el resumen:

```
> summary(stp)
Call:
lm(formula = bwt ~ lwt + race + smoke + ht + ui, data =
birthwt)
Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1842.14  -433.19    67.09   459.21  1631.03
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2837.264    243.676  11.644  < 2e-16 ***
lwt            4.242      1.675   2.532 0.012198 *
race2       -475.058    145.603  -3.263 0.001318 **
race3       -348.150    112.361  -3.099 0.002254 **
smoke1       -356.321    103.444  -3.445 0.000710 ***
ht1         -585.193    199.644  -2.931 0.003810 **
ui1         -525.524    134.675  -3.902 0.000134 ***
—
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 645.9 on 182 degrees of freedom
Multiple R-squared: 0.2404,  Adjusted R-squared: 0.2154
F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

y se nota que 24.04% de la variación en la variable de respuesta, *bwt*, está explicada por el conjunto de las variables explicativas, mientras que 75.96% de la variación no está explicada.

El *PV* de un término es:[4,5] $PV = \frac{t^2}{t^2 + df_{err}} = \frac{F}{F + df_{err}}$. Se puede calcular el *PV* de cualquiera de los términos con la función *PowerLM1*.

```
PowerLM1<- function(lm.obj,term){
  tmp <- anova(lm.obj)
  t2 <- tmp$»F value»[term]
  df2 <- lm.obj$df.residual
  cat(«PV:»,round(t2/(t2+df2),4),»\n»)
}
> PowerLM1(stp,1)
PV: 0.0434
> PowerLM1(stp,2)
PV: 0.0323
> PowerLM1(stp,3)
PV: 0.0764
> PowerLM1(stp,4)
PV: 0.0364
> PowerLM1(stp,5)
PV: 0.0772
```

Se puede utilizar la función *PowerLM2* para calcular el número de repeticiones para cualquier potencia, *power,* deseada, generalmente 0.8, pero nunca menor que 0.5.[4,5]

```
PowerLM2 <- function(PV,Power=0.80,Alpha=0.05){
  number <- power <- count <- 0
  i <- 1
  while(power <= Power){
    count <- count + 1
    power <- powerF(PV,i,alpha=Alpha)
    number <- i
    i <- i+1
  }
  tmp <- data.frame(number,round(power,4))
  names(tmp) <- c(«Number»,»Power»)
  print(tmp)
}
```

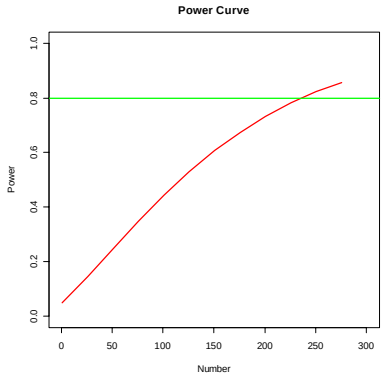
PowerLM2 utiliza la función *powerF* para calcular la distribución F no central:[7]

```
powerF <- function (PV, df2, df1 = 1, alpha = 0.05){
  ncp <- df2 * (PV/(1 - PV))
  power <- 1 - pf(qf(1 - alpha, df1, df2), df1, df2, ncp)
  return(power)
}
> PowerLM2(0.0434)
  Number  Power
1    175 0.8002
> PowerLM2(0.0323)
  Number  Power
1    238 0.8015
> PowerLM2(0.0764)
  Number  Power
1     97 0.8008
> PowerLM2(0.0364)
  Number  Power
```

```
1    210 0.8006
> PowerLM2(0.0772)
  Number  Power
1     96 0.8011
```

Finalmente, se puede levantar una curva de *power* de cualquier término entre *n1* y *n2,* con incrementos de *incr* con la función *PowerLM3.*

```
PowerLM3 <- function(PV,n1,n2,incr,Power=0.8){
  number <- power <- count <- 0
  i <- n1
  while(i <= n2){
    count <- count + 1
    power[count] <- powerF(PV,i)
    number[count] <- i
    i <- i+incr
  }
  tmp <- data.frame(number,round(power,4))
  names(tmp) <- c(«Number»,»Power»)
  print(tmp)
            plot(tmp$Number,tmp$Power,xlim=c(0,i/
1),ylim=c(0,1),type=»l»,
  col=»red»,xlab=»Number»,ylab=»Power»,main=»Power
Curve»,lwd=2)
   abline(h=Power,col=»green»,lwd=2)
  }
  > PowerLM3(0.0323,1,300,25)
    Number  Power
1       1 0.0508
2      26 0.1461
3      51 0.2489
4      76 0.3494
5     101 0.4439
6     126 0.5300
7     151 0.6067
8     176 0.6738
9     201 0.7317
10    226 0.7808
11    251 0.8221
12    276 0.8565
```



Se debe notar que, con la solución propuesta por Feinstein,[1] se necesitarán 20 repeticiones por variables o 100 repeticiones. Cálculo que será deficiente en aproximadamente 50%, en tres de las cinco variables explicativas (ver los resultados de *powerLM2*).

Referencias

1. Alvan R. Feinstein. Comunicación personal (carta del 5 noviembre de 2001).
2. John Concato and Peter Peduzzi and Theodore R. Holford and Alvan R. Feinstein. Importance of Events Per Independent Variable in Proportional Hazards Analysis I. Background, Goals, and General Strategy, Journal of Clinical Epidemiology, 1995, vol. 48, no. 12, pp. 1495-1501.
3. Jacob Cohen. (1988). Statistical Power Analysis for the Behavioral Sciences. Second edition. Lawrence Erlbaum Associates, Inc. Hillsdale, NJ, USA.
4. Kevin R. Murphy and Brett Myors. (2004). Statistical Power Analysis. A Simple and General Model for Traditional and Modern Hypothesis Tests. Second edition. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA.
5. David W. Hosmer and Stanley Lemeshow. (2000). Applied Logistic Regression. Second edition. Wiley

Series in Probability and Mathematical Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, NY, USA.

6. John Fox. I am grateful to Douglas Bates, David Firth, Michael Friendly, Gregor Gorjanc, Spencer Graves, Richard Heiberger, Georges Monette, Henric Nilsson, Derek Ogle, Brian Ripley, Sanford Weisberg, and Achim Zeileis for various suggestions and contributions. (2009). car: Companion to Applied Regression. R package version 1.2-14. http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/.
7. Thomas D. Fletcher. (2008). QuantPsyc: Quantitative Psychology Tools. R package version 1.3.