Informe Estandarización Perú Escala INDI, Parte 4: Análisis de Regresión

Muestra Nivel 3

Martín Vargas Estrada

2025-01-31 21:33:31.76167

Índice

Introducción

Informe de Exploración Psicométrica de los ítems de la prueba INDI obtenidas con muestra de Perú, Nivel 3.

Análisis de Regresión entre el valor de las Escalas y las Variables Demográficas

Introducción

El objetivo de esta sección realizar un análisis de regresión entre el valor de las escalas y las diferentes variables demográficas cuya información fue recabada durante la investigación.

Cabe especificar el sentido del análisis de regresión, ya que, a diferencia de los anteriores, es fácil malentenderlo y su significado no es evidente a primera vista.

El objetivo del análisis de regresión es encontrar una relación de causalidad entre una o más variables independientes, y una variable dependiente. Esta última es la "estrella" de nuestro análisis; en contra de lo que pudiera parecer, la variable dependiente es en realidad la principal, la que motiva toda la pesquisa estadística. Estamos interesados determinar si ciertas variables pueden ayudarnos a explicar y/o predecir a nuestra variable dependiente.

Una variable es "independiente" en el sentido de que, dentro de nuestra teoría y/o nuestro sentido común, ciertas variables podrían causar o predecir a nuestra variable dependiente, y esta relación, teóricamente, debería ser unidireccional. Por ejemplo, cuando decimos que la *variable independiente* es la Edad del participante y nuestra *variable dependiente* es el puntaje en la escala Cognitiva del INDI, estamos planteando la posibilidad de que la edad nos ayude a explicar y/o predecir el desempeño en el INDI, y plantearnos una relación inversa (¿el puntaje influenciando a la Edad?) o incluso recíproca (¿Edad y desempeño en el INDI influenciándose mutuamente?) no tendría sentido.

A veces, la lógica la define nuestro sentido común (claramente la edad, por definición, no es influenciada por ningún tipo de acción del participante); a veces es determinada por la teoría que nos guía para plantear toda la investigación en primer lugar.

Interpretación General de los Resultados de la Regresión

1. **Qué es la ecuación de regresión**. En primer lugar, es importante entender cuál es el sentido de la regresión. Recordemos que nuestro propósito es explicar/predecir los valores de una variable dependiente a partir de variables independientes. Explicar significa en este caso definir una ecuación que describa el comportamiento de la variable dependiente. En esa ecuación la variable dependiente será el resultado (es decir, irá antes del signo "=") y las variables independientes serán los sumandos que al agregarse resultan en el valor que la variable dependiente tomará según el caso.

En términos matemáticos:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Cada valor X representa una variable (numérica o categórica), y cada valor β un coeficiente que pondera el valor que cada variable asume para poder así determinar un valor dado que asumirá la variable dependiente Y. El subíndice se usa para especificar que el coeficiente será distinto para cada variable independiente.

El coeficiente β_0 es lo que se llama *intercepto*; es el valor que asume la variable dependiente cuando todas las variables independientes asumen un valor cero.

2. **Qué significan las variables *dummy* en el contexto del tratamiento estadístico de la regresión**. Cada vez que incluimos una variable categórica (es decir, una variable cuyos valores no son numéricos, sino más bien categorías; ejemplo: Región, cuyos valores serían las categorías "Costa", "Sierra", y "Selva"), nos vemos obligados a utilizar un algoritmo que nos permita incluir esas variables en

el cálculo. Ese algoritmo o método consiste en asumir como referencia la primera categoría de la variable, y luego pasar a incluir las demás categorías como si se tratase de variables distintas, convirtiéndolas en variable dicotómicas.

En el caso de nuestro ejemplo, si nuestra variable es "Región", y las categorías son Costa, Sierra y Selva, entonces podemos asumir la categoría "Costa" como valor de referencia; cuando un participante pertenezca a la Región "Sierra", lo traduciremos en términos de variables *dummy* como un "1" en la variable *RegiónSierra*. Si el participante no pertenece a la la Región Sierra, tendrá un "0" en la variable *RegiónSierra*. Igual haremos con todos los demás casos de variables categóricas.

Esto tiene una enorme ventaja: nos permite tratar a las variables categóricas como si fueran variables dicotómicas numéricas, haciendo posible incluirlas en la ecuación de regresión. Pero tiene una gran desventaja: nos obliga a tener mucho cuidado al momento de interpretar la ecuación.

Pongamos un ejemplo para clarificar: imaginemos que estamos tratando de explicar la variable Escala Cognitiva a partir de la variable Edad y la variable Región. Es decir,

$$Y = EscalaCognitiva, X_1 = Edad, X_2 = Región$$

Con la variable Edad no tendremos problema porque se trata de una variable numérica. No es el caso con la variable Región. Tomaremos la categoría "Costa" como referencia; por lo tanto, cuando tengamos un participante de la Costa sencillamente tomaremos el valor del intercepto (β_0) como el aporte de la variable *RegiónCosta*. Si el participante proviene, por el contrario, de la Sierra, entonces tendremos que la variable RegiónSierra=1, y la variable RegiónSelva=0. Todas las variables independientes con valor cero serán eliminadas de la ecuación porque tendrán valor nulo, y solo quedarán las variables independientes con valor distinto de cero. Para fines de nuestro ejemplo, asumamos que:

$$\beta_0 = 1.2; \beta_1 = 0.9; X_1 = Edad; \beta_2 = 2.7; X_2 = RegionSierra; \beta_3 = 4.1; X_3 = RegionSelva$$

Por lo tanto, nuestra ecuación de regresión quedará como sigue:

$$EscalaCognitiva = 1.2 + 0.9 * Edad + 2.7 * RegionSierra + 4.1 * RegionSelva$$

Si un participante pertenece a la Región Costa y tiene una edad de 40 meses, entonces nuestro modelo predice que:

$$EscalaCognitiva = 1.2 + 0.9 * 40 + 2.7 * 0 + 4.1 * 0$$

Es decir,

$$EscalaCognitiva = 37.2$$

Regresión y Correlación

La diferencia clave entre la regresión y la correlación es que, mientras la correlación solo mide asociación entre variables, la regresión trata de determinar la *causalidad* de una o más variables (llamadas independientes) sobre otra (llamada dependiente).

Otra diferencia importante es que mientras la correlación solo describe la asociación o co-ocurrencia entre dos variables, *la regresión está diseñada de tal modo que nos ayuda a determinar en qué medida una variable independiente explica el comportamiento de la dependiente*. Como luego veremos, es completamente posible, como resultado de un análisis de regresión, establecer en qué porcentaje una variable independiente X nos ayuda a explicar el comportamiento o variabilidad de los datos de una variable dependiente.

Finalmente, una tercera distinción fundamental es que mientras que la correlación se limita *describir* la co-ocurrencia de dos conjuntos de datos ya existentes, el análisis de regresión nos permite *predecir* valores de

la variable dependiente a partir de valores de las variables independientes, *incluso si tales valores no se dan en nuestra muestra.*

Regresión como descriptor de una relación lineal

Antes de pasar a ver los resultados, debemos tener en cuenta que el tipo de análisis de regresión que pasaremos a ejecutar se basa en el supuesto de que existe una relación lineal entre el conjunto de variables independientes y los puntajes en las escalas INDI. Es decir, que el efecto de las variables independientes es constante, va en una sola dirección y no cambia de sentido. Por ejemplo, si asumimos una relación lineal entre Edad en Meses y Puntaje en el INDI, estaremos asumiendo que el aumento de un mes en la Edad mejorará el desempeño en el INDI, y esto no cambiará y además es constante (el incremento de un mes en la edad debería tener el mismo efecto en el puntaje, sin importar la edad del participante).

Es posible que la relación entre una variable dependiente y una o más variables independientes sea no lineal; en ese caso, el resultado de una regresión lineal será nulo, sin que esto necesariamente implique falta de relación en general. Tan solo estaríamos hablando de falta de relación *lineal*.

Habiendo dicho esto, por lo general la regresión lineal es la que se analiza, ya que se trata del modelo más frecuente y sencillo de entender y de evaluar. Salvo que nuestra teoría establezca lo contrario, será el tipo de regresión que usaremos.

Variables Demográficas Consideradas (Variables Independientes)

Consideraremos las siguientes variables:

1. Edad en Meses
2. Fecha de Eval
3. Región
4. Área
5. Modalidad
6. Gestión
7. Departamento
8. Quintil
9. Inst. Previa al Nivel 3

Las variables dicotómicas de Incidencia (codificadas como VSS y similares) y Tratamiento (RSS y similares) no fueron consideradas para este análisis, ya que definen subgrupos extremadamente pequeños (menos del 5%,; en la mayoría de casos, menos del 3%) y tomarlas en cuenta distorsionaría los resultados, al obligar a tomar medidas de ajuste que a su vez introducirían nuevos sesgos. Lo mismo sucede con los casos de la variable Gestión, y el pequeño grupo perteneciente a la categoría “No escolarizado” dentro de la variable “Modalidad”.

Análisis 1: Escala Cognitiva como Variable Dependiente

A continuación determinaremos qué tan útil es un análisis de regresión para intentar explicar/predecir los puntajes en la escala Cognitiva del INDI a partir de las variables independientes ya mencionadas.

Regresión entre Escala Cognitiva y Edad en Meses, Región, Área, Modalidad, Departamento, Quintil, Instrucción Previa

Variable	Coeficiente	p-valor	Significancia
(Intercept)	46.706	0.0000	***
Edad en Meses	1.233	0.0000	***
RegiónSierra	5.213	0.2224	NS
RegiónSelva	11.987	0.0214	*
ÁreaRural	-3.265	0.0569	NS
ModalidadJardín	-3.574	0.0861	NS
DepartamentoLima Met.	29.297	0.0000	***
DepartamentoLoreto	0.951	0.7289	NS
DepartamentoPiura	2.021	0.6689	NS
Quintil2	-3.755	0.4396	NS
Quintil3	-2.572	0.5037	NS
Quintil4	-3.318	0.3490	NS
Quintil5	6.332	0.0993	NS
Instrucción PreviaSí	6.464	0.0000	***
Instrucción PreviaNS/NR	9.132	0.0007	***

La ecuación de regresión obtenida es (solo coeficientes significativos):

Escala Cognitiva = 46.706 + 1.233 * Edad en Meses + 11.987 * RegiónSelva + 29.297 * DepartamentoLima Met. + 6.464 * Instrucción PreviaSí + 9.132 * Instrucción PreviaNS/NR

R Cuadrado Ajustado: 0.391

La información anterior nos indica que:

1. Usar todas las variables que hemos recolectado para predecir el puntaje en la Escala Cognitiva del INDI nos genera un modelo que puede explicar casi el 40% de la variabilidad de los datos.
2. Las únicas variables que realmente aportan a la explicación del puntaje en la Escala Cognitiva INDI son:

a. Edad
b. Departamento (Lima)
c. Región Selva
d. Que el participante haya tenido Instrucción previa la Nivel 3
e. Que el participante no pueda determinar si tuvo Instrucción previa la Nivel 3

Análisis 2: Escala Motora como Variable Dependiente

A continuación determinaremos qué tan útil es un análisis de regresión para intentar explicar/predecir los puntajes en la escala Motora del INDI a partir de las variables independientes ya mencionadas.

Regresión entre Escala Motora y Edad en Meses, Región, Área, Modalidad, Departamento, Quintil, Instrucción Previa

Variable	Coeficiente	p-valor	Significancia
(Intercept)	15.863	0.0000	***
Edad en Meses	0.476	0.0000	***
RegiónSierra	0.736	0.6200	NS
RegiónSelva	3.207	0.0762	NS
ÁreaRural	-0.514	0.3875	NS
ModalidadJardín	1.172	0.1052	NS
DepartamentoLima Met.	5.794	0.0013	**
DepartamentoLoreto	-0.482	0.6126	NS
DepartamentoPiura	0.949	0.5630	NS
Quintil2	-0.445	0.7922	NS
Quintil3	1.184	0.3754	NS
Quintil4	1.136	0.3559	NS
Quintil5	3.334	0.0125	*
Instrucción PreviaSí	1.472	0.0070	**
Instrucción PreviaNS/NR	1.349	0.1502	NS

La ecuación de regresión obtenida es (solo coeficientes significativos):

Escala Motora = 15.863 + 0.476 * Edad en Meses + 5.794 * DepartamentoLima Met. + 3.334 * Quintil5 + 1.472 * Instrucción PreviaSí

R Cuadrado Ajustado: 0.2342

La información anterior nos indica que:

1. Usar todas las variables que hemos recolectado para predecir el puntaje en la Escala Motora del INDI nos genera un modelo que solo puede explicar menos del 24% de la variabilidad de los datos. Esto significa que más del 76% del comportamiento de la Escala Motora del INDI se debe a factores no considerados dentro de este estudio.
2. Las únicas variables que realmente aportan a la explicación del puntaje en la Escala Motora INDI son:

a. Edad
b. Departamento (Lima)
c. Que el participante provenga del Quintil 5
d. Que el participante haya tenido Instrucción previa la Nivel 3

Análisis 3: Escala Disposicional como Variable Dependiente

A continuación determinaremos qué tan útil es un análisis de regresión para intentar explicar/predecir los puntajes en la escala Motora del INDI a partir de las variables independientes ya mencionadas.

Regresión entre Escala Disposicional y Edad en Meses, Región, Área, Modalidad, Departamento, Quintil, Instrucción Previa

Variable	Coeficiente	p-valor	Significancia
(Intercept)	16.752	0.0000	***
Edad en Meses	0.136	0.0004	***
RegiónSierra	1.406	0.2370	NS
RegiónSelva	3.660	0.0116	*
ÁreaRural	-1.693	0.0004	***
ModalidadJardín	0.824	0.1547	NS
DepartamentoLima Met.	5.521	0.0001	***
DepartamentoLoreto	0.282	0.7122	NS
DepartamentoPiura	1.155	0.3798	NS
Quintil2	2.669	0.0485	*
Quintil3	3.157	0.0032	**
Quintil4	3.011	0.0023	**
Quintil5	4.707	0.0000	***
Instrucción PreviaSí	1.570	0.0003	***
Instrucción PreviaNS/NR	2.848	0.0002	***

La ecuación de regresión obtenida es (solo coeficientes significativos):

Escala Disposicional = 16.752 + 0.136 * Edad en Meses + 3.66 * RegiónSelva - 1.693 * ÁreaRural + 5.521 * DepartamentoLima Met. + 2.669 * Quintil2 + 3.157 * Quintil3 + 3.011 * Quintil4 + 4.707 * Quintil5 + 1.57 * Instrucción PreviaSí + 2.848 * Instrucción PreviaNS/NR

R Cuadrado Ajustado: 0.2106

La información anterior nos indica que:

1. Usar todas las variables que hemos recolectado para predecir el puntaje en la Escala Disposicional del INDI nos genera un modelo que solo puede explicar algo más del 20% de la variabilidad de los datos. Esto significa que casi el 80% del comportamiento de la Escala Disposicional del INDI se debe a factores no considerados dentro de este estudio.
2. Las únicas variables medidas que realmente aportan a la explicación del puntaje en la Escala Cognitiva INDI son:

a. Edad
b. Región Selva
c. Que el participante provenga de la Región Rural (en detrimento)
d. Que el participante provenga de Lima
e. Quintil (todos)
f. Que el participante haya tenido Instrucción previa la Nivel 3
g. Que el participante no pueda determinar si tuvo Instrucción previa la Nivel 3