Maestría en estadística aplicada Modelos lineales

Modelo lineal Relación de salario con varias variables para funcionarios del estado de Durango

Estudiante:

Carlos Daniel Hernández Valdez

Docente: Dr. Mario Alberto García Meza.

Junio - 2022

**Modelo lineal Relación de salario con varias variables para funcionarios del estado de Durango**

**Introducción**

Las necesidades de información clara y bien explicada son conceptos que deberían estar ligados a temas de transparencia permitir a las personas conocer en que se gasta sus impuestos para ellos el ISSSTE a través de su sitio [SINAVID Oficina Virtual (issste.gob.mx)](https://oficinavirtual.issste.gob.mx/) pone a disposición del publico una base de datos que mes por mes contiene una relación de detalles de los trabajadores en activo.

**Descriptivos de la base de Datos**

Es necesario mencionar que de la data presentada en este proyecto fueron eliminadas distintas variables que decisión propia se prestan para temas polémicos como lo es el nombre y apellidos de todos los trabajadores y que aunque lo que se publica es un salario base puede ser un tema de controversia tal base de datos pose los registros de la totalidad de la población, es importante mencionar esto ya que por ende no necesitamos estar datos desconocidos y únicamente serán utilizados para explicar los efectos que tiene cierta variables sobre el salario de la población de funcionarios ya sea en forma general o por grupos.

La data consiste en 53635 registro con 8 variables de la cuales 4 son de interés de estudio de este proyecto.

**Distribución de las variables en el modelo**

Captura de pantalla de un celular con texto

Descripción generada automáticamente

Es parte de la teoría de los modelos lineales que para poder utilizar una variable dentro de un modelo es necesario que este presente cierto nivel de varianza de manera que los cambios en esta permitan establecer un efecto según el valor de nuestra o nuestra variables regresoras, según el resumen de nuestra base de datos nuestras variables aparte de constar de dos tipos de datos diferentes (carácter y numérico) todas presentan un nivel de variación suficiente para ser utilizadas

**Correlación de variables**

Gráfico

Descripción generada automáticamenteAlgo que debemos tener siempre presente al momento de iniciar una análisis mediante modelos de regresión lineal es ver la forma en que la variables se relacionan entre si para esto tenemos el concepto de correlación que por desgracia r solo permite el análisis de variable numéricas.

Tabla

Descripción generada automáticamente con confianza baja

**Variables de interés**

Algo que ya se mencione es que tanto las variables regresoras como la variable independiente deben mostrar cierto nivel de varianza de los contrario nuestro modelo no arrojara mas que un valor constante inútil hasta cierto punto para dar una explicación lógica y ya se mencionó que todas las variables cumplen con este criterio

Como variables explicativas de nuestro modelo tomaremos tres las cuales serán de utilidad en distinto momento y a las cuales haremos distintas evaluaciones según su efecto.

**Sexo. -** algo muy importante en los últimos tiempos es poder explicar si el sexo de una persona influye sobre lo que puede ganar es por eso que se elige esta variable que por lógica puede tomar solo dos valores masculino y femenino o para el caso de la base de datos.

Gráfico, Gráfico circular

Descripción generada automáticamente

**Antigüedad.-** otra de nuestras variables de interés es demostrar si los años trabajados tiene influencia sobre el salario adquirido es por eso que verificamos la frecuencia y vemos con la población tiene un comportamiento decreciente es decir la mayor parte de la población tiene pocos años de experiencia lo cual es un comportamiento totalmente esperado.

Gráfico, Histograma

Descripción generada automáticamente

**Nombramiento. –** para le caso de nombramiento es un caso diferente de las otras dos variables ya que es una variable de texto que para el caso de un modelo lineal el objetivo es estimar que tanto afecta al salario que un funcionario pertenezca a cierto grupo según su tipo de contrato.

Texto

Descripción generada automáticamente

Es de notar que la mayoría de los funcionarios entran en dos categorías pero eso no impide poder usar la variable para calcular si algunas de las otras categorías tienen algún efecto.

**Relación grafica con el salario**

Como modelo inicial es de importancia demostrar gráficamente que nuestra variable regresora principal muestra algún comportamiento el cual podría tener alguna relación con el salario y dado que la data es muy grande un diagrama de dispersión simple no ayuda para conocer la tendencia quizá si se agrega una funcionalidad de densidad, pero creo que para este caso un boxplot es mejor opción y poner atención en la tendencia que tienen las medias.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Dado el boxplot anterior notamos una ligera tendencia hacia arriba sobre todo a partir de los 15 años de experiencia y dada la forma es posible que sea necesario algún termino polinómico, pero para no complicar la explicación del presente trabajo se continuara con un termino sencillo concluyendo en la información que este nos pueda proporcionar.

**Elección de unidades de medida**

Es conocido que en cuestión de salario existe cierto comportamiento donde la concentración de salarios tiene una mayor concentración de observaciones sobre la cantidades de ingresos bajas mientras que las cantidades altas tienen muy poca densidad, para el caso y dado que lo que se esta evaluando en la data es el salario base esta situación no se presenta en forma tan marcada considerando que las observaciones mas altas no pasan de los 10 salarios mínimos aun así y dado que los especialistas recomiendan hacer un ajuste logarítmico con el fin de que estas observaciones no tengan tanto pesos sobre los coeficientes de nuestro modelo la primera etapa de este análisis será la de utilizar ambas unidades para elegir cual será la mejor presentación de nuestra variable dependiente que como recordamos es el salario.

Unidad natural: La primera unidad disponible es la calculada a partir del salario mensual donde solo se hace una división entre 30 a fin de suavizar los coeficientes del modelo poder interpretarlos y así establecer una conclusión directa sobre el ingreso diario de los funcionarios en nuestra data

Imagen que contiene Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Texto

Descripción generada automáticamente

La segunda unidad es a partir de nuestra unidad anterior pero esta vez sometida a un calculo de logaritmo natural y de esta manera suavizar el efecto que las observaciones con valores altos tienen sobre los coeficientes de nuestro modelo en la figura anterior se muestra el resumen de los valores en ambas variables, así como el comportamiento de la variable original dejando ver el comportamiento ya mencionado sobre la densidad de los salarios sobre los valores menores.

**Relación simple**

Algunas investigaciones tratan de aumentar la bondad de ajuste del modelo. Sin embargo, para estimar el efecto de una sola variable no es necesario aumentar la bondad de ajuste sino simplemente cumplir con los supuestos del Modelo Lineal Clásico, tales como la media condicional cero, la linealidad en los parámetros y el resto de los supuestos que se describen en la sección correspondiente.

Sin embargo en esta primera iteración lo que tratamos de elegir es la mejor unidad para la variable dependiente de las dos unidades antes mencionas para esto utilizaremos un método de modelado de una sola variable a partir de la antigüedad del funcionario y evaluaremos los resultados predichos de cada modelo contra el valor de la variable original todo evaluado en la unidad original o de lo contrario sería lógico que aquella unidad con lo valores más bajos se llevaría la victoria, dicho esto procedemos con los modelos.

Imagen de la pantalla de un celular de un mensaje en letras blancas

Descripción generada automáticamente con confianza baja

Texto

Descripción generada automáticamente

Teniendo en cuenta estos modelos podemos ya ir haciendo alguna conclusión sobre todo por la simplicidad de nuestra primera unidad y el valor p que nuestro modelo muestra podemos ir concluyendo una influencia positiva de la experiencia del funcionario sobre el modelo como ya antes habíamos concluido a partir del bloxplot solo que en esta ocasión podemos cuantificar de forma preliminar que cada año de experiencia aumenta en 3.48 pesos el salario de un funcionario de forma general.

Procediendo con la evaluación de los resultados en esta primera iteración se obtuvo que la menor diferencia cuadrada entre las observaciones y los resultados predichos la obtuvo nuestra segunda unidad, aunque la diferencia fue menor a lo esperado según menciona la teoría consultada, cabe mencionar que esta evaluación se hizo de forma manual

**Relación múltiple y Análisis de Varianza**

Continuando ya únicamente con una de nuestras unidades podemos formular diferentes modelos ya sea de una o varias variables predictoras ya mencionadas considerando que dos de ellas son cuantitativas y

Tabla

Descripción generada automáticamenteTabla

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamenteTexto

Descripción generada automáticamente

Como aspecto importante en la evaluación delos resultados de los modelos debemos mencionar el ANOVA realizado donde se evaluaron los dos primeros modelos y aunque resulto mas significativo nuestro modelo con dos variables la diferencia de R2 no es tan importante como quisiéramos para este caso sería conveniente evaluar que tan importante es incluir la variable sexo en el modelo que si bien aporta información a la conclusiones no lo hace para la exactitud de nuestras predicciones, todo esto cambia al agregar los factores de nuestra tercera variable el cual no solo mejora la diferencia de los residuales sino que aporta información sobre el efecto que tiene el tipo de contrato sobre el cual se encuentra un funcionario .

**Diagnóstico y supuestos del modelo**

El modelado asume que la variable dependiente (y) tiene una relación lineal con la(s) variable(s) independiente(s) y el término de error (u). Es decir, un aumento de una unidad de x resulta en un efecto constante sobre la variable dependiente y. De ahí la forma funcional de la ecuación de regresión:

**Y=β0+β1x+u**

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Como la relación no es del todo lineal, entonces nos enfrentamos a un problema de especificación del modelo. Es decir, los valores predichos por nuestro modelo no encajan en la realidad de nuestros datos y, en consecuencia, las estimaciones estarán sesgadas. Por lo tanto, es fundamental evaluar si la relación que queremos estimar es lineal o si la forma funcional que caracteriza esa relación es otra

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Interfaz de usuario gráfica, Sitio web

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente