UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMATICAS

E.A.P. DE ESTADÍSTICA

"Rotación de personal: Predicción con modelo de regresión logística multinivel"

TESIS
Tesis para optar el título profesional de Licenciada en Estadística

AUTOR
Quispe Millones, Sandra Giovana

Lima-Perú
2014

ROTACIÓN DE PERSONAL: PREDICCIÓN CON MODELO DE REGRESIÓN LOGÍSTICA MULTINIVEL

Quispe Millones, Sandra Giovana

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el título profesional de Licenciada en Estadística.

Aprobada por:

……..
Mg. Antonio Bravo Quiroz
Presidente del jurado

……..
Mg. Ana María Cárdenas Rojas
Miembro del jurado

……..
Mg. Wilfredo Domínguez Cirilo
Miembro asesor del jurado

Lima- Perú
Julio – 2014

FICHA CATALOGRÁFICA

QUISPE MILLONES, SANDRA GIOVANA

Rotación de personal: Predicción con modelo de regresión logística multinivel, (Lima) 2014.
vii, 54 p., 29,7 cm. (UNMSM, Licenciada, Estadística, 2014)
Tesis, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas 1. Estadística
I. UNMSM/FdeCM.

DEDICATORIA

A mi familia por su gran cariño, apoyo y confianza.

A mi abuelita Etelvina, por ser la mayor inspiración para creer en los sueños.

R E S U M E N

ROTACIÓN DE PERSONAL: PREDICCIÓN CON MODELO
DE REGRESIÓN LOGÍSTICA MULTINIVEL

QUISPE MILLONES, SANDRA GIOVANA

Julio – 2014

Orientador: Mg. Wilfredo Domínguez Cirilo
Título Obtenido: Licenciada en Estadística

Se analiza la rotación del personal en una empresa privada a través de un modelo de regresión logística de 2 niveles, buscando establecer la relación entre las características del trabajador, el área en que trabaja y la desvinculación laboral durante el periodo de prueba de 6 meses establecido por la empresa.

Se introducen conceptos de modelos lineales generalizados, regresión logística y modelos multinivel que sirven como base para describir los aspectos más relevantes de la regresión logística multinivel y sus ventajas frente a los modelos de un solo nivel.

Se analizó la desvinculación de los trabajadores (primer nivel) anidados en áreas de la empresa (segundo nivel), identificando la variabilidad existente entre las áreas ($\rho = 0.28$) y el perfil del desertor. Los resultados se comparan con los obtenidos con un modelo de regresión logística múltiple de un solo nivel, se encontraron diferencias respecto al aporte de las variables estado civil, escala remunerativa del puesto y beneficios adicionales brindados por el área.

PALABRAS CLAVES: ROTACIÓN DE PERSONAL, GRUPO OCUPACIONAL, REGRESIÓN LOGÍSTICA, MODELO MULTINIVEL.

ABSTRACT

TURNOVER: PREDICTION WITH MULTILEVEL LOGISTIC REGRESSION MODEL

QUISPE MILLONES, SANDRA GIOVANA

July - 2014

Adviser : Mg. Wilfredo Domínguez Cirilo
Obtained degree: Bachelor in Statistic

It is a turnover analysis in a private company through a multilevel logistic regression model of 2 levels seeking to establish the relationship between worker characteristics, the area in which they work and the termination of employment during the period of six months test set by the company.

Concepts are introduced of generalized linear models, logistic regression and multilevel models that serve as a basis for describing the most important aspects of the multilevel logistic regression and its advantages over single-level models.

The decoupling of workers (first level) nested in areas of the business (second level), are analyzed, identifying the variability between areas ($\rho = 0.28$) and the profile of the deserter. The results are compared with those obtained with a multiple logistic regression model of one level. Differences were found with respect to input variables marital status, remuneration scale and fringe benefits.

Concepts are introduced generalized linear models, logistic regression and multilevel

KEY WORDS : TURNOVER, OCCUPATIONAL GROUP, LOGISTIC REGRESSION, MULTILEVEL MODEL.

INDICE

INDICE DE GRÁFICOS Y CUADROS

CAPÍTULO I

INTRODUCCIÓN

La rotación de personal, considerada como la desvinculación del personal de una empresa por diferentes motivos, tiene un gran impacto en las organizaciones debido a que genera gastos innecesarios y un ambiente de tensión al no poder crear una estabilidad necesaria para el desarrollo óptimo de objetivos y metas. Considerando que las desvinculaciones pueden ocurrir en cualquier momento, se incrementa el riesgo de dejar posiciones claves, funciones y proyectos en marcha que serían difíciles de cubrir en el tiempo estimado. Por ello las empresas fijan un periodo de prueba de 6 meses en el cual se evalúa al trabajador y se espera recuperar la inversión realizada en selección, capacitación y otros gastos que impliquen su incorporación.

El estudio se realiza en base a datos reales de una empresa peruana de servicios con varias áreas de negocio, la cual por motivos de confidencialidad llamaremos LA EMPRESA.

La empresa en estudio, cuenta con diversos negocios siendo uno de los más importantes el traslado de valores y recaudación. En los últimos meses la rotación de personal operativo y administrativo se ha incrementado por lo que es necesario para los líderes de la empresa reconocer el perfil del trabajador que no está dispuesto a permanecer laborando y con ello dirigir y mejorar el proceso de selección.

Todas las áreas dentro de la organización no necesitan el mismo nivel de rotación por lo que se busca reconocer la variabilidad en cada una de ellas.

El objetivo de esta tesis es identificar las características de los trabajadores que se retiran de la empresa antes del término del periodo de prueba, de modo que se pueda mejorar el proceso de selección, implementar programas de retención y comprobar la diferencia de perfiles de trabajadores que cesan durante el periodo de prueba, de acuerdo al área donde trabajaron.

Construir un modelo que permita pronosticar la permanencia de un trabajador después de los 6 meses, comprobando si el área a la cual pertenece explica la decisión de permanecer en el puesto de trabajo.

El trabajo se organiza de la siguiente forma: En el capítulo II se revisan los fundamentos teóricos, se comenta la importancia de la rotación de personal en los el área de recursos humanos de la empresa, se explica la problemática, los principales conceptos como rotación de personal, periodo de prueba y el comportamiento de la rotación de personal en el Perú. Presentación de los principales conceptos para el desarrollo de los modelos multinivel. Se inicia describiendo el modelo lineal general, los modelos lineales generalizados y se pone énfasis en el modelo lineal generalizado para respuesta binaria, en este punto se explican conceptos importantes de la regresión logística simple y múltiple. Finalmente se describen los modelos multinivel y el modelo de regresión logístico multinivel.

En el capítulo III se mencionan los materiales y métodos utilizados, el tipo de estudio, el diseño de investigación y la definición de las variables, se describen los resultados del análisis estadístico.

Finalmente se presentan las conclusiones y recomendaciones del trabajo.

CAPÍTULO II

FUNDAMENTOS TEÓRICOS

1. Administración de recursos humanos y rotación laboral

La administración de recursos humanos es el manejo integral del capital humano, implica diferentes funciones, desde el inicio hasta el fin de una relación laboral buscando el compromiso y la productividad máxima de los trabajadores[1]. (Ver figura 1.1).

Gráfico 1.1. Dirección Estratégica de Recursos Humanos

Fuente: Dirección estratégica de Recursos Humanos: Gestión por competencias, Alles Martha , Edición Granica 2000

Entre sus roles se encuentran:

- Desarrollar esquemas salariales acordes con el valor de cada puesto para la organización y las condiciones de mercado.

- Implementar mecanismos de competencia variable que provean incentivos adecuados al desempeño.
- Definir cuáles son los puestos críticos para el desarrollo y la ejecución de la estrategia de la empresa.
- Identificar el perfil que requiere cada puesto y reclutar a las personas que encajan en él.
- Identificar las competencias que deben tener los trabajadores que necesita la organización.

Los recursos humanos se consideran estratégicos cuando permiten marcar la diferencia entre una organización y otra[1]. Por ello el plan de trabajo del área de recursos humanos debe estar alineado con la estrategia de la empresa, agregando valor y definiendo la visión, misión y valores de la organización.

1.1 Rotación laboral

Uno de los aspectos más importantes de la dinámica organizacional es la rotación de personal.

La rotación de personal se define de diversas formas, una de ellas es la fluctuación de personal entre una organización y su ambiente la cual es definida por el volumen de personas que ingresan y que salen de la organización[3]. Otros autores la definen como la desvinculación del personal por diferentes motivos. Especialistas en recursos humanos indican que se debe considerar todo movimiento del personal, es decir los cambios de puesto y área dentro de la misma organización como también los retiros.

Para fines de este estudio se definirá la rotación de personal como el retiro o abandono voluntario de un trabajador de la empresa en que labora.

La organización como un sistema abierto, se caracteriza por el incesante flujo de recursos humanos que necesita para poder desarrollar sus operaciones y generar resultados.

Entre los insumos que la organización importa y los resultados que exporta debe existir cierto equilibrio dinámico capaz de mantener las operaciones del proceso de transformación

en niveles controlados. Si los insumos son más voluminosos que las salidas, la organización tiene sus procesos de transformación congestionados o sus reservas de resultados almacenados y paralizados. Si por el contrario, los insumos son menores que las salidas, la organización no tiene recursos para operar las transformaciones y continuar la producción de resultados. Así, tanto la entrada como la salida de recursos debe mantener entre sí mecanismos homeostáticos capaces de auto regularse y garantizar así un equilibrio dinámico[3].

Las desvinculaciones de personal tienen que ser compensadas a través de nuevas admisiones para que se mantenga el nivel de recursos humanos en proporciones adecuadas para la operación del sistema.

Es conveniente diferenciar el número de trabajadores desvinculados por voluntad de la empresa de aquellos que renuncian a permanecer en la misma. No es lo mismo la salida de una persona competente que la organización quisiera retener, que la expulsión de un individuo por necesidad o conveniencia de la organización.

El índice de rotación es utilizado en la proyección de la demanda de fuerza laboral, además de constituirse en uno de los indicadores de la gestión de personal.

El índice de rotación está determinado por el número de trabajadores que se desvinculan y salen en razón con la cantidad total promedio de personal en la organización en un cierto periodo de tiempo, permite realizar comparaciones, desarrollar diagnósticos o promover acciones[4].

1.1.1 Problemática

Cada persona que se retira de la organización conduce a la necesidad de seleccionar un nuevo trabajador para reemplazarlo, quien no siempre cuenta con las habilidades y capacitación necesarias para cubrir las expectativas del puesto vacante, requiriendo para ello de un periodo de adaptación y capacitación, prolongándose en algunas ocasiones dicho periodo por treinta días o más. Este es el costo más importante relacionado con la rotación dependiendo del tipo de negocio y de la complejidad e impacto del puesto dentro de la organización.

El proceso de reemplazar a un empleado es lento y costoso. Los costos se pueden dividir en tres categorías: los costos de separación del empleado que se va, los costos del reemplazo y los costos de capacitación para el nuevo empleado. Estos costos se estiman de manera conservadora en dos a tres veces el sueldo mensual del empleado que se va, y no incluyen los costos indirectos como la baja productividad del empleado antes de renunciar, el desánimo que experimentan los que se quedan y el sobre tiempo que generan al cubrir las funciones dejadas, afectando la productividad de otros trabajadores. Reducir la rotación de personal podría dar como resultado ahorros significativos para una organización[5].

Los costos asociados a la rotación están compuestos por:

- Costos de separación: Liquidación y acciones administrativas.
- Costos de reemplazo: Publicidad, entrevistas, pruebas.
- Costos de capacitación: Inducción, capacitación, curva de aprendizaje.
- Costos de riesgo: Riesgo que tiene una empresa al no cumplir con los tiempos y servicios establecidos con los clientes.

1.1.2 CAUSAS

La rotación de personal es el efecto de fenómenos internos y externos de la organización sobre la actitud del trabajador en su desempeño laboral y decisión de permanecer laborando o retirarse.

Dentro de los fenómenos externos podemos citar la situación de la oferta y demanda de recursos humanos en el mercado, la coyuntura económica, las oportunidades de empleo, etc[6].

Dentro de los fenómenos internos que ocurren en la organización podemos citar la política salarial, política de beneficios, oportunidades de crecimiento, cultura organizacional, política de reclutamiento y selección de recursos humanos, criterios y programas de entrenamiento.

Estas son algunas prácticas que utilizan algunas empresas para reducir la rotación de personal [7]:

- Mayor formación y recursos.
- Implicación de los agentes en distintos procesos dentro de la compañía.

- Ofrecer un plan de carrera.
- Dar la posibilidad de aportar información a sus mandos sobre su trabajo y clientes.
- Cambiar de campañas y de turno para evitar el desgaste que conlleva realizar una misma actividad.

1.1.3 Periodo de prueba

Se entiende por periodo de prueba aquel tiempo, libremente concertado por el trabajador y el empleador, durante el cual cualquiera de ellos, unilateralmente, puede dar terminada la relación laboral.

El periodo de prueba permite al empleador contrastar la aptitud profesional del trabajador para el desempeño de sus funciones y el perfil profesional requerido para el puesto[8].

La duración del periodo de prueba no podrá exceder de 6 meses, LA EMPRESA considera 3 meses de periodo de prueba para el personal operativo y 6 meses para el personal administrativo.

Para los fines de este trabajo se considerará un periodo de prueba de 6 meses para todos los puestos.

1.1.4 Rotación laboral en Perú

De acuerdo con el estudio de Indicadores Saratoga que realizó en el 2010 Pricewaterhouse Coppers, empresa de consultoría de negocios y financiera, entre 155 empresas de diversos rubros en 13 países, el índice promedio de rotación laboral en el Perú llega a 20.7%, mientras que en Latinoamérica es de 10.9%.
Las razones de este alto índice rotación de personal son: el alto despegue de la economía peruana y el que las personas ya no desean pasar mucho tiempo en una misma empresa, según explicó el consultor senior del área de Human Resource Consulting en Pricewaterhouse Coopers, a diferentes medios de comunicación en el año 2011.

Según la consultora, la rotación preocupa más cuando se trata de personal con alto desempeño, pues la media es de 5.3%, cuando la media latinoamericana llega a 3.1%. Por otro lado, la elevada rotación está generando altos costos a las empresas peruanas en

reclutamiento y contratación, que en promedio llegan a US$342 por persona, mientras que en Latinoamérica es de aproximadamente US$ 175 según el estudio realizado por la consultora Pricewaterhouse Coppers en el 2010.

Inés Temple, presidenta ejecutiva de DBM Perú, empresa de gestión de la Empleabilidad, Outplacement y Executive Coaching, afirmó a los medios en el 2011 que este nivel de rotación no solo se debe al avance de la economía nacional, sino también a que los sueldos en el mercado laboral se encuentran atrasados desde hace mucho tiempo y ante el surgimiento de una mejor oferta salarial no dudan en aceptar, frente al surgimiento de nuevas empresas hay una batalla abierta en la búsqueda de talento.

La empresa en estudio tiene un alto índice de rotación laboral, en el 2011 fue de 35% anual y solo en áreas operativas de 30%. Lo señalado genera gastos adicionales a lo presupuestado así como impacto en el clima laboral, por ello los ejecutivos de la empresa se encuentran interesados en conocer los motivos e iniciar proyectos de retención de personal.

2. Modelos estadísticos

2.1 Modelos

Un modelo es una representación simplificada de la realidad que se utiliza para entender mejor situaciones de la vida real [9].

Los **modelos matemáticos** se pueden clasificar en modelos deterministas y estocásticos.

a) **Modelo determinista:**

Es aquel en el que dado un conjunto de parámetros y variables de entrada produce siempre el mismo conjunto de variables de salida. El modelo no contempla la existencia del azar.

b) Modelo estocástico o estadístico:

Un modelo estadístico es una representación de un fenómeno mediante funciones matemáticas aplicadas a variables aleatorias. No es determinista, se puede identificar un componente **aleatorio** y un componente **sistemático**.

El interés es explicar el comportamiento de una variable respuesta, Y, a través de una o más variables explicativas, X, mediante una función matemática, en la cual se expresan las variables explicativas y los parámetros que la forman (componente sistemático).

La respuesta no es determinada en forma precisa a partir de las variables explicativas incluidas en el modelo, existe cierta variabilidad debida a factores aleatorios. Debido a que la explicación no es precisa se incluye un término denominado error ε (componente aleatorio).

Las etapas del modelado son las siguientes:

1. Identificación del modelo y selección de variables.
2. Estimación de parámetros.
3. Validación del modelo.

2.2 Modelo lineal general

Sea Y una variable aleatoria que fluctúa alrededor de un valor desconocido η, esto es Y= $\eta + \varepsilon$, donde ε es el error, de forma que η puede representar el valor verdadero e Y el valor observado.

Supongamos que η toma valores distintos de acuerdo con diferentes situaciones experimentales según el modelo lineal.

$$\eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j \ . \qquad (2.1)$$

Donde β_j son parámetros desconocidos y X_j son valores conocidos, cada uno de los cuales ilustra situaciones experimentales diferentes (j=1,2,.., k).

Si se tiene n observaciones de la variable Y. Diremos que y_1, y_2,…y_n observaciones independientes de Y siguen un modelo lineal si:

$$y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \ ; \ \text{i=1,…, n.} \qquad (2.2)$$

Donde Y es la **variable respuesta**, $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ es la **componente sistemátic**a o predictor lineal formada por las **variables regresoras o independientes** X_j y ε es el **componente aleatorio**.

Los supuestos del modelo lineal general son:

1. $E(e_i) = 0$
2. $Var(Y/X_1 = x_1, \dots, X_k = x_k) = \sigma^2 < \infty$, homocedasticidad; i=1,...,n
3. $Cov(Y_i, Y_j) = 0$, incorrelación ; $\forall\ i \neq j$
4. Linealidad de los parámetros del predictor lineal.
5. Aditividad del predictor.
6. $Y/X_1 = x_1, \dots, X_k = x_k \sim N(\mu, \sigma^2)$

Por el Teorema de Gauss Markov, el estimador de mínimos cuadrados es $\hat{\beta} = (X`X)^{-1}X`Y$, expresada matricialmente[11]. Este estimador es consistente y de mínima varianza entre todos los estimadores lineales insesgados, siempre que se verifiquen los supuestos.

Si no se verifica el supuesto de homocedasticidad, $Var(Y) = V\sigma^2$ donde V es una matriz simétrica definida positiva, se utiliza el estimador de mínimos cuadrados generalizados.

$$\hat{\beta} = (X`V^{-1}X)^{-1}\ X`V^{-1}\ Y \qquad (2.3)$$

Validación del modelo

Se utilizan diferentes métodos y herramientas para evaluar la significancia de los parámetros estimados, la bondad de ajuste del modelo y validación de los supuestos:

- Coeficiente de determinación.
- Pruebas F de bondad del ajuste del modelo.
- Prueba t para evaluar la significación de los coeficientes del modelo.
- Análisis de residuos.
- Análisis de influencia.

2.3 Modelos lineales generalizados

En 1972, Nelder y Wedderburn publican un artículo en el cual se propone unificar tanto los modelos con variables respuesta numérica como categórica, considerando las distribuciones binomial, Poisson, etc. denominándola "*Modelos lineales generalizados*" (GLM).

Esta generalización tiene varios aspectos importantes:

- Los modelos incluyen una variedad de distribuciones seleccionadas de una familia de distribuciones de probabilidad denominada "**familia exponencial**".
- Involucran transformaciones de la esperanza matemática de la variable respuesta (media poblacional) a través de una función denominada "**función enlace**".

El modelo lineal generalizado es una extensión del modelo lineal general que involucra tres componentes:

1. Una variable respuesta Y, cuyos elementos son independientes, tiene distribución de probabilidades la cual es miembro de la familia exponencial de distribución de probabilidades.

$$E(Y/X_1 = x_1, \dots, X_k = x_k)=\mu \ .$$
$$V(Y/X_1 = x_1, \dots, X_k = x_k)=V \ .$$

Suponer que Y es una variable aleatoria con función densidad de probabilidad $f_Y(y,\theta)$. Se dice que la distribución de probabilidades de Y es miembro de la familia exponencial, si se puede expresar en la forma:

$$f_Y(y, \theta) = \exp[\{y\, c(\theta) - b(\theta)\} + h(y)] \qquad (2.4)$$

Donde $c(\theta)$ es una función del parámetro θ; $\boldsymbol{y}$ es la estadística suficiente, $b(\theta)$ es una función únicamente del parámetro θ, y $h(y)$ es un función únicamente de la variable aleatoria.

O expresar mediante la forma canónica:

$$f_Y(y_i, \theta, \varphi) = \exp[\{y_i\theta - b(\theta)\}/a(\varphi) + c(y_i, \varphi)] \qquad (2.5)$$

Donde, θ es un p***arámetro de localización*** *(*parámetro canónico); y_i es la estadística suficiente, b(θ) es una función únicamente del parámetro canónico; c(y, φ) es una función del parámetro de perturbación φ y de la variable aleatoria , a(φ) , parámetro constante, es función de φ.

2. Un conjunto de variables explicativas, cuantitativas o cualitativas, que se combinan linealmente y dan lugar a la componente sistemática del modelo

$$\eta = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad .$$

3. Una función g(μ) que enlaza la componente aleatoria y la componente sistemática.

$$g(\mu) = \eta \quad .$$

Dado que es una función conocida, monótona y diferenciable de η , $\mu_i = g^{-1}(\eta_i)$; i=1,2,…n. La esperanza y varianza de las variables aleatorias con distribución de probabilidades de la familia exponencial se pueden obtener a partir de una función derivada de la teoría de verosimilitud, la función score S(y,θ,φ), que tiene una gran utilidad para derivar una serie de estadísticas útiles entre ellas la media y varianza de la variable aleatoria.

$$S(y,\theta,\varphi) = \frac{\delta ln}{\delta \Theta} f(y, \theta, \varphi) \qquad (2.6)$$

De la expresión anterior se pueden obtener como resultado:

$$E(Y) = \frac{\delta b(\Theta)}{\delta \Theta} \quad \text{y} \quad V(Y) = \frac{\delta^2 b(\Theta)}{\delta^2 \Theta} . \; a(\varphi) \qquad (2.7)$$

2.3.1 Estimación de parámetros

Entre los métodos de estimación, de los coeficientes del modelo, más utilizados están el método Score de Fisher el cual se basa en la función score obtenida a partir de la función de verosimilitud y el método de mínimos cuadrados generalizados, siendo el método de mínimos cuadrados ponderados un caso particular de este último. Nelder y Wedderburn mostraron que los métodos de mínimos cuadrados y máxima verosimilitud producen resultados muy similares.

2.3.2 Evaluación de la bondad de ajuste del modelo lineal generalizado

Para evaluar la bondad de ajuste del modelo se utiliza la función desvianza o función desvíos. Esta es una medida de la distancia entre el logaritmo de la función de verosimilitud del modelo saturado (con n parámetros) y el modelo que está siendo investigado (con p parámetros) evaluado en la estimación máximo verosímil de los parámetros del modelo.

$$D(\widehat{\mu}, y) = -2ln\left(\frac{\sum_{i=1}^{n} L(\widehat{\mu_i}; y_i)}{\sum_{i=1}^{n} L(y_i; y_i)}\right) = 2\ln\big(L(y_i; y_i)\big) - lnL(\widehat{\mu_i}; y_i)\,. \qquad (2.8)$$

Un valor pequeño de función desvianza indicaría que con un número menor de parámetros (p<n) se obtiene un ajuste tan bueno como el ajuste con el modelo saturado (p=n).

2.3.3 Validación del modelo ajustado

La estadística utilizada para evaluar los coeficientes del modelo, equivalente a la estadística de prueba t usada en el modelo lineal general, es la estadística de Wald, la cual tiene distribución χ^2 con un grado de libertad.

En un modelo de regresión lineal con k parámetros, la estadística t nos permite contrastar la hipótesis de que los coeficientes de regresión parcial son igual a cero. (H_0: β_j=0)

$$t = \left(\frac{\widehat{\beta}}{Se(\widehat{\beta})}\right)^2 \sim t_{n-k}\ .$$

Partiendo de $Y \sim N(X\beta, \sigma^2)$, donde $\hat{\beta} = (X'X)^{-1}X'Y$, $(X'X)^{-1} = \begin{pmatrix} c_{00} & \dots & c_{ji} \\ & c_{ii} & \\ c_{ij} & \dots & c_{kk} \end{pmatrix}$

La varianza de $\hat{\beta}$ es:

$Var(\hat{\beta}) = (X'X)^{-1}\hat{\sigma}^2$, $Se\big(\widehat{\beta_J}\big) = \sqrt[2]{c_{jj}\hat{\sigma}^2}$

En un modelo lineal generalizado la $Var(\hat{\beta})$ se estima utilizando la inversa de la matriz de información de Fisher I^{-1} donde I es una matriz hessiana definida positiva.

Las hipótesis son:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$, para algún j=1,2,…k

$$W = \left(\frac{\widehat{\beta}}{Se(\widehat{\beta})}\right)^2 \sim \chi_1^2 \ .$$

2.4 Modelo lineal generalizado para respuesta binaria

Cuando una variable respuesta Y tiene solo dos posibles resultados, éxito o fracaso, Y tiene distribución binomial B(n,π), donde n representa el número de ensayos independientes con probabilidad de éxito π , $0 \leq \pi_i \leq 1$.

$$P(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}; \ \text{i=1,…, } n_i \qquad (2.9)$$

Y_i corresponden a ensayos independientes, cuya esperanza y varianza son:

$E(Y_i) = \mu_i = n_i \pi_i$; i = 1,2,...,n_i

$V(Y_i) = \sigma_i^2 = n_i \pi_i (1-\pi_i)$; i =1,2,..., n_i

Cada Y_i puede estar afectado por un conjunto de variables explicativas combinadas linealmente, estas variables X_k pueden ser cuantitativas o cualitativas.

$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ (Predictor Lineal)

Al relacionar directamente la esperanza de la variable respuesta y el predictor lineal se presentan los siguientes casos:

$$E(Y/X_1 = x_1, \ldots, X_k = x_k) = n_i \pi_i = \eta_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad (2.10)$$

El predictor lineal puede tomar valores en el conjunto de números reales mientras que la esperanza de la variable respuesta toma valores solo en el intervalo [0,1]

Por otro lado, si se desea modelar la probabilidad de éxito π_i la cual varía en el intervalo [0,1], la función de enlace identidad no sería posible utilizarla.

Por ello se elige una función de enlace dependiendo de las características de la variable respuesta y considerando que debe cumplir con ser una función monótona y que permita que las predicciones varíen en el intervalo [0,1].

Existen tres funciones de enlace que cumplen las condiciones señaladas, la función de enlace logit, probit y log-log complementario. Estas funciones permiten relacionar la componente aleatoria y la componente sistemática.

- **Función de enlace logit**

Resulta de la transformación de una función logística, E(Y/X)=$\pi(X) = \frac{e^{\eta}}{1+e^{\eta}}$, realizando las transformaciones necesarias se obtiene la función de enlace logit que permite convertir la variable respuesta de modo que tome valores en el conjunto de números reales y permite linealizar la relación entre la esperanza de la variable respuesta y el predictor lineal:

$$\eta = ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$$=Logit (π (X)).

Se utiliza cuando la variable respuesta original en estudio tiene distribución binomial

- **Función de enlace probit**

Resulta de la transformación de una distribución normal estándar acumulada

E(Y/X)= $\pi(X) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} e^{\frac{z}{2}} dz = \phi(\eta)$, realizando las transformaciones necesarias se obtiene la función de enlace probit $\eta = \phi^{-1}(\pi(X))$

Se utiliza cuando la variable respuesta original en estudio es continua pero ha sido dicotomizada.

- **Función de enlace log-log complementario**

Resulta de la transformación de una distribución de Gompertz.

E(Y/X)= $\pi(X)$=1-$e^{-e^{(\eta)}}$ realizando las transformaciones necesarias se obtiene la función de enlace log log complementario $\eta = Ln(-Ln(1-\pi(X))$.

2.5 Regresión logística

2.5.1 Modelo de regresión logística múltiple

Sea Y la variable respuesta y X_k variables explicativas, los coeficientes β_k son los parámetros del modelo.

$$E(Y / X_1 = x_1 , \dots , X_k = x_k) = \pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} \quad . \tag{2.11}$$

De tal forma que $0 \leq \pi(X) \leq 1$.

Supuestos

1. Las respuestas Y_i son independientes ; i=1,2 …, n
2. Las variables explicativas son independientes
3. Cada Y_i tiene distribución Bernoulli B(π_i) ; i=1,2 …, n .
4. Varianza heterocedástica: V(Y_i)= π_i(1-π_i) ; i=1,2 …, n .

La función logit permite convertir la variable respuesta de modo que tome valores en $<-\infty, \infty>$ y linealiza la relación entre la variable respuesta y el predictor lineal

$$Logit(\pi(X)) = Ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \ . \tag{2.12}$$

2.5.1.1 Estimación de parámetros

A partir de una muestra de tamaño n, la función de verosimilitud es

$$L(\beta) = L(\beta, y, X) = \prod_{i=1}^{n} \pi_i^{y_i}(x_i)(1 - \pi_i(x_i))^{1-y_i} \quad . \tag{2.13}$$

$$l(\beta) = lnL(\beta, y, X) = \sum_{i=1}^{n} y_i x'\beta - \sum_{i=1}^{n} \ln\left[1 + \exp\left(x'\beta\right)\right] . \tag{2.14}$$

Al formar parte de los modelos lineales generalizados la estimación de los parámetros se realiza por el método de mínimos cuadrados iterativamente reponderados siendo el más utilizado el método de Newton –Raphson a través del Score de Fisher obtenida a partir de la función de verosimilitud y el método de mínimos cuadros iterativamente ponderados.

- ***Odds***

El cociente entre una probabilidad y su complemento se denomina odds, e indica cuantas veces es más probable que ocurra un evento respecto a que no ocurra.

Odds = $\frac{\pi(X)}{1-\pi(X)}$; oscila entre [0,∞]

Cuando X=x_i, Odds1 = $\frac{\pi(x_i)}{1-\pi(x_i)} = e^{\beta_0+\beta_1 x_i}$. (2.15)

Cuando X= x_{i+1}, Odds2 = $\frac{\pi(x_{i+1})}{1-\pi(x_{i+1})} = e^{\beta_0+\beta_1 x_{i+1}}$. (2.16)

- ***Razón de Odds (Odds Ratio)***

Relación entre los coeficientes del modelo de regresión logística y la razón de ventajas.

$$OR = \frac{Odds2}{Odds1} = \frac{\frac{\pi(x_{i+1})}{1-\pi(x_{i+1})}}{\frac{\pi(x_i)}{1-\pi(x_i)}} = e^{\beta_1} \quad . \tag{2.17}$$

El OR se interpreta como la chance de tener el evento estudiado cuando la variable regresora pasa de tomar el valor x_i a x_{i+1}.

2.5.2.2 Evaluación del modelo ajustado

- ***Función desvianza***

Permite medir que tan cerca se encuentra el modelo propuesto del modelo saturado, puede ser utilizada como una estadística para evaluar la bondad del ajuste del modelo propuesto.

$$D(\hat{\mu}, y) = -2Ln\frac{\sum_{i=1}^{n} L(\hat{\mu_i}; y_i)}{\sum_{i=1}^{n} L(y_i; y_i)} = 2\ln\big(L(y_i; y_i)\big) - lnL(\hat{\mu_i}; y_i) \quad . \tag{2.18}$$

Esta es una medida de la distancia entre el logaritmo de la función de verosimilitud del modelo saturado (con n parámetros) y el modelo que está siendo investigado (con p parámetros) evaluado en la estimación máximo verosímil de los parámetros del modelo.
Un valor pequeño de la función desvianza indicaría que con un número menor de parámetros (p<n) se obtiene un ajuste tan bueno como el ajuste con el modelo saturado (p=n).

A partir de las desvianzas se puede comparar un modelo con r parámetros y otro con p parámetros, para ello se calcula la estadística G la cual es la diferencia de las desvianzas de los dos modelos. G tiene una distribución chi cuadrado con p-r grados de libertad.

- ***Evaluación individual de los coeficientes***

La estadística de Wald permite la evaluación de cada coeficiente del modelo. La estadística de prueba es dada por:

$$W = \left(\frac{\widehat{\beta_J}}{Se(\widehat{\beta_j})}\right)^2 \sim X_1^2 \ .$$

- ***Estimación de las probabilidades***

Las probabilidades estimadas se obtienen mediante:

$$\hat{\pi}(x_i) = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}X_{i1}+\cdots+\widehat{\beta_k}X_{ik}}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}Xi_1+\cdots+\widehat{\beta_k}X_{ik}}} \quad . \tag{2.19}$$

Se debe definir un criterio para clasificar a los individuos de acuerdo a la variable respuesta estudiada. Este criterio puede ser un punto de corte 0.5, donde un valor menor o igual a 0.5 corresponde a Y=0 y un valor mayor a 0.5 corresponde a Y=1.

El análisis discriminante permite la predicción de pertenencia de la unidad de análisis a uno de los grupos pre establecido, pero se requiere que se cumplan los supuestos de normalidad de las variables regresoras y la igualdad de matrices de covarianzas de los grupos, para que la regla de predicción sea óptima.

2.6 Modelos multinivel

2.6.1 Estructuras multinivel

Los modelos multinivel o también llamados jerárquicos, son una respuesta a la necesidad de analizar la relación entre los individuos y el medio en el que se desenvuelven.

Una jerarquía de datos consiste en un conjunto de observaciones que conforman un nivel individual llamado primer nivel o micro nivel, que se encuentran anidadas dentro de un nivel superior llamado grupo, contexto o nivel macro.

Las estructuras jerárquicas o anidadas son comunes en datos de investigación de salud, económico y social, donde se encuentran agrupaciones reales de sujetos que hacen que aquellos que pertenecen al mismo grupo reciban una serie de influencias comunes que reducen la variabilidad natural del grupo, las observaciones dentro de un mismo grupo no son independientes entre sí, haciéndolo en cierta medida más homogéneo. Entre los

ejemplos más comunes se encuentran alumnos en escuelas, pacientes en hospitales, casas en vecindarios. La estructura jerárquica puede ser de 2 o más niveles.

Si se realizara un análisis de regresión clásico no se cumple uno de los supuestos básicos que es la independencia de los datos, lo que puede ocasionar la subestimación de los errores estándar de los coeficientes de regresión [15].

En los modelos multinivel los datos están estructurados en grupos y los coeficientes pueden variar por grupos. Estos modelos permiten tratar adecuadamente las varianzas producidas por distintos niveles de agregación.

Los primeros estudios con datos que presentan una estructura jerárquica fueron realizados en la década de los ochenta en el campo de la educación asociado a la eficacia de las escuelas y la evaluación de sistemas educativos [19].

2.6.1.1 Variables

Las variables se clasifican por la naturaleza de los datos pero para el análisis multinivel se considera además el propósito del estudio y la jerarquía de los datos. Por ello se cuenta con variables respuesta y variables explicativas. De acuerdo a la jerarquía con variables del primer nivel, nivel individual o micro nivel y las variables del nivel superior o nivel macro. En algunos casos las variables medidas en el nivel micro son similares a las del nivel macro.

En el análisis multinivel las variables respuesta se miden a nivel individual y las variables explicativas tanto a nivel individual como en el nivel superior[22].

2.6.1.2 Denominación de los modelos

Los distintos nombres utilizados para definir a los modelos difieren en el grado de generalidad. De todas las denominaciones la más genérica es la de modelos multinivel, debido a que refleja la naturaleza jerárquica de los datos.

Las denominaciones son varias: modelos jerárquicos lineales, modelos de efectos mixtos, modelos contextuales, de coeficientes aleatorios, etc.

Será multinivel porque asume que hay un conjunto de datos jerárquicos, con una sola variable dependiente que es medida en el nivel más bajo y variables explicativas en 2 o más

niveles. Se denomina modelo de coeficientes aleatorios debido a que los coeficientes del primer nivel son aleatorios y variarán aleatoriamente en los niveles superiores del modelo.
Puede ser mixto al incorporar efectos fijos asociados al impacto de las variables predictoras y efectos aleatorias que representan la variación entre contextos o grupos de los niveles superiores y considerarse como modelo de componentes de varianza al incluir estimaciones de la varianza y la covarianza.

2.6.1.3 Falacia atomística y falacia ecológica

Es importante considerar la unidad adecuada de análisis ya que se puede incurrir en uno de dos tipos de errores.
Se pueden analizar variables pertenecientes a diferentes niveles como si fueran de un único nivel común. Al asignar los valores de las variables de las unidades de contexto o grupo a cada unidad individual, se dice que se realiza desagregación de datos y se comete una "falacia atomística", debido a que los individuos del mismo grupo han compartido estímulos o influencias que hacen que sus valores en la variable respuesta sean más homogéneas, no se puede asumir el supuesto de independencia y se subestimaría el tamaño de error, se incrementa el riesgo de cometer error tipo I [20].
Al agregar los valores de las variables medidas en los individuos del primer nivel y considerarlas como variables del grupo, como calcular el valor de la media para cada grupo en las variables de estudio, realizando el análisis solo con las unidades del segundo nivel o grupo, se comete una "falacia ecológica" debido a la generalización que se realiza de las relaciones observadas en el nivel superior a los individuos. Se traduce en una pérdida de información y de potencia estadística, aumento del riesgo o probabilidad del error tipo II [21].
Al ignorar la estructura jerárquica de los datos eliminamos la varianza interna de los grupos en el caso de agregación y los efectos de las variables grupales no observadas quedarán recogidas en el error en caso se realice un análisis desagregado.
El modelo lineal clásico no considera esta estructura de datos en niveles, en cambio los modelos multinivel trabajan tanto con la agregación y la desagregación evitando analizar los datos en un nivel y extraer conclusiones a otro nivel.

2.6.1.4 Características de los datos en los modelos multinivel

- Datos anidados en estructuras jerárquicas.
- Los datos pueden provenir de una muestra multietápica de la misma estructura jerárquica.
- El grado de homogeneidad interna de los grupos viene expresado por la correlación intraclase o autocorrelación que habitualmente es ignorada por los modelos estadísticos clásicos. Si no hubiese correlación dentro de los grupos, no se estudiarían los modelos multinivel ya que no se presentaría la necesidad de realizar un análisis para datos anidados e incluir variables explicativas en el modelo porque no habría nada que explicar.
- Se pueden definir variables referidas a las unidades de análisis de cada nivel.

2.6.1.5 Ventajas y desventajas de los modelos multinivel

Ventajas

- Permite considerar las diferencias grupales o contextuales. Consideración de la heterogeneidad, interacción entre individuos y contextos, inclusión de conductas interrelacionadas y consideración de múltiples contextos.
- Permite obtener mejores estimaciones de los coeficientes de regresión y de su variación comparados con los modelos tradicionales. Una gran flexibilidad ofrecida por los modelos multinivel se da en términos de modelar la estructura de varianza de los datos en función de variables explicativas que permite analizar los datos en los cuales la varianza no es homogénea.

Desventajas

- La interpretación de los resultados es más compleja debido a la teoría que aborda y las estructuras complejas de variación. Mayor dificultad en la comunicación de resultados.
- Los programas computacionales disponibles no son muy populares y difundidos.

2.7 MODELOS DE REGRESION MULTINIVEL

2.7.1 Supuestos de los modelos de regresión multinivel

Modelo bien especificado: Relación entre variable respuesta y predictores sea lineal.
Varianzas de los errores no son constantes (heterocedasticidad), ni necesariamente distribuidos normalmente.
El modelo presentado considera como variable respuesta una variable continua con distribución normal y cuya medición se da en el primer nivel. Sin embargo, el modelo se puede extender para la familia de **modelos lineales generalizados**.

El supuesto de independencia de los datos no se aplica en estos modelos.

- Coeficientes: fijos y aleatorios. Los coeficientes aleatorios son coeficientes que se distribuyen según una función de probabilidad. Los coeficientes de regresión aleatorio solo pueden ser considerados en el nivel superior en el que fueron medidos.

Interceptos y pendientes aleatorios deben estar distribuidos normalmente, este supuesto puede relajarse con distribuciones distintas a la distribución normal, aunque se puede complicar el proceso de estimación.

2.7.2 Modelo de regresión de dos niveles

Conocido como modelo de coeficientes aleatorios, modelo de componentes de varianza o modelos lineales jerárquicos.

Se asume que hay un conjunto de datos jerárquicos con una sola variable dependiente Y medida en el nivel más bajo y variables explicativas que existen en todos los niveles.
Partiendo del modelo de regresión simple $y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$, donde $\varepsilon_i \sim N(0, \sigma^2)$
Considerando que se tienen J grupos se considera:

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \varepsilon_{ij} \quad ; \text{i} = 1, 2... \text{n}; \quad \text{j} = 1,2,...,\text{J} \quad . \qquad (2.20)$$

No es práctico estimar una ecuación de regresión para cada grupo, considerando que pueden existir muchos grupos y que en cada uno se deben estimar la pendiente y el intercepto, esto disminuye la eficiencia del modelo además de no existir un interés real en las diferencias de la variable dependiente respecto a la variable regresora.

Dentro de cada grupo, $\beta_0 + \beta_1 X_{1i}$ es la parte fija o sistemática y ε_i la parte aleatoria o residual. β_0 y β_1 son los coeficientes de regresión que representan el intercepto, valor medio de Y cuando X =0 , y la pendiente respectivamente.

Al ser β_{0j} y β_{1j} valores propios de cada grupo, se supone que existe variabilidad entre grupos, por lo que estos coeficientes se convertirán en coeficientes aleatorios.

$$\beta_{0j} = \beta_0 + \mu_{0j} \tag{2.21}$$

$$\beta_{1j} = \beta_1 + \mu_{1j} \tag{2.22}$$

Donde

$$\mathrm{E}(\beta_{0j}) = \beta_0 \quad , \mathrm{V}(\beta_{0j}) = \sigma^2_{\mu_0}$$

$$\mathrm{E}(\beta_{1j}) = \beta_1 \quad , \mathrm{V}(\beta_{1j}) = \sigma^2_{\mu_1}$$

$$\mathrm{Cov}(\beta_{0j,}\beta_{1j}) = \sigma^2_{\mu_0\mu_1}$$

β_{0j} y β_{1j} tienen una distribución normal bivariada y representan la media general de la población o intercepto y la pendiente, respectivamente.

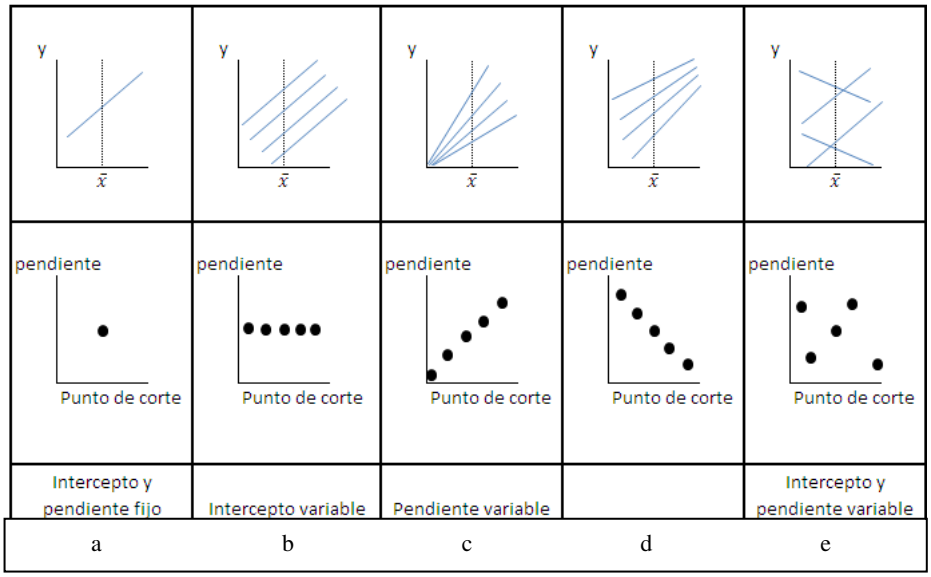
$\sigma^2_{\mu_0}, \sigma^2_{\mu_1}$ y $\sigma^2_{\mu_0\mu_1}$ corresponde a la varianza del intercepto, pendiente y la covarianza entre intercepto y pendiente, respectivamente.

La variabilidad entre grupos es la característica principal en el modelo multinivel. Si esta variación no existiera no sería necesario el empleo de estos modelos .En el gráfico 2.1 se ilustra las variaciones y relaciones entre interceptos y pendientes.

En el gráfico 2.1.a todos los grupos comparten la misma ecuación, es decir la relación entre X e Y es la misma. Comparten la misma recta de regresión y, por tanto, la variación entre puntos de corte ($\sigma^2_{\mu_0}$), pendientes ($\sigma^2_{\mu_1}$) y la covarianza entre ambos ($\sigma^2_{\mu_0\mu_1}$) será igual a cero. En el gráfico 2.1.b, todos los grupos comparten la misma pendiente, es decir no hay variación en la relación que se establece dentro de cada grupo entre la variable X e Y ($\sigma^2_{\mu_1} = 0$), en cambio los puntos de corte varían de grupo a grupo ($\sigma^2_{\mu_0} > 0$). En el gráfico 2.1.c se observa que cuanto mayor es la media del grupo mayor es su pendiente. Los grupos se diferencian en el punto de corte, pero también en la pendiente ($\sigma^2_{\mu_0}, \sigma^2_{\mu_1} > 0$).Además se puede observar que cuanto mayor es el punto de corte también es mayor la pendiente

($\sigma^2_{\mu_0\mu_1}$> 0). Por lo contrario en el gráfico 2.1.d donde se observa que cuanto mayor es la media del grupo, menor es la pendiente. Como en el caso anterior, $\sigma^2_{\mu_0}, \sigma^2_{\mu_1}$ adoptan valores mayores que cero, en cambio, al ser la relación entre punto de corte y pendiente negativa, el valor de la covarianza será negativo. El gráfico 2.1.e muestra la situación en la que no hay relación entre el punto de corte y la pendiente por lo que el valor de la covarianza tenderá a cero [13].

Gráfico 2.1 Variación de pendientes e interceptos



Fuente: Modelos jerárquicos lineales. Gaviria y Castro. Editorial La Muralla 2005.

Reemplazando (2.21) y (2.22) en (2.20), tenemos el **modelo completamente aleatorio**

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \mu_{1j} X_{1ij} + \mu_{0j} + \varepsilon_{ij} \qquad (2.23)$$

i = 1, 2... n_j , j = 1,2,...,J .

Donde

$\beta_0 + \beta_1 X_{1ij}$ es la parte fija del modelo y $\mu_{1j} X_{1ij} + \mu_{0j} + \varepsilon_{ij}$ la parte aleatoria.

Los parámetros a estimar son:

$\beta_0, \beta_1, \sigma^2_\varepsilon, \sigma^2_{\mu_0}, \sigma^2_{\mu_1}$ y $\sigma^2_{\mu_0\mu_1}$.

- ***Modelo de intercepto aleatorio***

Existe variación entre los puntos de corte $\sigma^2_{\mu_0} > 0$ pero no hay variación entre pendientes $\sigma^2_{\mu_1} = 0$ y β_1 se asume que es el mismo para cada grupo.

Si

$$\beta_{0j} = \beta_0 + \mu_{0j} \tag{2.24}$$

$$\beta_{1j} = \beta_1 \tag{2.25}$$

Reemplazando en (2.20) se tiene el modelo de intercepto aleatorio

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \mu_{0j} + \varepsilon_{ij} \tag{2.26}$$

- ***Modelo de pendientes aleatorias***

Existe variación entre los pendientes $\sigma^2_{\mu_1} > 0$ pero no hay variación entre los puntos de corte $\sigma^2_{\mu_0} = 0$.

Si

$$\beta_{0j} = \beta_0 \tag{2.27}$$

$$\beta_{1j} = \beta_1 + \mu_{1j} \tag{2.28}$$

Reemplazando en (2.20) se tiene el modelo de pendientes aleatorias.

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + (\mu_{1j} X_{1ij} + \varepsilon_{ij}) \tag{2.29}$$

En los modelos multinivel, el modelo micro representa la relación dentro de cada grupo entre la variable respuesta y la variable predictora, el modelo macro representa la relación entre grupos, en donde los parámetros del micro modelo son las variables respuesta de los macromodelos donde se reconoce la variación entre grupos no identificado por los modelos clásicos.

El especificar un modelo multinivel consiste en estimar:

La(s) media(s) que componen la parte fija. $\beta_0, \beta_1, ..$

Las varianzas de los interceptos y pendientes. $\sigma^2_{\mu_0}, \sigma^2_{\mu_1} .., \sigma^2_{\varepsilon}$

La covarianza entre interceptos y pendientes.$\sigma^2_{\mu_0 \mu_1}$,...

2.7.2.1 Modelo Nulo

Es el término de comparación de cualquier otro modelo alternativo debido a que no incluye variables predictoras. Si la varianza en el modelo nulo es cero, no tendría sentido incluir variables explicativas.

Modelo en el nivel micro: $y_{ij} = \beta_{0j} + \varepsilon_{ij}$

Modelo en el nivel macro: $\beta_{0j} = \beta_0 + \mu_{0j}$

Modelo completo: $y_{ij} = \beta_0 + \mu_{0j} + \varepsilon_{ij}$, donde $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$ y $\mu_{0j} \sim N(0,\sigma_{\mu_0}^2)$

Donde i = 1, 2... n_j , j = 1,2,..., J.

Este modelo no sirve para explicar la varianza de la variable dependiente, este solo descompone la varianza en dos componentes independientes: la varianza del error del primer nivel (σ_ε^2) y la varianza del error del segundo nivel$(\sigma_{\mu_0}^2)$ [23].

A partir de la las varianzas y covarianzas se puede calcular un coeficiente de correlación llamado correlación intraclase ρ.

$$\rho = \frac{\sigma_{\mu_0}^2}{(\sigma_{\mu_0}^2)+(\sigma_\varepsilon^2)} \tag{2.30}$$

La **correlación intraclase** es un estimador de la proporción de la varianza explicada en la población, es igual a la proporción estimada de la varianza del nivel grupo comparada con la varianza total estimada.

Se puede considerar como un indicador de homogeneidad interna de los grupos, una medida del grado de semejanza entre unidades de nivel inferior que pertenecen a la misma unidad de nivel superior[22].

Mide el grado en que la variable dependiente tiene valores similares en los individuos del mismo grupo. Puede definirse como la proporción de la varianza de la variable dependiente que corresponde a diferencias entre grupos o unidades de nivel superior.

El coeficiente de correlación intraclase toma valores entre 0 y 1, tal que si es 0 no hay diferencias entre los elementos del nivel superior $(\sigma_{\mu_0}^2 = 0)$ y si es 1, no hay diferencias

dentro de cada grupo $(\sigma_\varepsilon^2) = 0$. En el caso de variables dependientes no distribuidas normalmente, el cálculo es más complejo y no siempre es sencillo[22].

El cálculo del coeficiente de correlación intraclase permite conocer si hay diferencias entre grupos y si es válido construir un modelo multinivel a partir de ello.

2.7.2.2 Inclusión de predictores en el nivel macro

Al incluir un predictor propio del nivel macro Wj .Se plantean los modelos macro:

$$\beta_{0j} = \beta_{00} + \beta_{01}W_j + \mu_{0j} \tag{2.31}$$

$$\beta_{1j} = \beta_{10} + \beta_{11}W_j + \mu_{1j} \tag{2.32}$$

Donde las variables aleatorias $\mu_{0j}\ y\ \mu_{1j}$ son variables aleatorias con media cero y varianzas $\sigma_{\mu_0}^2, \sigma_{\mu_1}^2$.

Reemplazando (2.31) y (2.32) en (2.20)

$$y_{ij} = \beta_{00} + \beta_{01}W_j + \beta_{11}W_j x_{ij} + (\mu_{1j}x_{ij} + \mu_{0j} + \varepsilon_{ij}) \tag{2.33}$$

Generalizando para cualquier número de variables tendríamos en el nivel micro:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \cdots + \beta_{pj}x_{pij} + \varepsilon_{ij} \tag{2.34}$$

Donde ε_{ij} ~N $(0,\sigma_\varepsilon^2)$.

Para el nivel macro:

$$\beta_{0j} = \beta_{00} + \beta_{01}W_{1j} + \beta_{02}W_{2j} + \cdots + \beta_{0L}W_{Lj} + \mu_{0j}$$

$$\beta_{1j} = \beta_{10} + \beta_{11}W_{1j} + \beta_{12}W_{2j} + \cdots + \beta_{1L}W_{Lj} + \mu_{1j}$$

$$\cdots$$

$$\beta_{pj} = \beta_{p0} + \beta_{p1}W_{1j} + \beta_{p2}W_{2j} + \cdots + \beta_{pL}W_{Lj} + \mu_{pj}$$

Y la distribución de la variación entre contextos o grupos es:

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \cdots \\ \mu_{pj} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}, T\right] \tag{2.35}$$

Donde T es la matriz de varianzas y covarianzas definida como:

$$T = \begin{pmatrix} \sigma^2_{\mu_0} \sigma^2_{\mu_0\mu_1} \ldots \sigma^2_{\mu_0\mu_p} \\ \ldots \quad \sigma^2_{\mu_1} \quad \ldots \sigma^2_{\mu_1\mu_p} \\ \ldots \quad \ldots \quad \ldots \\ \qquad\qquad \sigma^2_{\mu_p} \end{pmatrix}$$

El modelo multinivel en general se puede escribir de la siguiente forma:

$$y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon_{ij} \tag{2.36}$$

donde ε_{ij} ~N(0,σ^2_ε)

$$\beta_{pj} = \beta_{00} + \sum_{l=1}^{L} \beta_{Pl} W_{lj} + \mu_{pj} \tag{2.37}$$

donde μ_j ~N(0,T)

Si algunos de los términos del modelo jerárquico son iguales a cero podemos obtener una serie de submodelos como el modelo ANOVA de un criterio con efectos aleatorios, modelo de regresión con medias como respuesta, modelo de análisis de covarianza ANCOVA con efectos aleatorios, modelo de regresión con efectos aleatorios, modelo con interceptos y pendientes como respuestas y otros.

Los modelos mencionados asumen que los errores en ambos niveles son homogéneos, sin embargo el modelo puede ser extendido para tratar estructuras más complejas en ambos niveles.

2.7.2.3 Ajuste del modelo

El objetivo es analizar la significancia de los coeficientes del modelo y analizar el ajuste global del modelo.

Si bien la significancia de un predictor se evalúa con la estadística t de student:

Estimador del parámetro / Error típico del parámetro, se debe tener en cuenta la teoría y el contexto en el que se desenvuelve el estudio.

Para analizar el ajuste global del modelo, comparamos el modelo propuesto versus el modelo nulo, el cual se encuentra anidado en el primero.

Para realizar la comparación se utiliza la función de verosimilitud $-2ln\frac{L_1}{L_2} \sim X^2_p$, p es el número de parámetros de diferencia.

El valor de la desvianza nos indicará si el modelo propuesto es mejor que el modelo nulo.

Al comparar las desvianzas de dos modelos propuestos con m_1 parámetros y m_2 parámetros, respectivamente, cuyas desvianzas son D_1 y D_2, la diferencia de ellas con distribución X^2_{m1-m2} se utiliza como estadística de prueba. Se elegirá el modelo que resulte significativo, esperando que sea el que mayor varianza explique y tenga el menor número de parámetros.

2.7.2.4 Estimación de parámetros

Los parámetros del modelo a estimar son los parámetros fijos y aleatorios. Los parámetros fijos corresponden a la pendiente y el intercepto. Los aleatorios corresponden a las varianzas y covarianzas de todos los niveles[12].

- ***Estimación de parámetros de efectos fijos***

Sea el modelo nulo

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \ , \text{ donde } \ \varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon) \ .$$

$$\beta_{0j} = \beta_0 + \mu_{0j}, \text{ donde } \ \mu_{oj} \sim N(0,\sigma^2_{\mu_o}) \ .$$

El parámetro a estimar es β_0 , la media general en la variable de respuesta para el conjunto de la población.

Partiendo de que hay j grupos de tamaño n_j, $\bar{y}_{.j} = \beta_{0j} + \bar{\varepsilon}_{.j}$, donde $\bar{\varepsilon}_{.j} \sim N\ (0,\frac{\sigma^2_\varepsilon}{n_j})$.

Reemplazando $\bar{y}_{.j} = \beta_{0j} + (\mu_{oj} + \bar{\varepsilon}_{.j})$ es el estadístico con el que se quiere estimar β_0 y es un estimador insesgado, $\mu_{oj} + \bar{\varepsilon}_{.j}$ es el término de error cuyas medias son iguales a cero.

La varianza de $\bar{y}_{.j}$ es $V(\bar{y}_{.j}) = V(\beta_{0j} + \mu_{oj} + \bar{\varepsilon}_{.j}) = V(\mu_{oj}) + V(\bar{\varepsilon}_{.j}) = \sigma^2_{\mu_o} + \frac{\sigma^2_\varepsilon}{n_j}$, donde $\sigma^2_{\mu_o}$ es la varianza entre los grupos y $\frac{\sigma^2_\varepsilon}{n_j}$ es la varianza dentro de los grupos.

Si se define $\Delta_j = \sigma^2_{\mu_o} + \frac{\sigma^2_\varepsilon}{n_j}$, la eficiencia del estimador $\bar{y}_{.j}$ será Δ_j^{-1} . Si se conociera cada Δ_j^{-1} una solución sería ponderar cada $\bar{y}_{.j}$. Entonces el estimador de mínimos cuadrados ponderados será:

$$\widetilde{\beta_0} = \frac{\sum \Delta_j^{-1} \bar{y}_{.j}}{\sum \Delta_j^{-1}} \tag{2.38}$$

Donde $\Delta_j^{-1} = \frac{n_j}{\sigma_\varepsilon^2}$ si todas las medias de los grupos son iguales, $\sigma_{\mu_o}^2 = 0$ y $\Delta_j^{-1} = \frac{1}{\sigma_{\mu_o}^2}$ si casi toda la varianza es varianza entre los grupos[13].

- ***Estimación conjunta de parámetros de efectos fijos y las varianzas***

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \mu_{0j} + \varepsilon_{ij}$$

Al no conocer los valores de los parámetros fijos se puede hacer uso de una de las propiedades del estimador de máxima verosimilitud: Si en cualquier función de los parámetros se sustituye a estos por los estimadores de máxima verosimilitud, entonces la función resultante es a su vez un estimador de máxima verosimilitud. Reemplazamos $\beta_0 \, y \, \beta_1$ por sus estimadores máximos verosímiles.

Otro método de estimación de σ_ε^2 es mediante el *estimador máximo verosímil iterativo.* Cuando no se conoce los valores de $\beta_0 \, y \, \beta_1$, ni la varianza de los residuos, se asigna un valor proporcional a la varianza de los residuos y se obtiene un valor provisional para los estimadores de la parte fija. Se toman los estimadores de la parte fija y se vuelve a estimar la varianza. Se repite el proceso hasta que $\left|\sigma_\varepsilon^{2(n+1)} - \sigma_\varepsilon^{2(n)}\right| < \epsilon$

Existen algoritmos de estimación como:

- Maximización esperada (EM): En caso se presenten datos perdidos. Encuentra estimadores máximo verosímiles.
- Mínimos cuadrados generalizados iterativo(IGLS)
- IGLS restringido (RIGLS)
- Cuasiverosimilitud marginal (MQL)
- Cuasiverosimilitud penalizada (PQL)

2.7.2.5 Explicación de la varianza

Se busca definir en los modelos multinivel de 2 niveles, una medida similar a la medida de correlación múltiple R^2 utilizada en los modelos de regresión.

Una definición de la proporción de la varianza explicada se define por Snijder y Bosker como "Reducción de la proporción del error de predicción" [16].

Se parte de una regresión simple, donde la estimación de Y es E(Y) y cuya varianza es V(Y). Si se conoce los valores de X para el sujeto i, la predicción de Y_i es $\sum \beta_h X_{hi}$. El error de predicción es la diferencia entre Y_i y $\sum \beta_h X_{hi}$ y la varianza del error de predicción es $E(Y_i - \sum \beta_h X_{hi})^2$, donde $Y_i - \sum \beta_h X_{hi} = \mu_{oj} + e_{ij}$, entonces $V(Y_i - \sum \beta_h X_{hi}) = \sigma^2_{\mu_o} + \sigma^2_e$.

Entonces la reducción de la proporción de la varianza del error de predicción al introducir predictores es:

$$R^2 = \frac{V(Y_i) - V(\varepsilon_i)}{V(Y_i)} = 1 - \frac{V(\varepsilon_i)}{V(Y_i)} = 1 - \frac{\sigma^2_{\mu_o} + \sigma^2_e}{V(Y_i)}.$$

En un modelo de regresión multinivel se puede buscar predecir el valor de Y de un individuo o predecir el valor medio de un grupo.

Partiendo de un modelo de intercepto aleatorio $y_{ij} = \beta_0 + \sum_{h=1}^{q} \beta_h x_{hij} + \mu_{ij} + \varepsilon_{ij}$

Se cumple:

$$\hat{R}_1^2 = \frac{\sigma^2_{\mu_0} + \sigma^2_{\varepsilon}}{V(Y_{ij})} = 1 - \frac{\left(\hat{\sigma}^2_{\mu_0} + \hat{\sigma}^2_{\varepsilon}\right)_{modelo\ nulo}}{\left(\hat{\sigma}^2_{\mu_0} + \hat{\sigma}^2_{\varepsilon}\right)_{modelo\ con\ predictores}}$$, para el nivel 1 o individual (2.40)

$$\hat{R}_2^2 = 1 - \frac{V(\overline{Y_{.j}} - \sum \beta_h \bar{X}_{hi\cdot})}{V(\overline{Y_{.j}})} = 1 - \frac{\left(\hat{\sigma}^2_{\mu_0} + \frac{\hat{\sigma}^2_{\varepsilon}}{n}\right)_{modelo\ nulo}}{\left(\hat{\sigma}^2_{\mu_0} + \frac{\hat{\sigma}^2_{\varepsilon}}{n}\right)_{modelo\ con\ predictores}}$$, para el nivel 2 o grupo (2.41)

Si los tamaños de cada grupo varían mucho, puede tomarse la media armónica, el valor $\frac{N}{\sum_j \frac{1}{n_j}}$ para n.

Respecto a R_1^2 y R_2^2, sus valores poblacionales no pueden ser menores de cero. En cambio sus estimaciones pueden aumentar su valor al eliminar un predictor o disminuir al incluir un nuevo predictor, esto puede ser debido al azar o por una mala especificación de la parte fija

del modelo. Si el modelo es de pendientes aleatorias los cálculos son más complejos y los resultados no difieren mucho de los valores para el modelo anidado de interceptos aleatorios[16].

2.8 Modelo logístico multinivel

Hasta el momento se ha trabajado con una variable respuesta continua y bajo el supuesto que los residuos tienen distribución normal. Sin embargo hay estudios, como el presente trabajo, en los que la variable respuesta es una variable discreta. Como se mencionó, este tipo de variables forman parte de la familia de *modelos lineales generalizados.*

Este trabajo se basa principalmente en una variable discreta dicotómica, donde la variable respuesta tiene distribución binomial. El modelo a explicar es un modelo de regresión logística multinivel, exactamente de dos niveles.

2.8.1 Modelo logístico multinivel

El modelo logístico multinivel incluye los efectos aleatorios de los grupos. Al considerar 2 niveles y P variables predictoras X_p (p=1,…,P) en el primer nivel y L variables predictoras W_l (l=1,…,L) en el segundo nivel, la probabilidad de éxito no solo dependerá del individuo sino también del grupo por lo que será denotado por π_{ij}.

El modelo de regresión logística multinivel será:

$$\text{logit}\,(\pi(x_{ij})) = y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon_{ij} \quad , \quad \text{j=1,…,J ; i=1,…, } n_j \tag{2.42}$$

$$\beta_{0j} = \beta_{00} + \sum_{l=1}^{L} \beta_{0l} W_{lj} + \mu_{0j} \tag{2.43}$$

$$\beta_{pj} = \beta_{00} + \sum_{l=1}^{L} \beta_{Pl} W_{lj} + \mu_{pj} \tag{2.44}$$

En el modelo presentado se consideran J grupos y dentro de cada grupo una muestra aleatoria n_j , j=1,..,J y N=$\sum n_j$. La variable respuesta Y_{ij} es dicotómica por lo que toma los valores 0 para éxito y 1 para fracaso.

La probabilidad de éxito denotada por π_{ij} es considerada como variable aleatoria.

Al igual que en un modelo de intercepto aleatorio para una variable respuesta continua, el intercepto consta de dos términos: una componente fija y una componente aleatoria μ_{oj} , la cual se asume que sigue una distribución normal con media cero y varianza $\sigma^2_{\mu_o}$. No se incluye los residuos del primer nivel porque es una ecuación para la probabilidad π_{ij} y no para Y_{ij} .

Debido a que la variable respuesta toma valores 0 y 1, el promedio de cada grupo, $\bar{Y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$ es la proporción de éxitos en el grupo j y la proporción total de éxitos es $\hat{\pi}_{.} = \bar{Y}_{..}$.

Para probar si en realidad existen diferencias entre los grupos, se puede utilizar el estadístico: $\chi^2 = \sum_{j=1}^{J} n_j \frac{(\bar{Y}.j - \hat{\pi}.)^2}{\hat{\pi}.(1-\hat{\pi}.)}$, con J-1 g.l. bajo los supuestos requeridos[16].

También se puede utilizar la prueba dada por COMENGES y JACQMIN [10]

$T = \frac{\sum_{j=1}^{J}\{n_j^2(Y_{.j} - \hat{\pi}.)^2\} - N\hat{\pi}.(1-\hat{\pi}.)}{\hat{\pi}.(1-\hat{\pi}.)\sqrt{2\sum_{j=1}^{J} n_j\ (n_j\ -1)}}$, con distribución normal estándar, la ventaja de este estadístico es que puede ser aplicado cuando se tienen muchos grupos (J <10 y $\frac{n_j}{N}$ <0.10).

2.8.2 Modelo nulo

El modelo nulo de dos niveles para una variable respuesta dicotómica, solo considera la variación entre grupos sin tomar en cuenta variables explicativas.

Haciendo uso de la función de enlace logit, la cual permite convertir la variable respuesta y linealizar la relación entre la variable respuesta y la parte sistemática del modelo

$\text{logit}(\pi_{ij}) = \beta_{0j} + e_{ij}$, $\beta_{0j} = \beta_{00} + u_{0j}$.

β_{00} es el promedio de los π_{ij} y μ_{oj} es un error aleatorio asociado con la j-ésima unidad del nivel 2 y se asume que tiene distribución normal con media 0 y varianza $\sigma^2_{\mu_o}$.

El modelo nulo completo es: $\text{logit}(\pi_{ij}) = \beta_{00} + u_{0j} + e_{ij}$

Entonces: Var $(\beta_{00} + u_{0j} + e_{ij}) = \sigma^2_{\mu_o} + \sigma^2$

El coeficiente de correlación intraclase es $\rho = \frac{\sigma^2_{\mu_o}}{\sigma^2_{\mu_o} + \sigma^2}$, el cual mide la proporción de la variación de la respuesta que es explicada por las variables del nivel 2.

2.8.3 Modelo de intercepto aleatorio

Al incluir variables explicativas en el modelo, p variables del primer nivel, l variables en el segundo nivel y considerando solo el intercepto aleatorio, tenemos el siguiente modelo:

$\text{logit}\,(\pi(x_{ij})) = y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon_{ij}$, j=1,…,J ; i=1,…, n_j

$\beta_{oj} = \beta_{00} + \sum_{l=1}^{L} \beta_{0l} W_{lj} + \mu_{0j}$

$\mu_{0j} \sim N(0, \sigma^2_{\mu_0})$

Donde β_{00} se puede interpretar como el intercepto en conjunto y $\beta_{00} + \mu_{0j}$ el intercepto para una determinada unidad del nivel macro.

Un modelo multinivel para respuestas binarias se puede derivar también a través de una variable latente de contextualización.

Asumimos que existe una variable continua Y* subyacente a Y, así se puede formular un modelo llamado umbral:

$$y_{ij} = \begin{cases} 1 & si\ y^*_{ij} \geq 0 \\ 0 & si\ y^*_{ij} < 0 \end{cases}$$

Teniendo en cuenta esta representación se puede escribir el siguiente modelo de regresión logística multinivel de 2 niveles para la variable Y*:

$\text{logit}\,(\pi(x_{ij})) = y*_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} x_{pij} + \varepsilon*_{ij}$, j=1,…,J ; i=1,…, n_j

Sin embargo, para que represente un modelo de regresión logística, los residuos del primer nivel considerando la variable Y* deben tener una distribución logística:

- P ($\varepsilon*_{ij} < x$) = log (x)
- La media de los residuos del primer nivel es 0.
- La varianza es $\frac{\pi^2}{3}$ =3.29.

Cuando se asume que $\varepsilon*_{ij}$ tiene esta distribución, el modelo logístico multinivel es equivalente al modelo umbral definido.

2.8.4 Estimación de los parámetros.

La estimación de parámetros se realizará por métodos iterativos.

Sea

$$E(Y\ /X = x\ ) =\ \pi(X) = \frac{e^{X_{ij}\beta+\mu_j}}{1+e^{X_{ij}\beta+\mu_j}},$$

Donde $X_{ij}\beta$ se refiere a la parte fija y μ_j a la parte aleatoria.

Para poder estimar los parámetros es necesario alinear la función exponencial y luego aplicar el método de estimación de cuasi-verosimilitud.

Actualmente se utilizan algoritmos computacionales incluidos en diversos software para poder realizar las estimaciones.

Estudios de simulación demuestran que cuando la variable respuesta es binaria, el número de unidades de nivel 1 en una unidad de nivel 2 es pequeño y la varianza entre grupos es grande, se pueden producir sesgos muy grandes en la estimación[18].

Un ajuste de las estimaciones se basa en el desarrollo de series de Taylor[17], se utiliza el desarrollo por series de Taylor de primer orden sobre las estimaciones actuales para la parte fija. Para la parte aleatoria se utiliza el desarrollo de segundo orden sobre 0, ambas son modificadas para obtener estimaciones más precisas. Aplicando el método de mínimos cuadrados generalizados se obtiene:

$$f(H_{t+1}) = f(H_t) + X_{ij}\left(\hat{\beta}_{t+1} - \hat{\beta}_t\right) + f'(H_t) +\ \mu_j f'(H_t) + \frac{\mu_j f''(H_t)}{2} \qquad (2.44)$$

Donde

$$f'(H) = f(H)(1 + \exp(H))^{-1}$$

$$f''(H) = f'(H)(1 - \exp(H))(1 +\ exp\ (H))^{-1}$$

Hay dos opciones para fijar H_t :

a. $H_t = X_{ij}\hat{\beta}_t$, utiliza la parte fija para el desarrollo por series de Taylor y es conocido como método de cuasi verosimilitud marginal (MQL). El procedimiento MQL de 1º orden, proporciona una primera aproximación y se pueden obtener estimaciones

sesgadas, especialmente si el tamaño dentro de las unidades del segundo nivel es pequeño.

b. $H_t = X_{ij}\hat{\beta}_t + \hat{\mu}_{tj}$, utiliza los residuos estimados y es conocido como método de cuasi verosimilitud penalizada (PQL).

Rodriguez y Goldman consideraron que el MQL con una corrección de segundo orden mejora las estimaciones[18], mientras que Goldstein y Rasbash consideraron la corrección de segundo orden PQL[17]. El procedimiento PLQ de 2º orden mejora la aproximación aunque es un método menos estable y puede dar problemas de convergencia.

Se considerará obtener valores de inicio con el procedimiento MQL de 1º orden los cuales servirán para la obtención de las estimaciones a través del procedimiento PQL de 2º orden.

CAPITULO III
ESTUDIO DE LA ROTACIÓN LABORAL

3.1 Método

3.1.1 Tipo y diseño de investigación

Se utilizó un diseño cuasi experimental, retrospectivo y de tipo transversal.

3.1.2 Población y muestra

La población está formada por los trabajadores de la empresa en estudio, que se desvincularon, cesaron o renunciaron, durante el tiempo de actividad de la empresa.

No fue necesario realizar una encuesta para la recolección de datos por la accesibilidad al sistema de información existente.

Se consideraron las desvinculaciones de trabajadores presentadas durante el periodo Enero 2010 - Agosto 2012. Los datos fueron obtenidos de una fuente secundaria, el sistema de información EXACTUS de donde se obtuvieron campos como fecha de nacimiento, edad, sexo, dirección, estado civil, número de hijos, motivo de cese, fecha de ingreso, fecha de cese, puesto y área asociados al trabajador que se desvinculó de la empresa.
Se clasificaron todos los casos de desvinculaciones en dos grupos, ceses y renuncias. En el primer grupo se consideraron los casos donde la empresa decidió la desvinculación, anulación de ingreso o cambio de modalidad de contrato. En el grupo renuncias, se consideraron las desvinculaciones a solicitud del trabajador por motivos asociados a estudios, otro trabajo, viaje, salud u otros.

La empresa cuenta con más de 20 áreas organizacionales, sin embargo se decidió considerar las 10 áreas relacionadas a los negocios más importantes, donde existe mayor rentabilidad y riesgo en la operación.

Para el desarrollo y análisis de resultados, se consideraron 2249 renuncias en el periodo Enero 2010 - Agosto 2012, distribuidas en 10 áreas.

Cuadro 1

Distribución de renuncias por área en el periodo Enero 2010 - Agosto 2012.

AREA	TOTAL
CANALES	391
DISTRIBUCION	368
BLINDADOS	267
SUCURSALES	267
MULTISER	251
ADMINISTRATIVOS	176
SEGURIDAD	170
PROCESAMIENTO	167
BPO	114
SEGURIDAD EXTERNA	78
TOTAL	**2249**

Fuente: Propia.

Si bien la rotación es un problema por los costos y pérdidas que genera, el nivel de rotación esperado es distinto para cada área. El servicio brindado por el área de Canales, por ejemplo, contempla un nivel de rotación alto ya que se espera que personal joven y que se encuentre estudiando cubra las vacantes disponibles. Por ello, además de identificar las características del personal que renuncia antes de cumplir el periodo de prueba de 6 meses, se buscará identificar las diferencias entre áreas.

3.1.3 Operacionalización de variables

Debido a que las variables obtenidas desde la base datos de la empresa no se podían utilizar directamente, se adecuaron algunas con la finalidad de obtener las variables necesarias. Estos cambios se muestran en el cuadro 2.

Cuadro 2

Definición de variables

CAMPO INICIAL	VARIABLE	OBSERVACIONES
FECHA DE NACIMIENTO	EDAD	
SEXO	SEXO	
DIRECCION	DISTANCIA	Distancia del domicilio del trabajador a LA EMPRESA. La clasificación se realizó tomando como criterio la distancia entre el distrito limeño de residencia y el distrito donde se ubica el centro de trabajo. Solo en el caso de provincias se consideró como sucursales.
ESTADO CIVIL	ESTADO CIVIL	Casado: Se consideran los casados y convivientes. Soltero: Se considera solo a los trabajadores que hayan declarado ser solteros a su ingreso.
NÚMERO DE HIJOS	NÚMERO DE HIJOS	Número de hijos registrados en el sistema de la empresa.
FECHA DE INGRESO VS FECHA DE CESE	RENUNCIÓ ANTES DEL PERIODO DE PRUEBA	Indica si el trabajador renunció pasado el periodo de prueba o no. El periodo de prueba es de 6 meses desde su ingreso.
PUESTO AREA HORARIO DE TRABAJO	AREA	Area de trabajo, sector de negocio.
	GRUPO OCUPACIONAL	Clasificación del puesto que desempeña el trabajador de acuerdo a la misión del puesto.
	ESCALA REMUNERATIVA	Intervalo remunerativo asignado para fijar el total de ingresos percibidos por el colaborador de acuerdo al área y puesto que desempeña.
EVALUACION DE DESEMPEÑO	EVALUACION DEL JEFE	Puntaje obtenido en la última evaluación de desempeño realizada al jefe del área.
NÚMERO DE TRABAJADORES	NÚMERO DE TRABAJADORES	Número de trabajadores por área.

Fuente: Propia.

Se utilizó la siguiente clasificación para indicar la DISTANCIA del domicilio del trabajador al centro de trabajo.

Cuadro 3

Definición de variables

Variable DISTANCIA

DISTRITOS	DISTANCIA
Barranco , Chorrillos, Miraflores, S.J.M.,Surco, Surquillo, Villa el salvador, V.M.T.	CERCA
Ate Vitarte, Callao, Carabayllo, Cercado de Lima, Comas, El Agustino, Independencia, Los Olivos, Puente Piedra, Rimac, Santa Anita, S.J.L, S.M.P, Ventanilla.	LEJOS
Breña, Jesús Maria, La Molina, La Victoria, Lince, Magdalena, Pueblo Libre, San Borja, San Isidro, San Luis, San Miguel.	MEDIO
Ciudades en otros departamentos.	SUCURSALES

Fuente: Propia.

Al establecer intervalos de remuneración, la variable ESCALA REMUNERATIVA se define como el valor asignado a cada rango de remuneración.

Uno de los objetivos del estudio es reconocer las características de los posibles trabajadores que se renuncian, por decisión propia, de LA EMPRESA antes del término del periodo de prueba de acuerdo al área donde laboran, haciendo uso del análisis de regresión logística multinivel.

Cuadro 4

Variables por nivel

					VALORES FINALES	
					CATEGORIA BASE	CATEGORIA DE CONTRASTE
RENUNCIÓ ANTES DE CULMINAR PERIODO DE PRUEBA		PERIODO DE PRUEBA		Y	Renunció despues de culminar periodo de prueba (0)	Renunció antes de culminar periodo de prueba (1)
NIVEL 1	TRABAJADOR	EDAD	Edad actual del trabajador.	X1	De 30 a mas años(0)	De 18 a 30 años (1)
		SEXO		X2	Femenino (0)	Masculino (1)
		DISTANCIA	Distancia del domicilio del trabajador a LA EMPRESA. La clasificación se realizó tomando como criterio la distancia entre el distrito limeño de residencia y el distrito donde se ubica el centro de trabajo. Solo en el caso de provincias se consideró como sucursales.	X3	Cerca(0)	Lejos(1) Medio(2) Sucursales(3)
		ESTADO CIVIL	Casado: Se consideran los casados y convivientes. Soltero: Se considera solo a los trabajadores que hayan declarado ser solteros a su ingreso.	X4	Casado (0)	Soltero (1)
		NÚMERO DE HIJOS	Número de hijos registrados en el sistema de la empresa.	X5		
NIVEL 2	ÁREA	GRUPO OCUPACIONAL	Clasificación del puesto que desempeña el trabajador de acuerdo a la misión del puesto.	W1	Administrativo(0)	Operativo(1)
		ESCALA REMUNERATIVA	Intervalo remunerativo asignado para fijar el total de ingresos percibidos por el colaborador de acuerdo al área y puesto que desempeña.	W2	Puntos	
		EVALUACIÓN DEL JEFE	Puntaje obtenido en la última evaluación de desempeño realizada al jefe del área.	W3	Puntos	
		NUMERO DE TRABAJADORES	Número de trabajadores en el área.	W4	Menos de 300 (0)	Mas de 300 (1)
		BENEFICIOS ADICIONALES	Beneficios brindados a los trabajadores por el área donde laboran , adicionales a los brindados por la empresa de manera general.	W5	Si tiene (0)	No tiene (1)

Fuente: Propia.

Las variables mencionadas han sido seleccionadas debido a que es la información inicial que se tiene cuando un trabajador se incorpora, se espera con estos datos iniciales identificar a los trabajadores que podrían renunciar durante el periodo de prueba**.**

3.2 Procesamiento y análisis de datos

El análisis se desarrolló en 2 fases, en la primera se realizó el estudio descriptivo de las variables involucradas.

Con la finalidad de validar la correcta selección de variables y analizar la asociación de las variables regresoras propuestas con el hecho de renunciar antes de culminar el periodo de prueba, se realizó el análisis de regresión logística.

Por último se planteó un modelo de regresión logístico multinivel, de intercepto aleatorio, y se evaluó la existencia de diferencias en las características de los renunciantes entre áreas. Se compararon los resultados del modelo de regresión logística con los obtenidos con el modelo de regresión logística multinivel.

Para la definición del modelo se consideraron 2249 (N=2249) retiros presentados en el periodo establecido agrupados en 10 áreas (J=10). Las variables en el primer nivel fueron 5 (P=5): $X_{1:}$ *Edad del trabajador,* $X_{2:}$ *Sexo del trabajador,* $X_{3:}$ *Distancia,* $X_{4:}$ *Estado civil,* $X_{5:}$ *Número de hijos.* Las variables dummy creadas a partir de la variable Distancia son: $X_{31:}$ *Distancia lejos,* $X_{32:}$ *Distancia medio,* $X_{33:}$ *Distancia sucursales.*

Las variables en el segundo nivel fueron 5 (L=5): $W_{1:}$ *Grupo ocupacional,* $W_{2:}$ *Escala remunerativa,* $W_{3:}$ *Evaluación del jefe,* $W_{4:}$ *Número de trabajadores,* $W_{5:}$ *Cuenta con beneficios adicionales.*

Modelo nulo:

Como parte del análisis multinivel se planteó previamente un modelo nulo, el cual no incluía variables explicativas. Se realizó la descomposición de la varianza total en la variación entre trabajadores de la misma área y la variación entre áreas.

$logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j}$, $\beta_{0j} = \beta_{00} + \mu_{0j}$ donde se asume que $\mu_{0j} \sim N(0, \sigma^2_{\mu_0})$.

j = 1,…,10

Se calculó la correlación intraclase,ρ_1, para medir la proporción de la varianza total explicada por las diferencias entre áreas. Considerando la definición de la variable latente subyacente se asume que los residuos del primer nivel tienen varianza $\pi^2/3$.

Modelo de intercepto aleatorio:

Se planteó un modelo de intercepto aleatorio el cual incluía todas las variables propuestas para el primer y segundo nivel.

$$logit(\pi_{ij}) = \beta_{0j} + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_{ij}$$

$$\beta_{0j} = \beta_{00} + \beta_{01} W_{1j} + \cdots + \beta_{0l} W_{lj} + \mu_{oj}$$

$$logit(\pi_{ij}) = \beta_{00} + \sum_{l=1}^{5} \beta_{0l} W_{lj} + \sum_{p=1}^{5} \beta_{pj} X_{pi} + \mu_{0j} + \varepsilon_{ij}$$

$\mu_{0j} \sim N(0, \sigma^2_{\mu_0})$, $\varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon})$

i = 1,…,2249; j = 1,…,10

$$\begin{aligned} logit(\pi_{ij}) = {} & \beta_{00} + \beta_{01} Grupo\ ocupacional_j + \beta_{02} Escala\ remunerativa_j \\ & + \beta_{03} Evaluación\ del\ jefe_j + \beta_{04} Número\ de\ trabajadores_j \\ & + \beta_{05} Beneficios\ adicionales_j + \beta_{1j} Edad\ del\ trabajador_i \\ & + \beta_{2j} Sexo\ del\ trabajador_i + \beta_{3j} Distancia_i + \beta_{4j} Estado\ civil_i \\ & + \beta_{5j} Número\ de\ hijos_i + \mu_{0j} + \varepsilon_{ij} \end{aligned}$$

Se empleó el método de estimación cuasi verosimilitud penalizada (PQL) para eliminar el sesgo cuando el número de trabajadores en cada área es reducido o la varianza del nivel área es grande.

Se calculó nuevamente la correlación intraclase ρ_2 para ver la variación del modelo al incluir variables explicativas.

- **Software**

Se utilizaron 2 paquetes estadísticos para el análisis: “IBM SPSS Statistics” versión 19 para el análisis descriptivo y el análisis de regresión logística y “MLwiN” versión 2.26 para el análisis multinivel.

Existen otros programas para analizar datos con estructura jerárquica, como HLM, STATA, LISREL entre otros. Se eligió MLwiN por ser un programa gratuito, amigable y más completo debido a que incluye métodos y conceptos que se explicaron en el capítulo anterior.

3.3 Resultados

- **Análisis descriptivo**

En el periodo de análisis de 2 años y medio, de los 2249 casos de retiro presentados, 1283 (57.05%) se retiraron antes de culminar el periodo de prueba de 6 meses. El rango de edades era entre 18 y 60 años. 1064 (83%) trabajadores eran jóvenes entre 18 y 30 años, 761 (59.31%) hombres, 831 (64.77%) solteros y sin hijos. Respecto a la distancia del domicilio al centro de trabajo 492 (38%) vivían cerca y 499 (39%) lejos.
El mayor porcentaje de retiros antes de culminar el periodo de prueba se presentó en el grupo ocupacional operativo (95.87%), en áreas con más de 300 trabajadores, cuyo jefe tiene una evaluación de desempeño menor al 80% y que brindan beneficios adicionales a los ofrecidos a todos los trabajadores por igual. El personal con puestos operativos participan directamente de la operación de los diferentes negocios, cuentan con horarios rotativos y jornadas extensas.

Las áreas operativas más críticas para el negocio presentaron un porcentaje de retiro durante el periodo de prueba mayor a 60%. Entre ellas se encuentran Seguridad, Blindados, Procesamiento, Canales. (Ver cuadro 6).

En el cuadro 5 se muestra la distribución de las frecuencias de retiros voluntarios en el periodo enero 2010- agosto 2012 según las variables investigadas.

Cuadro 5

Análisis descriptivo. Retiros antes de culminar periodo de prueba por variable.

	Retiros antes de culminar periodo de prueba		
	No (%)	Si (%)	Total (%)
Variable independiente			
Retiro antes de culminar PP	42,95	57,05	100
Variables del trabajador			
Edad			
18-30 años	39,73	60,27	78,49
30 - mas años	54,77	45,23	21,51
Sexo			
Femenino	44,23	55,77	36,69
Masculino	42,04	57,96	63,31
Estado civil			
Soltero	39,81	60,19	33,02
Casado	44,53	55,47	66,98
Distancia			
Cerca	41,22	58,78	37,22
Lejos	39,92	61,08	36,33
Medio	56,90	43,10	7,74
Sucursales	48,460	51,540	18,72
Número de hijos promedio	1	0	1
Variables del area			
Grupo ocupacional			
Administrativo	69,89	30,11	11,98
Operativo	40,67	59,33	88,02
Escala remunerativa promedio	4	3	4
Evaluación del jefe promedio	0,75	0,73	0,73
Número de trabajadores			
Menos de 300	36,79	63,21	34,75
Mas de 300	46,22	53,78	65,26
Beneficios adicionales			
Si	40,91	59,09	55,74
No	45,27	54,73	44,26

n=2249

Fuente: Propia.

Cuadro 6

Porcentaje de trabajadores que se retiraron antes de culminar el periodo de prueba por área.

AREA	RETIRO DESPUES DEL PERIODO DE PRUEBA %	RETIRO DURANTE EL PERIODO DE PRUEBA (Antes de culminar P.P.) %
SEGURIDAD	**14,12%**	**85,88%**
SEGURIDAD EXTERNA	**29,49%**	**70,51%**
BLINDADOS	**34,83%**	**65,17%**
PROCESAMIENTO	**35,93%**	**64,07%**
CANALES	**36,83%**	**63,17%**
BPO	**41,23%**	**58,77%**
SUCURSALES	**47,94%**	**52,06%**
DISTRIBUCION	51,90%	48,10%
MULTISER	52,99%	47,01%
ADMINISTRATIVOS	69,89%	30,11%
Total (%)	42,95%	57,05%

Fuente: Propia.

- ***Modelo de regresión logística***

La finalidad de realizar un modelo de regresión logística fue identificar las variables significativas para la predicción de los retiros durante el periodo de prueba. Sirvió de referencia para el análisis multinivel realizado.

Si bien se trabajaron con todas las variables definidas, no se identificaron si pertenecen a diferentes niveles, si correspondían a variables del trabajador o del área.

Considerando solo la constante, el modelo nulo clasifica correctamente el 57% de los casos. La desviación disminuye con la incorporación de las variables explicativas, esto indica que el modelo se ajusta mejor al incluir las variables. La clasificación de los casos mejora de 57% a 62%.

Se consideraron en el modelo las variables edad, sexo, distancia, estado civil, número de hijos, grupo ocupacional, escala remunerativa, evaluación del jefe de área, cantidad de trabajadores en el área y beneficios adicionales. Utilizando la estadística de Wald se encontró que la mayoría de variables consideradas fueron significativas, sin embargo el

ajuste global del modelo no es muy bueno según nos indica las medidas de R^2 de Cox y Snell ($R^2 = 0.5$).

Las variables que no resultaron significativas fueron: distancia, escala remunerativa, evaluación del jefe y beneficios adicionales.

El modelo de regresión logística obtenido es el siguiente:

$$\pi(X) = \frac{1}{1 + e^{-\eta}}$$

Donde

$$\begin{aligned} logit(\pi(X)) = {} & 1.386 + 1.180 * Grupo\ ocupacional\ operativo_i + 0.020 \\ & * Escala\ remunerativa_i + 0.0638 * Evaluación\ del\ jefe_i - 0.295 \\ & * Número\ de\ trabajadores\ mayor\ a\ 300_i - 0.013 \\ & * Beneficios\ adicionales\ no\ recibidos_{\ i} + 0.583 \\ & * Edad\ del\ trabajador\ de\ 18\ a\ 30_i + 0.169 \\ & * Sexo\ del\ trabajador\ masculino_{\ i} + 0.099 * Distancia\ lejos_i - 0.224 \\ & * Distancia\ medio_i - 0.315 * Distancia\ sucursal_i - 0.376 \\ & * Estado\ civil_i - 0.068 * Número\ de\ hijos_i + \varepsilon_i \end{aligned}$$

Los resultados mostrados en el cuadro 7 señalan que los trabajadores entre 18 y 30 años presentan mayor posibilidad de retirarse antes de culminar el periodo de prueba comparados con los mayores de 30 años.

Los trabajadores de sexo masculino presentan mayor posibilidad de retirarse antes de culminar el periodo de prueba comparados con los trabajadores de sexo femenino.

Los trabajadores que viven lejos del centro de trabajo presentan mayor posibilidad de retirarse antes de culminar el periodo de prueba comparados con los que viven cerca.

Los casados tienen mayor posibilidad de retirarse antes de culminar el periodo de prueba comparados con los solteros.

Al incrementarse la cantidad de hijos disminuye la posibilidad de retirarse durante el periodo de prueba.

Los trabajadores operativos tienen 3 veces mayor posibilidad de retirarse durante el periodo de prueba comparados con los trabajadores administrativos.

Al incrementarse en un punto la evaluación de desempeño del jefe del área, incrementa el riesgo de que los trabajadores se retiren antes del periodo de prueba, es decir, mientras el

jefe tenga un puntaje mayor por su desempeño , los trabajadores de su área estarían dispuestos a renunciar a la empresa.

Los trabajadores de áreas que brindan beneficios adicionales tienen mayor posibilidad de retirarse durante el periodo de prueba comparada con los trabajadores de áreas que solo cuentan con los beneficios generales. Se tiene en cuenta que los programas de beneficios adicionales están orientados al personal operativo con la finalidad de compensar las extensas jornadas de trabajo, horario de ingreso y salida y riesgo en la operación.

Los trabajadores de áreas con menos de 300 trabajadores tienen mayor posibilidad de retirarse durante el periodo de prueba comparados con los trabajadores de áreas con más de 300 trabajadores.

Estos resultados dieron una primera aproximación a los resultados obtenidos por el modelo de regresión logística multinivel.

- ***Modelo de regresión logística multinivel***

Los resultados de la regresión logística multinivel se presentan en el cuadro 7.
El modelo nulo muestra una varianza entre áreas de 1.506, un error estándar (SE) de 0.195 y un ρ_1=0.314. Es decir el 31.4% de la variación en la decisión de retirarse durante el periodo de prueba es atribuible a las áreas, independientemente de las variables individuales, existe el efecto del área a la cual pertenece el trabajador. Al introducir las variables del primer y segundo nivel la variabilidad disminuye a ρ_2=0.28, indica que el 28.4% de la variación en la decisión de retirarse durante el periodo de prueba es atribuible a las áreas.

<u>Modelo nulo:</u>

$$logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j}\ ,\ \beta_{0j} = -2.414 + \mu_{0j}$$

$\mu_{0j} \sim N(0, \sigma^2_{\mu_0})$.

j = 1,…,10

Modelo de intercepto aleatorio:

Se planteó un modelo de intercepto aleatorio el cual incluía todas las variables propuestas para el primer y segundo nivel.

$$logit(\pi_{ij}) = \beta_{00} + \sum_{l=1}^{5} \beta_{0l} W_{lj} + \sum_{p=1}^{5} \beta_{pj} X_{pi} + \mu_{0j} + \varepsilon_{ij}$$

$$\mu_{0j} \sim N(0, \sigma^2_{\mu_0}) \; , \; \varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon})$$

i = 1,…,2249; j = 1,…,10

$$\begin{aligned} logit(\pi_{ij}) = {} & -2.414 + 1.286 * Grupo\ ocupacional\ operativo_j - 0.031 \\ & * Escala\ remunerativa_j - 2.17 * Evaluación\ del\ jefe_j - 0.43 \\ & * Número\ de\ trabajadores\ mayor\ a\ 300_j + 1.19 \\ & * Beneficios\ adicionales\ no\ recibidos_j + 0.562 \\ & * Edad\ del\ trabajador\ de\ 18\ a\ 30_i - 0.027 \\ & * Sexo\ del\ trabajador\ masculino_i + 0.181 * Distancia\ lejos_i - 0.224 \\ & * Distancia\ medio_i - 0.295 * Distancia\ sucursal_i - 0.354 \\ & * Estado\ civil\ soltero_i - 0.088 * Número\ de\ hijos_i + \mu_{0j} + \varepsilon_{ij} \end{aligned}$$

Cuadro 7

Modelo de regresión logística multinivel - intercepto aleatorio y modelo de regresión logística múltiple.

Variable dependiente : Retiro antes de culminar periodo de prueba									
	Modelo nulo	Modelo logistico multinivel				Modelo de regresión logistico múltiple			
		Parámetro	SE	OR	IC	Parámetro	SE	OR	IC
Constante		-2,414	1,499	0,09		-1,386	0,868	0,25	
Variables del trabajador									
Edad									
18-30 años		0,562	0,135	1,75	(1,35-2,29)	0,583	0,116	1,79	(1.43-2.25)
30 - mas años									
Sexo									
Femenino									
Masculino		-0,027	0,132	0,97	(0,75-1,26)	0,169	0,105	1,18	(0,96-1,45)
Estado civil									
Soltero		-0,354	0,134	0,70	(0,54-0,91)	-0,376	0,117	0,69	(0,55-0,86)
Casado									
Distancia									
Cerca									
Lejos		0,181	0,125	1,20	(0,94-1,53)	0,099	0,105	1,10	(0,90-1,36)
Medio		-0,224	0,21	0,80	(0,53-1,21)	-0,244	0,182	0,78	(0,55-1,12)
Sucursales		-0,295	0,241	0,74	(0,46-1,19)	-0,315	0,177	0,69	(0,55-0,86)
Numero de hijos		-0,088	0,085	0,92	(0,78-1,08)	-0,068	0,074	0,93	(0,81-1,08)
Variables del area									
Grupo ocupacional									
Administrativo									
Operativo		1,286	0,5	3,62	(1,36-9,64)	1,18	0,312	3,25	(1,77-5,99)
Escala Remunerativa		-0,031	0,059	0,97	(0,86-1,09)	0,02	0,034	1,02	(0,95-1,09)
Evaluación del jefe		2,17	1,545	8,76	(0,42-18,95)	0,638	0,8	1,89	(0,40-9,07)
Número de trabajadores									
Menos de 300									
Mas de 300		-0,43	0,228	0,65	(0,42-1,02)	-0,295	0,156	0,745	(0,55-1,01)
Beneficios adicionales									
Si									
No		0,171	0,212	1,19	(0,78-1,80)	-0,013	0,149	0,99	(0,73-1,32)
$\sigma^2_{\mu_0}$	1,506	1,305							
Correlación intraclase : ρ (%)	31%	28%							

α = 0.05 . Fuente: Propia.

Las estimaciones realizadas indicaron que las variables relacionadas al trabajador son significativas. Los jóvenes entre 18 y 30 años tienen mayor posibilidad de retirarse durante el periodo de prueba que los mayores de 30. Los solteros tienen menor posibilidad de retirarse. El hecho de ser hombre no incrementa la posibilidad de retirarse respecto a las mujeres , esto es discutible debido a que el IC del OR obtenido fue 0.75-1.26 , por otro lado el aporte que hace la variable sexo al modelo es mínimo, lo señalado no se observa bajo el modelo de regresión logística de un nivel . El hecho de que un trabajador viva lejos del centro de labores aumenta la posibilidad de que se retirare durante el periodo de prueba que un trabajador que vive cerca. De las variables contextuales o de las áreas, se observó que las áreas operativas tienen 3 veces mayor posibilidad de tener trabajadores que se retiren durante el periodo de prueba comparado con las áreas administrativas.

Al incrementar la escala remunerativa disminuye la posibilidad de retiro, por otro lado el pertenecer a un área que no tenga beneficios adicionales aumenta la posibilidad de retiro. Lo último señalado es diferente a lo encontrado en el modelo de regresión logística de un nivel, esto se debería a la variación entre áreas no considerado en el modelo.

CONCLUSIONES Y RECOMENDACIONES

- Los resultados obtenidos por el modelo de regresión logístico multinivel son similares al del modelo de regresión logística múltiple, sin embargo es importante considerar la agrupación por áreas ya que explican un porcentaje alto de la variabilidad (28.4%) de la decisión de renunciar antes de culminar el periodo de prueba, es decir durante los 6 primeros meses.
- El perfil de la persona que renunciaría durante este periodo son jóvenes entre 18 y 30 años, sin hijos, solteros, que viven lejos del centro de labores, que ingresan a laborar a áreas operativas donde no hay beneficios adicionales. La evaluación de desempeño del jefe no influye en la decisión de renunciar al igual que el sexo del trabajador.
- Los resultados muestran que hay un efecto de las áreas sobre la decisión de retirarse antes de culminar el periodo de prueba. Si bien los resultados son similares al ignorar esta agrupación, es recomendable utilizar un modelo multinivel por la variabilidad de las variables consideradas en el tiempo y porque permite identificar variables que son significativas en la decisión de retirarse que no se logran identificar en un modelo clásico. Esto lleva a sugerir que el modelo debe ser revisado cada 1 o 2 años con la finalidad de actualizar la realidad de cada área y teniendo en cuenta que la estrategia de LA EMPRESA implica que se realicen cambios organizacionales en periodos más cortos.
- Debido a que la finalidad era identificar un modelo que permita predecir la decisión de retirarse de la empresa durante el periodo de prueba considerando variables del trabajador , en función de postulante, y del área a la cual pertenece , se puede sugerir incorporar variables como tiempo transcurrido desde el último trabajo, formación, estudios actuales, actividades adicionales (hobbies) . Las variables que actualmente se tienen de un ingresante son básicas. Respecto a las áreas, se puede considerar el tipo de liderazgo del jefe, las características y funciones del puesto.

Las variables señaladas no han sido incluidas en el modelo debido a que actualmente no se cuenta con los datos al no ser solicitados al trabajador al ingresar a laborar o no se encuentran registradas en los sistemas de información que se utilizan.

La variabilidad encontrada entre las áreas nos indicó que es correcto considerar esta agrupación para analizar la decisión de retiro o permanencia.

- Se sugiere a las áreas de recursos humanos, como el área encargada de la elaboración de perfiles, alinear los actuales a la realidad de la población, realizando este tipo de análisis con mayor frecuencia. Al área de selección, realizar filtros durante la etapa de reclutamiento en base a los resultados obtenidos y el perfil mejorado con la finalidad de reducir costos.

REFERENCIAS BIBLIOGRÁFICAS

[1] ALLES, M. Dirección estratégica de Recursos Humanos: Gestión por competencias. Edición Granica 2000.

[2] AMADOR, M. y LOPEZ-GONZALES, E. Una aproximación bibliométrica a los modelos multinivel. RELIEVE, v. 13.

[3] CHIAVENATO,I. Administración de Recursos Humanos. Quinta Edición, Editorial Mc Graw Hill 1999.

[4] CASTILLO,J. Administración de Personal: Un enfoque hacia la calidad. Segunda Edición 2006 pág. 69.

[5] BOHLANDER,G y SNELL,S. Administración de recursos humanos,14ª. Edición, 2008.

[6] MILLAN R,J. Rotación de personal, México , Universidad autónoma metropolitana. 2006.

[7] SARRIES,L. y CASARES,E. Buenas prácticas de recursos humanos . Pág. 145.

[8] FERRER,M. Casos prácticos sobre el contrato de trabajo 2010 . Pág. 35.

9] KRUGMAN, P y WELLS, R. Introducción a la economía: Macroeconomía. Editorial Worth Publishers 2006.

[10] COMENGES, D y JACQMIN, H. The intraclass correlation coefficient: distribution-free definition and test. 1994. Pág 517-526.

[11] CARMONA,F. Modelos Lineales. Electronic-University Mathematical Books.

[12] MOUTINHO,G. y DEMETRIO,C. Modelos Lineares Generalizados e Extensões . 2010.

[13] GAVIRIA, J. y CASTRO,M. Modelos Jerárquicos Lineales. Editorial La Muralla 2005.

[14] DE LA CRUZ, F. Modelos Multinivel. Revista peruana de epidemiología .Vol.12 Nº3 Diciembre 2008.

[15] GOLDSTEIN.H. Multilevel Statistical Models. London: Institute of Education, Multilevel Models Project 1999.

[16] SNIJDERS,T y BOSKEL,R. Multilevel Analysis. SAGE Publications 1999.

[17] GOLDSTEIN.H y RASBASH,J. Improved Approximations for Multilevel Models with Binary Responses. Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 159, No. 3. (1996), pág. 505-513.

[18] RODRIGUEZ, G y GOLDMAN, N. An assesment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society 1995.

[19] BRYK, A y RAUNDENBUSH,S. Hierarchical Linear Models: applications and data analysis methods.1992.

[20] MOERBEEK M. The consequence of ignoring a level of nesting in multilevel analysis. Multivariate Behavioral research. 2004.

[21] ROBINSON WS. Ecological Correlations and the Behavior of Individuals. American Sociological Review. 1950.

[22] Diez Roux A V. A glossary for multilevel analysis. J Epidemial Community Health.2002.

[23] Hox JJ. Applied multilevel analysis. 1995.