Análisis de encuestas basado en diseño y modelos muestrales

Una comparación entre métodos de inferencia aplicados al estudio de la vocación emprendedora en alumnos universitarios

Lic. Natacha Liseras
UNC, Abril de 2004

Miembros del Comité Asesor

Dr. Raúl Macchiavelli (UPR)

Dra. Mónica Balzarini (UNC)
Dr. José Vila Gisbert (UV)

Problema

- Estimar la proporción de alumnos universitarios con vocación emprendedora.
- Población objetivo:
  - Alumnos que cursan el último año
  - Economía, administración e ingeniería
  - Universidades públicas y privadas
  - Ciudad Autónoma de Bs. As. – Gran Bs. As. (Zona 1) y resto de la Provincia de Bs. As. (Zona 2)
- Encuestas obtenidas mediante un diseño por conglomerados (clusters) en dos etapas.

Notación

- y_{ij} es la respuesta del j-ésimo alumno del i-ésimo cluster:
  - $i = 1, \ldots, k$
  - $j = 1, \ldots, m_i$ → Clusters de tamaño desigual.
- Variable respuesta binaria:
  - $y_{ij} = 1$ si el alumno posee VE.
  - $y_{ij} = 0$ caso contrario.

Objetivos

- Comparar la aplicación de métodos de inferencia basados en diseño muestral y en modelos.
- Estimar la proporción de alumnos universitarios con vocación emprendedora en la población objetivo.
- Aplicar metodologías modernas de modelación en el área de las ciencias sociales.

MARCO CONCEPTUAL

- Inferencia basada en diseño muestral (clásica)
- Inferencia basada en modelos

Inferencia clásica

- Principio de representatividad.
- Estricta aleatoriedad en la recolección de los datos, respondiendo al diseño muestral.
- Buenos marcos muestrales.
- Muestreo por conglomerados → la potencia depende del número de clusters más que del tamaño de la muestra.

7

Inferencia basada en modelos

- La muestra puede ser obtenida con un proceso de selección aleatorio o no, en tanto pueda pensarse en su comportamiento como aleatorio.
- Con los datos recolectados se ajusta un modelo y se hace inferencia en base a los parámetros estimados.
- Modelos más comunes:
  - Lineal clásico (ML), lineal generalizado (MLG) o extensiones de los MLGs.
  - Efectos fijos y/o aleatorios.

8

Modelos lineales y modelos lineales generalizados (MLGs)

Modelo lineal

$$E(y_{ij}) = \mu_{ij} = x_{ij}'\beta$$

- Las y_{ij} son independientes.
- Las y_{ij} tienen distribución normal.

MLGs

$$g(\mu_{ij}) = x_{ij}'\beta$$
$$\eta_{ij} = x_{ij}'\beta$$

- Las y_{ij} son independientes.
- Las y_{ij} provienen de una misma distribución en la familia exponencial uniparamétrica. (MLGs)

9

Ejemplo de MLG para variables binarias

- El logit de la media o log de las chances es función lineal de los parámetros:

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 x_{ij}$$

- La pendiente se interpreta en términos de un cociente de chances (OR):

$$\log(OR) = \beta_1 \Leftrightarrow OR = \exp(\beta_1)$$

10

Observaciones binarias correlacionadas

- Ignorar la dependencia puede subestimar los errores estándares y hacer inconsistente la varianza de los estimadores.
- Primera solución propuesta: aplicar un modelo jerárquico beta-binomial. Problema: dificultad de incluir covariables para interpretar efectos.
- Alternativa: formular modelos que incorporen covariables, además de contemplar la dependencia entre las observaciones.

11

Modelo marginal

- Se modela la esperanza marginal de la variable respuesta mediante covariables y se especifica una estructura de dependencia entre las observaciones.
- La inferencia se refiere a promedios poblacionales (*population average inference*).
- Los estimadores obtenidos mediante GEE son consistentes y asintóticamente normales. La varianza de los estimadores se estima en forma robusta. (GEE)

12

Modelo mixto

- Valores compartidos de variables aleatorias no observables (U_i) generan dependencia entre las observaciones. U_i puede representar un efecto de cluster.
- Dado U_i, las observaciones del i-ésimo cluster se consideran independientes entre sí y se modelan con un MLG.
- La inferencia es específica para cada cluster (*cluster specific inference*).

13

Modelo mixto con verosimilitud completa

- Se modela la esperanza de la variable respuesta condicional a parámetros aleatorios específicos para cada cluster, incorporando el efecto aleatorio en el predictor lineal.
- Se debe asumir una distribución de probabilidad para los efectos aleatorios. Usualmente,

$$U_i \sim N\left(0, \sigma_u^2\right)$$

- Este modelo combina la información provista por las comparaciones entre y dentro de los clusters.

14

Modelo mixto con verosimilitud condicional

- Se modela la esperanza de la variable respuesta, considerando a los U_i como parámetros auxiliares (nuisance).
- Al condicionar sobre los U_i, los efectos aleatorios no aparecen en la función de verosimilitud.
- Este modelo sólo puede estimar parámetros asociados con covariables que varían dentro de los clusters.

15

Relación entre los modelos

Modelo marginal	Modelo mixto
Se modela: $Pr(y_{ij} = 1)$	Se modela: $Pr(y_{ij} = 1 \mid U_i)$
$\exp(\beta)$ describe el cociente de chances en la población.	$\exp(\beta)$ describe el cociente de chances para un cluster específico.
La dependencia intra-cluster se modela por separado.	La correlación intra-cluster se induce mediante efectos aleatorios (U_i).

16

APLICACIÓN DE LOS MÉTODOS DE INFERENCIA

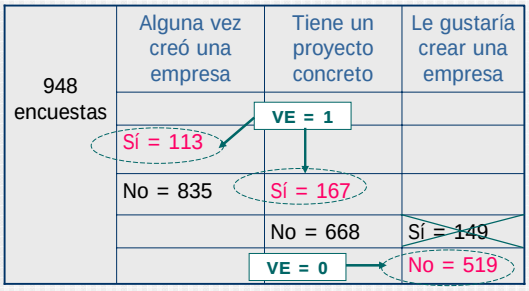
- Diseño muestral
- Inferencia clásica
- Inferencia basada en modelos
- Comparación de resultados

Diseño muestral

- Diseño por conglomerados en dos etapas:
  - Selección al azar de facultades (clusters).
  - Selección al azar de alumnos dentro de los clusters.
- Observaciones correlacionadas dentro de cada cluster:
  - Variables que influyen en la elección de la universidad.
  - Variables que influyen sobre todos los alumnos de una misma universidad.
  - Vínculos entre los alumnos de una misma universidad.

18

Definición de VE



- Se procesan 723 encuestas (261 VE=1 y 462 VE=0).

Inferencia clásica

- La media estimada es de **0.399** (MEDIA)
- La varianza estimada es de 0.0007 (VARIANZA)
- IC = (0.348, 0.452), amplitud = 0.104
- Para comparar proporciones correspondientes a distintas subpoblaciones, es necesario particionar la muestra.

Proporción de alumnos con VE por tipo de universidad

- La proporción de alumnos con VE es mayor en instituciones privadas que en públicas. (Z)

$$z = \frac{0.374 - 0.674}{\sqrt{0.001 + 0.004}} = -4.24$$

Públicas — Privadas

$$\Pr(z < -4.24) < 0.001$$

Proporción de alumnos con VE por género

- La proporción de alumnos con VE es mayor entre los hombres que entre las mujeres.

$$z = \frac{0.316 - 0.469}{\sqrt{0.001 + 0.001}} = -3.42$$

Mujeres — Hombres

$$\Pr(z < -3.42) < 0.001$$

Variables seleccionadas

Variable	Modalidad	% de 1
VE	1=sí; 0=no	36%
GENERO	1=hombre; 0=mujer	66%
OCUPADO	1=ocupado; 0=desocupado/inactivo	54%
ACTITUD	1=empresarial; 0=no	39%
VISION	1=favorable; 0=desfavorable	76%
RIESGO	1=propenso; 0=adverso	29%
CREATIV	1=alta; 0=media/baja	27%

- Los modelos se estimaron con SAS v. 8.2.
- No hay multicolinealidad (matriz de correlación, índices de condición, cocientes de chances marginales entre covariables).

Modelo marginal

- SAS PROC GENMOD
- Sobredispersión estimada = 1.03
- Estructuras de dependencia:
  - TYPE=IND
  - TYPE=EXCH → $Corr(y_{ir}, y_{is}) = \alpha$
  - LOGOR=EXCH
  - LOGOR=NEST1 subcluster=subcluster
  - LOGOR=LOGORVAR(CARRERA)
  - LOGOR=LOGORVAR(ZONA) (ALR)

Modelo marginal

```
proc genmod data=base descending;
class cluster genero ocupado actitud
      vision riesgo creativ;
model ve=genero ocupado actitud vision riesgo creativ
      / dist=bin link=logit type3;
repeated subject=cluster / type=exch;
output out=predicciones pred=predichos;
run;
```

25

Modelo marginal (TYPE=EXCH)

	Beta	e.e.	Valor p test de Wald	Valor p test de score	exp(β)
INTERCEPTO	−3.804	0.363	<0.001		0.02
GENERO	0.930	0.181	<0.001	0.027	2.23
OCUPADO	1.003	0.243	<0.001	0.022	2.87
ACTITUD	1.040	0.128	<0.001	0.006	2.93
VISION	1.503	0.329	<0.001	0.009	4.96
RIESGO	0.766	0.172	<0.001	0.006	2.26
CREATIV	0.520	0.232	0.037	0.100	1.62

$$\text{logit}\left(\hat{\mu}_{ij}\right) = \hat{\eta}_{ij} = -3.80 + 0.93GENERO_{ij} + 1.00OCUPADO_{ij} + \\ +1.04ACTITUD_{ij} + 1.50VISION_{ij} + 0.77RIESGO_{ij} + 0.52CREATIV_{ij}$$

Modelo mixto con verosimilitud completa

- SAS PROC NLMIXED

```
proc nlmixed data=base;
parms beta0=−3.9  beta1=0.8  beta2=1.0
beta3=1.1  beta4=1.6  beta5=0.8  beta6=0.5
sigma=0.05;
pred=beta0+beta1*genero+beta2*ocupado+beta3*
actitud+beta4*vision+beta5*riesgo+beta6*creativ+u;
prob=exp(pred)/(1+exp(pred));
model ve ~ binary(prob);
random u ~ normal(0, sigma*sigma) subject=cluster;
predict exp(pred)/(1+exp(pred)) out=predicciones;
predict u out=efectoscluster;
run;
```

27

Modelo mixto con verosimilitud completa

	Beta	e.e.	Valor p test de Wald	exp(β)
INTERCEPTO	−3.937	0.409	<0.001	0.02
GENERO	0.985	0.230	0.001	2.68
OCUPADO	1.035	0.193	<0.001	2.81
ACTITUD	1.085	0.186	<0.001	2.96
VISION	1.598	0.291	<0.001	4.94
RIESGO	0.787	0.197	0.002	2.20

$$\text{logit}\left(\hat{\mu}_{ij}/U_i\right) = \hat{\eta}_{ij} = -3.94 + 0.99GENERO_{ij} + 1.03OCUPADO_{ij} + \\ +1.09ACTITUD_{ij} + 1.60VISION_{ij} + 0.79RIESGO_{ij} + 0.55CREATIV_{ij} + U_i$$

28

Modelo mixto con verosimilitud condicional

- SAS PROC PHREG

```
data base;
time=2-ve;
run;
proc phreg data=base;
model time*ve(0)=genero ocupado actitud vision
      riesgo creativ / ties=discrete rl;
strata=cluster;
run;
```

29

Modelo mixto con verosimilitud condicional

	Beta	e.e.	Valor p test de Wald	exp(β)
GENERO	1.047	0.237	<0.001	2.85
OCUPADO	1.003	0.196	<0.001	2.73
ACTITUD	1.078	0.187	<0.001	2.94
VISION	1.586	0.297	<0.001	4.88
RIESGO	0.756	0.199	<0.001	2.13
CREATIV	0.572	0.209	0.006	1.77

$$\text{logit}\left(\hat{\mu}_{ij}/U_i\right) = \hat{\eta}_{ij} = 1.05GENERO_{ij} + 1.00OCUPADO_{ij} + \\ +1.08ACTITUD_{ij} + 1.59VISION_{ij} + 0.76RIESGO_{ij} + 0.57CREATIV_{ij}$$

30

Interpretación coeficiente GENERO

- Modelo marginal (TYPE=EXCH) → cocientes de chances promedio para la población:

  "Controlando por las restantes covariables, un hombre tiene 2.2 veces más chances de poseer VE que una mujer"

- Modelo mixto con verosimilitud completa → cocientes de chances específicos para cada cluster:

  "Controlando por las restantes covariables y para una facultad determinada, un hombre tiene 2.7 veces más chances de poseer VE que una mujer"

31

Regresión logística ordinaria

- SAS PROC GENMOD
- Supone observaciones independientes.
- Alternativa 1: Sólo efectos fijos de covariables en el predictor lineal
- Alternativa 2: Se adicionan clusters como efectos fijos

```
proc genmod data=base descending;
class cluster genero ocupado actitud
      vision riesgo creativ;
model ve=genero ocupado actitud vision riesgo
      creativ / dist=bin link=logit type3;
output out=predicciones pred=predichos;
run;
```

32

Regresión logística ordinaria

	Beta	e.e.	Valor p test de Wald	Valor p test del LR	exp(β)
INTERCEPTO	-3.919	0.352	<0.001		0.02
GENERO	0.801	0.202	<0.001	<0.001	2.28
OCUPADO	1.056	0.186	<0.001	<0.001	2.87
ACTITUD	1.074	0.181	<0.001	<0.001	2.93
VISION	1.602	0.285	<0.001	<0.001	4.96
RIESGO	0.818	0.191	<0.001	<0.001	2.27
CREATIV	0.484	0.202	0.016	0.016	1.62

$$\text{logit}\left(\hat{\mu}_{ij}\right) = \hat{\eta}_{ij} = -3.92 + 0.80GENERO_{ij} + 1.06OCUPADO_{ij} + \\ +1.07ACTITUD_{ij} + 1.60VISION_{ij} + 0.82RIESGO_{ij} + 0.48CREATIV_{ij}$$

33

Probabilidades estimadas

- Categoría modal sobre 32 tablas parciales:

GENERO	Hombre
OCUPADO	Ocupado
ACTITUD	Actitud no empresarial
VISION	Favorable
RIESGO	Adverso
CREATIV	Media-baja

$$g^{-1}(\hat{\eta}_{ij}) = \hat{\mu}_{ij} = \frac{\exp\left(\hat{\eta}_{ij}\right)}{1 + \exp\left(\hat{\eta}_{ij}\right)}$$

- La función de enlace inversa describe la relación entre el predictor lineal y la media de la variable respuesta.

34

Probabilidades estimadas bajo el modelo marginal (TYPE=EXCH)

Nivel covariables	Pr(VE=1)	RR
Categoría modal	0.409	
✓ Si es mujer	0.21	0.52
✓ Si no está trabajando	0.20	0.50
✓ Si posee actitud empresarial frente al desempleo	0.66	1.62
✓ Si posee una visión desfavorable de la actividad empresarial	0.13	0.33
✓ Si es propenso al riesgo	0.60	1.46
✓ Si posee alta creatividad	0.54	1.32

35

Probabilidades estimadas bajo el modelo mixto con verosimilitud completa

- Modelo mixto con verosimilitud completa (Ui)
- Categoría modal y efecto aleatorio nulo → **0.421**

Cluster	Pr(VE=1/Ui)	RR	Cluster	Pr(VE=1/Ui)	RR
U1	0.49	1.18	U8	0.45	1.07
U2	0.35	0.82	U9	0.25	0.60
U3	0.63	1.49	U10	0.38	0.90
U4	0.57	1.36	U11	0.42	1.00
U5	0.30	0.72	U12	0.52	1.23
U6	0.38	0.90	U13	0.37	0.88
U7	0.51	1.21	U14	0.33	0.78

36

Probabilidades estimadas bajo el modelo mixto con verosimilitud condicional

- Modelo mixto con verosimilitud condicional

Cluster	Pr(VE=1/U_i)	Cluster	Pr(VE=1/U_i)
U1	0.39	U8	0.46
U2	0.23	U9	0.13
U3	0.81	U10	0.27
U4	0.60	U11	0.27
U5	0.21	U12	0.45
U6	0.29	U13	0.27
U7	0.47	U14	0.22

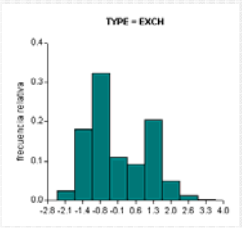
- ✓ Aprovechando la analogía con el análisis de datos de sobrevida:

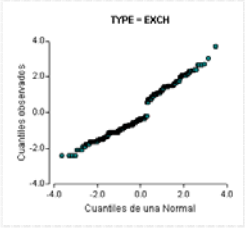
$$\beta_0 = \log\{-\log S_0(t_s)\}$$

37

Diagnóstico del modelo

- Residuos de Anscombe no resultan adecuados para variables respuesta binarias.





38

Poder predictivo del modelo

- Coeficiente de correlación entre valores observados y valores ajustados de la variable respuesta → 0.498 a 0.533.
- Tasa de error aparente con la que se calculan probabilidades condicionales por fila → z=0.4

Valor observado	Pr(VE=1) ≥ 0.4	Pr(VE=1) < 0.4	Total
VE = 1	Sensibilidad	Error I	100%
VE = 0	Error II	Especificidad	100%

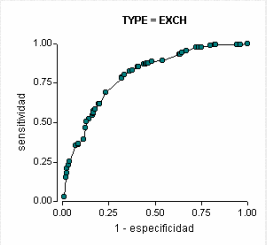
39

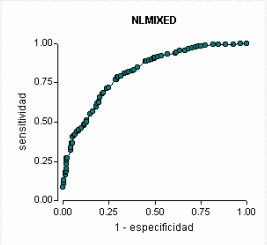
Poder predictivo del modelo

- Mayor sensibilidad → modelo marginal 78% (especificidad = 68%).
- Mayor especificidad → modelo mixto 78% (sensibilidad = 70%) y regresión logística ordinaria 80% (sensibilidad = 62%).
- Curvas ROC → sintetizan la relación entre sensibilidad y 1-especificidad para distintos valores de z.
- Cuanto mayor sea el área bajo la curva, mejor es la predicción.

40

Poder predictivo del modelo





- Áreas bajo la curva entre 0.80 y 0.82.

41

Poder predictivo del modelo

- Tasa de error por validación cruzada "leave-one-out" → el clasificador se construye con (n-1) datos (722) y se evalúa usando el dato restante. El proceso se repite n (723) veces.
- Modelo marginal y regresión logística ordinaria → la tasa de error CV es idéntica a la tasa de error aparente.
- Modelo mixto → la tasa de error CV es ligeramente mayor que la tasa de error aparente.

42

Comparación entre métodos de inferencia

Hombres vs mujeres con VE

- Los modelos incorporan información acerca de otras covariables, además del GENERO.
- Para comparar los resultados entre ambos métodos, las restantes covariables incluidas en el modelo deben reemplazarse por su valor esperado (proporciones muestrales en cada grupo).
- Los modelos utilizan los 723 datos en la estimación mientras que con la inferencia clásica se particiona la muestra.

43

Comparación entre métodos de inferencia

Proporción de hombres con VE=1

Inferencia	Proporción estimada	Amplitud IC
Inferencia clásica	0.469	0.137
Modelo marginal	0.414	0.143
Modelo mixto $U_i=0$	0.424	0.226
Modelo mixto U_i=prom	0.392	0.220
Regresión logística	0.386	0.102

- Media estimada con la inferencia clásica es mayor que la estimada con los modelos.

44

Comparación entre métodos de inferencia

Proporción de mujeres con VE=1

Modelo ajustado	Proporción estimada	Amplitud IC
Inferencia clásica	0.316	0.130
Modelo marginal	0.188	0.107
Modelo mixto $U_i=0$	0.184	0.169
Modelo mixto U_i=prom	0.165	0.154
Regresión logística	0.187	0.106

- Al disminuir el tamaño de muestra a 248, la inferencia clásica es menos precisa que el uso de modelos marginales.

45

Comparación entre métodos de inferencia

- Dos vías de clasificación: GENERO y OCUPADO.
- Inferencia clásica:
  - Se resiente al condicionar sobre más de una covariable y se hace más imprecisa que el uso de modelos en todos los casos
  - Se pierde información sobre algunos clusters en los que no se observa variabilidad suficiente para estimar la varianza.
- Los IC para la proporción más estrechos corresponden a los modelos marginales y de regresión logística ordinaria.

46

Comparación entre métodos de inferencia

Inferencia basada en modelos

- Desventaja:
  - Computacionalmente intensiva.
- Ventajas:
  - Gran flexibilidad de análisis.
  - No requiere conocer el tamaño total de los clusters.
  - Utiliza la totalidad de las observaciones para estimar los parámetros.

47

Comparación entre métodos de inferencia

Inferencia basada en diseño

- Ventaja:
  - No requiere cálculos iterativos.
- Desventajas:
  - Requiere un marco muestral completo y conocer el tamaño total de los clusters.
  - Al condicionar se reduce el tamaño muestral.
  - Es posible perder información para algunos clusters.
  - Fórmulas para un diseño complejo.

48

Comparación entre métodos de inferencia

Problemas de ambos métodos

- Relacionados con el cumplimiento de los supuestos.
- Naturaleza asintótica de la inferencia.

49

Conclusiones

- Respuestas binarias captadas por un muestreo por conglomerado en dos etapas → considerar la asociación entre observaciones del mismo cluster.
- Enfoque marginal → supuesto de equi-correlación resulta preferible si la correlación intra-cluster es baja.
- Enfoque mixto → modelo con verosimilitud completa permite estimar las probabilidades individuales.

50

Conclusiones

- Regresión logística ordinaria → ignora la dependencia intra-cluster y los e.e. estimados pueden ser inconsistentes.
- Inferencia clásica → aún disponiendo de información adicional, la reducción del tamaño muestral desalienta condicionar sobre más de una covariable.

51

Conclusiones

- Estrategia de análisis propuesta:
  - Definir correctamente la población objetivo.
  - Elegir el enfoque en base a las preguntas de investigación.
  - Seleccionar una muestra probabilística si los marcos de información son adecuados.
  - Formular y validar un modelo en base al cual inferir.
- La inferencia basada en modelos se adapta perfectamente a las ciencias sociales.

52

Futuras investigaciones

- Métodos de selección de modelos, medidas de bondad del ajuste y técnicas de diagnóstico para modelos marginales.
- Técnicas formales e informales de diagnóstico para variables de naturaleza binaria.
- Estimar la capacidad predictiva mediante bootstrap manteniendo la estructura de clusters (resampling within clusters).

53

FIN

Bibliografía principal

- Agresti, A. (2002) *Categorical data analysis.* 2nd ed. New York: Wiley.
- Brewer, K. (1999) "Design-based or prediction-based inference? Stratified random vs stratified balanced sampling". *International Statistical Review,* 67 (1): 35-47.
- Cochran, W. (1980) *Técnicas de muestreo.* México: CECSA.
- Diggle, P. *et al.* (2002) *Analysis of longitudinal data.* 2nd ed. New York: Oxford University Press.
- Fahrmeir, L. & Tutz, G. (2001) *Multivariate statistical modelling.* New York: Chapman & Hall.
- Liang,K. & Zeger, S. (1986) "Longitudinal data analysis using generalized linear models". *Biometrika*, 73 (1): 13-22.
- McCullagh, P. & Nelder, J. (1989) *Generalized linear models.* 2nd ed. New York: Chapman & Hall.

55

Bibliografía principal

- Pendergast, J. *et al.* (1996) "A survey of methods for analyzing clustered binary response data". *International Statistical Review,* 64 (1): 89-118.
- SAS Institute Inc. (1999) *SAS OnlineDoc* [en cd-rom], v.8. NC: SAS Institute Inc.
- Scheaffer, R. ; Mendenhall, W. & Ott, L. (1987) *Elementos de muestreo.* México: Grupo Editorial Iberoamérica.
- Zeger, S. & Liang, K. (1986) "Longitudinal data analysis for discrete and continuous outcomes". *Biometrics*, 42: 121-130.
- Zeger,S. ; Liang, K. & Albert, P. (1988) "Models for longitudinal data: a generalized estimating equation approach". *Biometrics*, 44: 1049-1060.

56

Media global

(*Sheaffer et al., 1987*)

$$\hat{\mu}_{..} = \frac{\sum_{i=1}^{k} M_i \hat{\mu}_{i.}}{\sum_{i=1}^{k} M_i}$$

- ✓ M_i = tamaño total del cluster.
- ✓ Media global ponderada.

(VOLVER)

57

Varianza de la media global

(*Sheaffer et al., 1987*)

$$Var\left(\hat{\mu}_{..}\right) = \left(\frac{K-k}{K}\right)\left(\frac{1}{k\overline{M}^2}\right)S_1^2 + \frac{1}{kK\overline{M}^2}\sum_{i=1}^{k} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{\hat{\mu}_{i.}\left(1-\hat{\mu}_{i.}\right)}{m_i - 1}\right)$$

0.0005

0.0002

- ✓ m_i = tamaño de la muestra por cluster.

58

Varianza entre medias de cluster

(*Sheaffer et al., 1987*)

$$S_1^2 = \frac{\sum_{i=1}^{k} M_i^2\left(\hat{\mu}_{i.} - \hat{\mu}_{..}\right)^2}{k-1}$$

(VOLVER)

59

Diferencia de proporciones

Ho) Las proporciones son iguales.

$$z = \frac{\hat{\mu}_{..}^{(1)} - \hat{\mu}_{..}^{(2)}}{\sqrt{Var\left(\hat{\mu}_{..}^{(1)}\right) + Var\left(\hat{\mu}_{..}^{(2)}\right)}} \sim N(0,1)$$

- ✓ Prueba asintótica.

(VOLVER)

60

Tests de hipótesis

- Estadístico de Wald : $z_{Wald} = \left(\frac{\hat{\beta}}{ASE\left(\hat{\beta}\right)} \right) \sim N(0,1)$
  - ✓ Utiliza el ASE no nulo, calculado a partir de la curvatura de la log-verosimilitud en su máximo.
- Estadístico de score: $\chi^2_{Score} = \left(\frac{L_0'}{ASE\left(L_0'\right)} \right)^2 \sim \chi^2_q$
  - ✓ Se basa en la pendiente y curvatura de la función de log-verosimilitud evaluadas en la hipótesis nula.

61

Tests de hipótesis

- Estadístico del cociente de verosimilitud:

$$\chi^2_{LR} = -2\log\left(\frac{\ell_0}{\ell_1}\right) = -2\left[\log(\ell_0) - \log(\ell_1)\right] = -2(L_0 - L_1) \sim \chi^2_q$$

- ✓ L_0 y L_1 representan las funciones de log-verosimilitud maximizadas sobre un conjunto de valores bajo la hipótesis nula y sobre un conjunto más amplio.
- ✓ $L_0 \leq L_1$ → una gran diferencia entre ambas conduce a rechazar la hipótesis nula.

62

Selección y diagnóstico del modelo

- Modelos mixtos → AIC, BIC, LR

$$AIC = -2\ell\left(\hat{\beta}\right) + 2p$$

$$BIC \text{ o } SBC = -2\ell\left(\hat{\beta}\right) + p\log(n)$$

- Residuos de Anscombe:

$$R_{Anscombe} = \frac{\left(\phi(y_{ij}) - \phi\left(\hat{\mu}_{ij}\right)\right)}{\hat{\mu}_{ij}^{1/6}\left(1-\hat{\mu}_{ij}\right)^{1/6}}$$

63

Modelos lineales generalizados (MLGs)

- Las y_{ij} son independientes y provienen de una misma distribución en la familia exponencial uniparamétrica.

$$f(y_{ij} \;/\; \phi) = \exp\left[\sum_{i=1}^{k}\sum_{j=1}^{m} h(y_{ij})\,\theta_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{b_{ij}(\theta_{ij})}{a(\phi)} + c_i\,(y_{ij}, \phi)\right]$$

Tiene la forma canónica si $h(y_{ij}) = y_{ij}$
$\mu_{ij} = E(y_{ij}) = b'(\theta_{ij})$
$V(\mu_{ij}) = b''(\theta_{ij})$
$Var(y_{ij}) = a(\phi)\, b''(\theta_{ij}) = a(\phi)\, V(\mu_{ij})$

64

Modelos lineales generalizados (MLGs)

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}'\beta$$

- ✓ g es una función conocida, monótona y diferenciable llamada función de enlace, que liga la media de la variable respuesta con las covariables.
- ✓ η_{ij} es un predictor lineal en los parámetros que indica la relación entre las covariables. (VOLVER)

65

Modelo marginal

- La esperanza marginal depende de las covariables a través de la función de enlace.

$$E(y_{ij}) = \mu_{ij} \;;\; g(\mu_{ij}) = x_{ij}'\beta$$

- La varianza marginal depende de la media marginal a través de la función de varianza.

$$Var(y_{ij}) = \phi V(\mu_{ij}) = \phi\mu_{ij}(1-\mu_{ij})$$

- La correlación entre las observaciones es función de la media marginal y de otros parámetros adicionales.

$$Corr(y_{ir}, y_{is}) = \rho(\mu_{ir}, \mu_{is}; \alpha)$$

66

Ecuaciones de estimación generalizadas (GEE)

- Extiende la cuasi-verosimilitud al análisis de observaciones dependientes *(Liang & Zeger, 1986; Zeger & Liang, 1986)*.

$$S_\beta(\beta,\alpha)=\sum_i^k\left(\frac{\partial\mu_i}{\partial\beta}\right)'V_i^{-1}(y_i-\mu_i)=0$$

- Incorpora una matriz de correlación de trabajo dentro de las ecuaciones de estimación, que se resuelven iterativamente.
- $R_i(\alpha)$ es una matriz m_i x m_i que refleja la estructura asumida.

Ecuaciones de estimación generalizadas (GEE)

Matriz de covarianzas de trabajo:

$$V_i=\phi A_i^{1/2}R_i(\alpha)A_i^{1/2}$$

- ϕ es el parámetro de sobredispersión.
- A_i = diag{μ_{ij} (1-μ_{ij})}, contiene las funciones de varianza correspondientes a cada observación.
- $R_i(\alpha)$ matriz de correlación de trabajo m_i x m_i que refleja la estructura asumida:

Ecuaciones de estimación generalizadas (GEE)

Estimador robusto de la varianza:

$$V_{\hat\beta}=\left[\sum_{i=1}^k\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)'\hat V_i^{-1}\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)\right]^{-1}$$

$$\left[\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)'\hat V_i^{-1}(y_i-\hat\mu_i)(y_i-\hat\mu_i)'\hat V_i^{-1}\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)\right]\left[\sum_{i=1}^k\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)'\hat V_i^{-1}\left(\frac{\partial\hat\mu_i}{\partial\beta}\right)\right]^{-1}$$

- La evidencia empírica de correlación queda en medio de las matrices de covarianzas propias del modelo.

Ecuaciones de estimación generalizadas (GEE)

- La matriz "sandwich" protege contra la elección de una estructura de correlación incorrecta. Las estimaciones son robustas a la elección de $R_i(\alpha)$.
- GEE ofrece estimadores consistentes y asintóticamente normales. Los intervalos de confianza son asintóticamente correctos.
- Sólo requiere que la función de enlace y el predictor lineal elegidos sean adecuados. (VOLVER)

Estructuras de correlación

- ✓ Independencia (TYPE=IND)

$$Corr(y_{ir},y_{is})=0$$

- ✓ Simetría compuesta (TYPE=EXCH)

$$Corr(y_{ir},y_{is})=\alpha$$

- ✓ Sin estructura (TYPE=UN)

$$Corr(y_{ir},y_{is})=\alpha_{rs}$$

Alternating logistic regression (ALR)

- Se parametriza la asociación en términos de logaritmos de cocientes de chances marginales.
- El algoritmo alterna entre actualizar el modelo para la media y actualizar los logaritmos de los cocientes de chances.
  - ✓ Intercambiables (LOGOR=EXCH)
  - ✓ Anidados a un nivel (LOGOR=NEST1)
  - ✓ Por cluster (LOGOR=LOGORVAR) (VOLVER)

Modelo mixto con verosimilitud completa

- Bajo el modelo de intercepto aleatorio, los valores realizados de U_i son una cantidad por la cual las mediciones en el i-ésimo cluster se ven incrementadas o disminuidas con relación a un cluster típico.

$$g\left[E\left(y_{ij}/U_i\right)\right] = \text{logit}\,\Pr\left(y_{ij}=1/U_i\right) = x_{ij}'\beta + U_i$$

- Los efectos fijos se estiman integrando sobre los efectos aleatorios.
- Los parámetros estimados son consistentes bajo el supuesto de que los efectos aleatorios son independientes de las covariables.

Modelo mixto con verosimilitud completa

- Con enlace logístico, suponiendo la distribución normal de los U_i y bajo la formulación de intercepto aleatorio, la función de verosimilitud resulta:

$$\prod_{i=1}^{k} \int \exp\left[\beta'\sum_{j=1}^{m_i} x_{ij}y_{ij} + U_i'\sum_{j=1}^{m_i} y_{ij} - \sum_{j=1}^{m_i}\log\left\{1+\exp\left(x_{ij}'\beta+U_i\right)\right\}\right] \times (2\pi)^{-1}|G|^{-q/2}\exp\left(\frac{-U_i'G^{-1}U_i}{2}\right)dU_i$$

- Se estiman los efectos fijos integrando o promediando sobre los efectos aleatorios.

Modelo mixto con verosimilitud condicional

- La inferencia es consistente aún si los efectos aleatorios no son independientes de las covariables.
- La función de verosimilitud condicional se expresa como:

$$L\left[\beta/\sum y_{ij}\right] = \prod_{i=1}^{k}\frac{\exp\left(\sum_{j=1}^{m_i} y_{ij}x_{ij}'\beta\right)}{\sum_{R_i}\exp\left(\sum_{\ell}^{y_{i\cdot}} x_{ij}'\beta\right)}$$

Proporciones por carrera

- La proporción de alumnos con VE no difiere por carrera.

$$z = \frac{0.405 - 0.373}{\sqrt{0.001+0.002}} = 0.58$$

Economía y adm. — Ingeniería

$$\Pr\left(|z| > 0.58\right) = 0.562$$

Covariables seleccionadas

GENERO	*Prevalencia de hombres entre los emprende-dores (Reynolds *et al.*, 2000).
OCUPADO	*Interés generado por la experiencia ocupacio-nal; *Aprendizaje en el puesto de trabajo actúa como incubadora (Côtè, 1991).
ACTITUD	*Condiciones del mercado laboral desfavora-bles; *Retornos superiores trabajando por cuenta propia (Henrekson & Rosenberg, 2001)
VISION	*Valoración favorable como opción de carrera
RIESGO	*Desafío de crear una empresa propia; *Menos adversos al riesgo son más propensos.
CREATIV	*Hobbies técnicos son más propensos (Scott & Twomey, 1988); *Realización de actividades creativas

Cocientes de chances marginales

Covariable	OR	IC exactos al 95%	
GENERO	2.16	1.52	3.09
OCUPADO	2.55	1.83	3.56
ACTITUD	3.84	2.75	5.35
VISION	7.39	4.31	13.34
RIESGO	2.93	2.07	4.13
CREATIV	1.44	1.01	2.04

Modelos marginales

- TYPE=IND → correlación baja.
- LOGOR=EXCH → cociente de chances común estadísticamente significativo.
- LOGOR=NEST1 → cocientes de chances estadísticamente significativos y prácticamente iguales.
- LOGOR=LOGORVAR(ZONA) → hay asociación entre las observaciones de la zona 1.
- LOGOR=LOGORVAR(CARRERA) → hay asociación entre las mediciones de ingeniería. (VOLVER)

Modelo mixto con verosimilitud completa

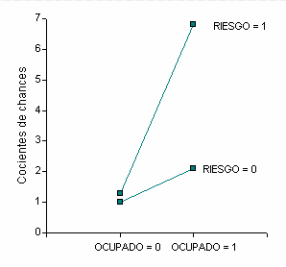
- Efectos de cluster → interceptos aleatorios. (VOLVER)

Cluster	U_i	$\beta_0 + U_i$	Cluster	U_i	$\beta_0 + U_i$
U1	0.299	–3.639	U8	0.119	–3.818
U2	–0.315	–4.252	U9	–0.775	–4.712
U3	0.836	–3.101	U10	–0.176	–4.113
U4	0.617	–3.320	U11	–0.007	–3.944
U5	–0.522	–4.459	U12	0.383	–3.554
U6	–0.178	–4.115	U13	–0.216	–3.959
U7	0.361	–3.576	U14	–0.404	–4.341

Interacciones dobles

- Son estadísticamente significativas:
  - GENERO * CREATIV
  - OCUPADO * RIESGO

	AIC	BIC
S/interac.	**757.1**	**762.2**
G*C	754.6	760.3
O*R	754.3	760.0
Ambas	751.2	757.5



Algunos resultados al comparar los modelos

- Los parámetros estimados difieren menos del 10%. Los efectos asociados a GENERO y CREATIV son menores bajo el enfoque marginal que bajo el enfoque mixto.
- Los IC para los β son, en promedio, más estrechos con el modelo marginal TYPE=EXCH y más amplios con el modelo mixto con verosimilitud completa.
- La amplitud relativa para estos IC es más variable bajo los modelos marginales.
- La estimación de los e.e. es robusta entre los modelos marginales y entre los modelos mixtos.
- OCUPADO se estima más eficientemente con los modelos mixtos y de regresión logística ordinaria, al contrario de GENERO y ACTITUD.

Probabilidades estimadas

Ajuste	Probabilidad estimada	Amplitud IC
TYPE=EXCH	0.409	0.176
NLMIXED - $U_i = 0$	0.421	0.259
NLMIXED - U_i prom.	0.390	0.253
Regresión logística ordinaria	0.387	0.158

Proporciones muestrales

	Hombres	Mujeres
OCUPADO	0.5116	0.6008
ACTITUD	0.4211	0.3427
VISION	0.7979	0.6855
RIESGO	0.3284	0.2298
CREATIV	0.2358	0.3387